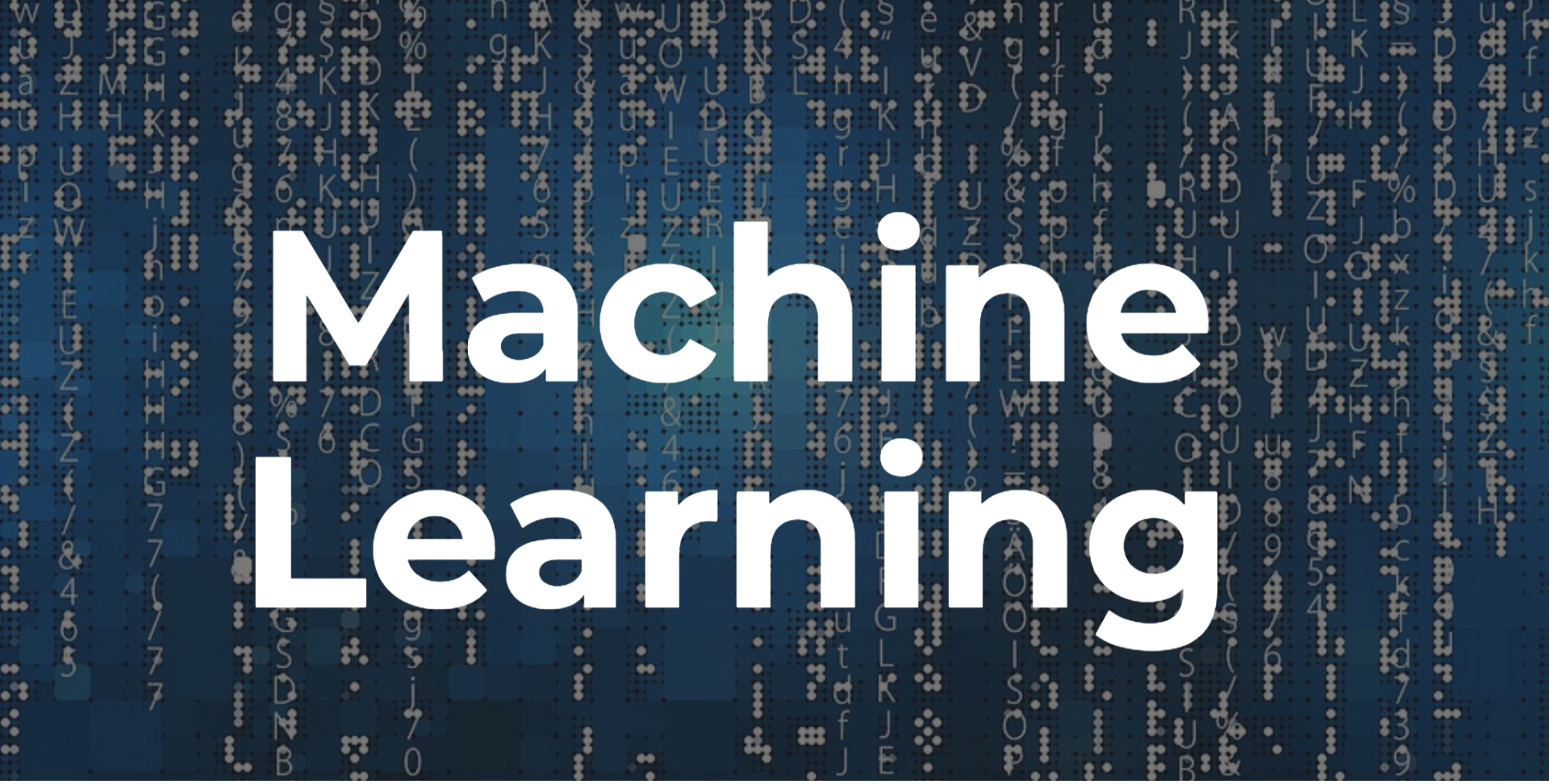
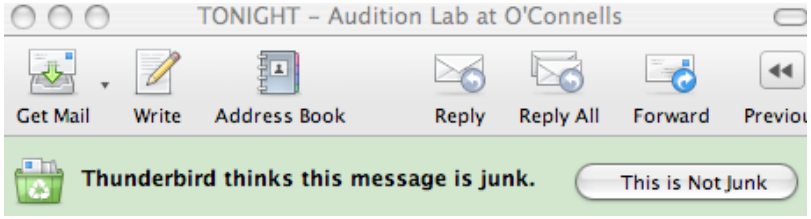


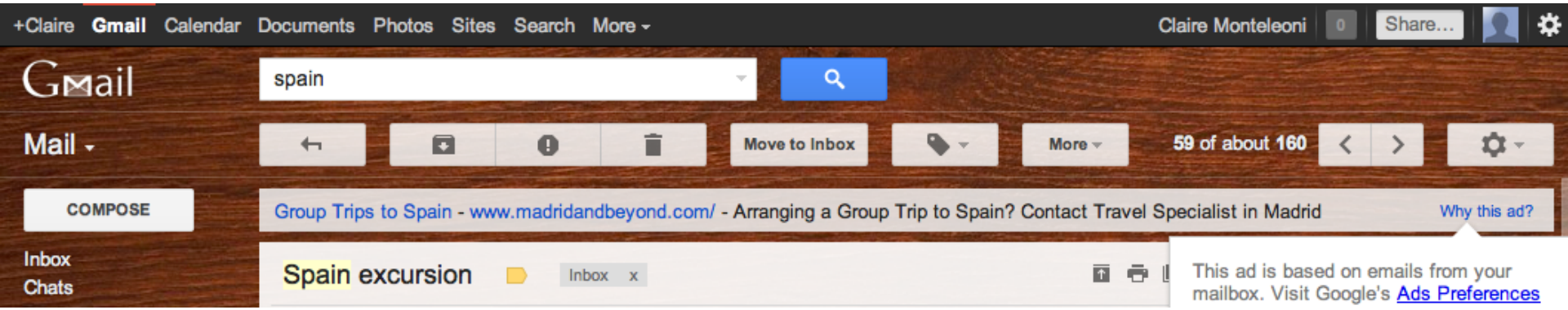
CSCI 4622 Fall 2019

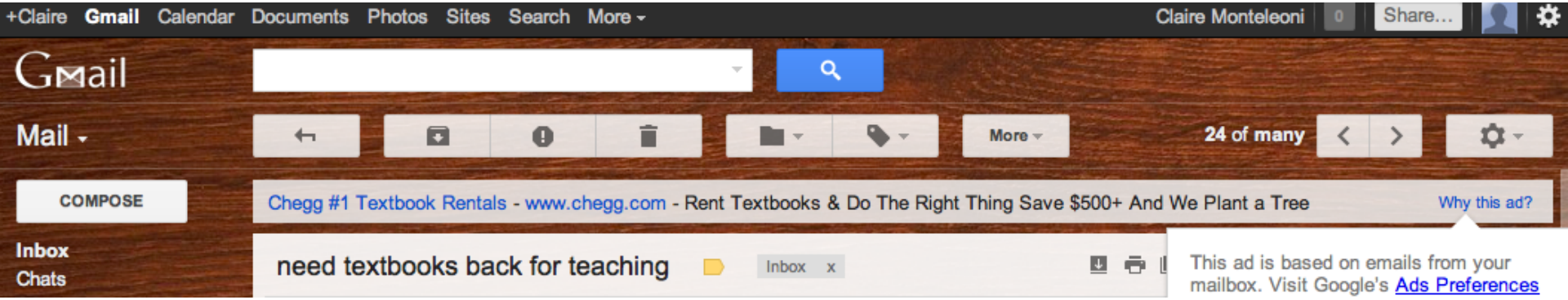
Prof. Claire Monteleoni

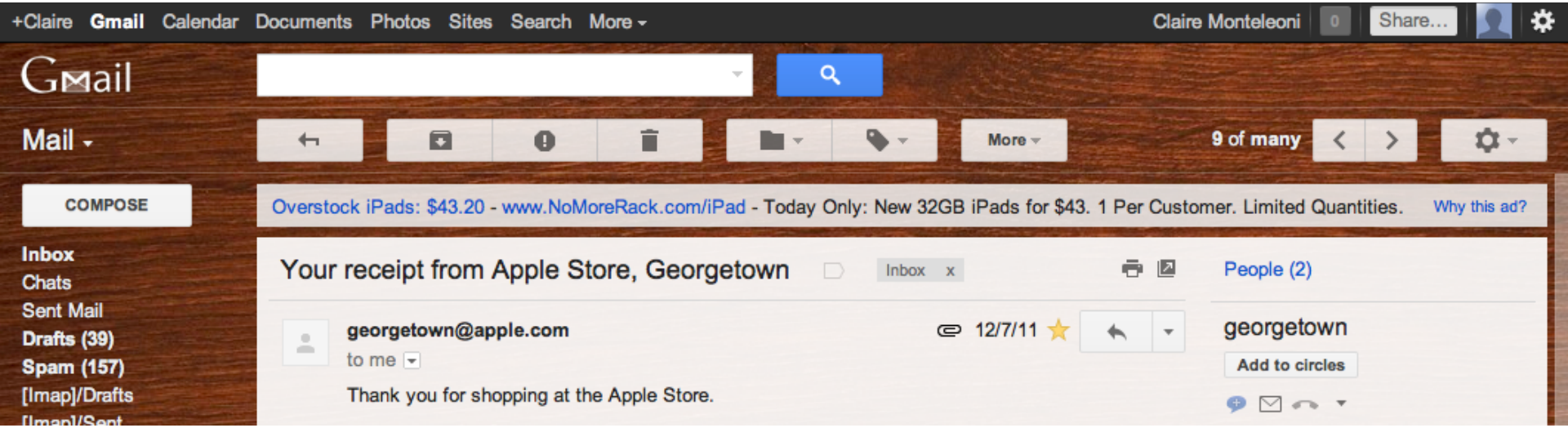
Machine learning is all around us…

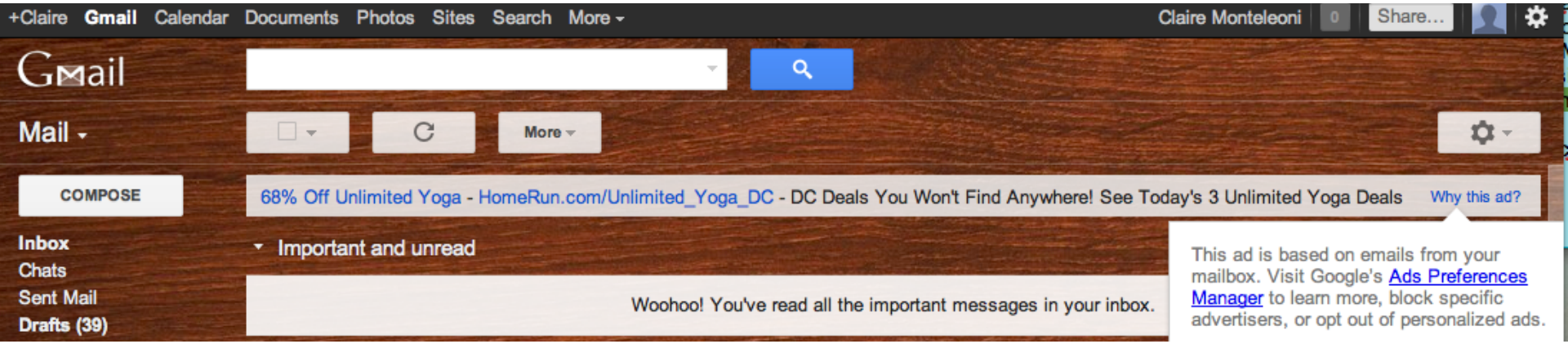
Machine learning is all around us…

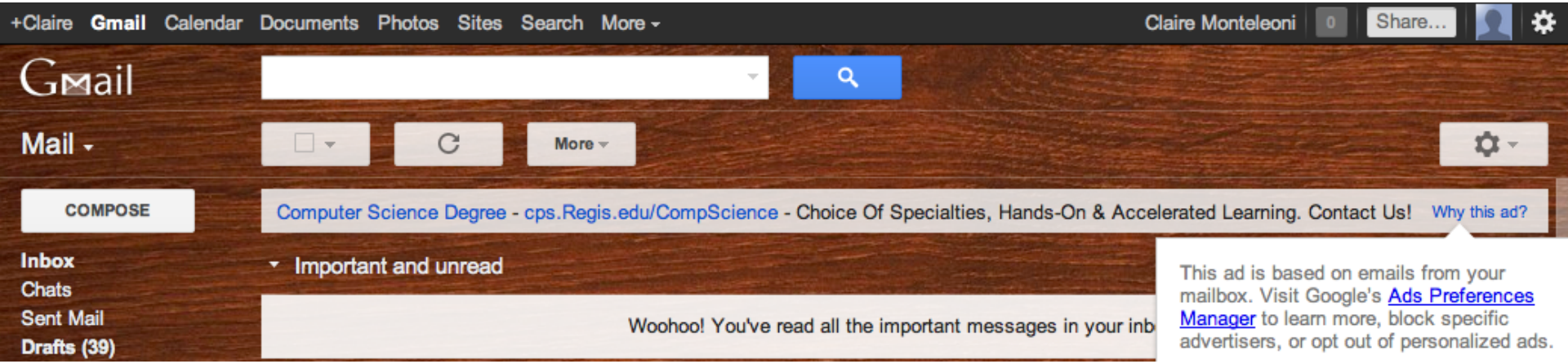



+Claire Gmail Calendar Documents Photos Sites Search More
Claire Monteleoni
Share...
Gmail
spain
Mail
Move to Inbox
More
59 of about 160
COMPOSE
Group Trips to Spain - www.madridandbeyond.com/ - Arranging a Group Trip to Spain? Contact Travel Specialist in Madrid
Why this ad?
Inbox
Chats
Spain excursion
Inbox x
This ad is based on emails from your mailbox. Visit Google's Ads Preferences


+Claire Gmail Calendar Documents Photos Sites Search More
Claire Monteleoni
Share...
Gmail
Mail
More
24 of many
COMPOSE
Chegg #1 Textbook Rentals - www.chegg.com - Rent Textbooks & Do The Right Thing Save $500+ And We Plant a Tree
Why this ad?
Inbox
Chats
need textbooks back for teaching
Inbox x
This ad is based on emails from your mailbox. Visit Google's Ads Preferences


+Claire Gmail Calendar Documents Photos Sites Search More
Claire Monteleoni
Share...
Gmail
Mail
More
9 of many
COMPOSE
Overstock iPads: $43.20 - www.NoMoreRack.com/iPad - Today Only: New 32GB iPads for $43. 1 Per Customer. Limited Quantities.
Why this ad?
Inbox
Chats
Sent Mail
Drafts (39)
Spam (157)
[Imap]/Drafts
Your receipt from Apple Store, Georgetown
Inbox x
People (2)
georgetown@apple.com
12/7/11
to me
Thank you for shopping at the Apple Store.
georgetown
Add to circles


+Claire Gmail Calendar Documents Photos Sites Search More
Claire Monteleoni 0 Share...
Gmail
Mail
More
COMPOSE
68% Off Unlimited Yoga - HomeRun.com/Unlimited_Yoga_DC - DC Deals You Won't Find Anywhere! See Today's 3 Unlimited Yoga Deals
Why this ad?
Inbox
Chats
Sent Mail
Drafts (39)
Important and unread
Woohoo! You've read all the important messages in your inbox.
This ad is based on emails from your mailbox. Visit Google's Ads Preferences Manager to learn more, block specific advertisers, or opt out of personalized ads.


+Claire Gmail Calendar Documents Photos Sites Search More
Claire Monteleoni 0 Share...
Gmail
Mail
More
COMPOSE
Computer Science Degree - cps.Regis.edu/CompScience - Choice Of Specialties, Hands-On & Accelerated Learning. Contact Us!
Why this ad?
Inbox
Chats
Sent Mail
Drafts (39)
Important and unread
Woohoo! You've read all the important messages in your inb
This ad is based on emails from your mailbox. Visit Google's Ads Preferences Manager to learn more, block specific advertisers, or opt out of personalized ads.

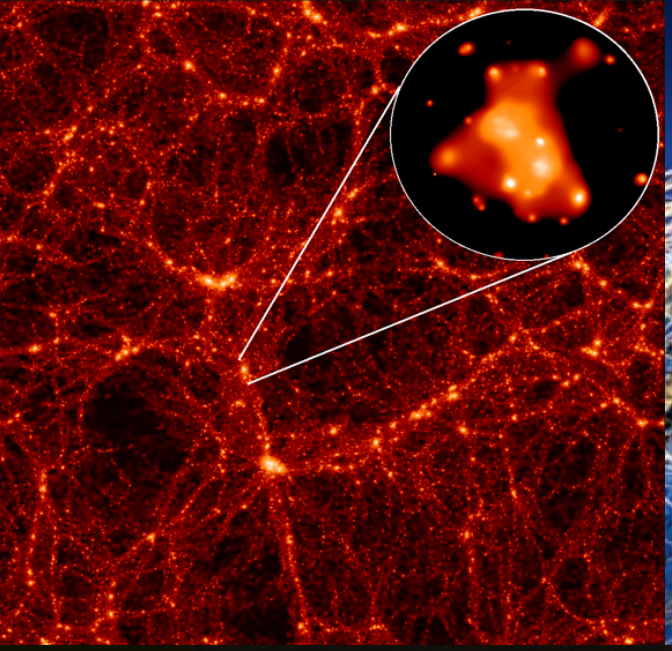
Machine learning

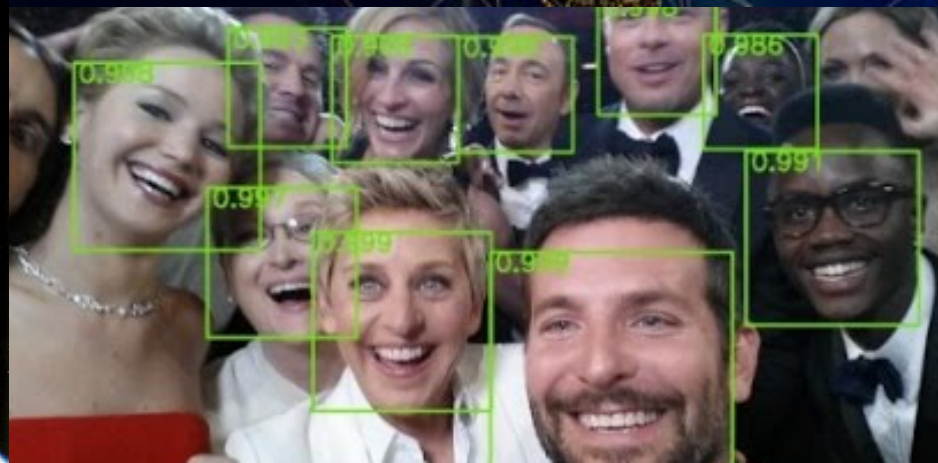
Machine learning is a cutting-edge technology with wide applications, e.g.:

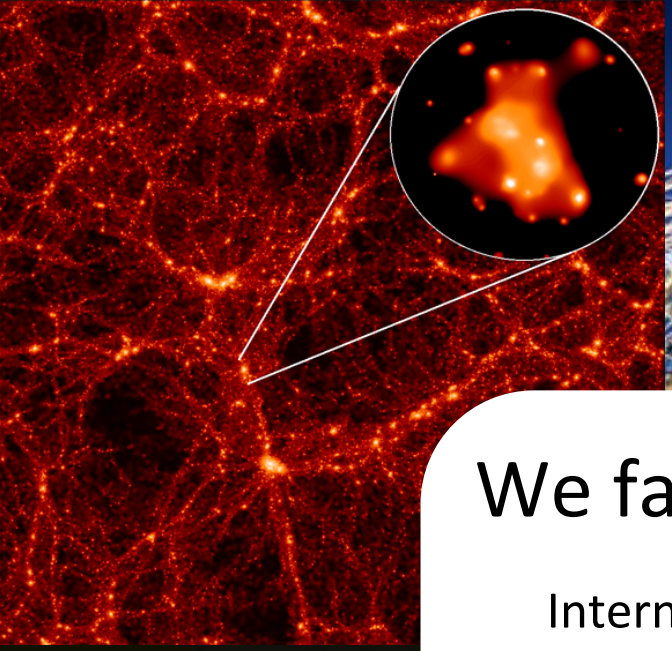
- Web search: Google, Yahoo!, Bing, etc.
- Recommendation systems for books/movies/music, e.g. Amazon, Netflix, Pandora, iTunes
- Personalized internet advertising, e.g. Facebook, Gmail
- Spam filtering
- Autopilot in planes, cruise control in cars
- Finance: Algorithmic trading
- Biology: Bioinformatics
- Chemistry: Computational chemistry
- Big Data – applications in any field

Take-home messages of course

- Machine learning / AI is not **magic**!
- Yes, ML/AI can be applied to a range of applications!
- **Our goal**: gain conceptual understanding of machine learning algorithms and how to apply them







We face an explosion of data!

Internet transactions, social networks

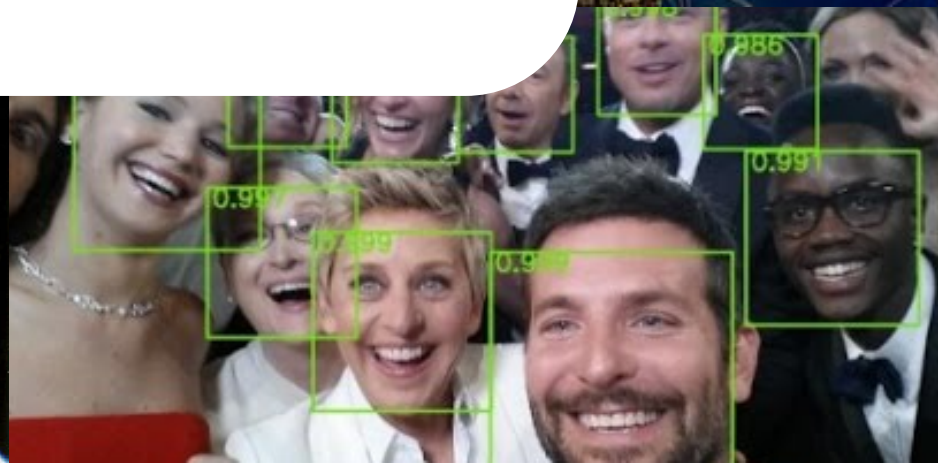
DNA sequences

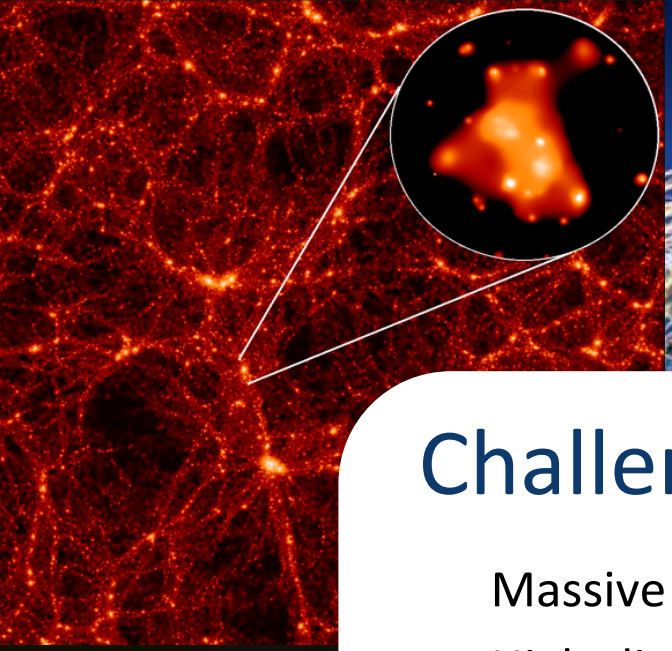
Financial indicators

Satellite imagery

Environmental sensors

….





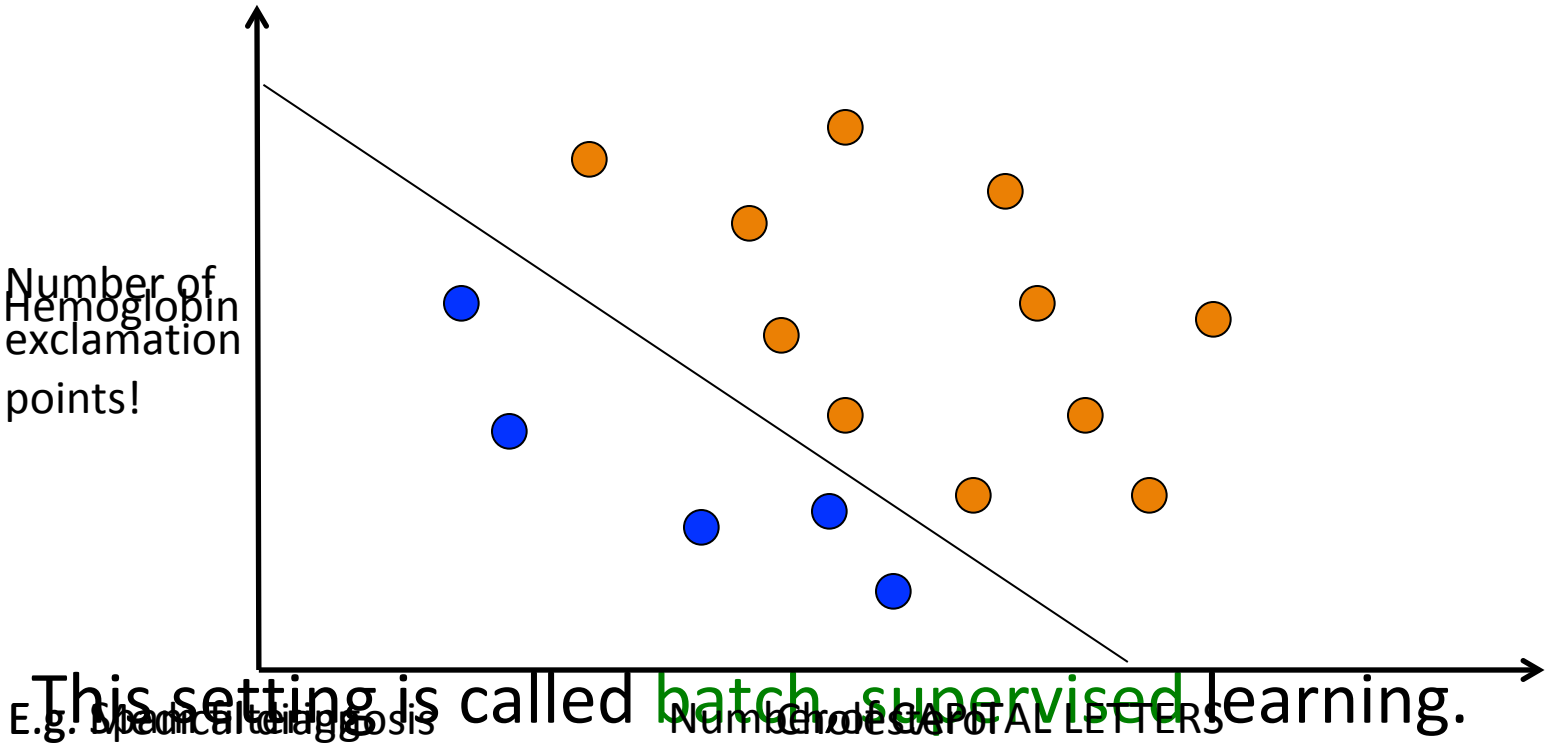
Challenges of real data

Massive	many data points
High-dimensional	many features
Noisy, raw	missing data, unlabeled
Sparse	low-dimensional structure
Streaming	time-varying
Sensitive/private	financial, health, social

Machine Learning 101

Given labeled data points, find a good classification rule.

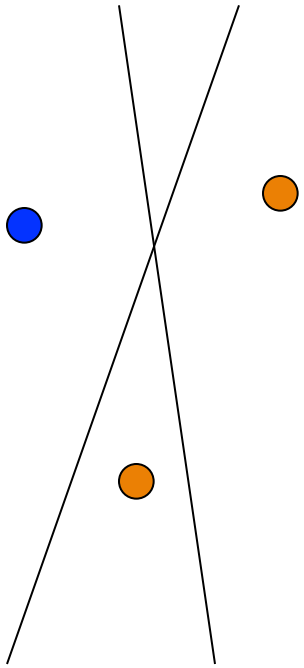
- Describes the data
- Generalizes well



This setting is called batch supervised learning.

Learning from data streams

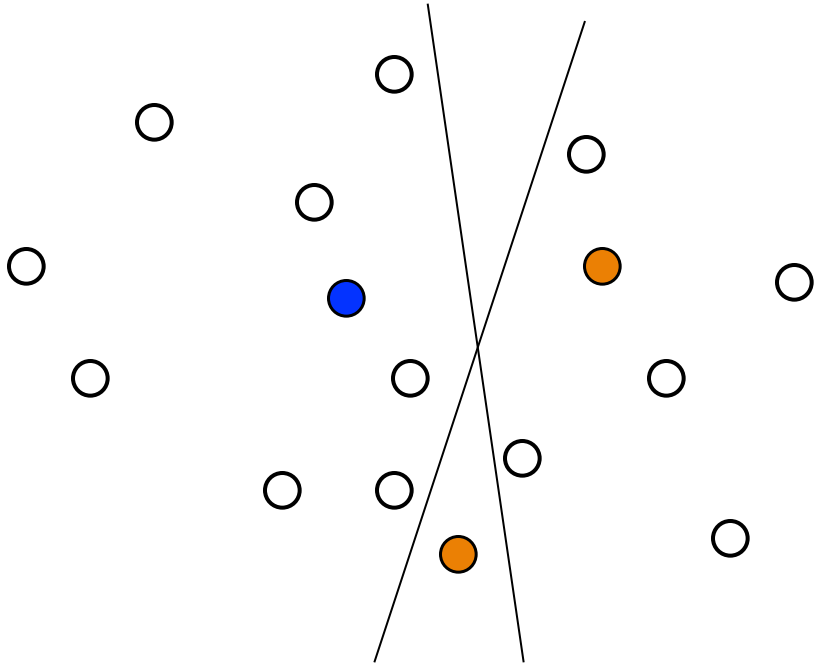
Data arrives in a stream over time.



Learning from raw data

Semi-supervised learning:

Learning from labeled and unlabeled data

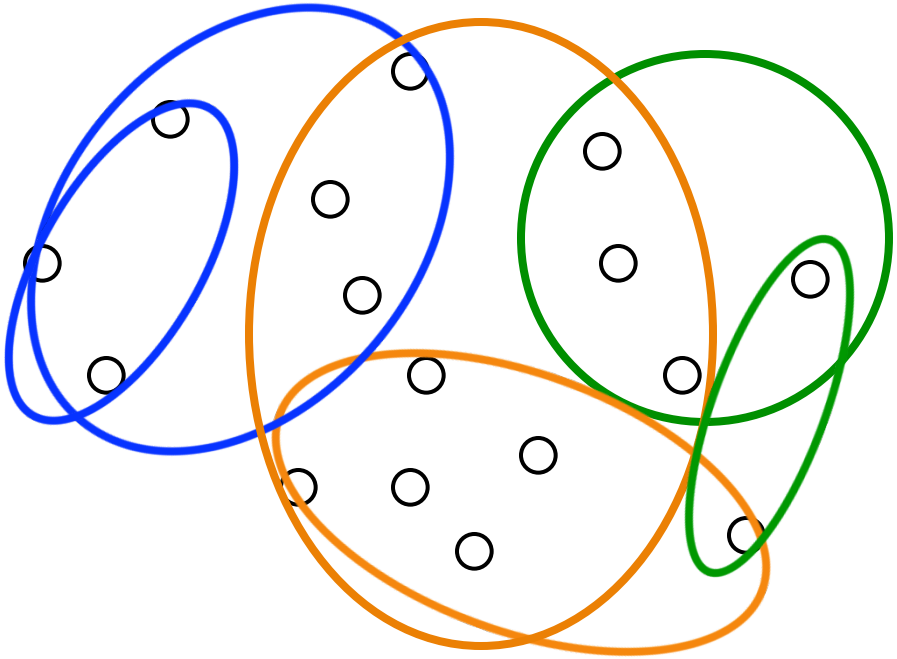


E.g. Active Learning: learner interacts with teacher to query labels.

Learning from raw data

What can be done without any labels?

Unsupervised learning



E.g. Clustering

Topics this semester

Supervised learning

- Introduction to supervised learning (e.g. k-NN, perceptron)
- Generative learning (e.g. HMM)
- Discriminative learning (e.g. SVM, neural networks)
- Regression

Intro. to learning theory

- Bias-variance tradeoff
- VC dimension

Unsupervised learning

- Unsupervised deep learning
- Dimensionality reduction / embedding
- Clustering

Other learning settings

- Online learning
- Active learning

Course information

Schedule (linked from both Canvas and Piazza)
sites.google.com/colorado.edu/csci4622fall19/schedule

- Please keep up with the readings on the schedule above.
- Please follow Piazza.com for important announcements
- Piazza is the preferred mode of communication in this course

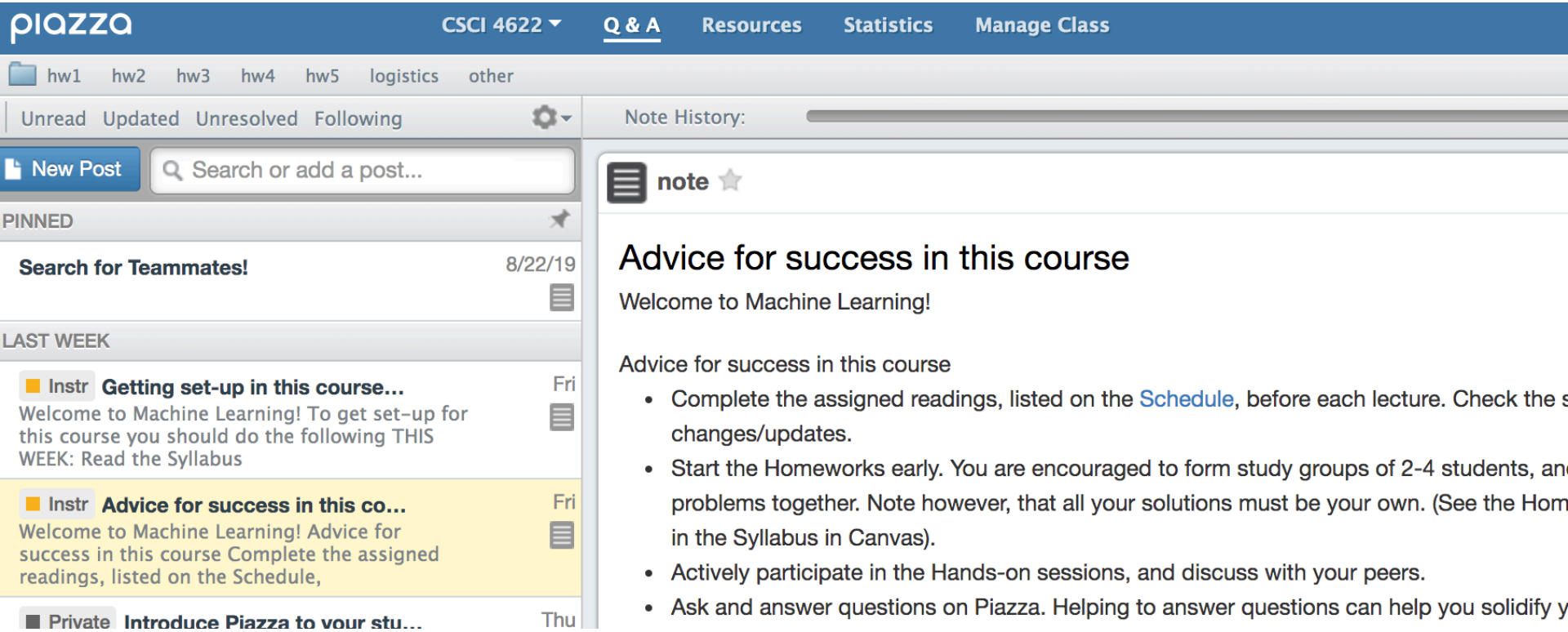
TA: Chris Bate

Graduate student staff: Bhalachandra Naik, Aman Satya

My office hours: Tuesdays 2:30-30:30pm, ECES 132

Piazza

piazza.com/colorado/fall2019/csci4622/home
(linked from Canvas)



Course Schedule

sites.google.com/colorado.edu/csci4622fall19/schedule
(linked from both Canvas and Piazza)

CSCI 4622: Machine Learning: Fall 2019

Home

Schedule

Undergrad Machine Learning Fall 2019: Schedule : Sheet1

Week	Date	Lecture	Reading	Assignments (due 5pm)
Week 1	8/26	Introduction, collaborative filtering, nearest neighbor (NN)	[CIML Chapters 1, 3] [AIMA 18.1, 18.2; Mitchell: 8.1, 8.2] [D1]	Intro questionnaire posted.
	8/28	Intro Quiz (on prerequisites: analysis of algorithms, probability, linear algebra). Introduction to linear classification.	[CIML Chapter 4, 6] [Duda, Hart, and Stork: 5.4, 5.5] [D3, J1, J2]	
	8/30	Hands-on 1: Computing Environment Orientation, k-NN [Chris Bate]		**Intro questionnaire due.**
Week 2	9/4	Perceptron, convergence.	[CIML Chapter 4] [HTF 4.5] [Mitchell: Chapter 3; AIMA: 18.3]	
	9/6			HW 1 posted (on Weeks 1-2)
Week 3	9/9	Decision trees, k-d trees. Overfitting. Bias-variance tradeoff	[CIML Chapters 1, 2][Dasgupta course: Decision Trees] [Description of CART, a canonical DT alg.] [HTF 9.2, Chapters 2, 7]	
	9/11	Bias-variance tradeoff. Model evaluation, cross-validation.	[CIML Chapter 2] [HTF Chapter 7][CIML Chapter 16]	
	9/13	Hands-on 2: Perceptron, Validation.		

Grading

- 10% Class participation, including on Piazza

- 45% Homeworks
  - 6 assignments: approximately every other week
  - 9% each: your best-scoring 5 out of 6 assignments

- 45% Exams
  - 3 exams (including Final Exam)
  - 25% your top scoring exam, 20% your 2nd-top scoring exam
  - Your lowest exam score is dropped
  - Therefore Final Exam is OPTIONAL

Intro Quiz

- On Wednesday (8/28), class will start with a quiz on the prerequisites:
  - Analysis of Algorithms (CSCI 3104 or equivalent)
  - Probability (CSCI 3022 or equivalent)
  - Linear Algebra (CSCI 2820 or equivalent)

- This will **not** count towards your grade. It is meant to help you decide if you have the preparation.

- If you have taken the prerequisites, the quiz will be straightforward.

- The actual content in this course will be more advanced.

Other set-up tasks

- Set-up a free account on www.Piazza.com and enroll in the course. All announcements will be there!
- Intro Questionnaire: please fill this out within the week of the semester (linked from Piazza and Canvas).
- Make sure you have access to Python and Jupyter notebooks, and brush up on basic programming.
- Look at the list of recommended textbooks and make sure you have access. Some are free online, others in the library.
- Do the readings listed in the schedule, before each lecture.
- Form a study group (2-4 students).

Electronic device policy

- No cell phone / smartphone use (please turn it off before entering the classroom)
- Tablets may be used when kept flat on the desk
- Laptops or tablets with vertical or slanted screens are restricted to the back 3 rows of the lecture hall, to avoid distracting other students. NOTE: This was a specific student request to CU CEAS faculty.
- When in doubt, if your screen could be viewed by other students, and cause a distraction, then please sit towards the back of the room.
- Exceptions:
  - **Exams**: no electronic devices allowed
  - **Hands-on days**: please bring your laptop!

Homework collaboration policy

Collaboration is allowed, and encouraged, to the following extent:

- you are welcome to discuss problems (both written and programming) with each other and to take your own notes during these discussions.
- However, you must write the solutions and implement code on your own;
- **Copying is not permitted**.
- You must write, on the assignment, the names of students you discussed each problem with, and any external sources you used in solving the problem.

Lateness policy

- Assignments are due Friday at 5pm.
- Late submissions received from 5:01PM Friday until 11:59PM Sunday night will be accepted, with a deduction of 25% credit.
- After that, late submissions will not be accepted. Recall: your lowest HW score will be dropped.

Advice for success in this course

- Complete the assigned readings, listed on the Schedule, before each lecture. Check the schedule often for changes/updates.

- Start the homeworks early. You are encouraged to form study groups of 2-4 students, and discuss the homework problems together. Note however, that all your solutions must be your own. (See the Homework Collaboration Policy).

- Ask and answer questions on Piazza. Helping to answer questions can help you solidify your own understanding.

Collaborative Filtering and Nearest Neighbor

amazon.com

Hello, Claire Monteleoni. We have recommendations for you. (Not Claire?)

Claire's Amazon.com | Today's Deals | Gifts & Wish Lists | Gift Cards

Jade Yoga Elite Yoga Mat

by Jade Yoga

(1 customer review) | Like (0)

Customers Who Bought This Item Also Bought

Aurorae Organic Yoga Mat Wash-Cleans, Restores, Refres... by Aurorae

(27)

$11.95

Jade Harmony Professional 3 / 16-Inch Yoga Mat by Jade Yoga

(162)

$62.95 - $105.00

Jade Fusion 74-inch Yoga Mat by Jade Yoga

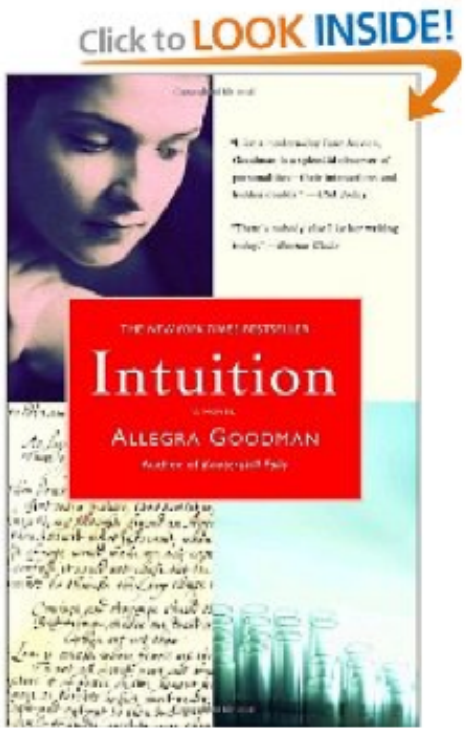
(36)

$116.78 - $169.95

Yoga Mat Bag Extra Large Easy Open Zipper 100% C... by Bean Products

(44)

$18.95 - $21.95



Intuition [Paperback]

Allegra Goodman (Author)

★★★★☆ (69 customer reviews) | Like (4)

This intimate portrait of life in a research institute in Cambridge, Massachusetts, revolves around a scientific mystery….

Customers Who Bought This Item Also Bought

Kaaterskill Falls by Allegra Goodman

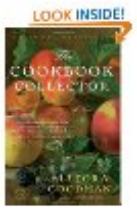
★★★½☆ (62)

$10.40

Paradise Park by Allegra Goodman

★★★☆☆ (36)

$11.12



The Cookbook Collector: A Novel by Allegra Goodman

★★★☆☆ (135)

$10.20

Statistics for the Life Sciences (4th Edition) by Myra L. Samuels

★★★☆☆ (20)

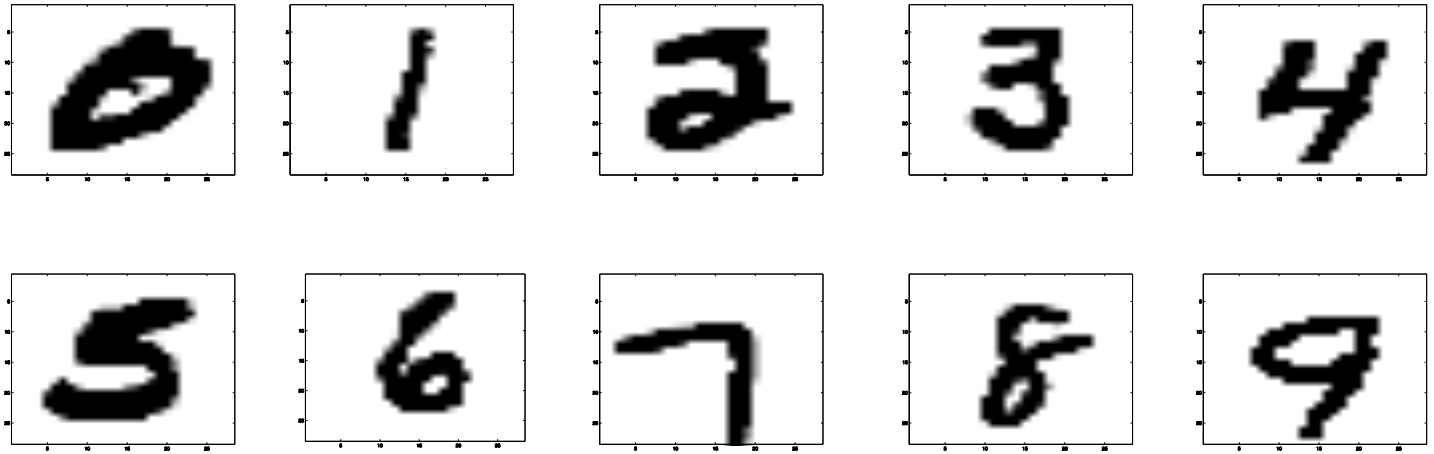
$113.99

Nearest Neighbor Classification

[With much credit to S. Dasgupta]

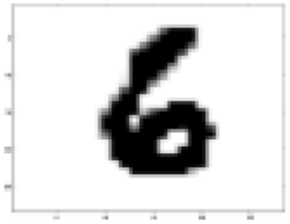
A first example

Database of 20,000 images of handwritten digits, each labeled by a human



[28 x 28 greyscale; pixel values 0-255; labels 0-9]

Use these to learn a classifier which will label digit-images automatically…



Data representation

Image: size 28 x 28 = 784 pixels, each pixel is grayscale, in 0-255.

- Unfold each image (28 x 28 pixels) into a vector of length 784.

- Now we have vectors in 784-dimensional Euclidean space:

Data space X = $\{0, 1, \ldots, 255\}^{784} \subset \mathbb{R}^{784}$

- Each image comes with a label, which is a digit.

Labels are in the set Y = {0,1, …, 9}

Training data set

- Each image comes with a label, which is a digit.

  Labels are in the set Y = {0,1, …, 9}

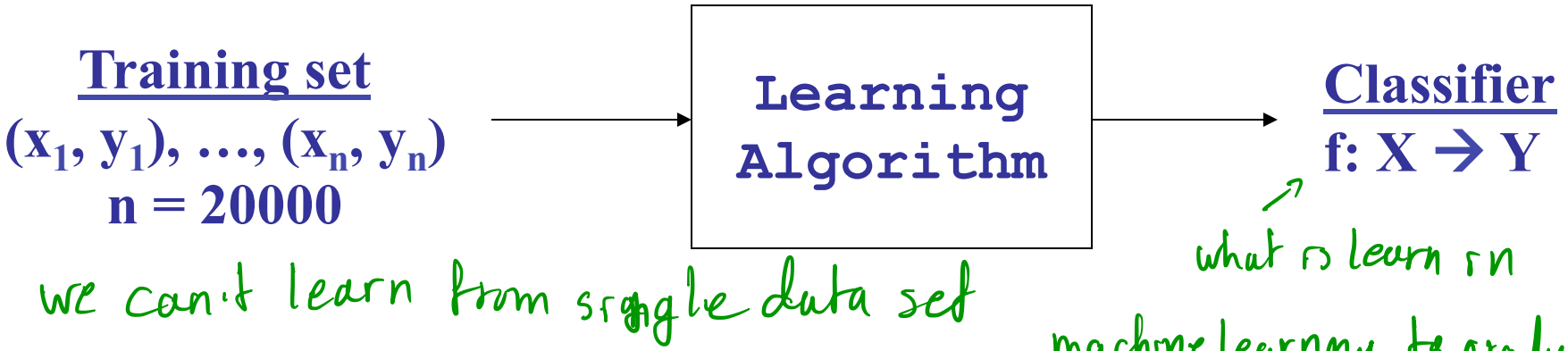
The learning problem

Input space $X = \{0,1,\ldots,255\}^{784}$

Output space $Y = \{0,1,\ldots,9\}$

scalar

f is function for training set



what is learn in machine learning to produced output digit (0 → 9)

we can't learn from single data set

To measure how good f is: use a test set.

We have a test set with 100 instances of each digit.

A possible strategy

Input space $X = \{0,1,\ldots,255\}^{784}$

Output space $Y = \{0,1,\ldots,9\}$

Treat each image as a point in 784-dimensional Euclidean space

To classify a new (unlabeled) image:
find its nearest neighbor in the database (training set) and return that label

f = search engine + entire training set

Euclidean distance in high dimension

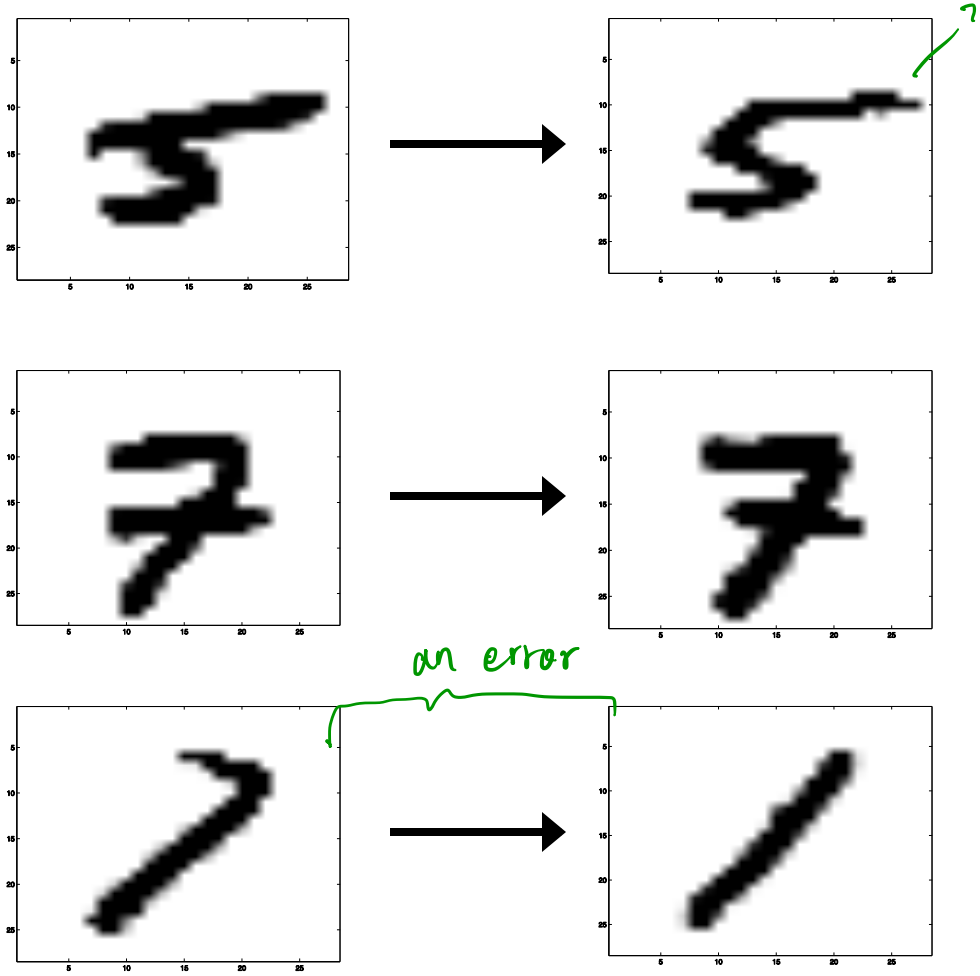
Euclidean distance between 784-dimensional vectors x, z is

$$\|x - z\| = \sqrt{\sum_{i=1}^{784} (x_i - z_i)^2}$$

Here x_i is the ith coordinate of x.

Nearest neighbor

Image to label **Nearest neighbor**



Overall:

error rate = 6%

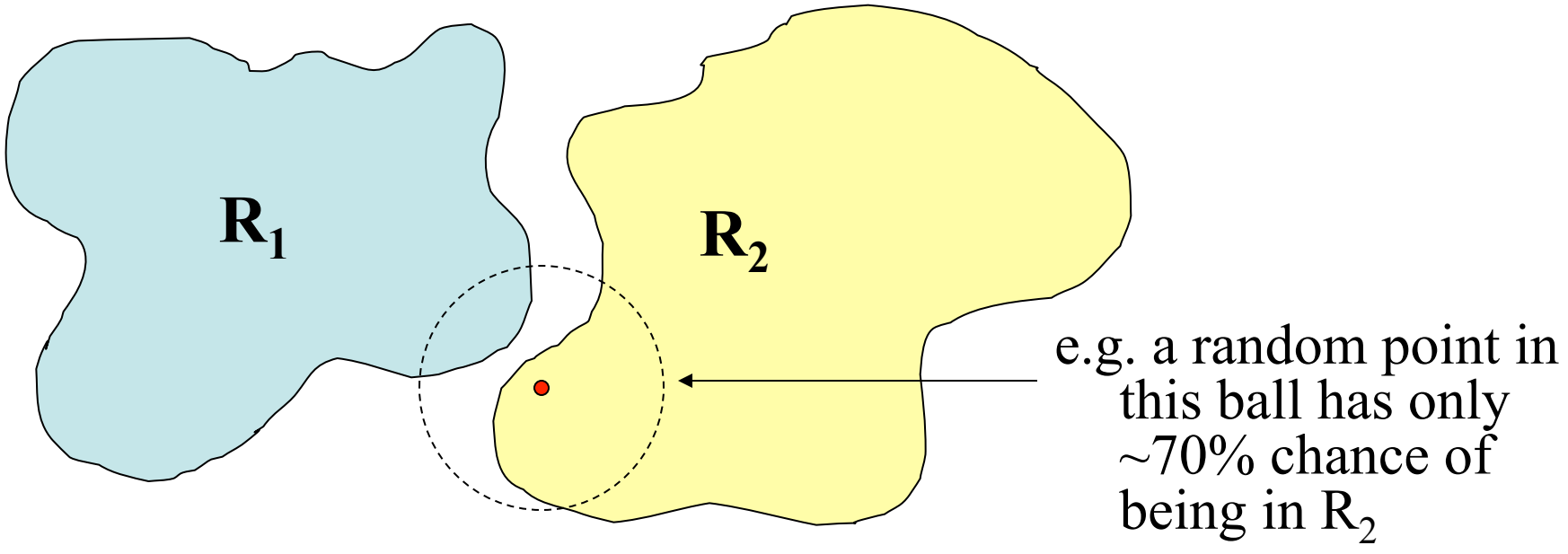
(on test set)

Question: what is the error rate for random guessing?

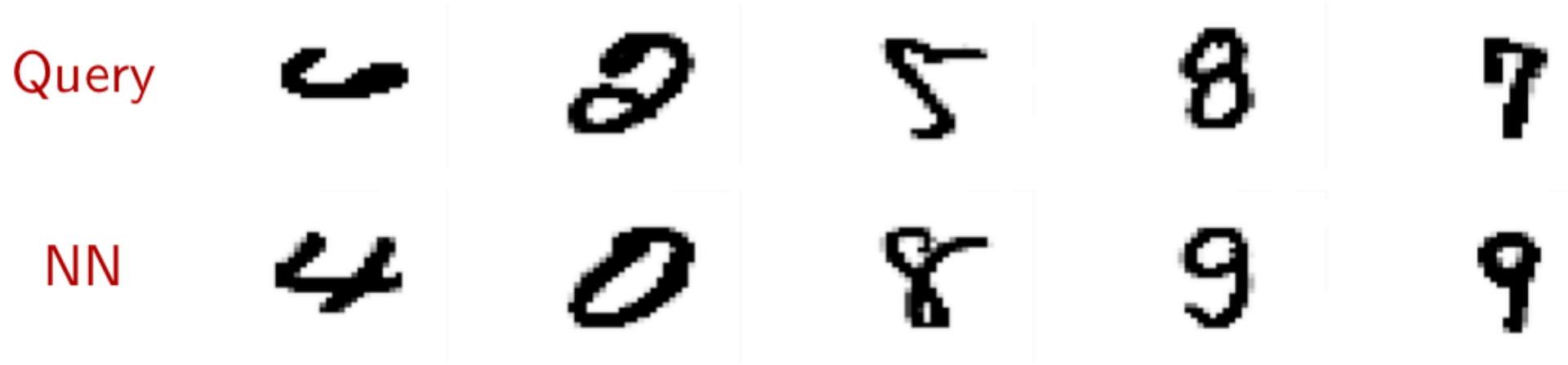
What does it get wrong?

Who knows… but here's a hypothesis:

Each digit corresponds to some connected region of R^{784}. Some of the regions come close to each other; problems occur at these boundaries.



Examples of errors



"Boost" the probability of success

Analogy: suppose a (biased) coin has
Pr(heads) = 0.70
Flip it 11 times and return the majority vote:
Pr(heads) = 0.92

Therefore: to classify x, find its $\boldsymbol{k}$ nearest neighbors (in the training set) and return their majority vote

[Large deviation theory: the foundation of machine learning…]

Time complexity

To classify a new (unlabeled) image:

find its nearest neighbor in the database (training set) and return that label

Given n training points in $\mathbb{R}^d$, what is the time complexity to label a new image, q?

For each image, x, in the database: $= O(n)$

Compute the Euclidean distance between x and q.

Return the $x_{\min}$ that minimizes this distance.

$$d(q,x) = \sqrt{\sum_{i=1}^{d}(q_i - x_i)^2} \quad = O(d)$$

$= O(nd)$

$$d(q,x) = \sqrt{\sum_{i=1}^{d}(q_i - x_i)^2} = O(d)$$

(Simple) Nearest neighbor: pros and cons

Pros

- Simple
- Flexible
- Excellent performance on a wide range of tasks

Cons

- Algorithmic: - Time consuming: with n training points in R^d, time to label a new point is O(nd).

- Statistical: - This is just memorization, not learning!
  - No insight is gained about the problem.
  - Would prefer a compact classifier.

Beyond Simple Nearest Neighbor

Learning problem:

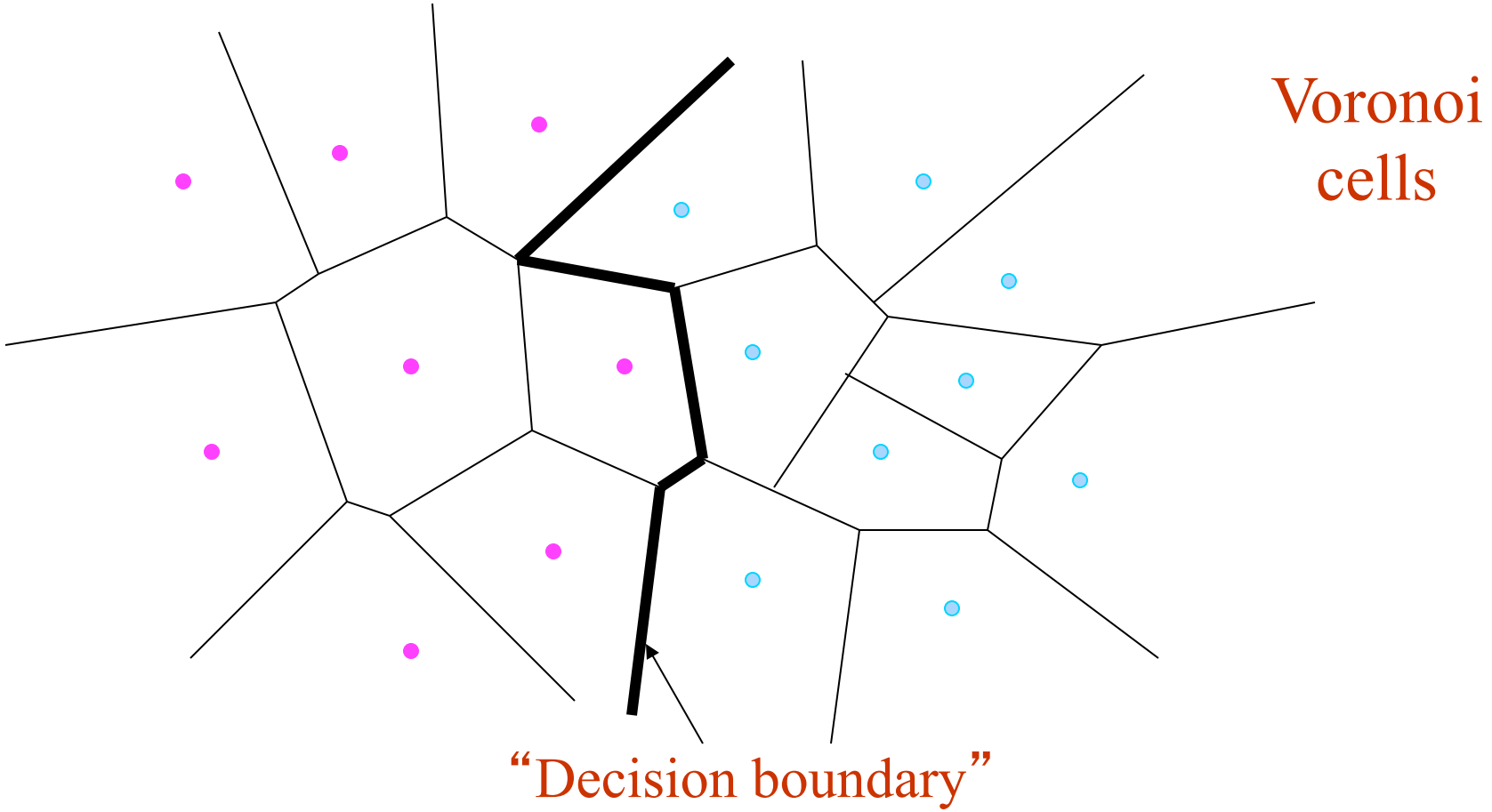
Design algorithms to perform nearest neighbor classification:

- with low error
- and low memory: do not store all the training examples!

- we want low training set

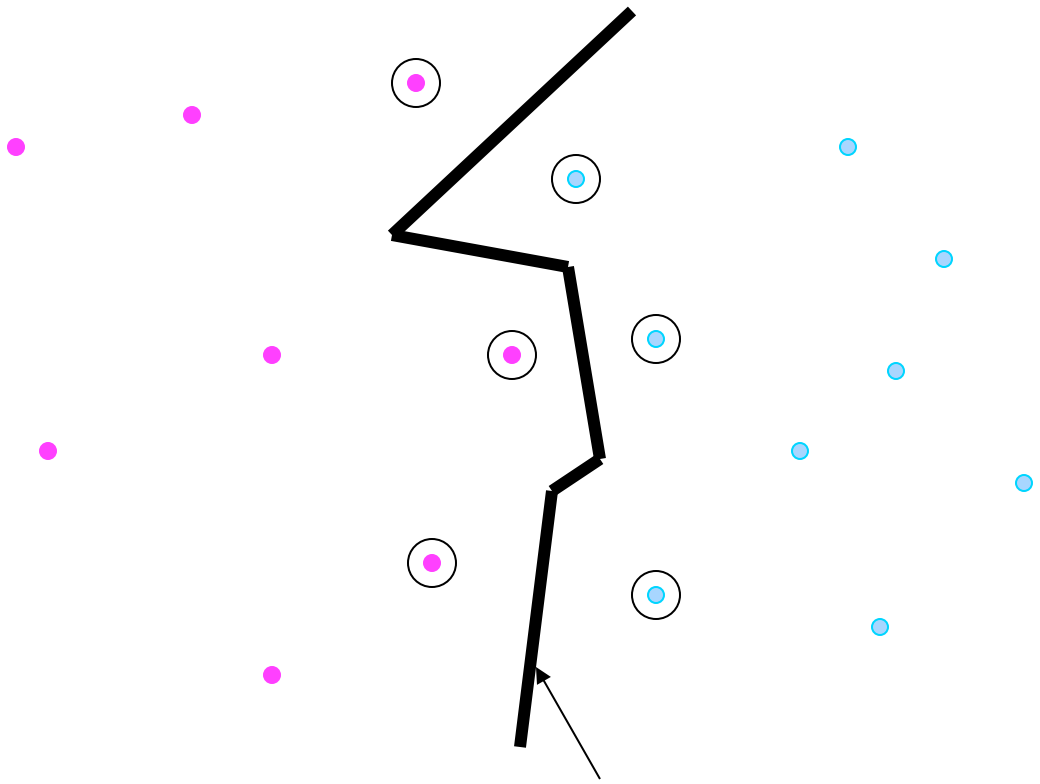
Prototype selection

A possible fix: instead of the entire training set, just keep a "representative sample"



Prototype selection

A possible fix: instead of the entire training set, just keep a "representative sample"



How to pick prototypes?

Idea 1: sample uniformly at random from training data

# prototypes	% error rate
20000	6.0
10000	7.4
5000	8.5
2000	10.7
1000	14.3
500	17.8
250	20.8
100	32.3
50	43.0

it's nice algorithm to pick random label dataset.

using over 100 prototypes, you get 32.3 error rate

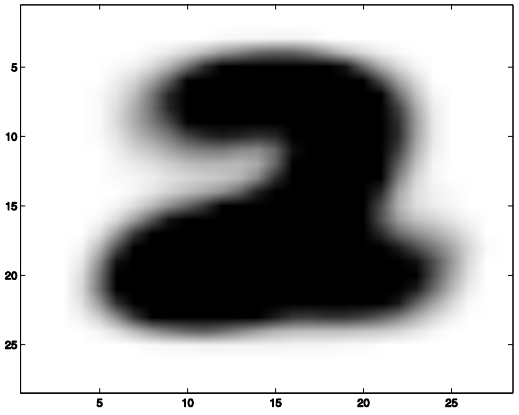
How to pick prototypes?

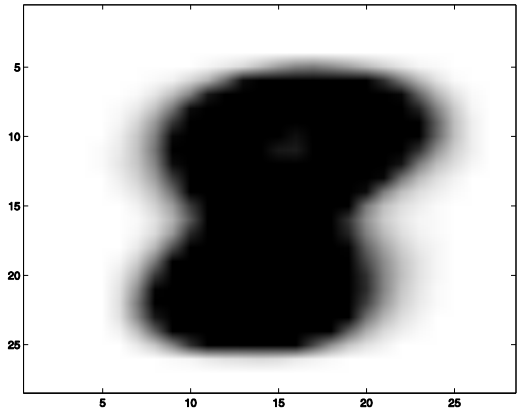
NOTE: They do not have to be actual data points!

Idea 2: one prototype per class: mean of training points

has to digit

Examples:





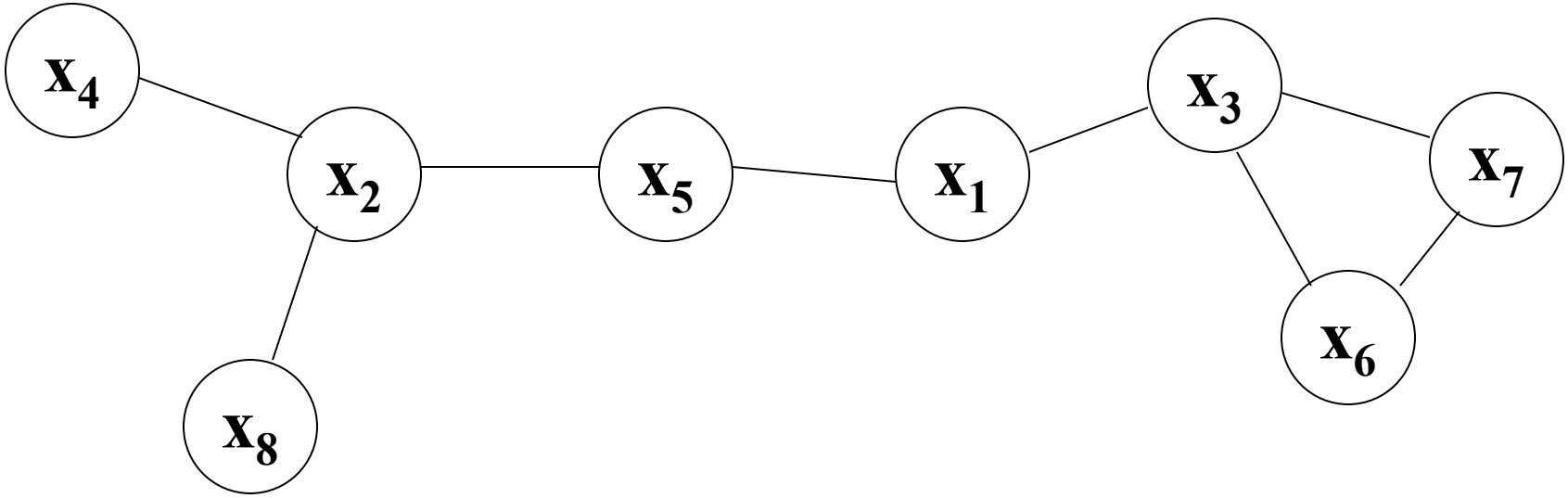
Error = 23%. Number of prototypes =

Picking prototypes

Other methods? e.g. Gain geometric insight from the neighborhood graph

Node = training point

Edges = connect close neighbors (small distance)



This is an area of machine learning research.

Generalization

Suppose a set of prototypes performs well on the training data.

==> Will it do well on future test data?

What kinds of classifiers generalize well?

less classifiers generalize well

Answer: the simplest (most compact) classifiers generalize the best. [cf. Occam's razor]

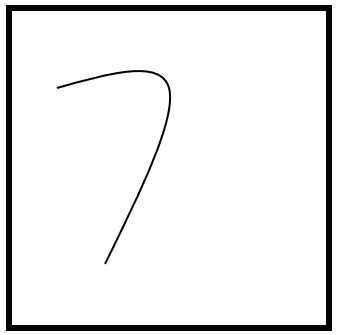
Measure of simplicity:

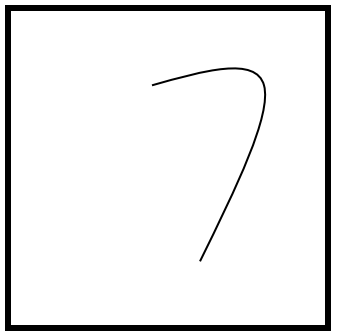
"Vapnik-Chervonenkis dimension" (later in course)

Postscript: representation I

The data lie in R^{784}, but our particular choice of Euclidean distance was pretty arbitrary.

Also suboptimal, eg.:





… are pretty far apart! (when computing Euclidean distance between vectors of pixels… Can you see why?)

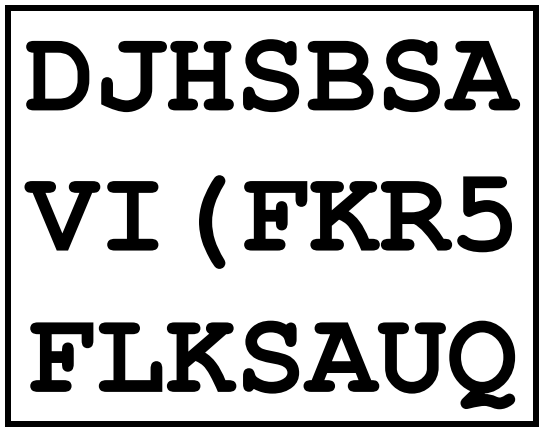
Is there a better distance measure for this application?

Postscript: representation II

Our 784 features: *pixels, are not very relevant to this task.*

Often data looks like this (*figuratively* speaking):

digit 5:

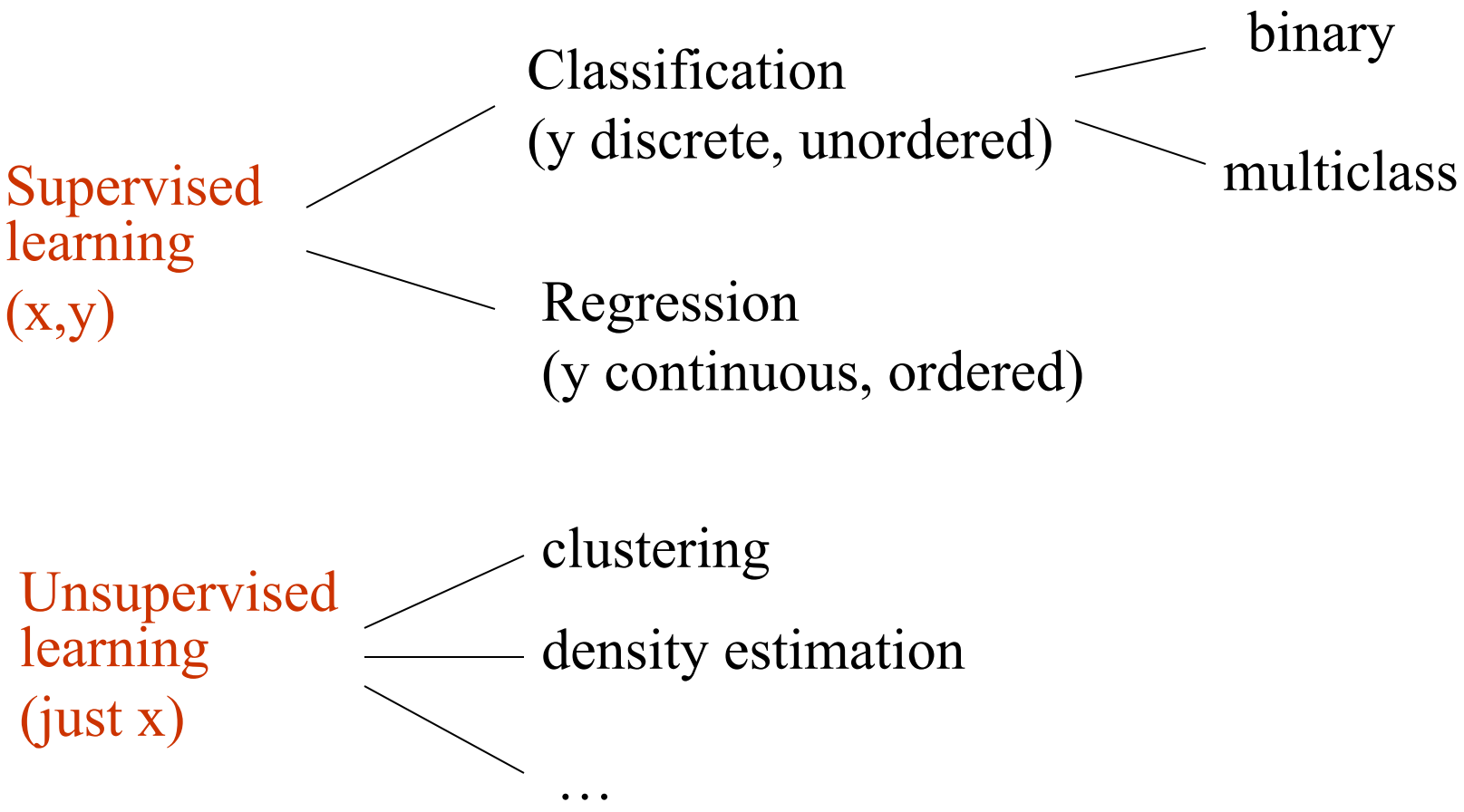


Information buried in a sea of irrelevant features.

Danger for nearest neighbor: match based on irrelevant features.

Feature selection is very important.

Postscript: learning tasks



All these tasks have a nearest-neighbor solution.