# CSC 503/SENG 474
# Data Mining

Nishant Mehta

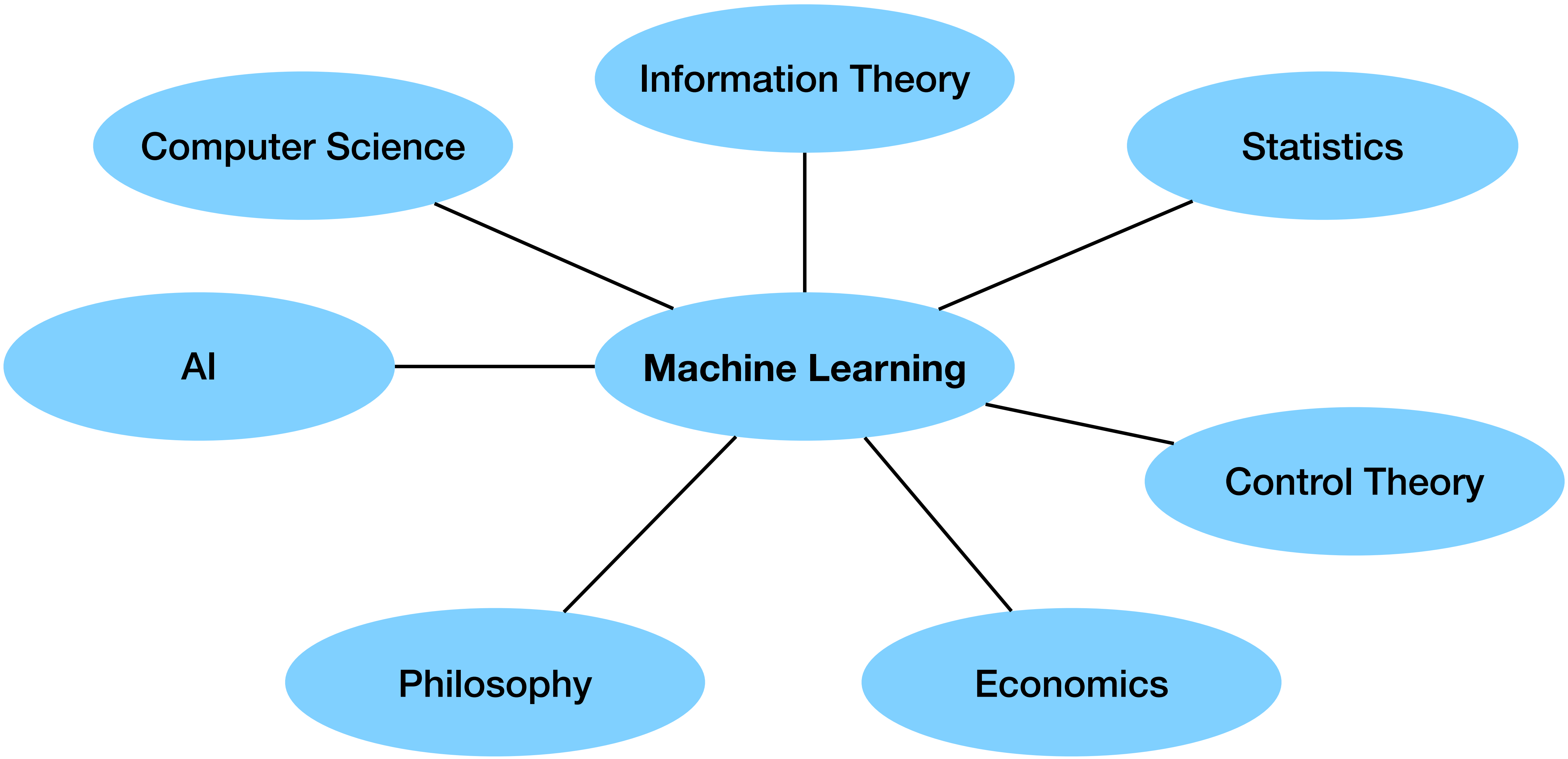
# Data Mining

- What is Data Mining?
  - Data Mining is roughly about finding patterns in large data sets. Data mining as a term has its origins more from the database community. A related term is “Knowledge discovery in databases”. Things like exploratory data analysis fit better within data mining than machine learning.
- Many of the tools are similar to tools from machine learning, but the purpose is somewhat different. Data Mining typically is more about obtaining some understanding/knowledge
- Draws from statistics, machine learning, and databases

# Machine Learning

- Machine learning is more about performing well on some learning task, as measured by minimizing prediction error. There can be a pattern, but we might not care about having some compact description of it. Prediction error is everything.
- There is a tremendous amount of overlap between statistics and machine learning. The label “statistician” or “machine learning person” says more about the types of problems you care about and how you study those problems.
- Things like reinforcement learning and online learning (especially against adversarial opponents) fit better with machine learning than data mining

# Connections to other fields



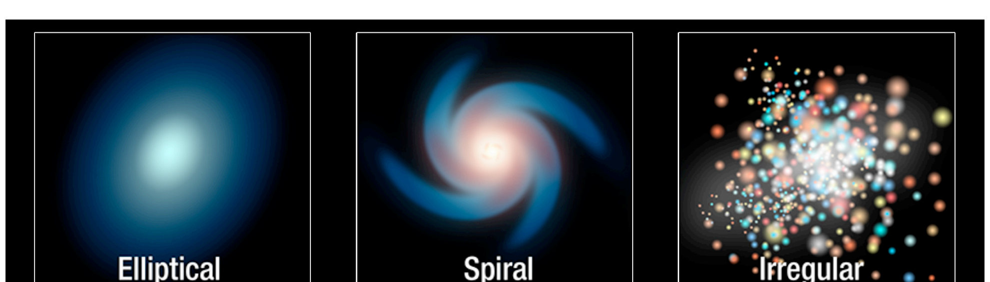
# Motivation for Machine Learning

- A lot of tasks that are tedious or less interesting to humans could be handled by machines

- Some tasks are too challenging for humans. Why?

    - Patterns are too hard to find (spotting forgeries)

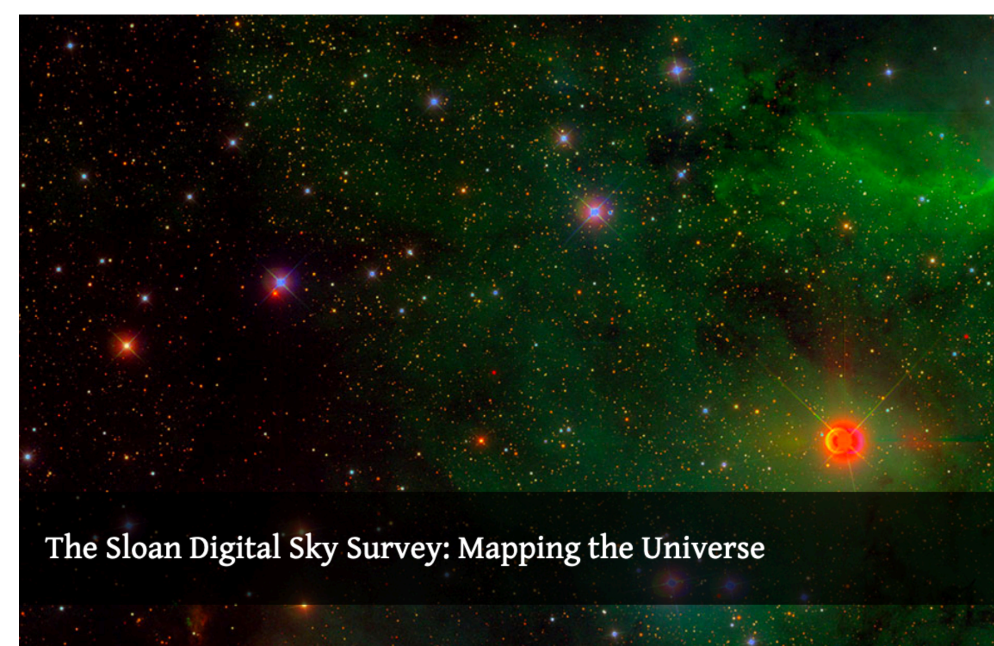
    - Volume of data is too large (classify all galaxies present in a telescope capture of the sky)




**The new tool in the art of spotting forgeries: artificial intelligence**

Instead of obsessing over materials, the new technique takes a hard look at the picture itself - specifically, the thousands of tiny individual strokes that compose it

▲ Johannes Vermeer's Girl with a Pearl Earring, circa 1665. Photograph: Mauritshuis, The Hague



# Anti-motivation for machine learning

# Anti-motivation for machine learning

- The machines will take over and eliminate us
- *OR* the machines will take over and we will not even be worthy of consideration; so, we'll exist but simply be irrelevant and live at the whimsy of the machines

# Anti-motivation for ~~machine learning~~ artificial intelligence

- The machines will take over and eliminate us
- *OR* the machines will take over and we will not even be worthy of consideration; so, we'll exist but simply be irrelevant and live at the whimsy of the machines
- BUT, Machine Learning is about excelling at particular tasks, while AI is about getting general, problem-solving agents
- A deep philosophical question: In the short-term, the pros might outweigh the cons. In the long term:

I think that is the single biggest existential crisis that we face and the most pressing one.

# Formally…

A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P*, if its performance at tasks in *T*, as measured by *P*, improves with experience *E*.

# Formally…

A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P*, if its performance at tasks in *T*, as measured by *P*, improves with experience *E*.

- Class of tasks?
  - Handwriting classification, face recognition, product recommendations, predicting stock prices, playing checkers/Go/Starcraft
- Performance measure?
  - Accuracy, Percentage of Games won
- Experience?
  - Data!

# Examples of Learning

| Task | Performance Measure | Experience |
|---|---|---|
| Classifying images of cats and dogs  | Percent of images in test set that are correctly classified | Training set of images with ground truth (i.e. correct) classifications |
| Playing Go  | Percent of games won against opponents | Records of previous games between humans, as well as playing practice games with itself |

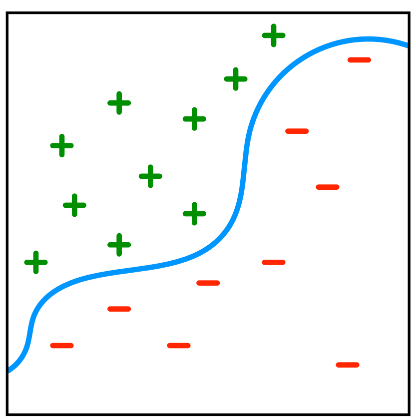
# Three main types of machine learning

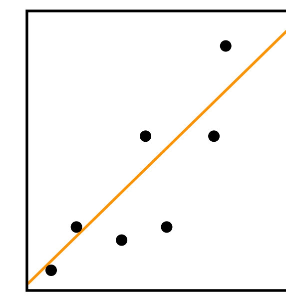
- Supervised learning
  - Prediction: Classification, Regression
- Unsupervised learning
  - Clustering, Density Estimation
- Reinforcement learning
  - Playing Games

# Three main types of machine learning

## Supervised Learning

- Prediction tasks, like classification and regression
- Input: a training set consisting of labeled examples consisting of both
  - *input features* describing the example
  - a special feature we are trying to predict, the *class label* (classification) or *target* (regression)



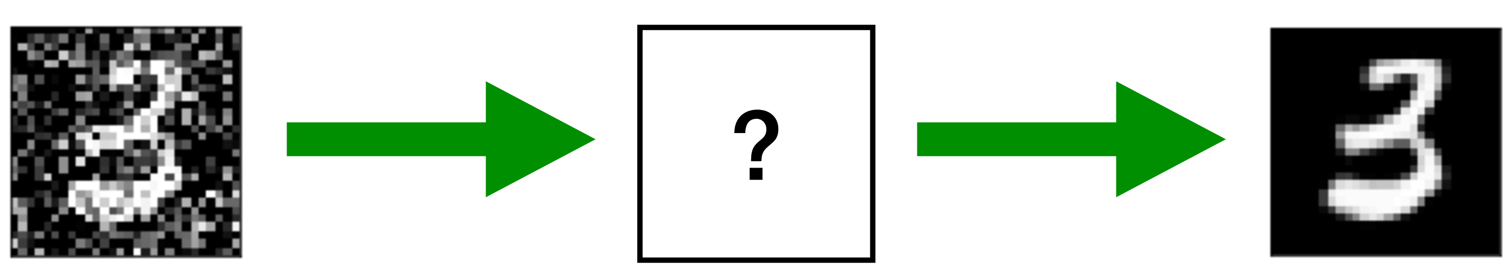


- Goal: learn (and output) a hypothesis that accurately predicts the label given the input features

# Three main types of machine learning

## Unsupervised Learning

- Input: a training set of examples (without labels)
- Goal: find a new representation of the data
- Examples:
  - Clustering: Group each example into one of *k* clusters
  - Density estimation: Model the underlying probability distribution that generates the data
  - Auto-encoding: Find new representation of input example to approximately reconstruct original example



# Three main types of machine learning

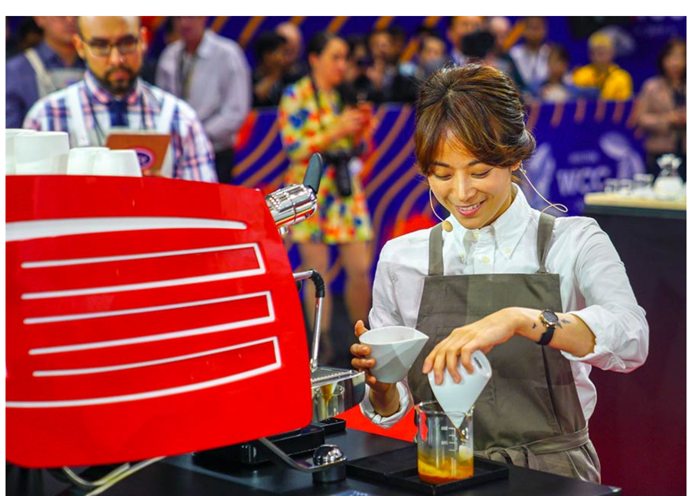
## Reinforcement Learning

- More on this later. Be sure do to the first assigned reading (Chapter 1 of Mitchell)

# Real-world example of supervised learning

- You go into a coffee shop and want to predict whether or not the barista will make a good espresso.

- Suppose that you’ve collected some features about previous baristas and the resulting expresso quality



Baristas

Homo Hipstericus


HOMO HIPSTERICUS

# Generalization

| Hipster | Australian | Sleepy | Espresso Label |
|---|---|---|---|
| No | Yes | No | **Good** |
| No | No | No | **Bad** |
| Yes | No | No | **Good** |
| No | Yes | Yes | **Bad** |
| No | Yes | No | **Good** |

Let's infer a rule.

# Generalization

| Hipster | Australian | Sleepy | Espresso Label |
|---|---|---|---|
| No | Yes | No | Good |
| No | No | No | Bad |
| Yes | No | No | Good |
| No | Yes | Yes | Bad |
| No | Yes | No | Good |

Let’s infer a rule.

It looks like being a not-sleepy Australian
is sufficient for making a good espresso.

# Generalization

| Hipster | Australian | Sleepy | Espresso Label |
|---|---|---|---|
| No | Yes | No | **Good** |
| No | No | No | **Bad** |
| Yes | No | No | **Good** |
| No | Yes | Yes | **Bad** |
| No | Yes | No | **Good** |

Let's infer a rule.

It looks like being a not-sleepy Australian is sufficient for making a good espresso.

Also, being a hipster also seems to result in a good espresso.

# Generalization

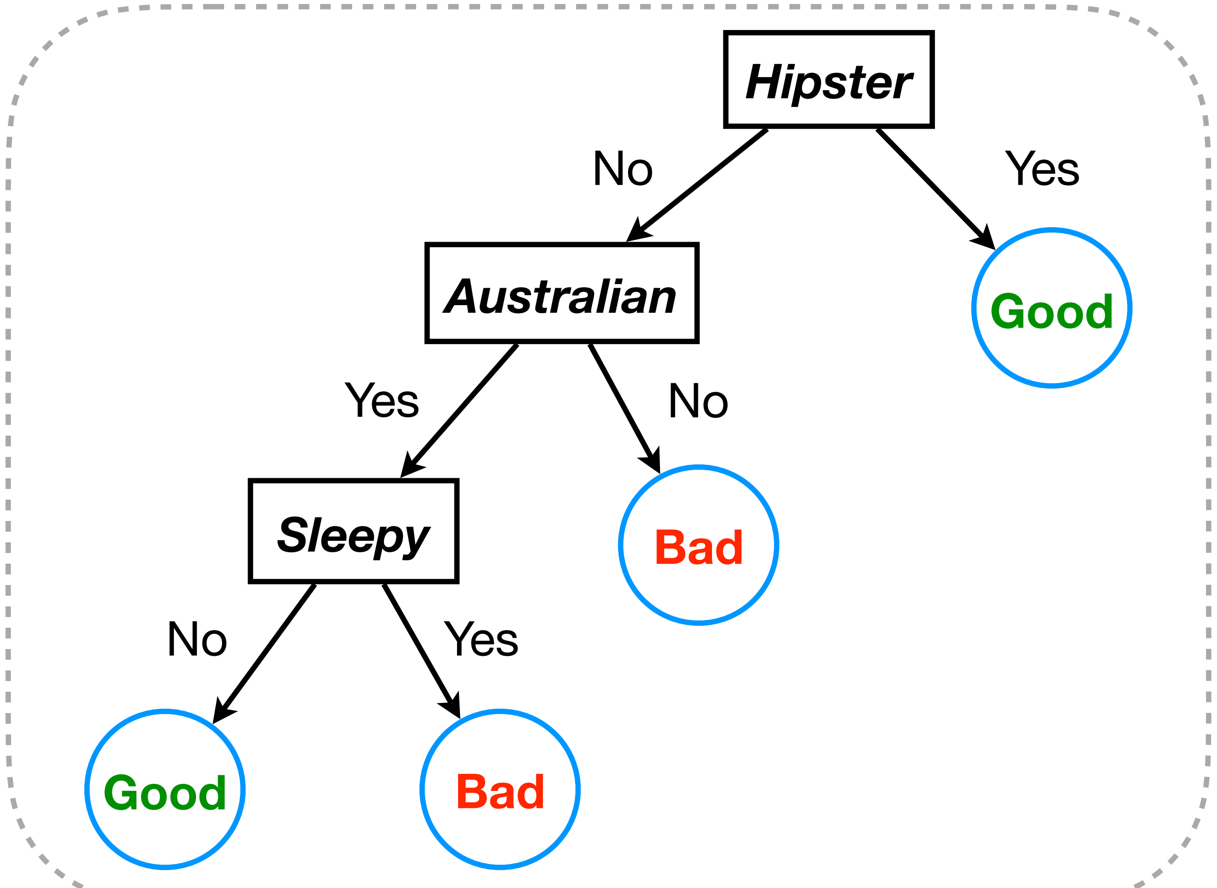
| Hipster | Australian | Sleepy | Espresso Label |
|---|---|---|---|
| No | Yes | No | Good |
| No | No | No | Bad |
| Yes | No | No | Good |
| No | Yes | Yes | Bad |
| No | Yes | No | Good |

Let's infer a rule.

It looks like being a not-sleepy Australian is sufficient for making a good espresso.

Also, being a hipster also seems to result in a good espresso.

A not-sleepy non-Australian that is not a hipster seems to make a bad espresso.



# Generalization

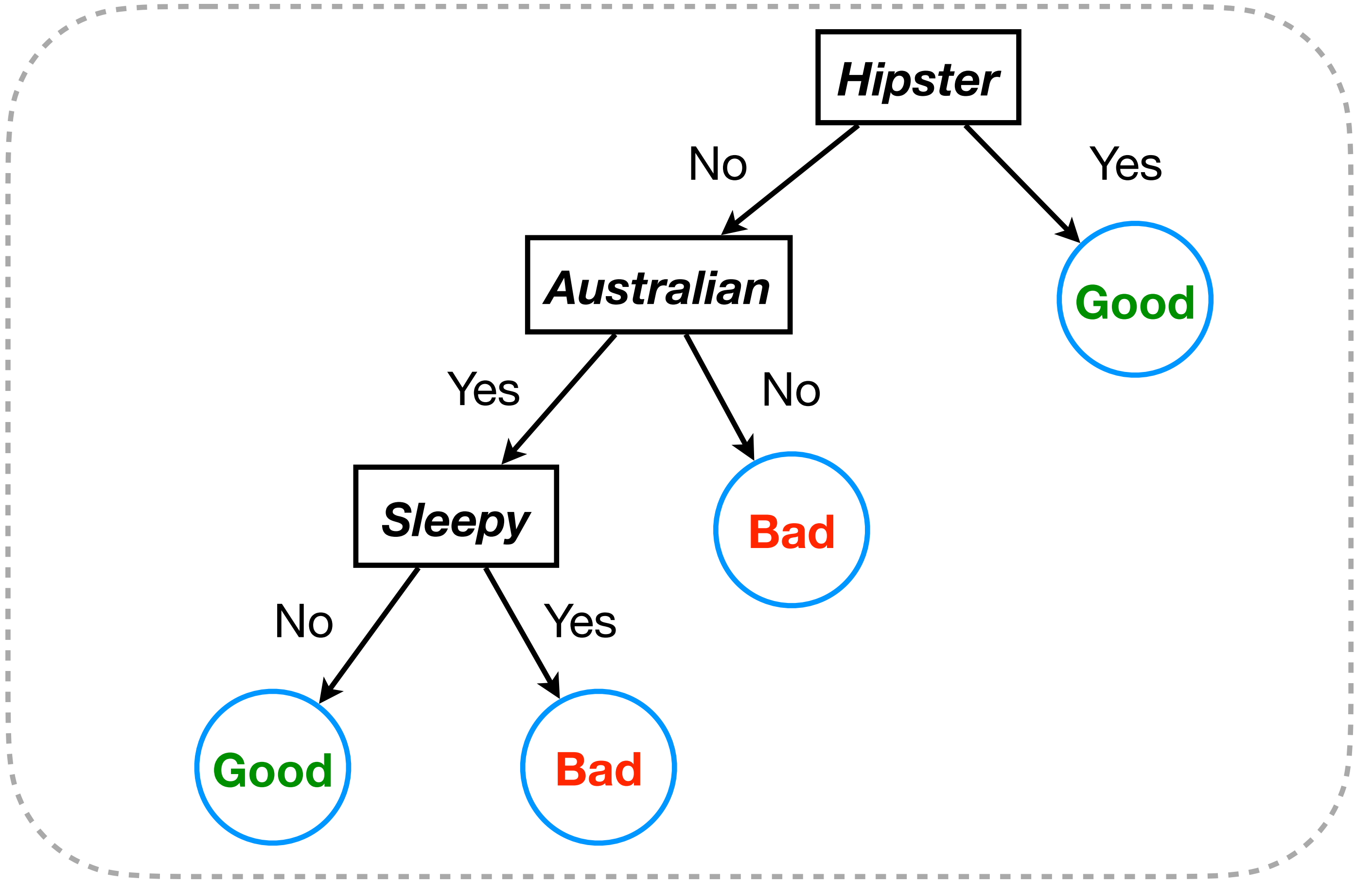
| Hipster | Australian | Sleepy | Espresso Label |
|---|---|---|---|
| No | Yes | No | Good |
| No | No | No | Bad |
| Yes | No | No | Good |
| No | Yes | Yes | Bad |
| No | Yes | No | Good |
| Yes | Yes | Yes | ? |

How about this espresso?

Our classifier predicts **Good**

But we are *generalizing*

(we never saw this example before)



# Data can be noisy

| Hipster | Sleepy | Espresso Label |
|---|---|---|
| No | No | **Good** |
| No | No | **Bad** |
| Yes | No | **Good** |
| No | Yes | **Bad** |
| No | No | **Good** |
| No | No | **?** |

What if we are missing the feature “Australian” and we run into this barista?

# Data can be noisy

| Hipster | Sleepy | Espresso Label |
|---|---|---|
| No | No | **Good** |
| No | No | **Bad** |
| Yes | No | **Good** |
| No | Yes | **Bad** |
| No | No | **Good** |
| No | No | **?** |

What if we are missing the feature “Australian” and we run into this barista?

For the same features, we have conflicting labels! This is a noisy label situation.

How to predict?

# Data can be noisy

| Hipster | Sleepy | Espresso Label |
|---|---|---|
| No | No | Good |
| No | No | Bad |
| Yes | No | Good |
| No | Yes | Bad |
| No | No | Good |
| No | No | ? |

What if we are missing the feature “Australian” and we run into this barista?

For the same features, we have conflicting labels! This is a noisy label situation.

How to predict?

Idea: Go with the class label which seems more likely, conditional on the features.

Predict Good

# Course webpage

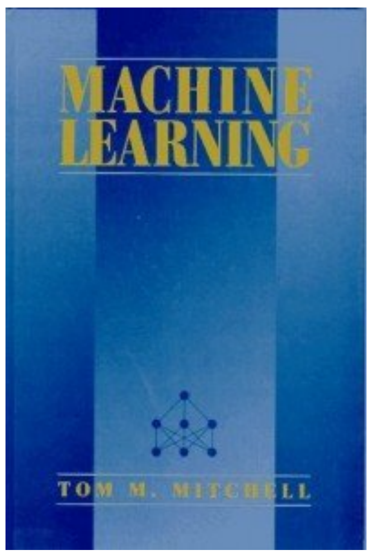
- https://web.uvic.ca/~nmehta/data_mining_spring2025
- I'll update this webpage frequently, so check it often. On the webpage, you can find:
  - The schedule, including planned topics
  - Required readings and some optional readings
  - Lecture slides and written notes

# This course is math-intensive

- We'll use statistics
  - We'll take expectations and play with conditional probability.
- We'll use calculus (rarely integration though):
  - We'll often take gradients of univariate functions
- We'll use linear algebra: [linear algebra review]
  - We'll multiply matrices and vectors
  - We'll see singular values and eigenvalues of matrices
- All this will combined with programming; you'll implement math-based algorithms in Python (or your favorite language)

# Textbooks

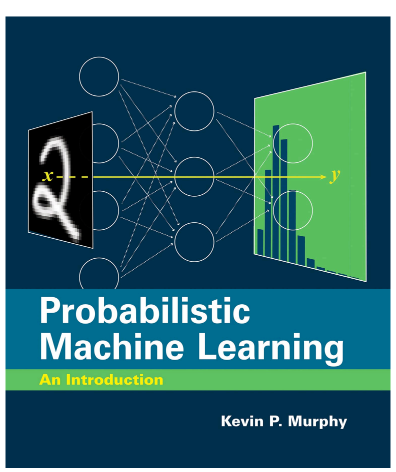
- Many required readings will be from:




**Machine Learning**
Tom Mitchell. McGraw Hill (1997)

- Yes, that's right, 1997. However, this is still THE best book for an introduction to machine learning

- Another book you could use is:




**Probabilistic Machine Learning: An Introduction**
Kevin Murphy. MIT Press (2022)

- A lot of the topics we cover are also in this book.

# Grading

## Undergrads

- 3 Assignments: 30% total
- Midterm: 20%
- Final exam: 25%
- Project: 25%

## Grad students

- 2 Assignments: 20% total
- Midterm: 20%
- Final exam: 25%
- Project: 25%
- Advising an undergrad group: 10%

# The Project

- The project is a major component of the course
- This is a group project. Each group is of between 4 to 6 people. It is better to have a group of 6, in case any group member drops the course.
- This is a machine learning course, so everyone working in a group needs to have a machine learning contribution
(e.g. one person cannot have the sole task of gathering data).

# The Project

- Initial Proposal - *Friday, Jan 17th*
- Group Formation - *Friday, Jan 24th*
- Formal Proposal - *Monday, Feb 10th*
- Progress Report - *Monday, Mar 10th*
- Final Presentation - *tentatively Apr 1st, Apr 2nd, and Apr 4th*
- Final Report - *tentatively Friday, Apr 11th*

# TAs and Labs

- The TAs are Ali Mortazavi and Mohamed Mouhajir
- The labs will be based on Jupyter Notebooks
- Tentative plan: first lab on Wednesday/Thursday, Jan $15^{th}/16^{th}$