Data mining & Machine Learning

CS 373
Purdue University

Dan Goldwasser
dgoldwas@purdue.edu

Today's Lecture

What is machine learning?

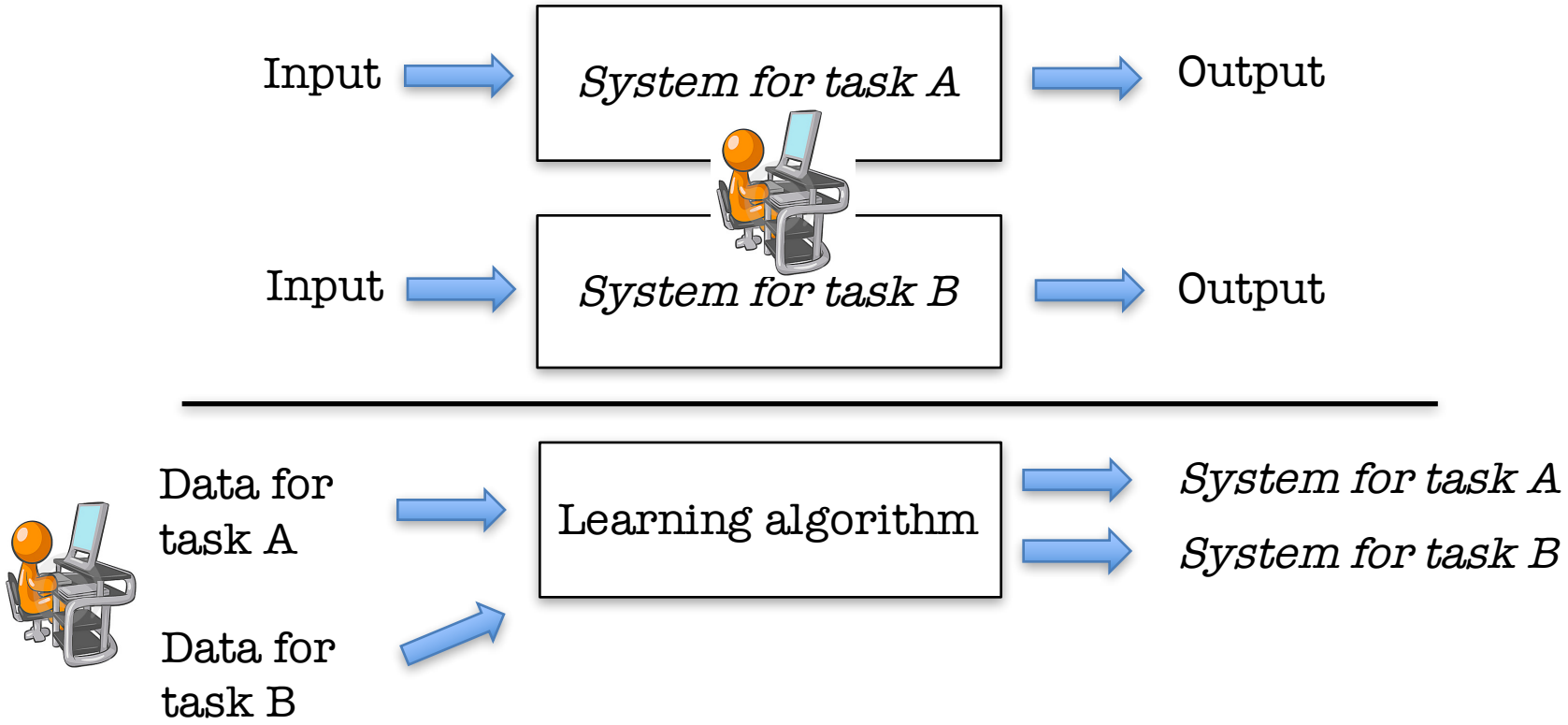
Why should I care about it?

- *Also, some class policies, grading, etc*

Machine Learning – short version

- A new paradigm for *telling a computer what to do!*
- **Traditionally:**
  - Write lines of code for different tasks
- **Machine learning**
  - Write code once (learning algorithm)
  - Supply **data** for different tasks
- Key issue: **Generalization**
  - How to generalize from training examples to new data
  - Identify patterns in the data and abstract from ***training*** examples to ***testing*** examples

Today's Lecture



The Badges Game

- The year was 1994, the conference was CoLT
  - *CoLT = Computational Learning Theory*
- Participants got a badge with a +/- label
- Only organizer knew the function generating the labels
- Label is determined only by the participant's name
- **Task**: look at many examples as you want in the conference, and induce the unknown function
  - *A great ice breaker for parties!*

The Badges Game

Can you figure out the function?

Name	Label
Claire Cardie	-
Peter Bartlett	+
Eric Baum	
Haym Hirsh	
Shai Ben-David	
Michael I. Jordan	

Do you need more examples?

The Badges Game

Can you figure out the function?

Name	Label
Claire Cardie	-
Peter Bartlett	+
Eric Baum	-
Haym Hirsh	+
Shai Ben-David	-
Michael I. Jordan	+

The Badges Game

How did you do it?

- Not a trivial process, even for humans.
- Requires raising hypotheses, validating those on data, accepting or rejecting them.

- **Can you find an algorithm to do the same process?**
  - **A form of search..**
  - *What are you searching over?*
  - *How would you bias the search?*
  - *When would you stop?*

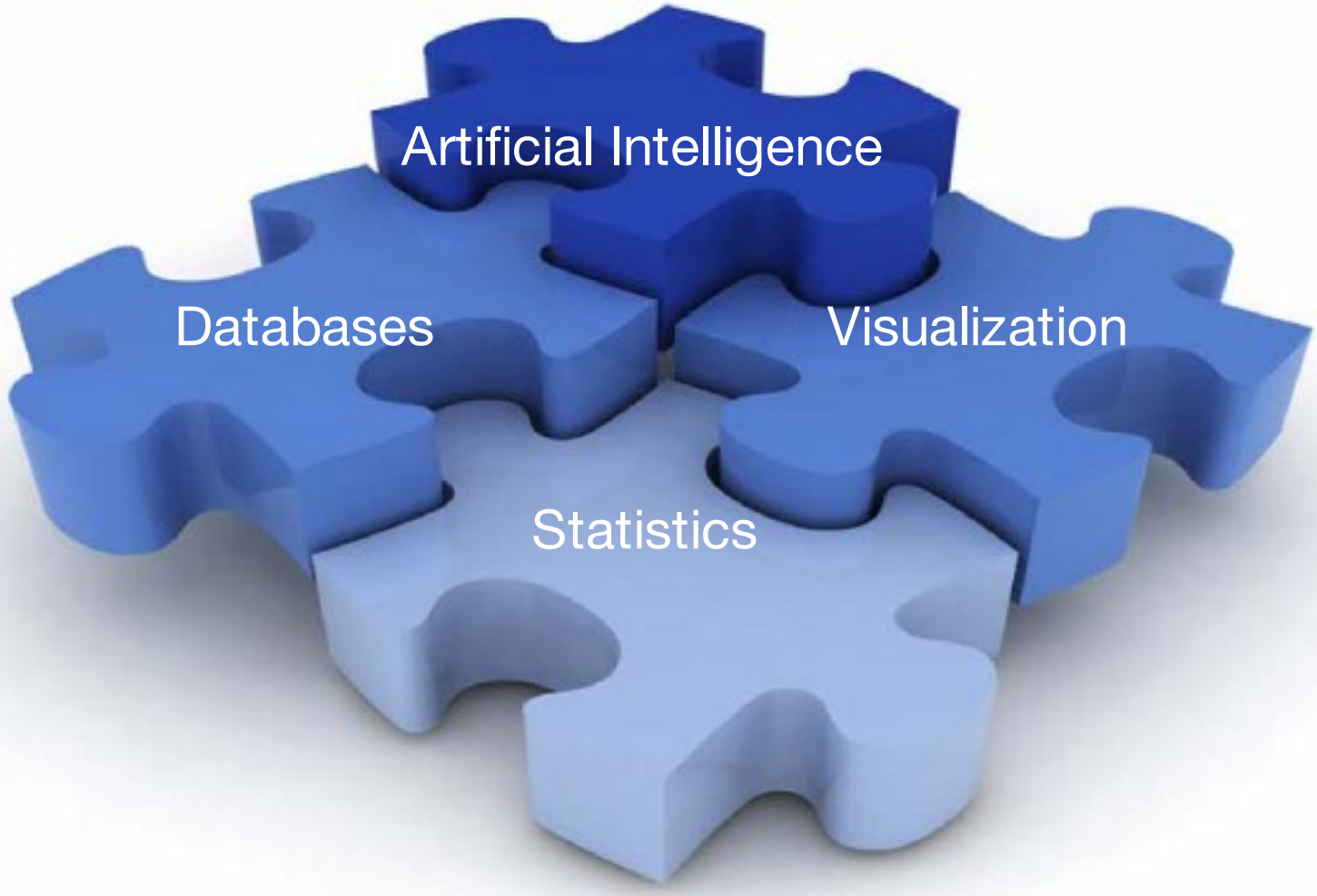
These are machine learning questions!

Definition

"Field of study that gives computers the ability to learn without being explicitly programmed"
Arthur Samuel (1959)

"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."
Tom Mitchell (1999)

The Big Picture

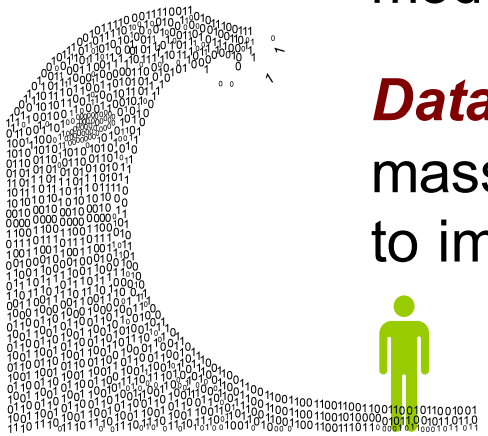


Today's Lecture

- The last 35 years of research in ML/DM has resulted in wide spread adoption of predictive analytics to automate and improve decision making.
- As "big data" efforts increase the collection of data… so will the need for new data science methodology. Data today have more volume, velocity, variety, etc.

Machine learning research develops statistical tools, models & algorithms that address these complexities.

Data mining research focuses on how to scale to massive data and how to incorporate feedback to improve accuracy while minimizing effort.



Data is everywhere

NATIONAL GEOGRAPHIC

THE WALL STREET JOURNAL.
WSJ

Top Rated Movies

SHARE

Top 250 as rated by IMDb Users

Showing 250 Titles

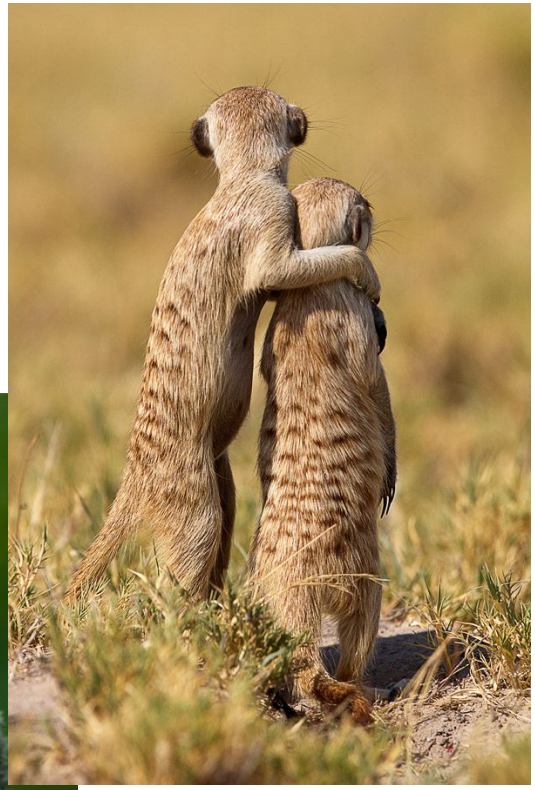
Sort by: Ranking

Rank & Title	IMDb Rating	Your Rating
1. The Shawshank Redemption (1994)	9.2	
2. The Godfather (1972)	9.2	
3. The Godfather: Part II (1974)	9.0	
4. The Dark Knight (2008)	8.9	
5. Schindler's List (1993)	8.9	

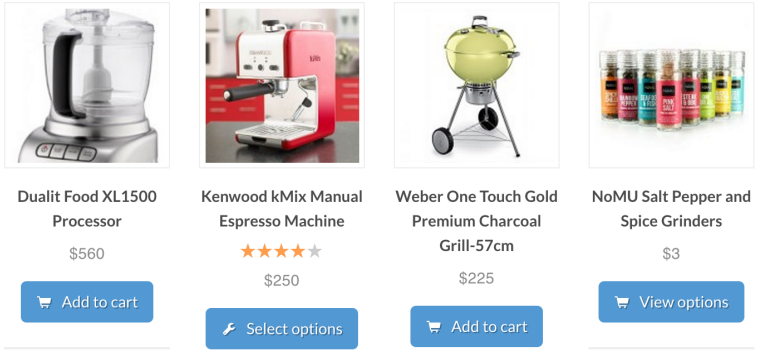
Data is everywhere

Data is **everywhere**

Data is **everywhere**



And it used for **everything**!

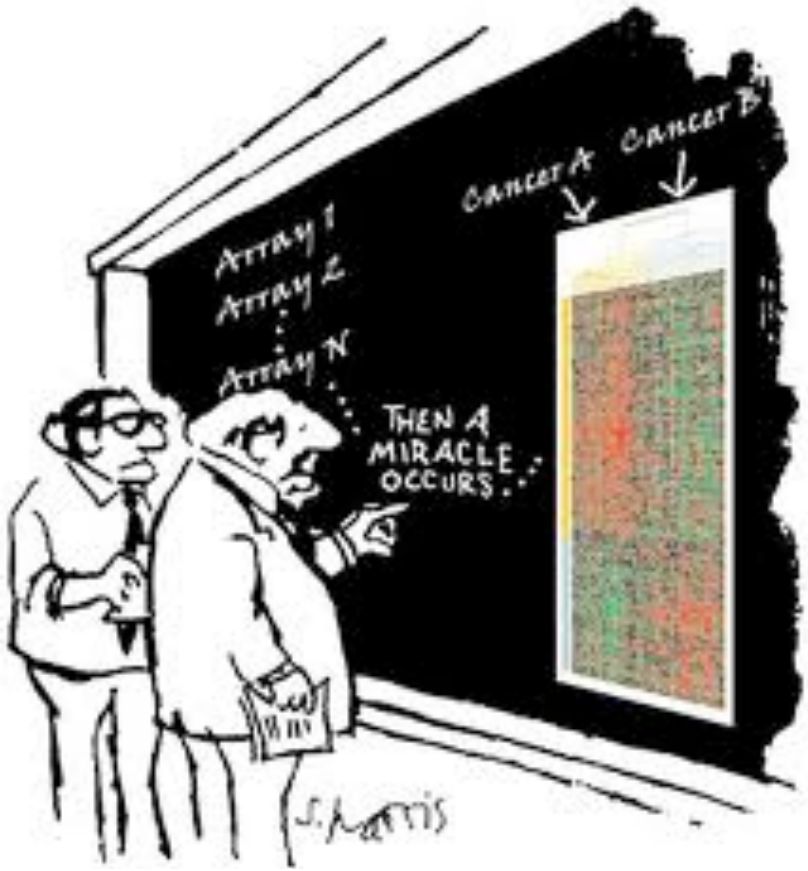


Google Translate

Universal Approach

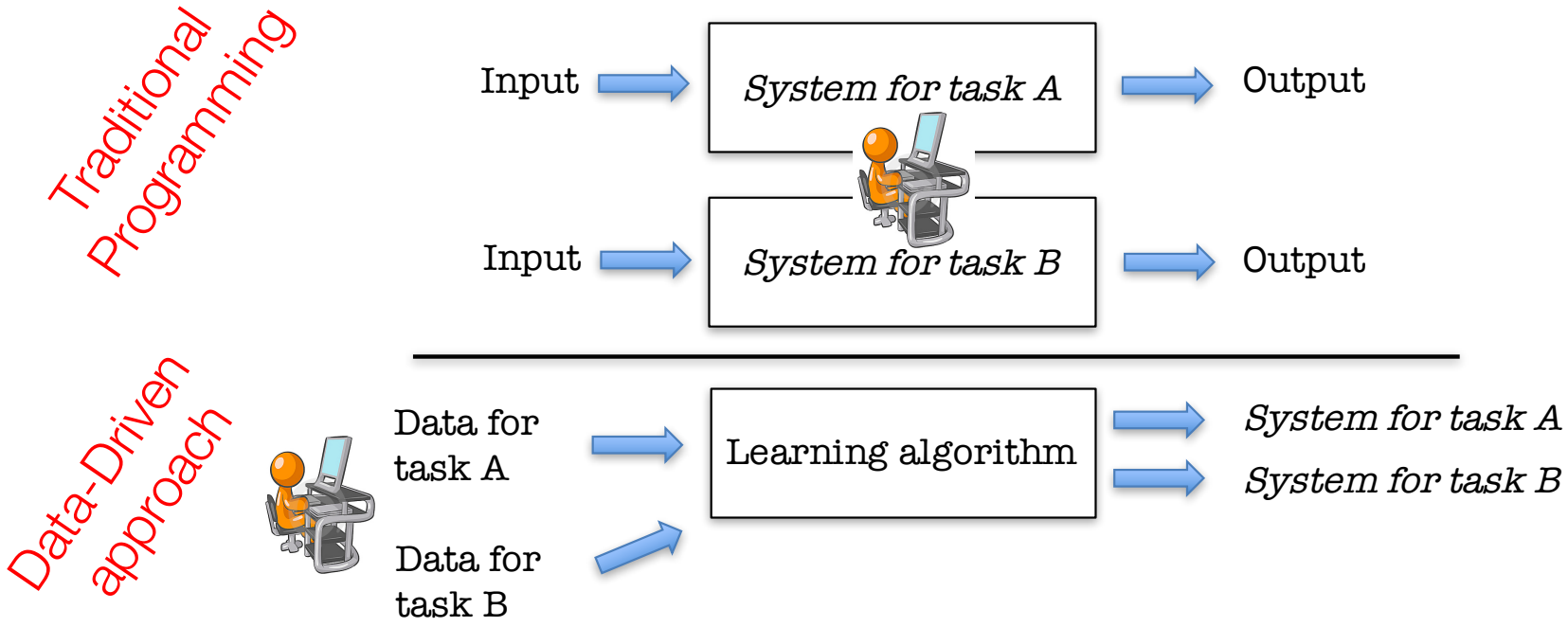
- Similar principles apply to all data-driven applications.
  - Automatic pilot
  - Personal assistants
  - Translation
  - Automatic trading
  - Content recommendation
  - .. *Your application?*

Let's demystify the process



"I THINK YOU SHOULD BE MORE EXPLICIT HERE IN STEP TWO."

The 10,000 feet view



- **In reality this process is much messier..**

The Process

Data

Selection

Target data

Preprocessing

Processed data

Knowledge

Interpretation evaluation

Machine Learning

Patterns

Mining

The process

1. **Application setup:**
   - Acquire relevant domain knowledge
   - Assess user goals
2. **Data selection**
   - Choose data sources
   - Identify relevant attributes
   - Sample data
3. **Data preprocessing**
   - Remove noise or outliers
   - Handle missing values
   - Account for time or other changes
4. **Data transformation**
   - Find useful features
   - Reduce dimensionality

The process

5. **Data mining/ml:**
   - Choose task (e.g., classification, regression, clustering)
   - Choose algorithms for learning and inference
   - Set parameters
   - Apply algorithms to search for patterns of interest

6. **Interpretation/evaluation**
   - Assess accuracy of model/results
   - Interpret model for end-users
   - Consolidate knowledge

7. **Repeat…**

Machine learning is at the heart of this process!

Key Principles of ML

- *We will look into several learning **protocols**, using different learning **algorithms**, for learning different **models**.*

- **Model**: function mapping inputs to outputs
- **Algorithm**: used for constructing a model, based on data
- **Protocol**: the settings in which the algorithm learns.

Learning Model

- We think about learning as producing a function mapping input to outputs, based on data
  - E.g., **spam: email → Boolean**

- A **model** is the type of function the learner uses
  - Linear functions, non-linear functions
  - Decision trees
  - Ensembles of classifier

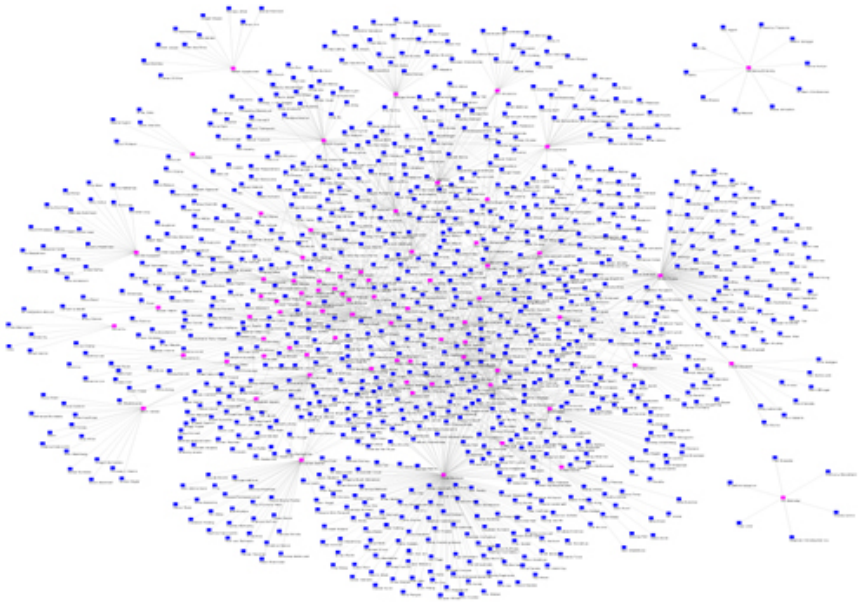
- The key question is **expressivity -** what can the model represent

Learning Models: Representation

- We think about learning as producing a function mapping input to outputs, based on data
  - E.g., **spam: email → Boolean**

What is the domain and range of this function?

- *The **output** space* defines the learning task.
  - *Binary, multiclass, continuous (regression), structure..*
- *The **input** space defines the data representation*



$$\begin{bmatrix}
0&0&1&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0\\
1&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0\\
0&1&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0\\
1&0&0&0&1&0&0&0&0&1&0&0&0&1&0&0&0&0\\
0&1&0&0&0&1&0&0&0&0&1&0&0&0&1&0&0&0\\
0&0&1&1&0&0&0&0&0&0&0&1&1&0&0&0&0&0\\
0&0&0&1&0&0&1&0&0&0&0&0&0&0&0&0&0&1\\
0&0&0&0&1&0&0&1&0&0&0&0&0&0&0&1&0&0\\
0&0&0&0&0&1&0&0&1&0&0&0&0&0&0&0&1&0\\
0&0&0&0&0&0&1&0&0&0&0&0&0&1&0&0&0&1\\
0&0&0&0&0&0&0&1&0&0&0&0&0&0&1&1&0&0\\
0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&1&0
\end{bmatrix}$$

?

Luke, I am your father.

(0,0,1,0,0,2,0,..,0)

Learning Protocol

- **Supervised learning**
  - Human (*teacher)* supplies a labeled examples
  - *Learner* has to learn a model using this data
- **Unsupervised learning**
  - No *teacher*, *learner* has only unlabeled examples
  - Data mining: finding patterns in unlabeled data
- **Semi-supervised learning**
  - *Learner* has access to *both* labeled and unlabeled examples
- **Weakly-supervised learning**
  - We have access to noisy labels

Learning Protocol

- **Active learning**
  - *Learner* and *teacher* interact
  - *Learner* can ask questions
- **Reinforcement learning**
  - Learner learns by interacting with the environment

- Why is Reinforcement learning **different/similar** to Supervised learning? Active learning?

Learning Algorithm

- Learning Algorithms generate a **model**, they work under the settings of a specific **protocol**
  - **Supervised vs. Semi-Supervised vs. Unsupervised**
  - **Online vs. Batch**
  - Online algorithm: learning is done one example at a time
    - Winnow, Perceptron,..
  - Batch algorithm: learning is done over entire dataset
    - SVM, Logistic Regression, Decision Trees, ...

- How can we compare learning algorithms?

Classification

- **Classification**: mapping data into categories
  - *Write a face recognition program*
  - *Determine if an English sentence is grammatical*
  - *Distinguish between normal and cancerous cells*

- **Can't we just write code?**

  - Provide **labeled examples** and let a classifier distinguish between the two classes
  - What are the labeled examples in each case?

Classification

Class	Outlook	Temperature	Windy?
Play	Sunny	Low	Yes
No play	Sunny	High	Yes
No play	Sunny	High	No
Play	Overcast	Low	Yes
Play	Overcast	High	No
Play	Overcast	Low	No
No play	Rainy	Low	Yes
Play	Rainy	Low	No

- Predict the **class label** (*play tennis)*
- **Features**: outlook, temperature, windy
- **Feature values**: can be binary, categorical, continuous

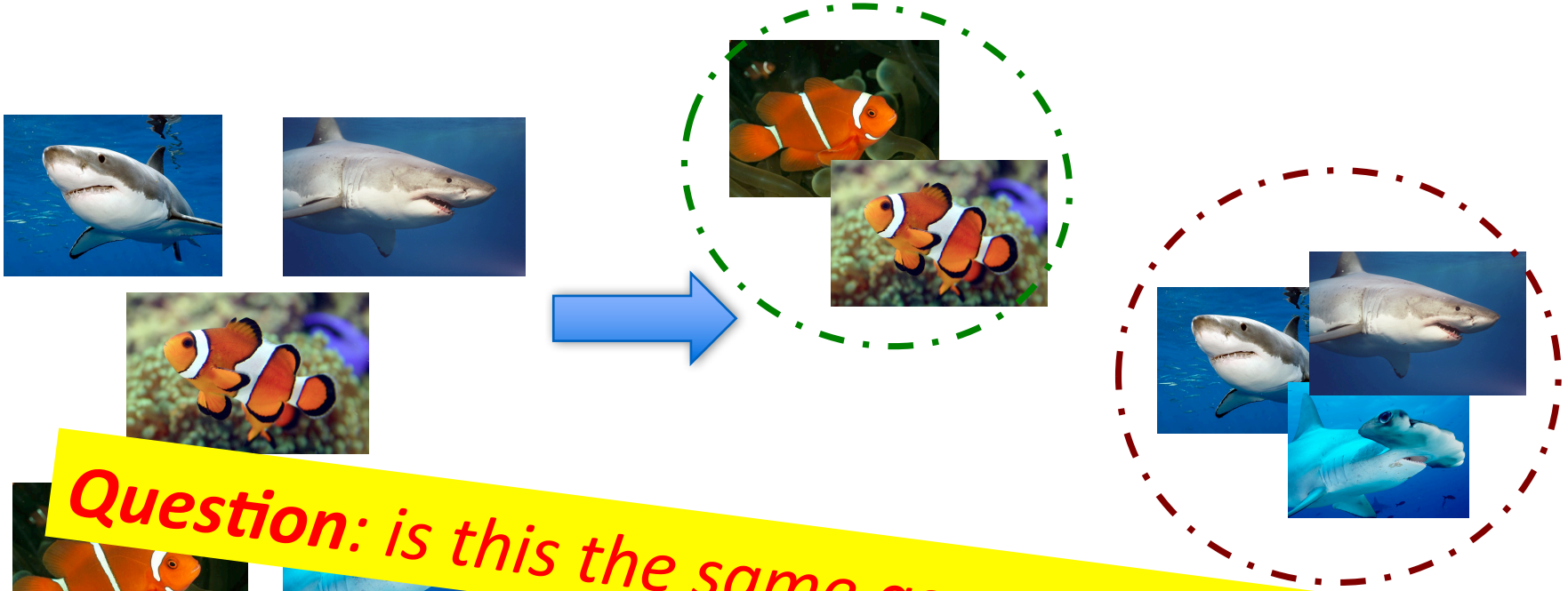
Classification

- A labeled dataset is a collection of (x,y) pairs
  - x refers to input examples
  - y refers to output labels
- Our goal is to build a model to predict new examples
- **Generalization**: can we make reliable predictions?

Class	Outlook	Temperature	Windy?
???	Sunny	Low	No

Supervised (Predictive) vs. Unsupervised (Descriptive)

- Unsupervised Learning: Learn properties of the data
- **Clustering:** group similar instance
- Define a similarity metric between instances



***Question**: is this the same as prediction?*

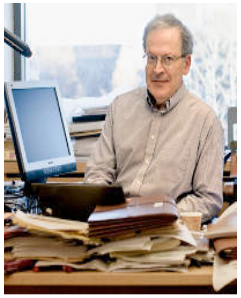
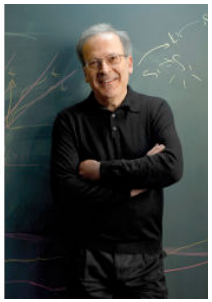
Clustering



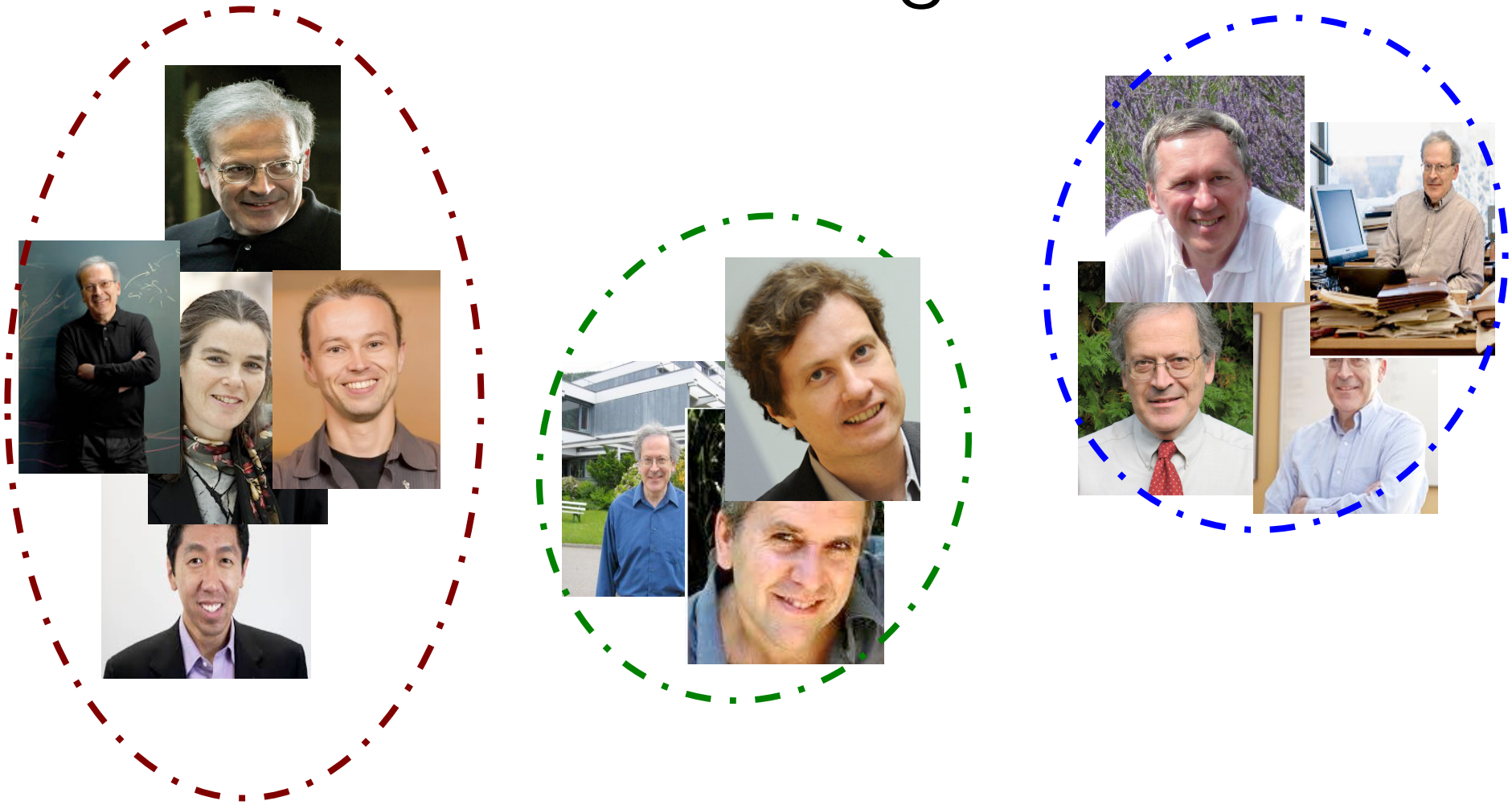




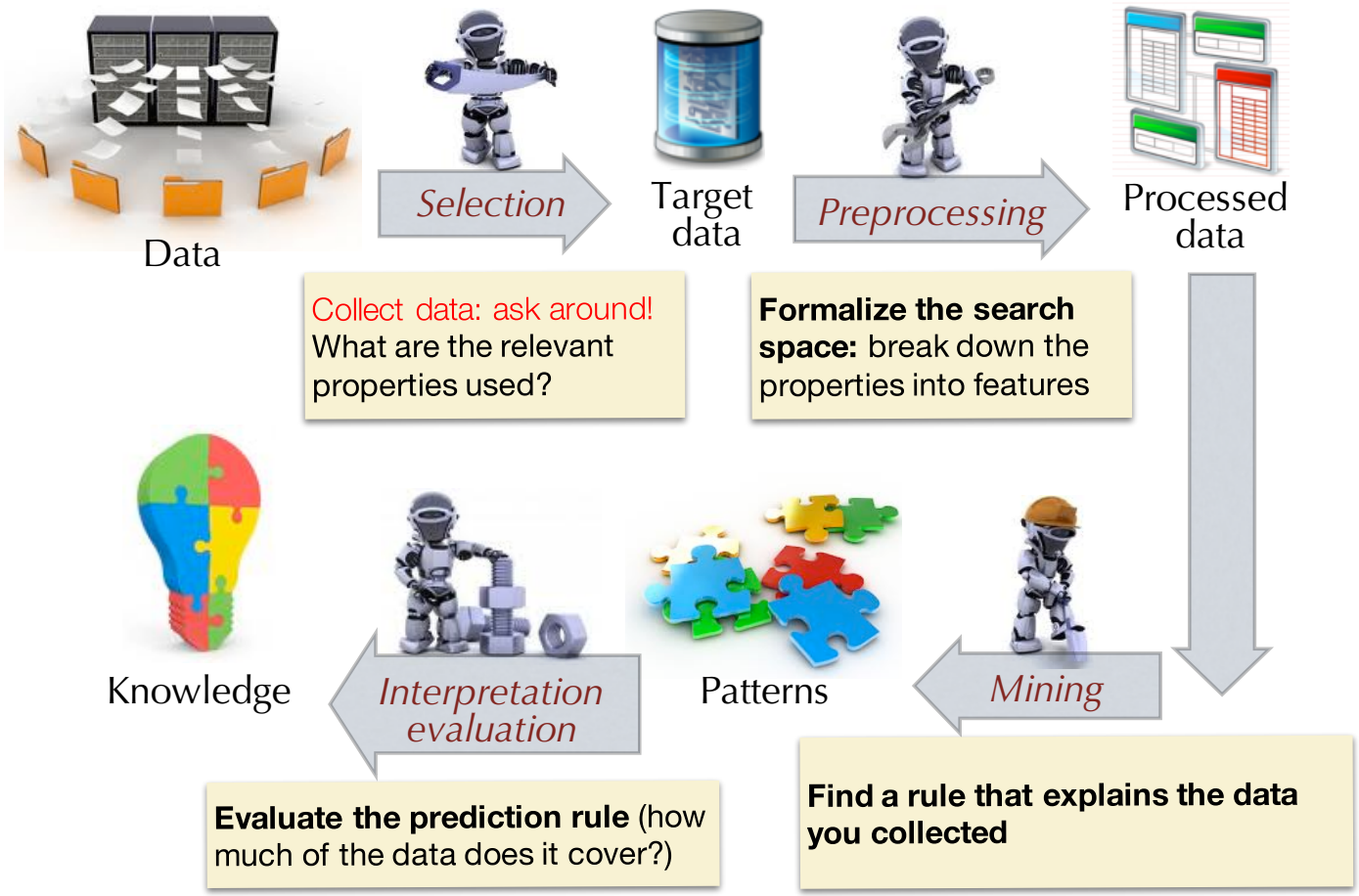


Clustering



The process



Important lesson:

it's not just about the best algorithm, understanding the data and your task is critical!

Course Policies

Course Learning Objectives

- Identify key elements of data mining and machine learning algorithms
- Understand how algorithmic elements interact to impact performance
- Understand how to choose algorithms for different analysis tasks
- Analyze data in both an exploratory and targeted manner
- Implement and apply basic algorithms for supervised and unsupervised learning
- Accurately evaluate the performance of algorithms, as well as formulate and test hypotheses

Topics

- Elements of data mining algorithms
- Statistical basics and background
- Data preparation and exploration
- Descriptive modeling
- Predictive modeling
- Methodology, evaluation
- Advanced Topics (subject to change):
  - Pattern mining and anomaly detection
  - Deep Learning and Reinforcement Learning

Logistics

- Instructor: Dan Goldwasser
  dgoldwas@purdue.edu, LWSN 2142A, office hours: After class
- Teaching assistant: details and office hours will be online
- Webpage: http://www.cs.purdue.edu/~dgoldwas/courses/DM373SP18
- Piazza: email me if you are not enrolled!
- https://piazza.com/purdue/spring/cs373
- Prerequisites: CS182, CS251
  Concurrent prerequisite: STAT350 or STAT511

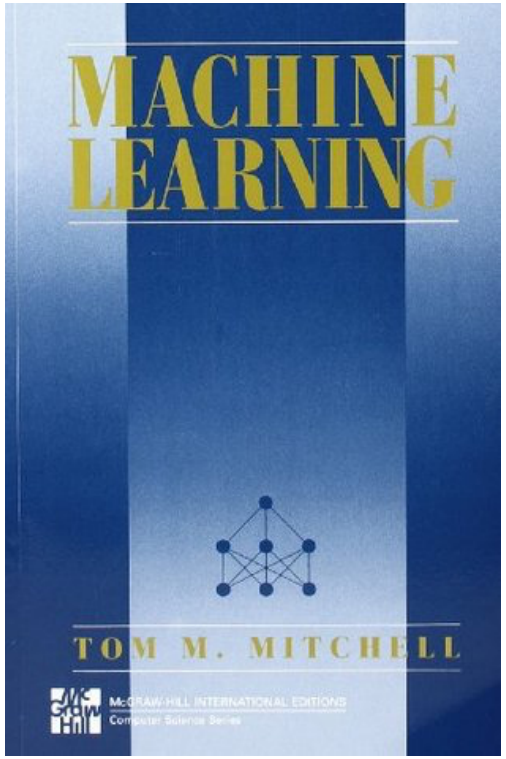
Piazza

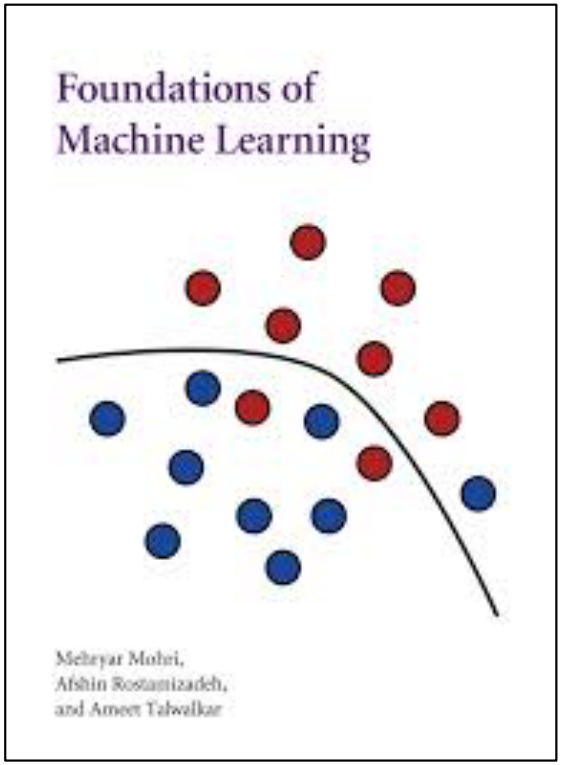
- We will rely on Piazza for posting materials and class discussions.

- Class forums are intended to allow students discuss academic issues
- DO NOT:
  - Ask personal questions, discuss unrelated topics, etc

Readings: Data mining perspective

- No required text, readings will be announced/ distributed on course webpage.





Very helpful (and free!) e-book: **http://ciml.info/**

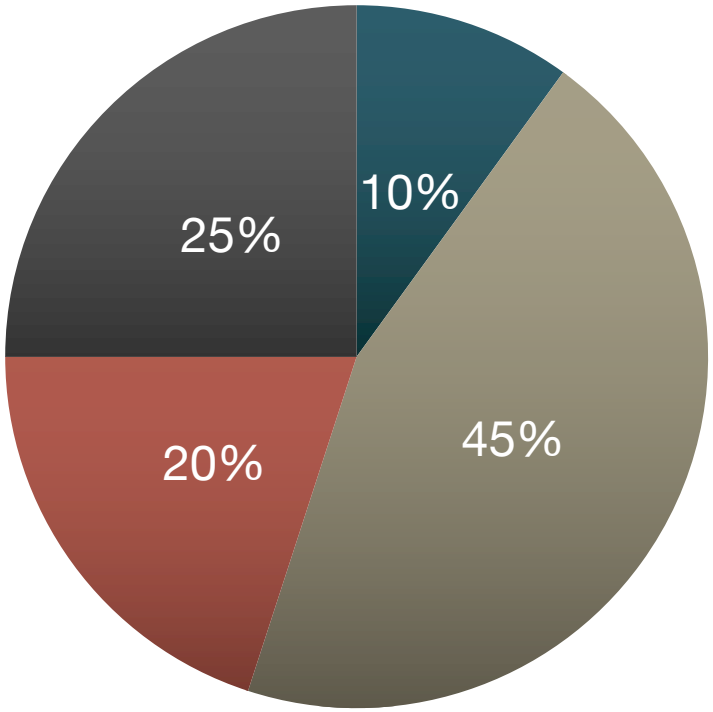
Workload

- **Homeworks**
- Five assignments including written/math exercises, programming assignments in python, analysis in R

- **Late policy**: 15% off per day late, maximum of 5 days
  - Five *extension* days can be applied anytime (no explanation needed)
  - Use must be stated explicitly (either in the submission header or by accompanying email to the TA)
  - No Fractional days
  - Five days **total,** can be used any way you want.

- **Exams**
  - Midterm and final exam

Grading



Quizzes and participation

Homework

Midterm

Final