

# CSE/ISyE 6740

## Computational Data Analytics

## Fall 2025

Kai Wang, Assistant Professor in Computational Science and Engineering  
[kwang692@gatech.edu](mailto:kwang692@gatech.edu)

Based on materials by Anqi Wu, B. Aditya Prakash, Le Song, Mahdi Roozbahani,  
Carlos Guestrin

# Course Information

- CSE/ISyE 6740 Computational Data Analytics
  - **Date and time:** Monday and Wednesday, 12:30pm – 1:45pm
  - **Classroom:** Boggs B9
  - **Format:** in-class lectures, no recording
- Instructor: Kai Wang
  - Assistant Professor, CSE, College of Computing
  - **Office:** CODA S1309
  - **Email:** kwang692@gatech.edu
  - **Research Interests:** AI for social impact, machine learning, optimization, reinforcement learning, online learning, multi-agent systems, diffusion models
  - **Office hours:** Wednesday 3-4pm ET (tentatively)



# Teaching Assistants



Po-Han Huang (head TA)  
[phuang322@gatech.edu](mailto:phuang322@gatech.edu)



Tianyi Chen  
[tchen667@gatech.edu](mailto:tchen667@gatech.edu)



Neeraj Kumar  
[nkumar355@gatech.edu](mailto:nkumar355@gatech.edu)



Tanish Patwa  
[tpatwa6@gatech.edu](mailto:tpatwa6@gatech.edu)

They will hold office hours and answer questions on Piazza and Gradescope for you. Both have been activated on Canvas.

# Piazza and Gradescope

- This term, we will be using Piazza for class discussion. The system is highly catered to getting you help fast and efficiently from classmates, the TA, and myself. Rather than emailing questions to the teaching staff, I encourage you to post your questions on Piazza
- Piazza link: <https://piazza.com/class/mcxkmyyva5h1ri>
- Gradescope link: <https://www.gradescope.com/courses/1067808>
- You can also find the Piazza link on your Canvas page.

# Outline

- Instructor, TAs, and course information
- Course overview
- Basics and prerequisites (with linear regression as an example)
- Assignment and grading
- Homework 0

# Course Overview

# What is Machine Learning (ML)?

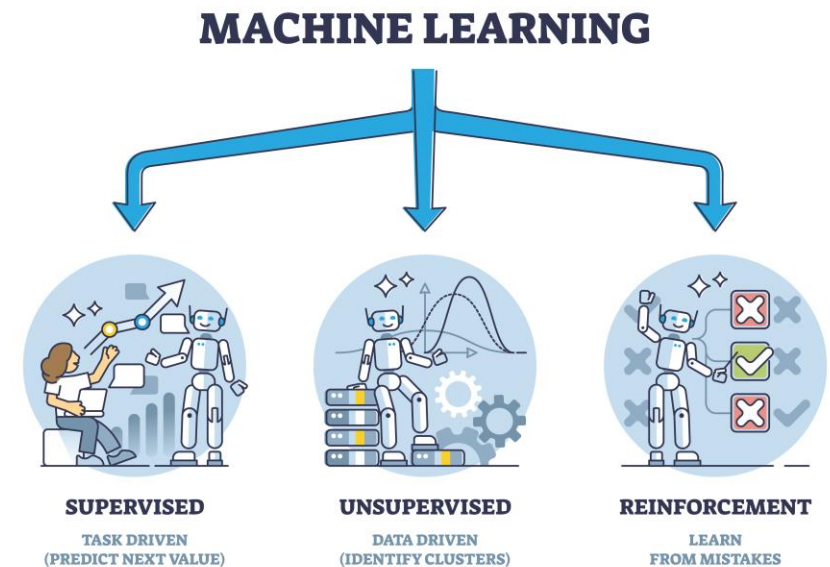
- Study of algorithms that improve their performance at some tasks based on experience



MACHINE LEARNING

# Syllabus

- Cover a range of most commonly used machine learning algorithms and their mathematical foundations.
- **Outline**
  - **Unsupervised learning (week 1 – 4)**
    - Learning without labels or without optimizing for prediction tasks
  - **Supervised learning (week 4 – 9)**
    - Learning with labels, focusing on predictive performance
  - **Advanced models (week 10 – 15)**
    - Advanced neural networks, diffusion models, reinforcement learning





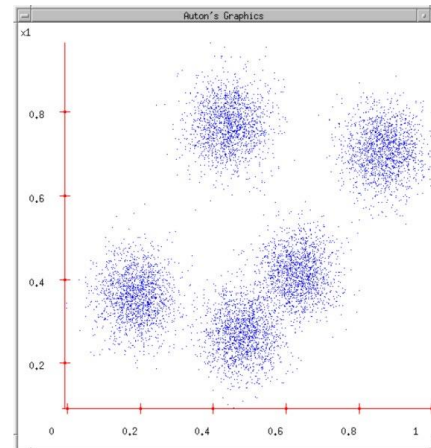
# Syllabus: Unsupervised Learning

- Learning without labels or without optimizing for prediction tasks
  - Clustering vectorized data (week 1)
    - K-mean clustering
    - Hierarchical clustering
    - Graph clustering
  - Dimensionality reduction (week 2)
    - Principal component analysis
    - Nonlinear dimensionality reduction
  - Density estimation (week 3-4)
    - Feature selection
    - Novelty/abnormality detection
    - Gaussian mixture models, EM algorithm

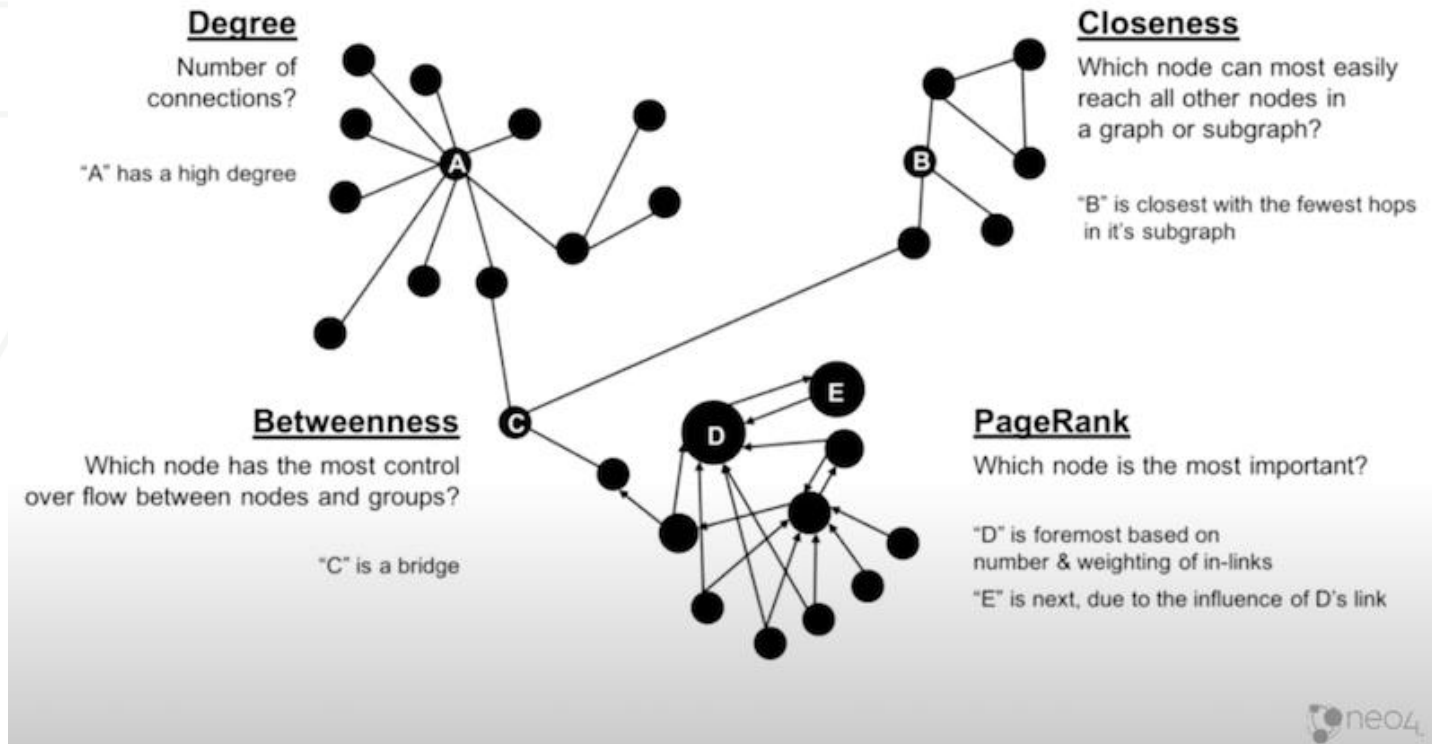
# Organizing Images



- What are the desired outcome?
- What are the input (data)?
- What are the learning paradigms?



# Find Communities in Social Network



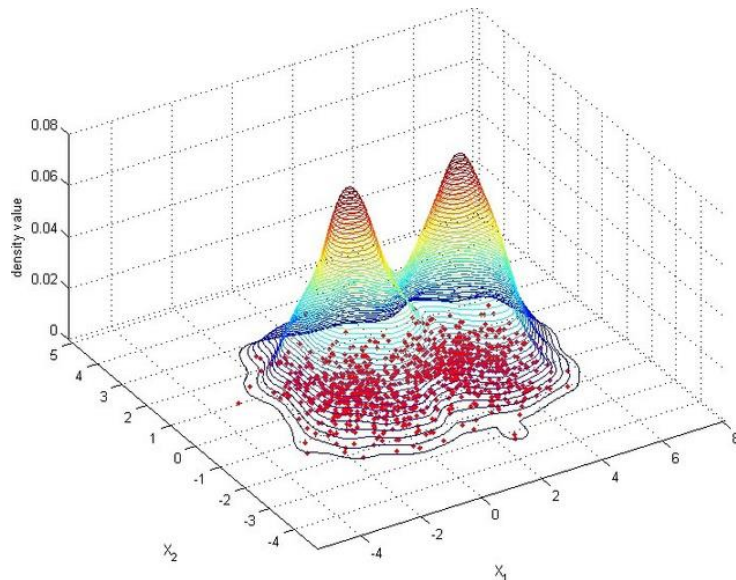
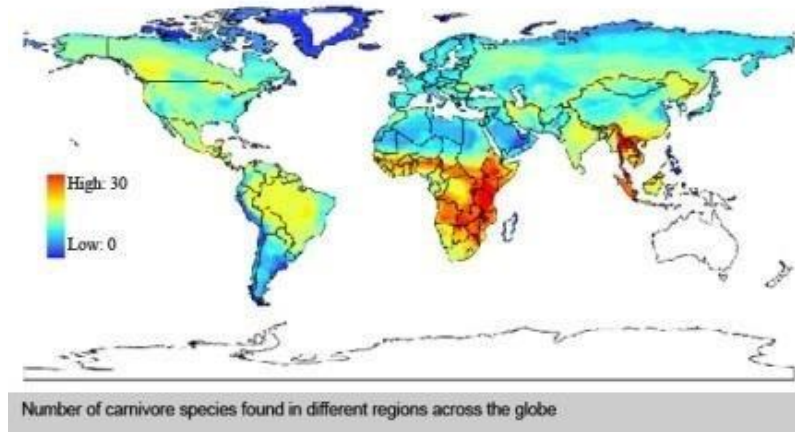
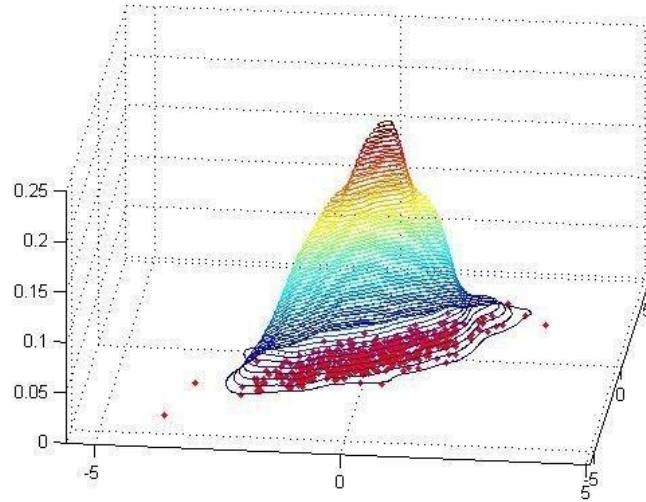
- What are the desired outcome?
- What are the input (data)?
- What are the learning paradigms?

# Visualize Image Relations



- What are the desired outcome?
- What are the input (data)?
- What are the learning paradigms?

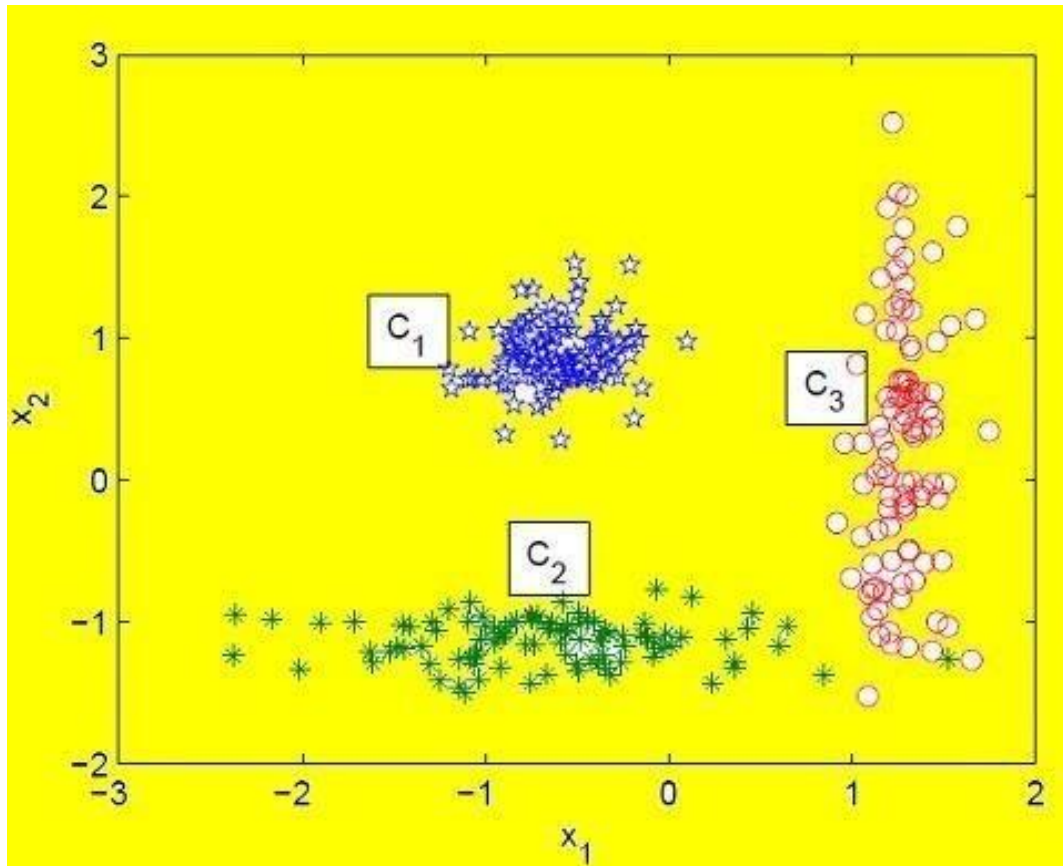
# Shape of Data



- What are the desired outcome?
- What are the input (data)?
- What are the learning paradigms?



# Feature Selection



- What are the desired outcome?
- What are the input (data)?
- What are the learning paradigms?

# Novelty/anomaly Detection

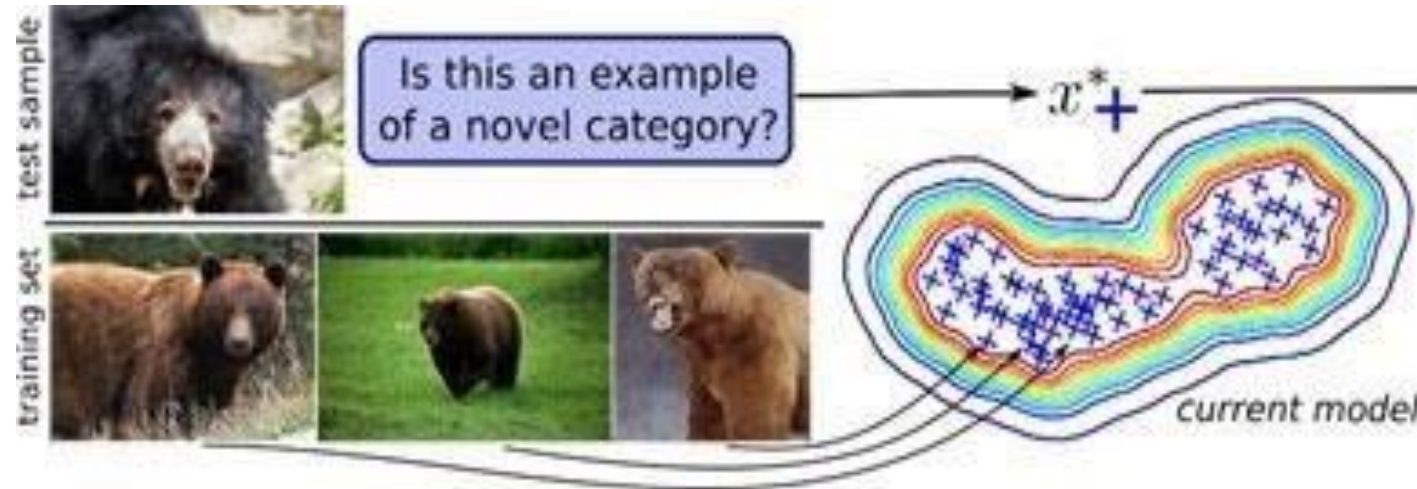


FIGURE 1 EXAMPLE OF NOVELTY DETECTION, BODESHEIM (2012).



- What are the desired outcome?
- What are the input (data)?
- What are the learning paradigms?

# Syllabus: Supervised Learning

- Learning with labels, focusing on predictive performance
  - Classical classification models (week 4)
    - K-nearest neighbors, Naive bayes
  - Decision tree (week 5)
    - Feature selection, entropy, random forest
  - Logistic regression and support vector machine (week 6)
    - Convex analysis, primal and dual, SVM
  - Kernel methods (week 7)
    - Kernel regression, Bayesian regression, Gaussian process regression
  - Neural networks (week 8)
    - Gradient descent, neural networks, overfitting, bias-variance tradeoff

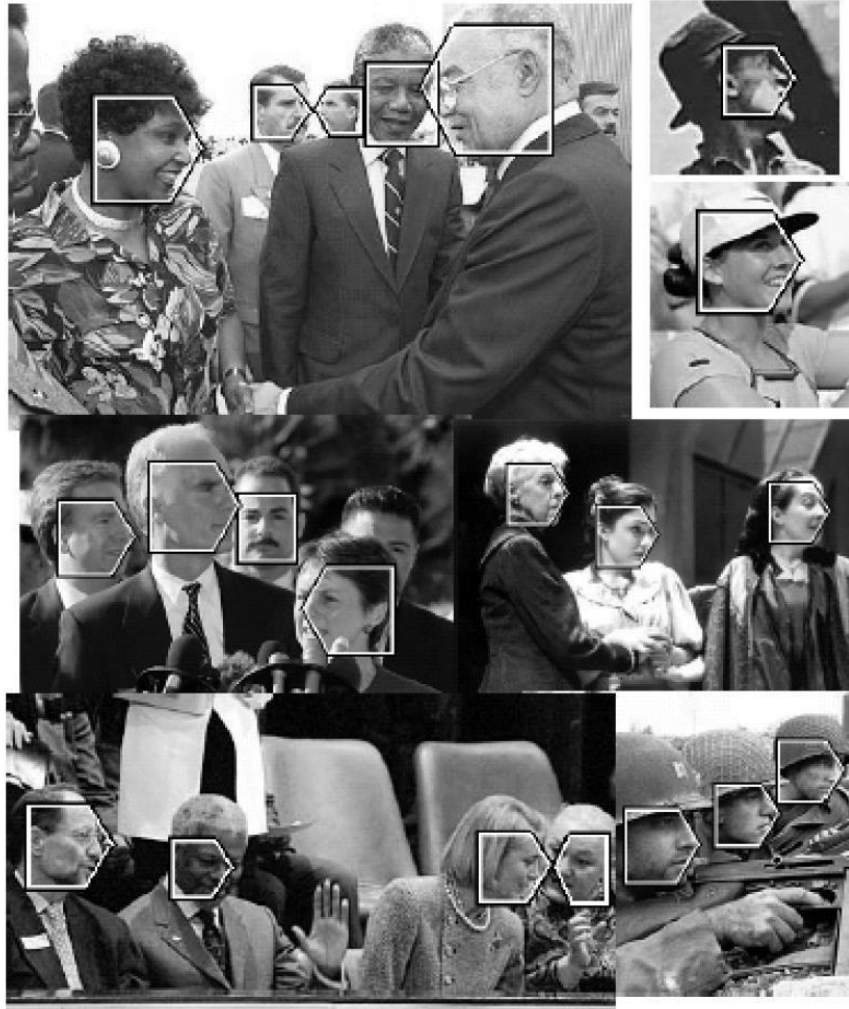


# Image Classification



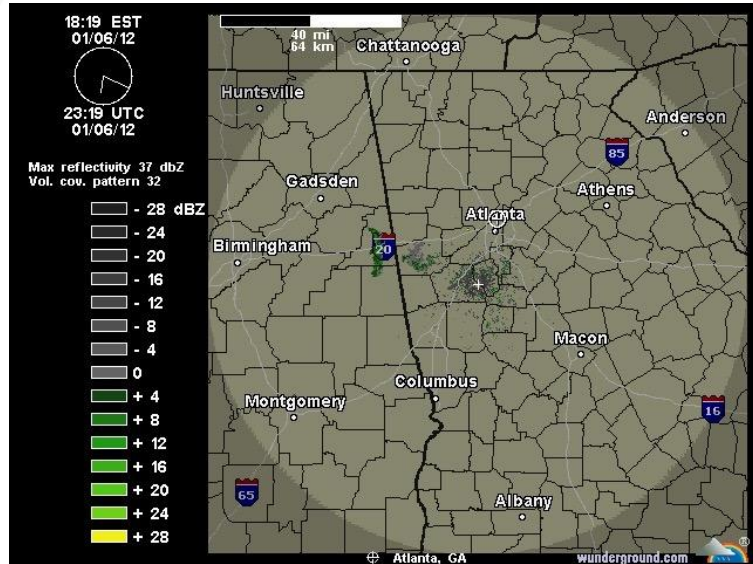
- What are the desired outcome?
- What are the input (data)?
- What are the learning paradigms?

# Face Detection



- What are the desired outcome?
- What are the input (data)?
- What are the learning paradigms?

# Weather Prediction



Predict

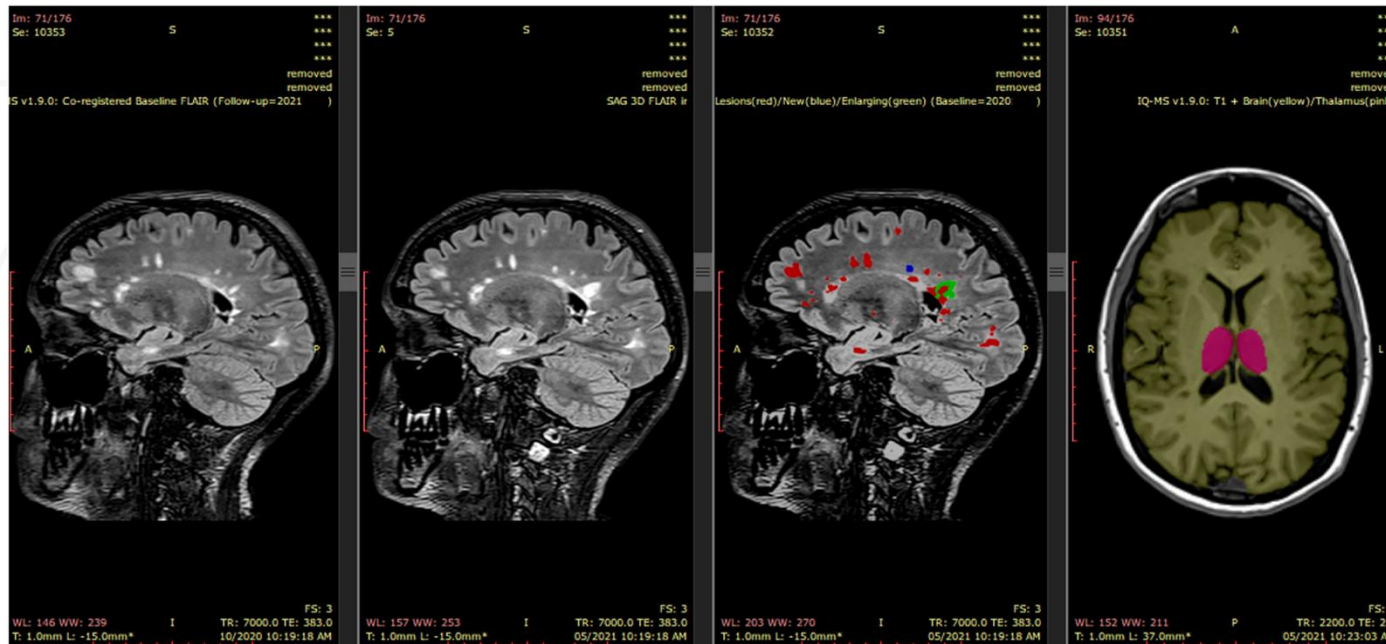
Numeric values: 40F  
Wind: NE at 14km/h  
Humidity: 83%

Predict



- What are the desired outcome?
- What are the input (data)?
- What are the learning paradigms?

# Understanding Brain Activity



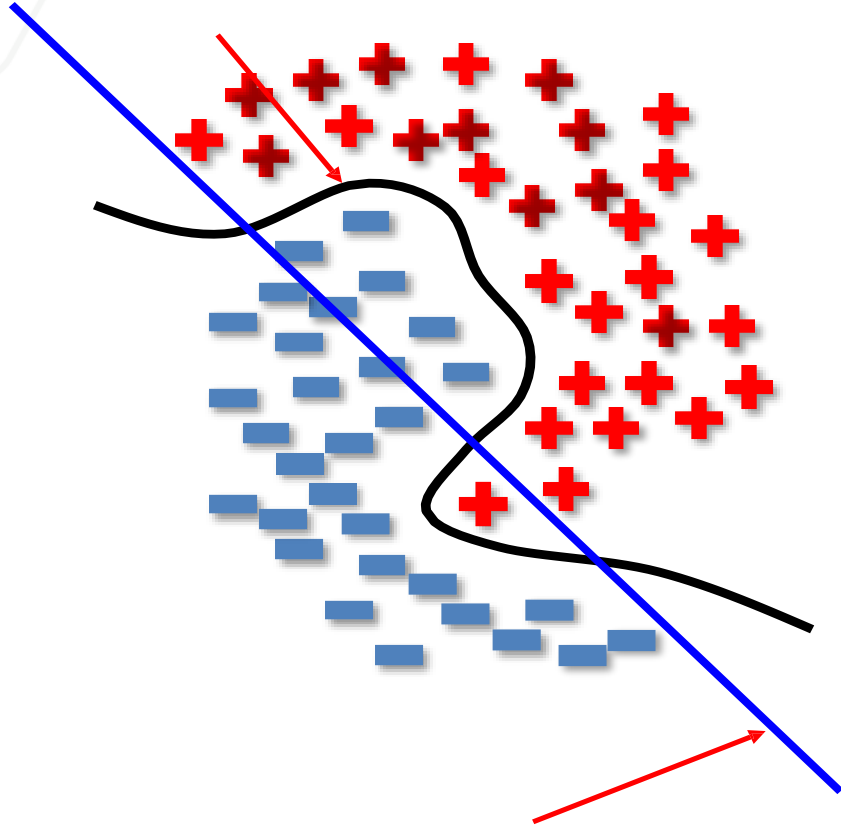
- What are the desired outcome?
- What are the input (data)?
- What are the learning paradigms?

“A real-world clinical validation for AI-based MRI monitoring in multiple sclerosis”  
Barnett et al., Nature 2023

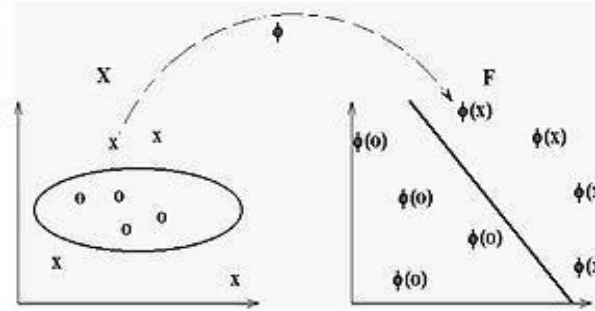


# Non-linear Classifier

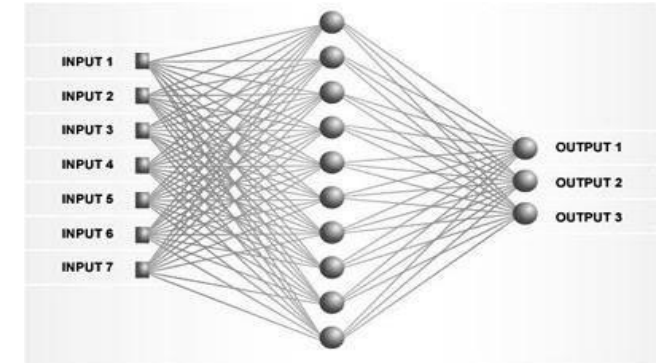
Non-linear decision boundary



Linear SVM decision boundary



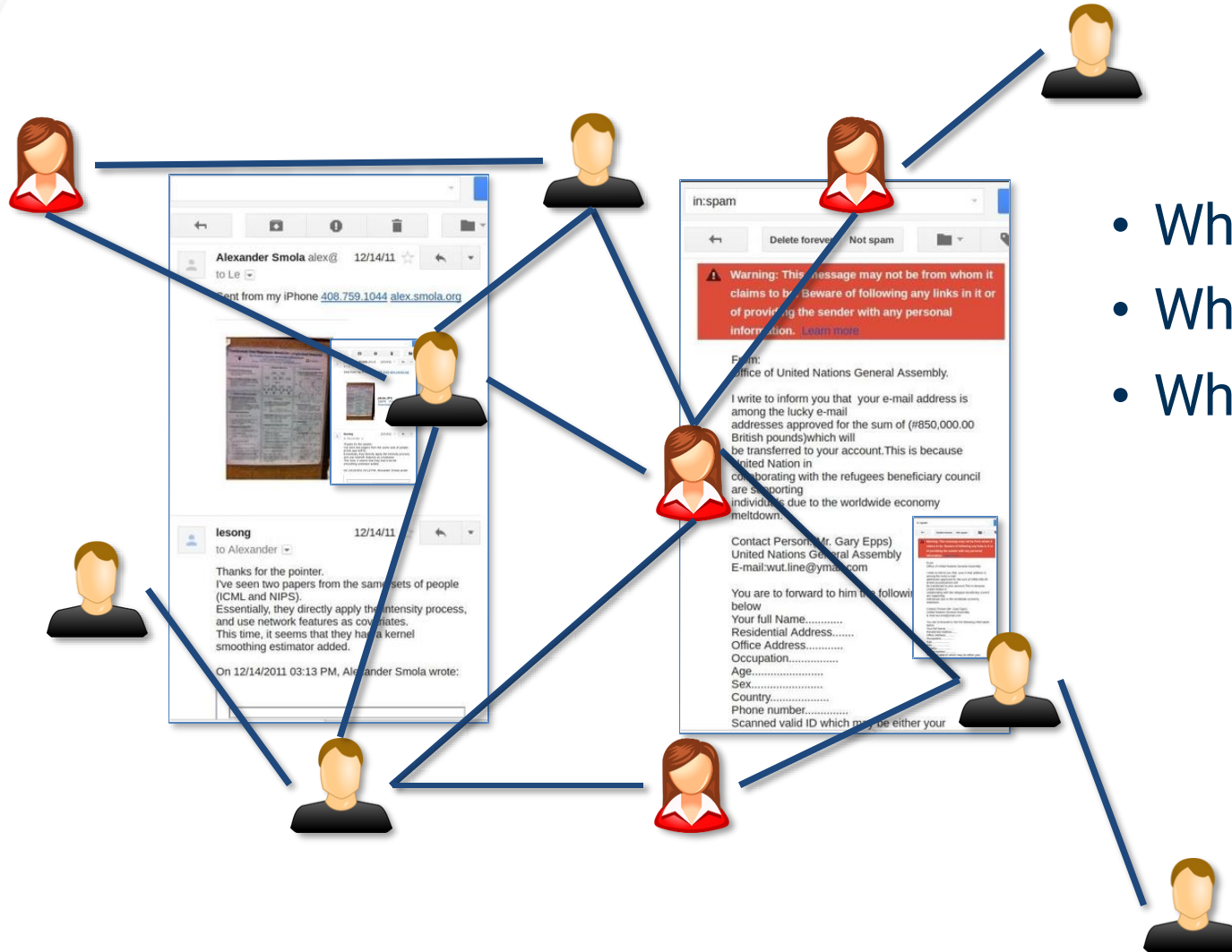
Kernel methods



Neural networks

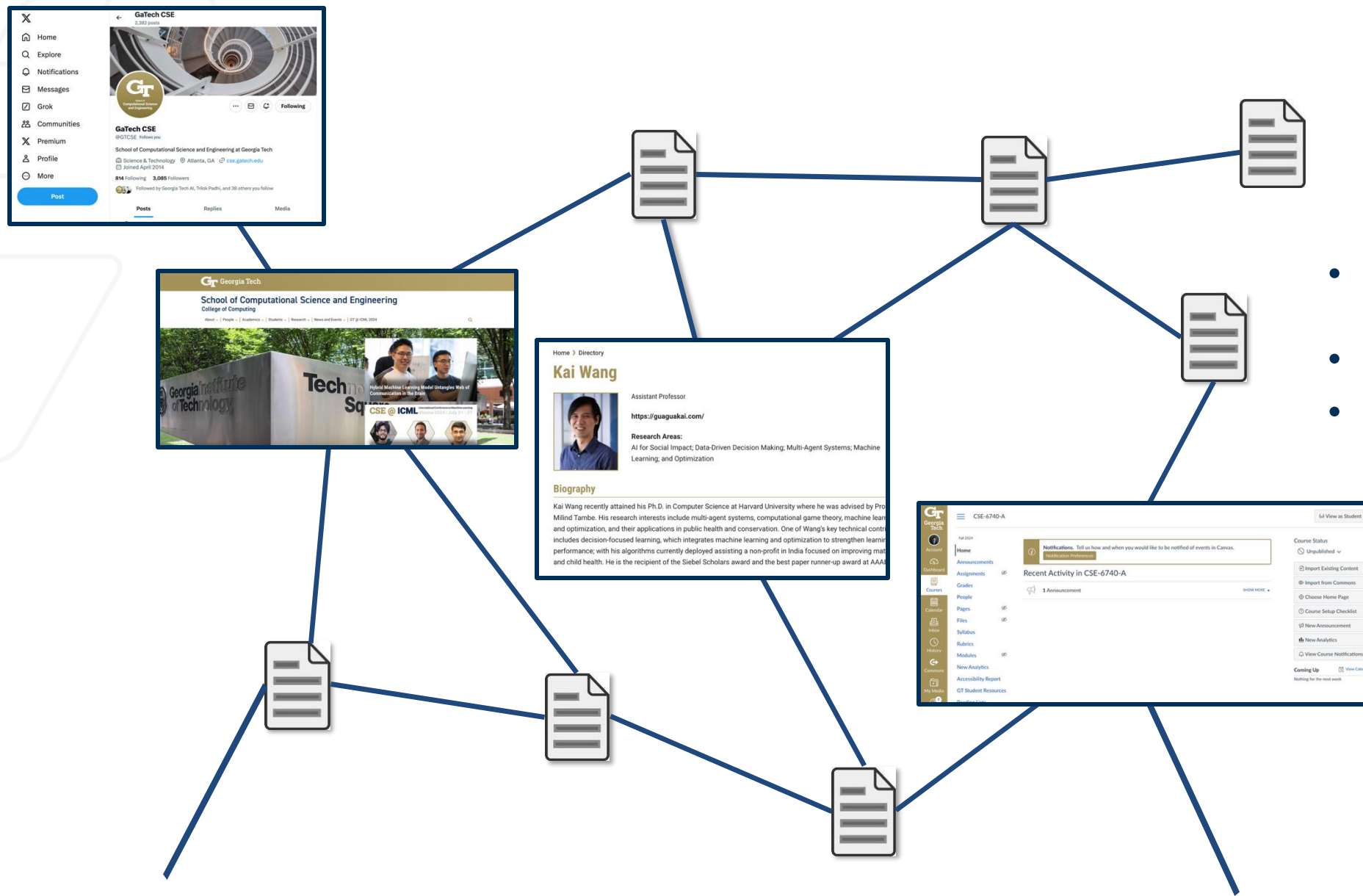
- What are the desired outcome?
- What are the input (data)?
- What are the learning paradigms?

# Spam Filtering



- What are the desired outcome?
- What are the input (data)?
- What are the learning paradigms?

# Webpage Classification



- What are the desired outcome?
- What are the input (data)?
- What are the learning paradigms?

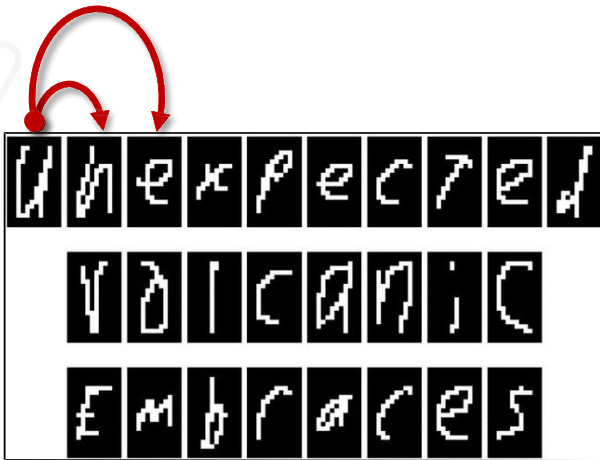
# Syllabus: Advanced Topics

- Advanced neural networks (week 10-12)
  - Convolutional neural networks
  - Deep learning
  - Autoencoders, variational autoencoders
  - Diffusion models
- Reinforcement learning (week 13-14)
  - Dynamic programming
  - Bellman equation
  - Q-learning and SARSA
  - Deep RL

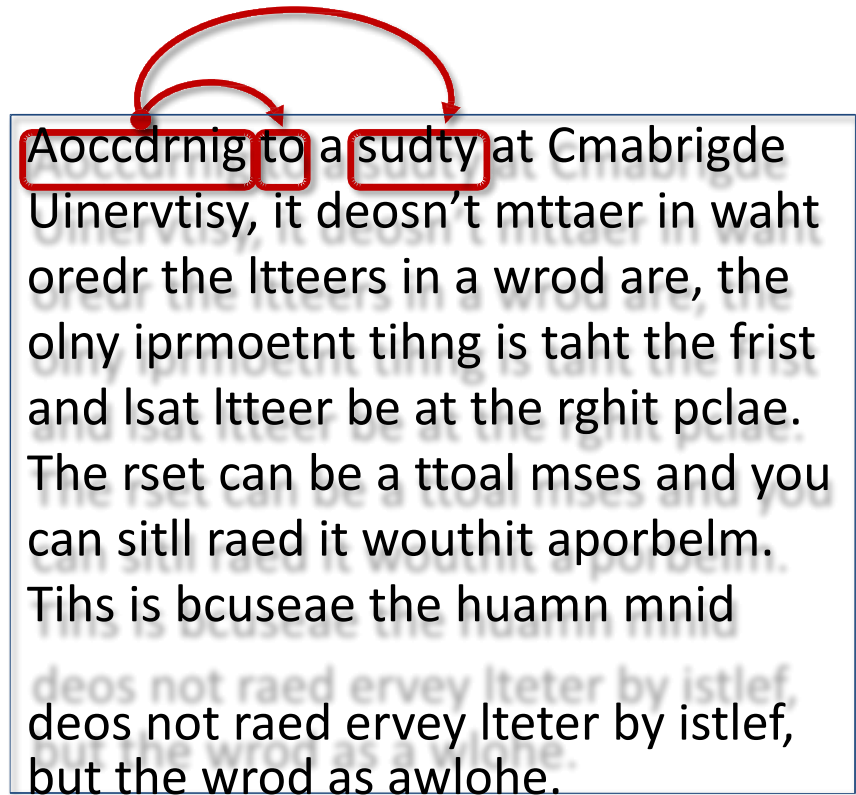


# Handwritten Digit Recognition / Text Annotation

Inter-character dependency

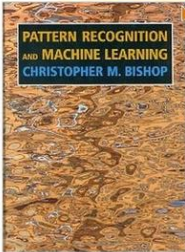


Inter-word dependency



# Product Recommendation

Click to **LOOK INSIDE!**



Share your own customer images

Search inside this book

Tell the Publisher!

I'd like to read this book on Kindle

Don't have a Kindle? [Get your Kindle here](#), or download a **FREE Kindle Reading App**.

Pattern Recognition and Machine Learning (Information Science and Statistics) [Hardcover]

Christopher M. Bishop (Author)

★★★★☆ (69 customer reviews) | Like (74)

List Price: **\$94.95**

Price: **\$67.98** & this item ships for **FREE** with Super Saver Shipping. [Details](#)

You Save: **\$26.97 (28%)**

[Special Offers Available](#)

**In Stock.**

Ships from and sold by Amazon.com. Gift-wrap available.

**Want it delivered Monday, January 9?** Order it in the next **21 hours and 41 minutes**, and choose **One-Day Shipping** at checkout. [Details](#)

**42 new** from \$67.98    **23 used** from \$69.97

**FREE Two-Day Shipping for Students.** [Learn more](#)

Formats

	Amazon Price	New from	Used from
Hardcover	\$67.98	\$67.98	\$69.97

Book Trade-In

Sell Back Your Copy for \$56.97

Whether you buy it new on Amazon for \$67.98 or somewhere else, you can sell it back through our Book Trade-In Program at the current price of **\$56.97**.

New Price	\$67.98
Trade-In Price	\$56.97
Price after Trade-in	\$11.01

Quantity: 1

Add to Cart

or

Sign in to turn on 1-Click ordering, or

Add to Cart with FREE Two-Day Shipping

Amazon Prime Free Trial required. Sign up when you check out. [Learn More](#)

Add to Wish List

Sell Back Your Copy

For a \$56.97 Gift Card

Trade in

[Learn more](#)

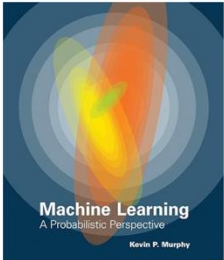
More Buying Choices

65 used & new from \$67.98

Have one to sell? [Sell on Amazon](#)


Share

Books › Science & Math › Mathematics



Roll over image to zoom in

Follow the author

 Kevin P. Murphy

Follow

Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning series) [Illustrated Edition]

by Kevin P. Murphy (Author)

4.4 ★★★★★ 341 ratings    4.4 on Goodreads 510 ratings    [See all formats and editions](#)

**A comprehensive introduction to machine learning that uses probabilistic models and inference as a unifying approach.**

Today's Web-enabled deluge of electronic data calls for automated methods of data analysis. Machine learning provides these, developing methods that can automatically detect patterns in data and then use the uncovered patterns to predict future data. This textbook offers a comprehensive and self-contained introduction to the field of machine learning, based on a unified, probabilistic approach.

The coverage combines breadth and depth, offering necessary background material on such topics as probability, optimization, and linear algebra as well as discussion of recent developments in the field, including conditional random fields, L1 regularization, and deep learning. The book is written in an informal, accessible style, complete with pseudo-code for the most important algorithms. All topics are cogently illustrated with color images and worked examples drawn from such application domains as biology, text processing, computer vision, and robotics. Rather than providing a cookbook of different heuristic methods, the book stresses a principled model-based approach, often using the language of graphical models.

[Read more](#)

[Report an issue with this product or seller](#)

ISBN-10	ISBN-13	Edition	Publisher	Publication date
0262018020	978-0262018029	# Illustrated	The MIT Press	August 24, 2012

Best of 2024 so far

[Learn more](#)

Kindle

\$68.99 (Earn 200 pts)

Available instantly

Hardcover

\$64.51 - \$74.50

Available instantly

Other Used and New from \$34.71

Delivery

Pickup

Buy new:

**-32%** **\$74.50**

List Price: \$110.00

[FREE Returns](#)

**FREE delivery August 27 - 29** for Prime members

Deliver to Kai - Atlanta 30309

Quantity: 1

Add to Cart

Buy Now

Ships from

Sold by

Returns

Gift wrap

[See more](#)

Amazon.com

Amazon.com

30-day refund/replacement

Available at checkout

## Frequently Bought Together

Price For All Three: **\$191.05**


[Add all three to Cart](#)    [Add all three to Wish List](#)

[Show availability and shipping details](#)


- ✓ **This item:** Pattern Recognition and Machine Learning (Information Science and Statistics) by Christopher M. Bishop Hardcover **\$67.98**
- ✓ The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics) by Trevor Hastie Hardcover **\$63.79**
- ✓ Machine Learning: An Algorithmic Perspective (Chapman & Hall/Crc Machine Learning & Pattern Recognition) by Stephen Marsland Hardcover **\$59.28**

## Customers Who Bought This Item Also Bought


Page 1 of 20




Machine Learning: An Algorithmic Perspective... by Stephen Marsland  
★★★★☆ (16)  
**\$59.28**




Probabilistic Graphical Models: Principles and T... by Daphne Koller  
★★★★☆ (8)  
**\$72.68**



Data Mining: Practical Machine Learning Tools an... by Ian H. Witten  
★★★★☆ (17)  
**\$44.07**



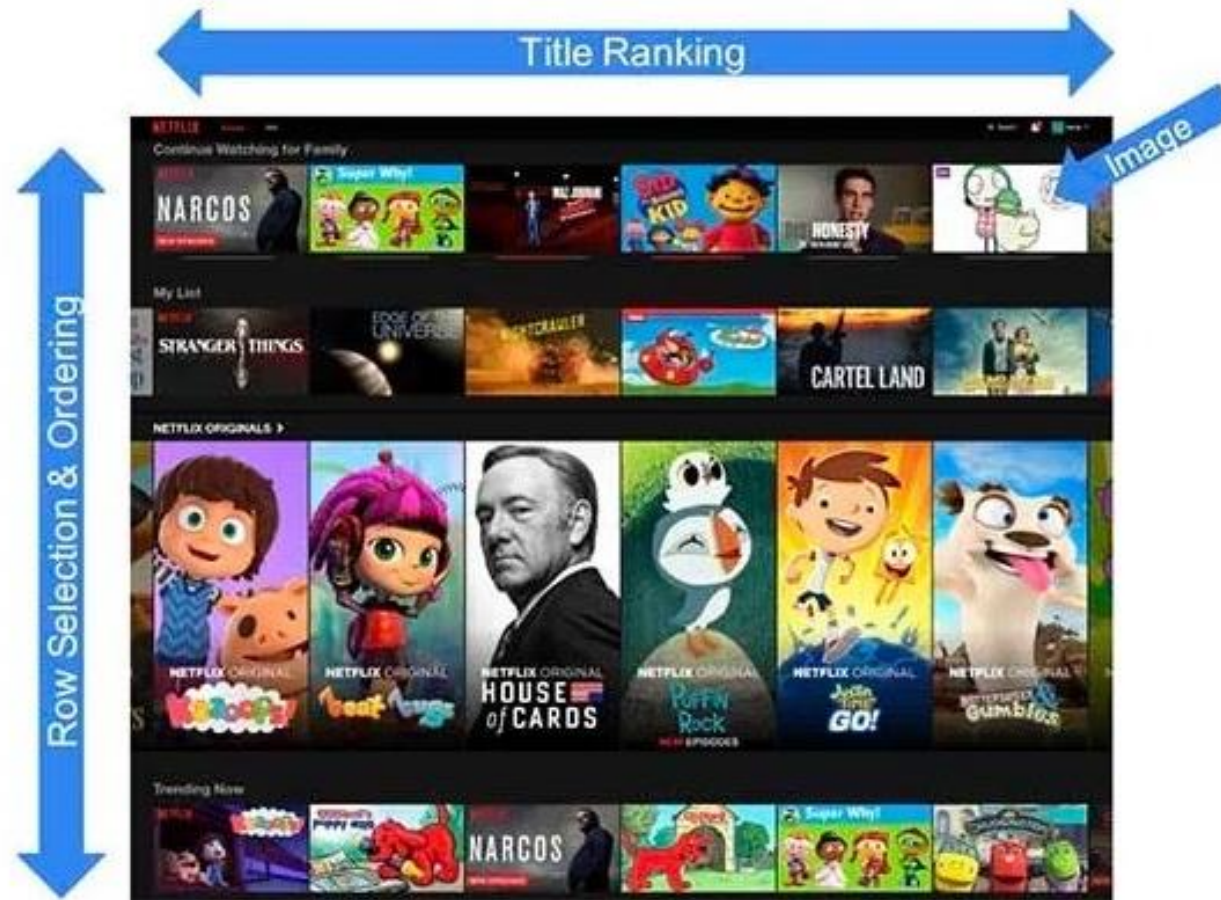
The Elements of Statistical Learning: Data Minin... by Trevor Hastie  
★★★★☆ (45)  
**\$63.79**



Pattern Classification (2nd Edition) by Richard O. Duda  
★★★★☆ (32)  
**\$91.41**


Georgia Tech

# Recommendation with Human Feedback



# Language Models

ChatGPT 4o ▾



Quiz me on ancient civilizations

Python script for daily email reports

Message to comfort a friend

Write a story in my favorite genre

ChatGPT is now available for macOS—Plus users get early access  
Get faster access to ChatGPT with the Option + Space shortcut and the floating companion window. [Learn more.](#) **Download** ×

Message ChatGPT

ChatGPT can make mistakes. Check important info.

Gemini ▾

Try Gemini Advanced



Hello, Kai

How can I help you today?

Find hotels in Recoleta in Buenos Aires, and things to do

Suggest a Python library to solve a problem

As a social trend expert, explain a term

Flights to Tokyo and Seoul, and things to do

Humans review some saved chats to improve Google AI. To stop this for future chats, turn off Gemini Apps Activity. If this setting is on, don't enter info you wouldn't want reviewed or used. [How it works](#)

[Manage Activity](#) [Dismiss](#)

Enter a prompt here



Gemini may display inaccurate info, including about people, so double-check its responses. [Your privacy & Gemini Apps](#)



# Image Generation



# AlphaGo





# Robotics



# Basics / Prerequisites



# Basics / Prerequisites

- **Probability**
  - Distributions, densities, marginalization, conditioning
- **Statistics**
  - Mean, variance, maximum likelihood estimation
- **Linear algebra**
  - Vector, matrix, multiplication, inversion, eigen-value decomposition
- **Algorithm, programming, and optimization**

# Homework 0

- We highly recommend you checking out **Homework 0** as soon as possible to see if you feel comfortable with the prerequisites!!
- Homework 0 is graded by SAT/UNSAT. As long as you submit, you get full credit. But please use this opportunity to reassess if you meet all the prerequisites of CSE6740.
- The amount of time spent on each homework (expected): ~20 hours
  - This varies a lot depending on your background. It can easily go over 40+ hours if you don't have the right prerequisites, and we also don't want you to feel overwhelmed by homework.

Question Text	N	RR	Interpol. Median	3	6	9	12	15	18	18+
Hours per week spent on course	63	53%	14.25	0	2	7	15	10	14	15

# CSE 8801 (1 credit)

- “CSE 8801 Linear Algebra, Probability, Statistics” is a good refresher course for you!

**Sections Found**

Linearalgebra,probability,sta - 91883 - CSE 8801 - LAS

**Long Title:** Linearalgebra,probability,sta

**Associated Term:** Fall 2025

**Registration Dates:** Apr 14, 2025 to Aug 22, 2025

**Levels:** Graduate Semester, Undergraduate Semester

Georgia Tech-Atlanta \* Campus


Lecture\* Schedule Type

1.000 Credits

**Grade Basis:** ALP

[View Catalog Entry](#)

**Scheduled Meeting Times**

Type	Time	Days	Where	Date Range	Schedule Type	Instructors
Class	2:00 pm - 3:15 pm	TR	Guggenheim Aerospace 246	Aug 18, 2025 - Dec 11, 2025	Lecture*	Raphaël Pestourie (P) 

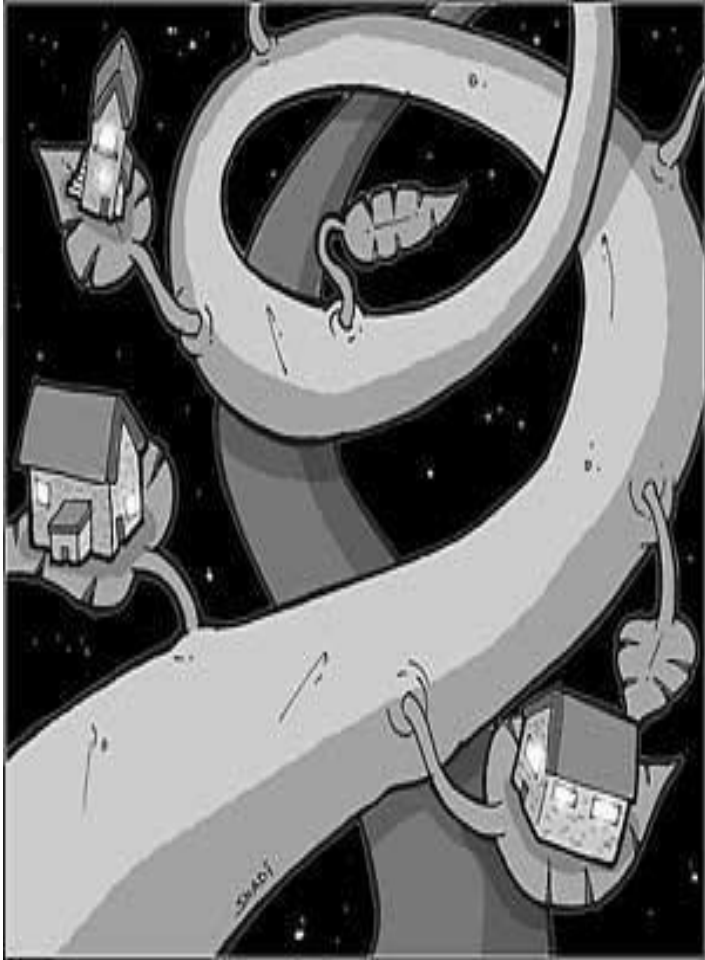


# Textbooks

- “[Pattern Recognition and Machine Learning](#)”, Christopher M. Bishop
- “[The Elements of Statistical Learning](#)”, Trevor Hastie, Robert Tibshirani, Jerome Friedman

# **Linear Regression (Prerequisite Example)**

# Machine Learning for Apartment Hunting



- Suppose you want to move to Atlanta, and you want to find the most **reasonably priced** apartment satisfying your **needs** (I know it is hard and probably too late):

Living area (ft <sup>2</sup> )	# bedroom	Monthly rent (\$)
230	1	900
506	2	1800
433	2	1500
190	1	800
...		
150	1	?
270	1.5	?

# Linear Regression Model

- Assume  $y$  is a linear function of  $x$  (features) plus noise  $\epsilon$

$$y = \theta_0 + \theta_1 x_1 + \cdots + \theta_d x_d + \epsilon$$

where  $\epsilon$  is an error modeled as Gaussian  $N(0, \sigma^2)$

Probability

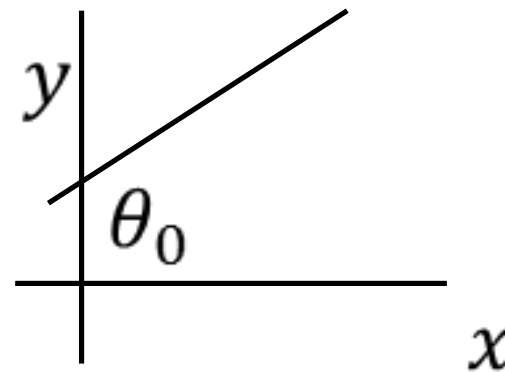
- Let  $\theta = (\theta_0, \theta_1, \dots, \theta_d)^\top \in \mathbb{R}^{d+1}$ , and augment data by one dimension

Linear algebra

$$x \leftarrow (1, x)^\top$$

Then  $y = \theta^\top x + \epsilon$

Linear algebra



# Least Mean Square (LMS) Method

- Given  $m$  data points, find  $\theta$  that minimizes the mean square error

$$\hat{\theta} = \operatorname{argmin}_{\theta} L(\theta) := \frac{1}{n} \sum_{i=1}^n (y^i - \theta^{\top} x^i)^2$$

Optimization

Statistics

- Set gradient to 0 to find the optimal parameter:

$$\frac{\partial L(\theta)}{\partial \theta} = -\frac{2}{n} \sum_{i=1}^n (y^i - \theta^{\top} x^i) x^i = 0$$

Calculus

$$\Leftrightarrow -\frac{2}{n} \sum_{i=1}^n y^i x^i + \frac{2}{n} \sum_{i=1}^n x^i x^{i\top} \theta = 0$$

Linear algebra



# Matrix Version of the Gradient

- Define  $X = (x^1, x^2, \dots, x^n) \in \mathbb{R}^{(d+1) \times n}$ ,  $y = (y^1, y^2, \dots, y^n)^\top \in \mathbb{R}^n$ , the gradient becomes

Linear algebra

$$\frac{\partial L(\theta)}{\partial \theta} = -\frac{2}{n} Xy + \frac{2}{n} XX^\top \theta = 0$$

Linear algebra

$$\Rightarrow \hat{\theta} = (XX^\top)^{-1} Xy$$

Algorithms  
Programming

- Matrix inversion in  $\hat{\theta} = (XX^\top)^{-1} Xy$  is often **expensive** to compute.
  - Alternatively, we can use gradient descent:

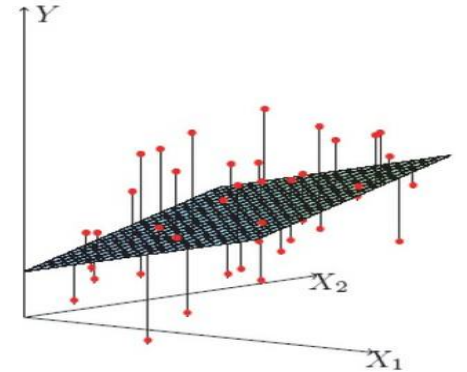
$$\hat{\theta}^{t+1} \leftarrow \hat{\theta}^t + \frac{\alpha}{n} \sum_{i=1}^n (y^i - \hat{\theta}^{t\top} x^i) x^i$$

Optimization

# Probabilistic Interpretation of LMS

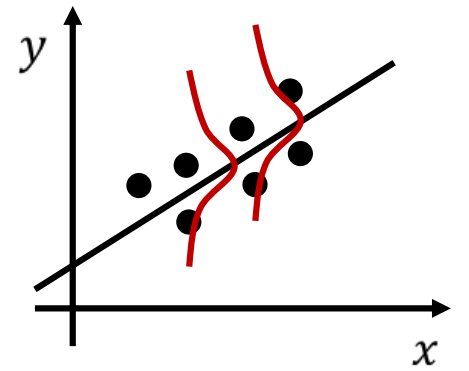
- Assume  $y$  is linear in  $x$  plus noise  $\epsilon$

$$y = \theta^\top x + \epsilon$$



- Assume  $\epsilon$  follows a Gaussian  $N(0, \sigma^2)$

$$p(y^i | x^i; \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^i - \theta^\top x^i)^2}{2\sigma^2}\right)$$



- By independence assumption, likelihood is:

$$L(\theta) = \prod_{i=1}^n p(y^i | x^i; \theta) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{\sum_{i=1}^n (y^i - \theta^\top x^i)^2}{2\sigma^2}\right)$$

Probability

# Probabilistic Interpretation of LMS

- Hence, the log-likelihood can be written as:

$$MLE: \log L(\theta) = n \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{i=1}^n (y^i - \theta^\top x^i)^2$$

- Therefore, LMS is equivalent to MLE of  $\theta$

Statistics

$$LMS: \frac{1}{n} \sum_{i=1}^n (y^i - \theta^\top x^i)^2$$

- How to make it work in real data?

Algorithms  
Programming

# Assignment and Grading

# Assignment

- There will be four homework (hw1 – hw4) + one homework letting you familiarize yourself with Gradescope (hw0).
- Each homework will come with both a written part and a programming part.
- **Deadline:** Two weeks after the release date (Sunday 11:59pm).
- **Coding language:** Python
- **Collaboration policy:** we think peer discussion is the key to learn and grow, and we strongly recommend discussing the course materials and homework with your peers. Please **mention your collaborator(s) (and on what topics)** in the beginning of your homework when submitting it.

However, all work you submit must be your own. You should never include in your assignment anything that was not written directly by you without proper citation (including quotation marks and LLM generated texts).

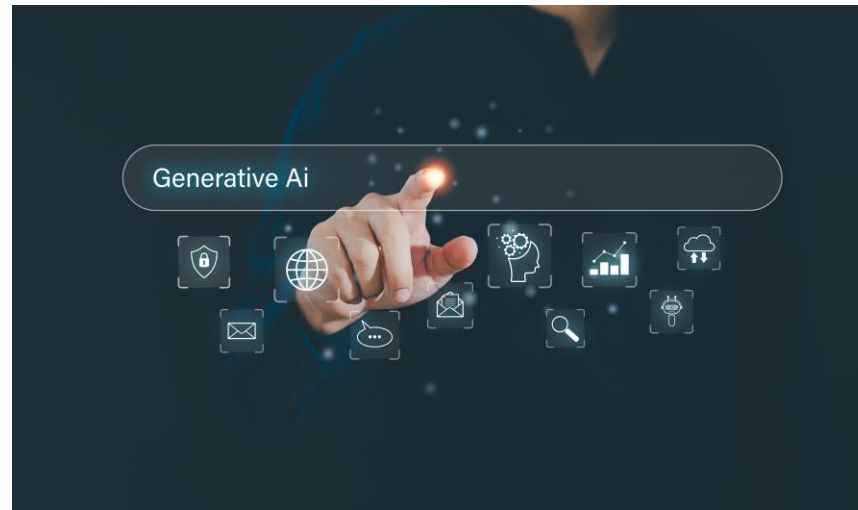
# Academic Integrity

- Any evidence of cheating or other violations will be referred to the Dean of Students with a recommendation that the penalty be an award of zero points for the graded requirement, and a one letter grade reduction in the course.
- Cheating includes, but is not limited to:
  - Using unauthorized references or notes
  - Copying directly from any source, including friends, classmates, tutors, or a solutions manual
  - Allowing another person to copy your work
  - Taking an exam or handing in a graded requirement in someone else's name, or having someone else take an exam or hand in a graded requirement in your name; or asking for a re-grade of a paper that has been altered from its original form.
- Any student suspected of cheating or plagiarizing on a quiz, exam, or assignment will be reported to the Office of Student Integrity, who will investigate the incident and identify the appropriate penalty for violations.



# The Use of Generative AI

- This course is about growing in your ability to write, communicate, and think critically. Generative AI agents such as ChatGPT, DALL-E 2, and others present great opportunities for learning and for communicating.
- However, AI cannot learn or communicate for you, and so cannot meet the course requirements for you.
- In this course, using generative AI tools in the work of the course (including assignments, discussions, ungraded work, etc.) is allowed only in instances specified by your instructor.



# Gradescope: Homework Submission

- Please use Gradescope to submit your homework.
  - You should submit both answers to the written and the programming questions in **one single PDF**.
  - You should also **submit your programming solution** in a single Jupyter notebook file.
  - We will use autograder to grade most of the programming questions. You should be able to see your scores on the public test cases.
- Homework late submission penalty
  - **20% penalty per late day**. Maximum 5 late days.
  - So please start working on your homework earlier!!

# Grading

Assignment	Weight (Percentage, points, etc)	Date
Homework	65% in total <ul style="list-style-type: none"><li>5% for hw0 (graded SAT/UNSAT)</li><li>15% for each hw1-hw4</li></ul>	See course schedule on Canvas
Midterm (in person)	15%	10/08/2025
Final (in person)	20%	12/05/2025

Extra credit opportunity	Weight (Percentage, points, etc)	Date
<b>LaTeX class notes (individual)</b>	3% (graded SAT/UNSAT) – first come first serve, notes in latex	Voluntary, starting from Week 2
<b>Advanced homework questions</b>	There will be additional challenging questions in the homework that you can solve and get bonus.	To be assigned in each homework

Please see the announcement on Piazza and Canvas for signing up for the note taking.

# Grading

- **Homework: 65%**
  - Homework 0:
    - 5%, graded SAT/UNSAT.
    - Familiarizing yourself with how to use Gradescope and some background knowledge!
  - Homework 1-4:
    - 15% each, due two weeks after the homework is released (no later than the end of Sunday). Due time is on **Sunday 11:59pm** (unless further notice).
- **Midterm** (in person on **Oct 8<sup>th</sup> 2025**, during regular class schedule): 15%
- **Final** (in person on **Dec 5<sup>th</sup> 2025, 11:20am – 2:10pm**): 20%
- **Bonus:**
  - LaTeX note taking: 3%
    - limited and first come first serve. See the head TA's announcement on Piazza.
    - Graded SAT/UNSAT
  - Homework bonus: TBD at each homework

# Final Grade

- Your final grade will be assigned as a letter grade according to the following scale. We will round it to the nearest integer.
  - A 90-100%
  - B 80-89%
  - C 70-79%
  - D 60-69%
  - F 0-59%

# Midterm and Final

- Both will be closed book exams.
- Midterm will be in class on **Oct 8<sup>th</sup> 2025 (12:30pm– 1:45pm)**
- Final exam: **Dec 5<sup>th</sup> 2025 (11:20am – 2:10pm)**
- Please plan your schedule earlier!



# Questions