

A red speech bubble with a white border and a white arrow pointing downwards towards the main title.

Introduction to Machine Learning

WQD7006: Machine Learning for Data Science

Semester II, Session 2021/2022

Data Science

Multidisciplinary blend of **data inference, algorithm development, and technology** in order to solve **analytically** complex problems.



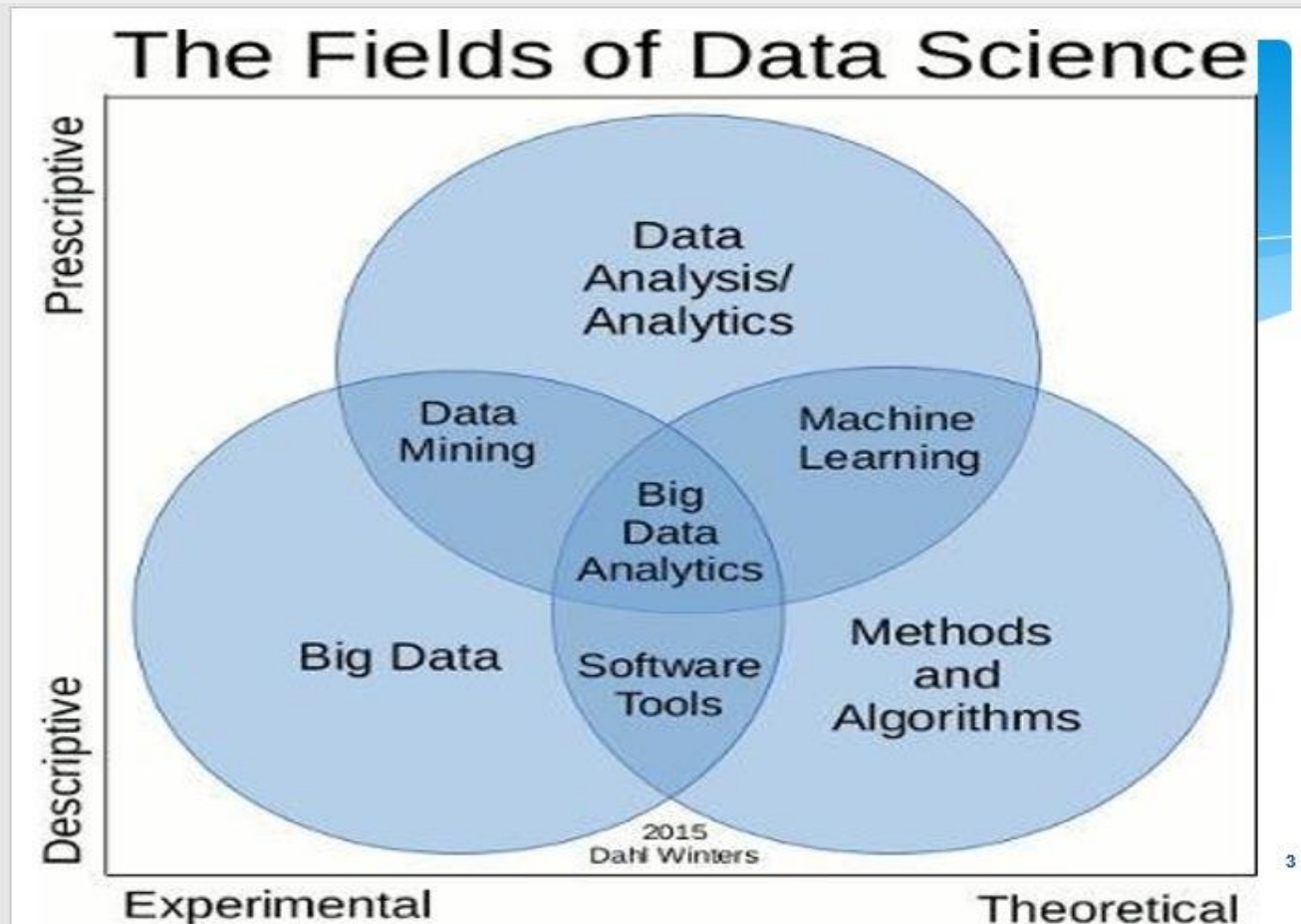
Multidisciplinary field that uses **scientific methods, processes, algorithms and systems** to extract **knowledge and insights** from data in various forms, **both structured and unstructured**

Data Scientist

“Someone who can **obtain, scrub, explore, model and interpret data, blending hacking, statistics and machine learning**”

-- Hilary Mason, chief scientist at bit.ly

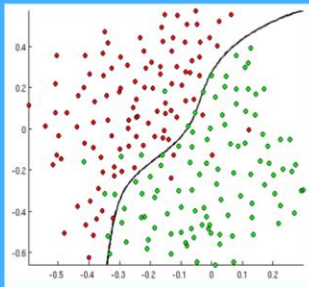
Data Science & Machine Learning?



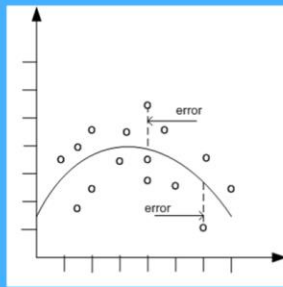
Data Science & Machine Learning?

Advanced Analytics Questions Answered by "Data Science"

Classification



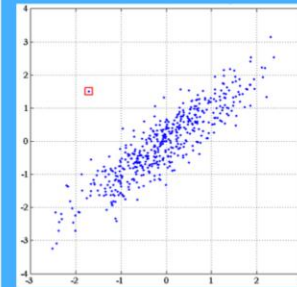
Regression (Forecasting)



Reinforcement Learning



Anomaly Detection



Clustering



Is it A or B?

e.g. Will the HDD fail next month? Yes/No.

How much/how many?

e.g. what is the temperature next Tuesday?

Which option?

e.g. do the car stop or go on an orange light.

Is it weird?

e.g. Fraud Detection.

Which groups? e.g. which viewers like the same type of movie.

Machine Learning algorithms are used in Advanced Analytics by Data Scientists.

Data Science & Machine Learning?

Data is cheap and abundant,
knowledge is expensive and scarce.

'Where' is Machine Learning in Data Science?

Business Understanding

Getting more data

- Asking the sharp question.
- Determining the base ML algorithms.
- Sourcing and Selecting data sources.
- Determining the features and label.

Data Understanding

Importing Data.

Exploring & Visualizing Data

- Summary statistics.
- Univariate plotting.
 - Histogram.
 - Box Plot.
 - Histograms (categorical).
- Plotting 2 Variables.
 - Scatter plots.
- Scatter plot matrices.
- Faceting scatter plots.
- Conditional histogram plots.
- Joining data into a single table.
 - SQL.
 - Programming R, Python.
- Feature Selection – numeric features
 - Features correlated with label.
 - Remove redundant collinear features
 - Feature importance.

Data Preparation

Cleansing:

- Missing value.
- Removing Repeated value.

Handling Error and Outliers:

- Censor
- Trim.
- Interpolate.
- Substitute.

Transforming data:

- Transform string to categorical.
- Group/aggregate categories.
- Quantizing continuous variables.
- Convert to indicator values.
- Rename features if required.
- Text Frequency Hashing.

Scale numeric features:

- Z-Score scaling.
- Min-max scaling.
- De-trend data.

Feature Engineering:

- Feature Creation:
 - Log of features
 - Feature difference/Differencing.
 - Add features.
 - Feature multiplication
- Refining Feature Selection
 - Removing marginal impact features.

Handle imbalance categorical labels for classification.

- Coding R/Python.
- SMOTE

Modeling

Classification (A or B)

- Logistic regression.
- Two-Class Boosted Decision Trees.
- Two-Class Decision Forest.
- Two-Class Neural Network.
- Two-Class [Locally-Deep] SVM

Regression (How many)

- K-Nearest Neighbors.
- Linear regression.
- Boosted Decision Tree Regression.
- Forest Regression.
- Ridge Regression.
- AdaBoost

Neural Net (Reinforcement Learning)

Clustering (Which Group)

- K-Means.
- Hierarchical Agglomerative Clustering.
- Recommenders.

Evaluation

Classification (A or B)

- Accuracy.
- AOC.
- Confusion matrix.
- Recall.
- Feature interpretation.
- Tuning/Sweeping Model Parameters.
- Cross validation.
- Nested cross validation to tune parameters.
- Comparison between models.

Regression (How many)

- Root Mean Square Error.
- Standard Error (ARMA).
- Coefficient of determination.
- Residual visualization:
 - Time series plots of actual versus predicted.
 - Box-Plot
 - Time series plots of residuals (line, histograms) .

- Regularization.
- Tuning/Sweeping Model Parameters.
- Cross validation.
- Nested cross validation to tune parameters.
- Comparison between models.

Clustering (Which Group)

- Maximal distance to cluster center.
- Principal Component Analysis (PCA)

Deployment

Deployment :

- Deployment lifecycle.
- Web Services APIs
- Web Interface.
- [Deployment coding.]

Advanced Analytics:

- Business Intelligence.
- Predictive Analytics.

Artificial Intelligence:

- AI Applications.
- Cognitive Services.

MACHINE LEARNING

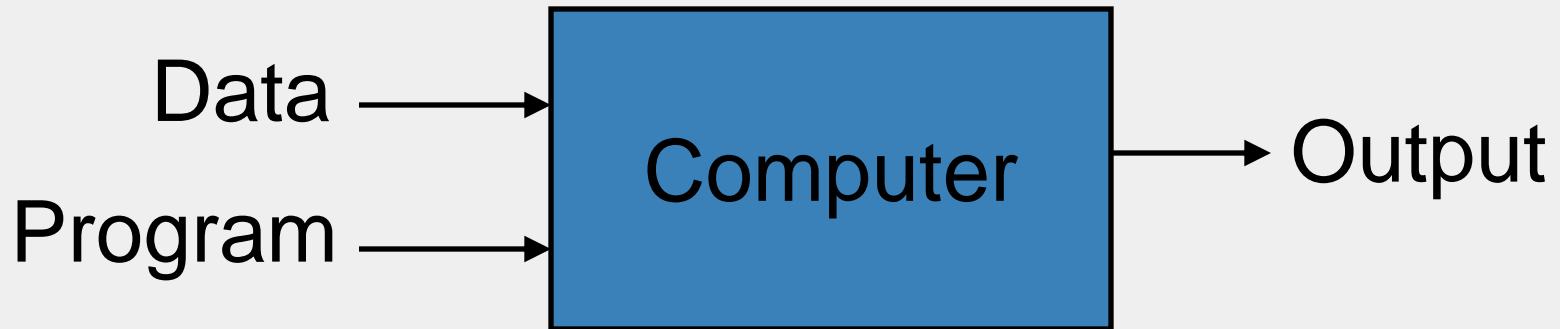
2. What is Machine Learning?



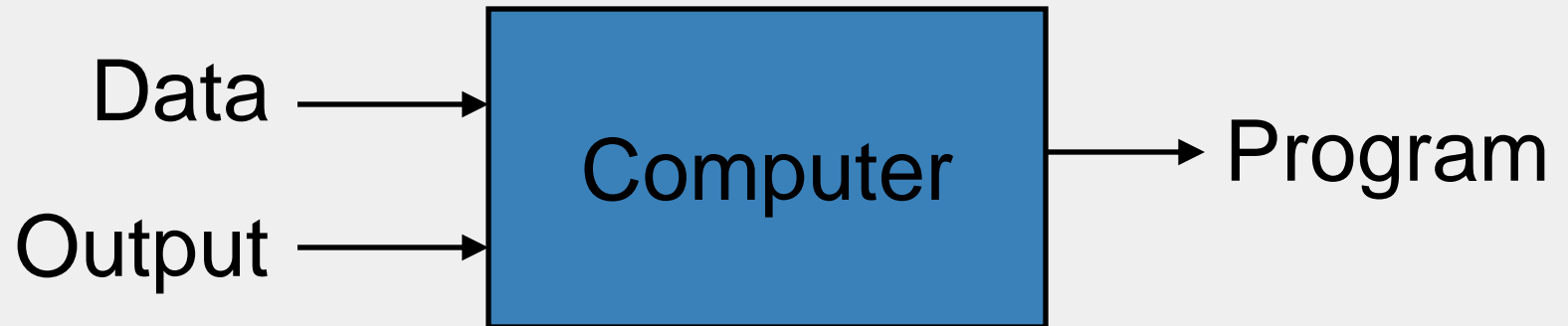
Generally and specifically

Machine Learning

Traditional Programming



Machine Learning



Applications of Machine Learning

- ***Database mining***

Large datasets from growth of automation/web.

E.g., Web click data, medical records, biology, engineering

- ***Applications & automation***

E.g., Autonomous helicopter, handwriting recognition, most of Natural Language Processing (NLP)

- ***Self-customizing programs***

E.g., Amazon, Netflix product recommendations

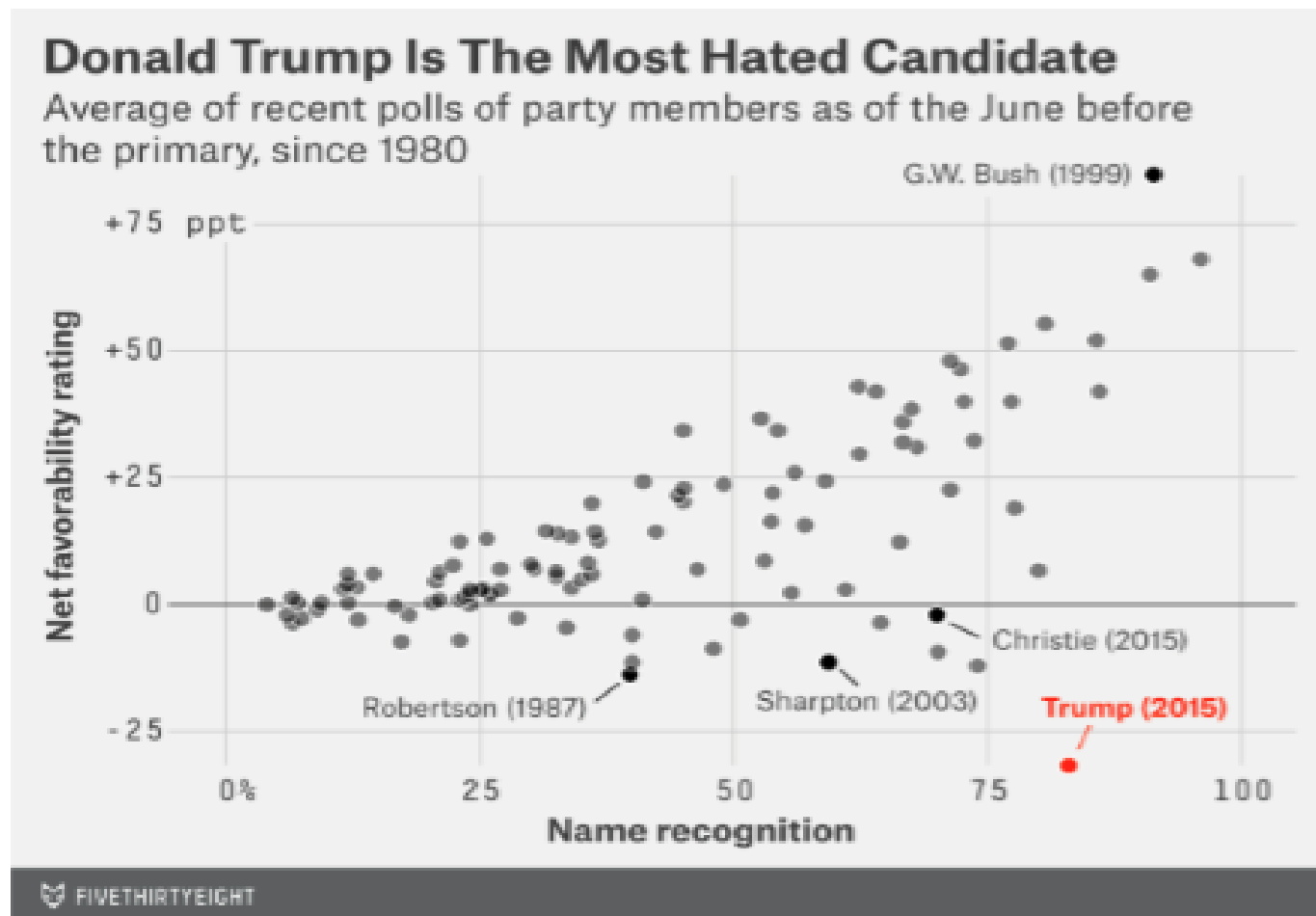
- ***Understanding human learning*** (robotics).

Applications of Machine Learning



ELECTION 2016

“Why Donald Trump Isn’t A Real Candidate, In One Chart” (June 2015)



Definition of Machine learning

■ Arthur Samuel (1959). Machine Learning:

*Field of study that gives **computers** the ability to **learn** without being explicitly programmed.*

■ Tom Mitchell (1998)

Well-posed Learning Problem:

*A computer program is said to learn from **experience E** with respect to some **task T** and some **performance** measure **P**, if its performance on T, as measured by P, improves with experience E.*

#Checkpoint

“A computer program is said to learn from experience E with respect to some task T and some performance measure P , if its performance on T , as measured by P , improves with experience E .”

Suppose your email program watches which emails you do or do not mark as spam, and based on that learns how to better filter spam. What is the task T in this setting?

- ☐ Classifying emails as spam or not spam.
- ☐ Watching you label emails as spam or not spam.
- ☐ The number (or fraction) of emails correctly classified as spam/not spam.
- ☐ None of the above—this is not a machine learning problem.

Machine learning in a nutshell

- Every machine learning algorithm has three components:

Representation – decision trees, rules, neural networks, ensemble methods etc.

Evaluation – precision, recall, AUC, likelihood etc.

Optimization –

- Combinatorial optimization

E.g.: Greedy search

- Convex optimization

E.g.: Gradient descent

- Constrained optimization

E.g.: Linear programming

Types of Learning Algorithms

Machine learning algorithms:

- ☐ Supervised (inductive) learning (task driven)
 - ☐ Training data with desired outputs
 - ☐ Prediction
 - ☐ Classification (discrete labels), Regression (real values)

- ☐ Unsupervised learning (data driven)
 - Training data without desired outputs
 - Clustering
 - Probability distribution estimation
 - Finding association (in features)
 - Dimension reduction

Types of Learning Algorithms

Machine learning algorithms:

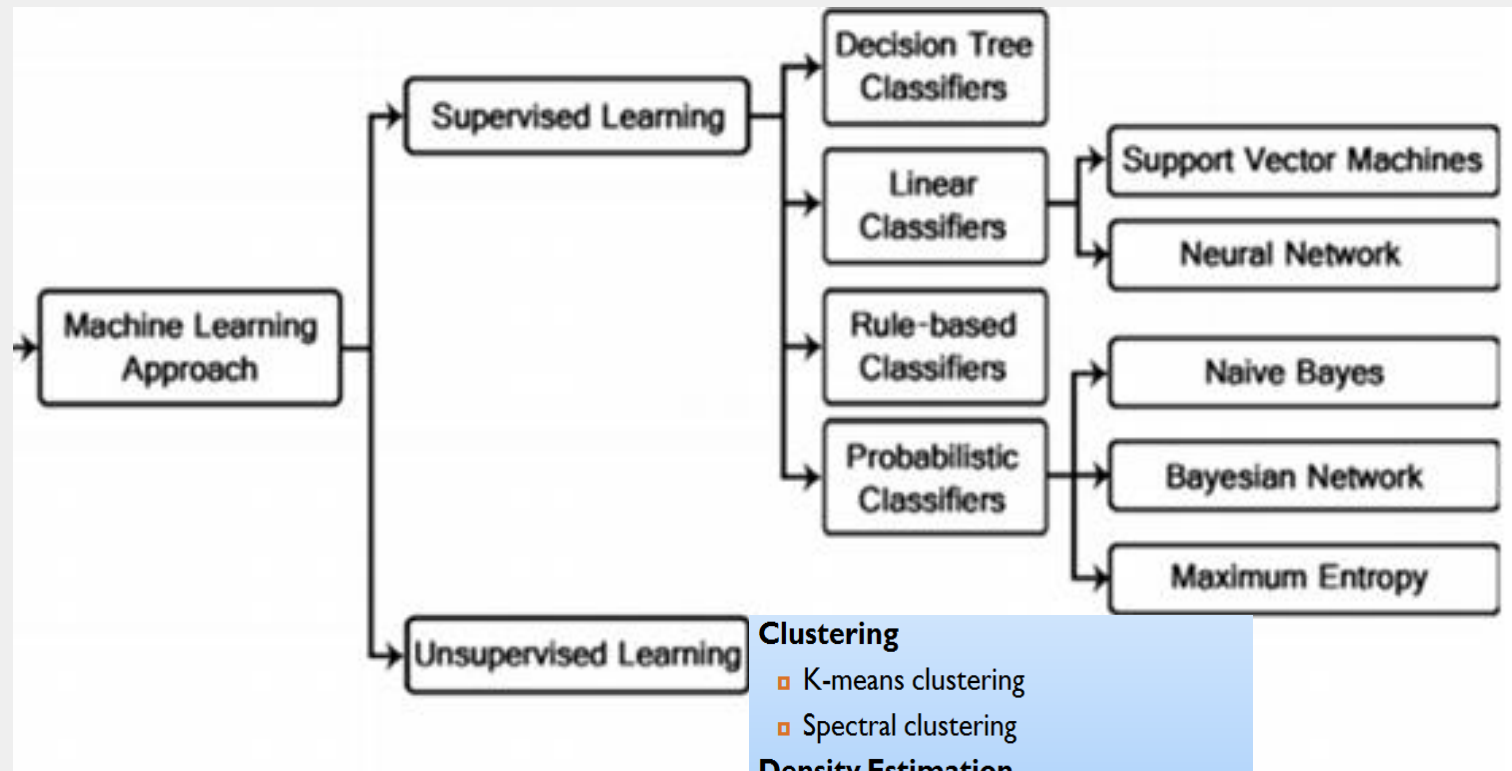
- ☐ Semi-supervised

- ☐ Training data with a few desired outputs

- ☐ Others:

- ☐ Reinforcement learning (delayed reward), decision making (robot, chess game)
- ☐ Recommender systems.

Types of Learning Algorithms



3. Introduction to Supervised Learning



Supervised?

Supervised Learning and Regression

Given examples of a function $(X, F(X))$

Predict function $F(X)$ for new examples X

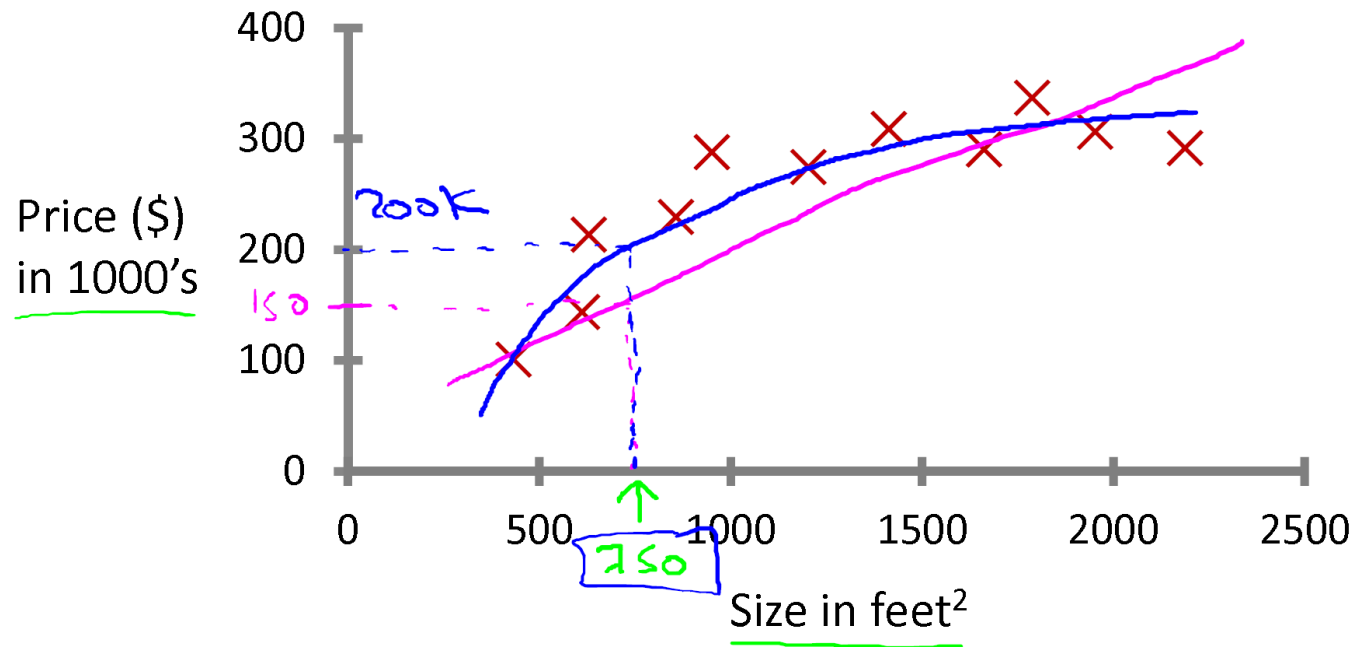
Discrete $F(X)$: Classification

Continuous $F(X)$: Regression

$F(X) = \text{Probability}(X)$: Probability estimation

Supervised Learning and Regression

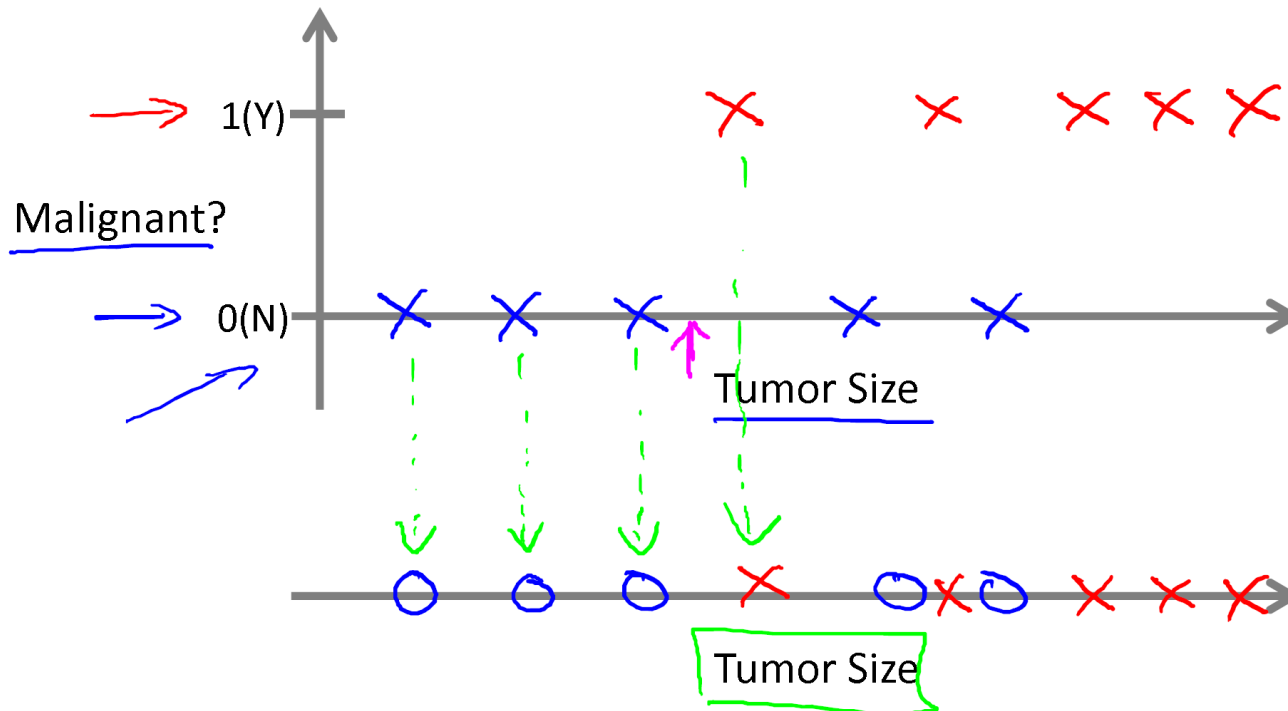
Housing price prediction.



Regression: Predict *continuous* valued output (price)

Supervised Learning for Classification

Breast cancer (malignant, benign)



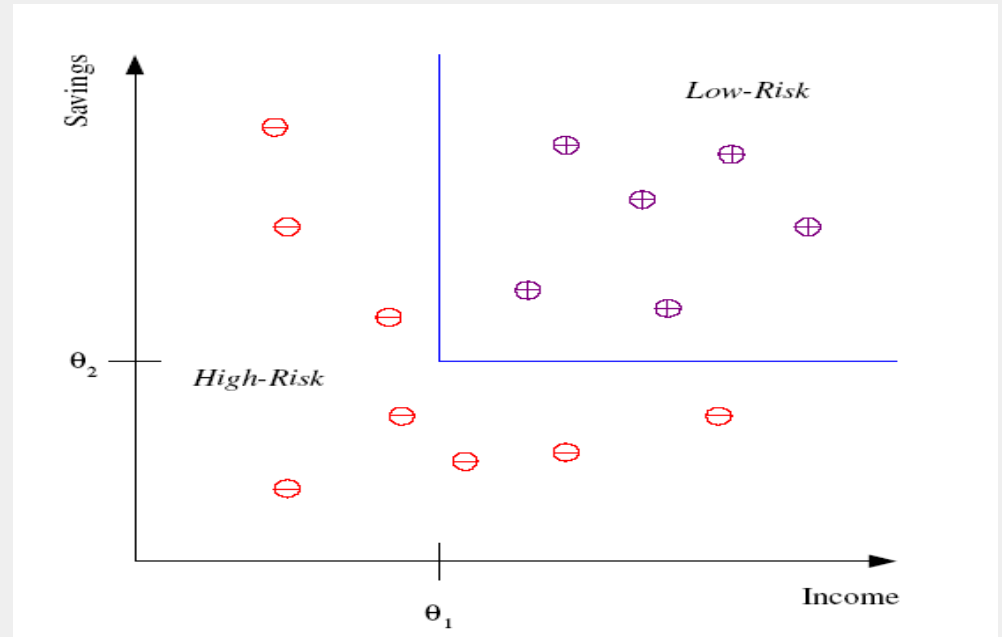
Classification

Discrete valued output (0 or 1)

0, 1, 2, 3
↓
benign type 1
cancer

Supervised Learning for Classification

- Example: Credit scoring
- Differentiating between **low-risk** and **high-risk** customers from their *income* and *savings*



Discriminant: IF $income > \theta_1$ AND $savings > \theta_2$
THEN **low-risk** ELSE **high-risk**

Model

#Checkpoint

You're running a company, and you want to develop learning algorithms to address each of two problems.

- ❑ Problem 1: You have a large inventory of identical items. You want to predict how many of these items will sell over the next 3 months.
- ❑ Problem 2: You'd like software to examine individual customer accounts, and for each account decide if it has been hacked/compromised.

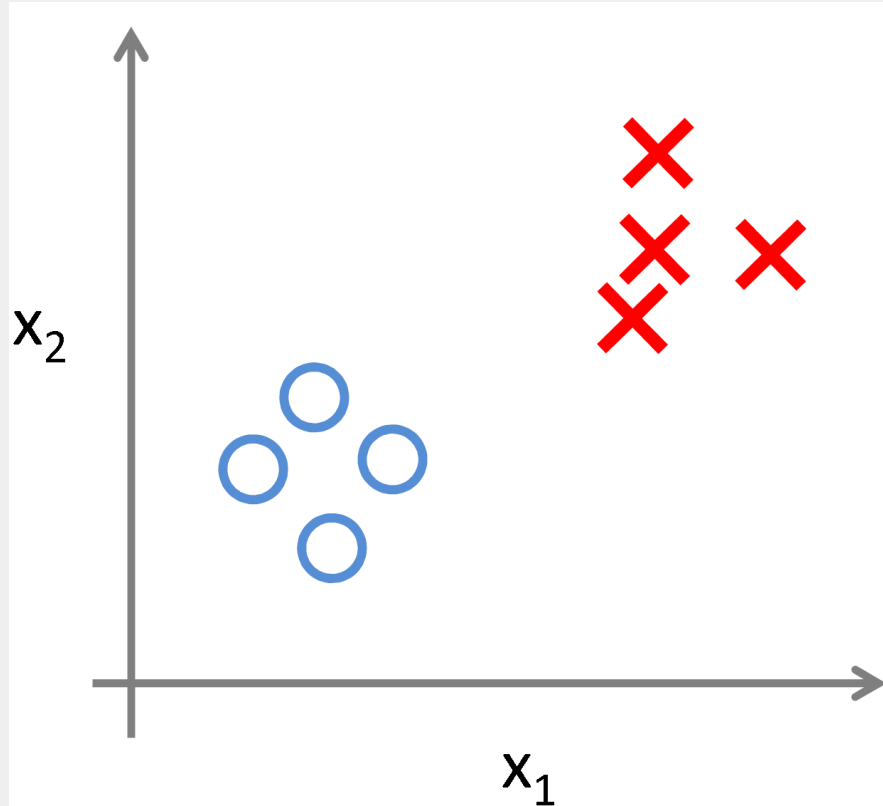
Should you treat these as classification or as regression problems?

4. Introduction to Unsupervised Learning

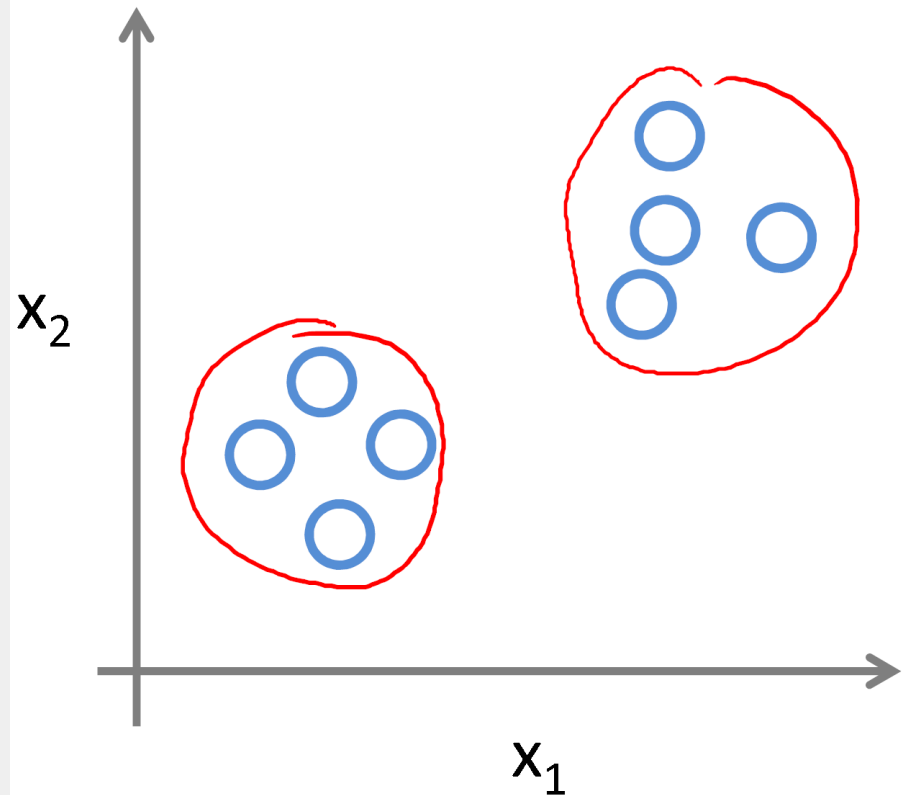


Unsupervised?

Supervised Learning vs Unsupervised Learning

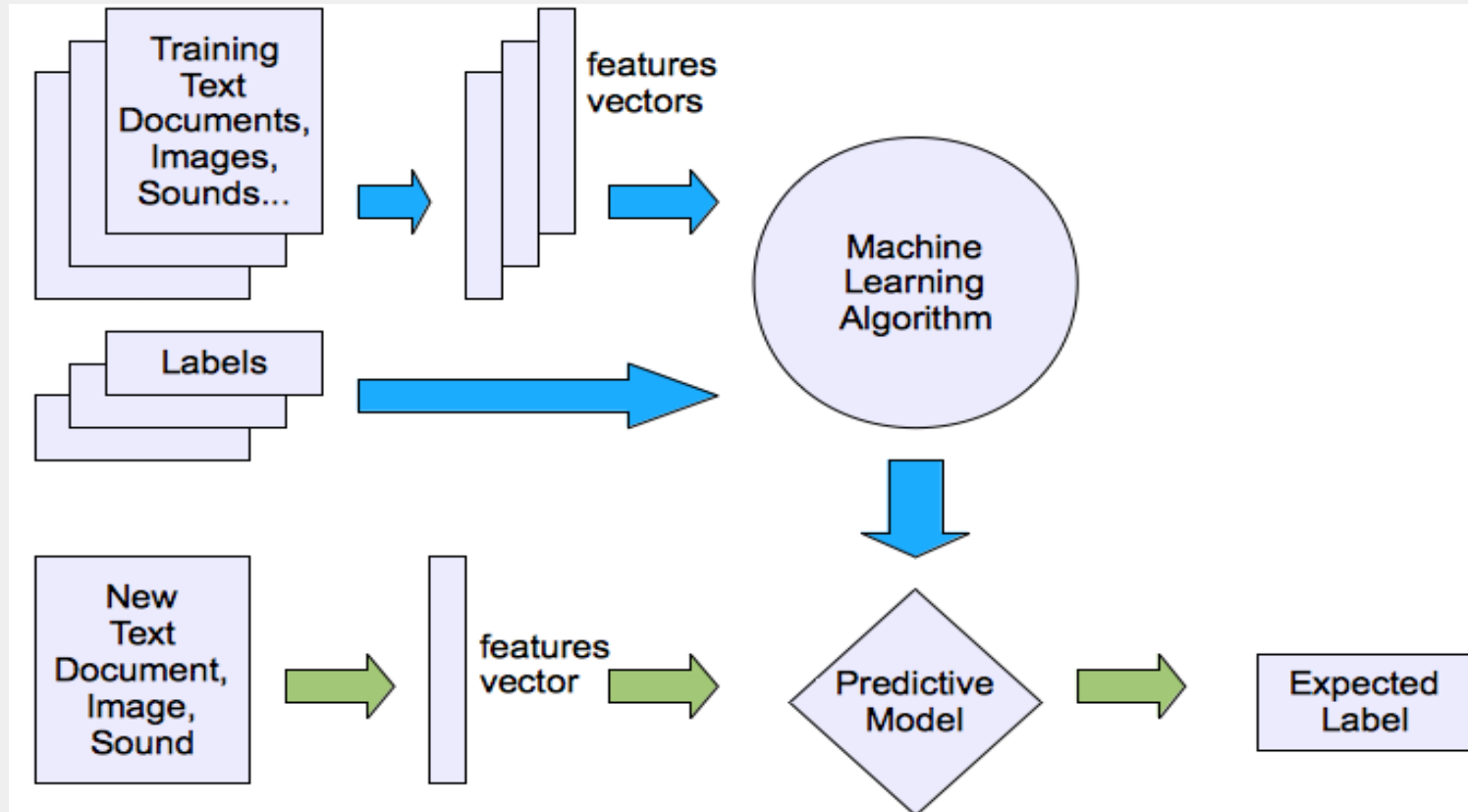


Supervised Learning



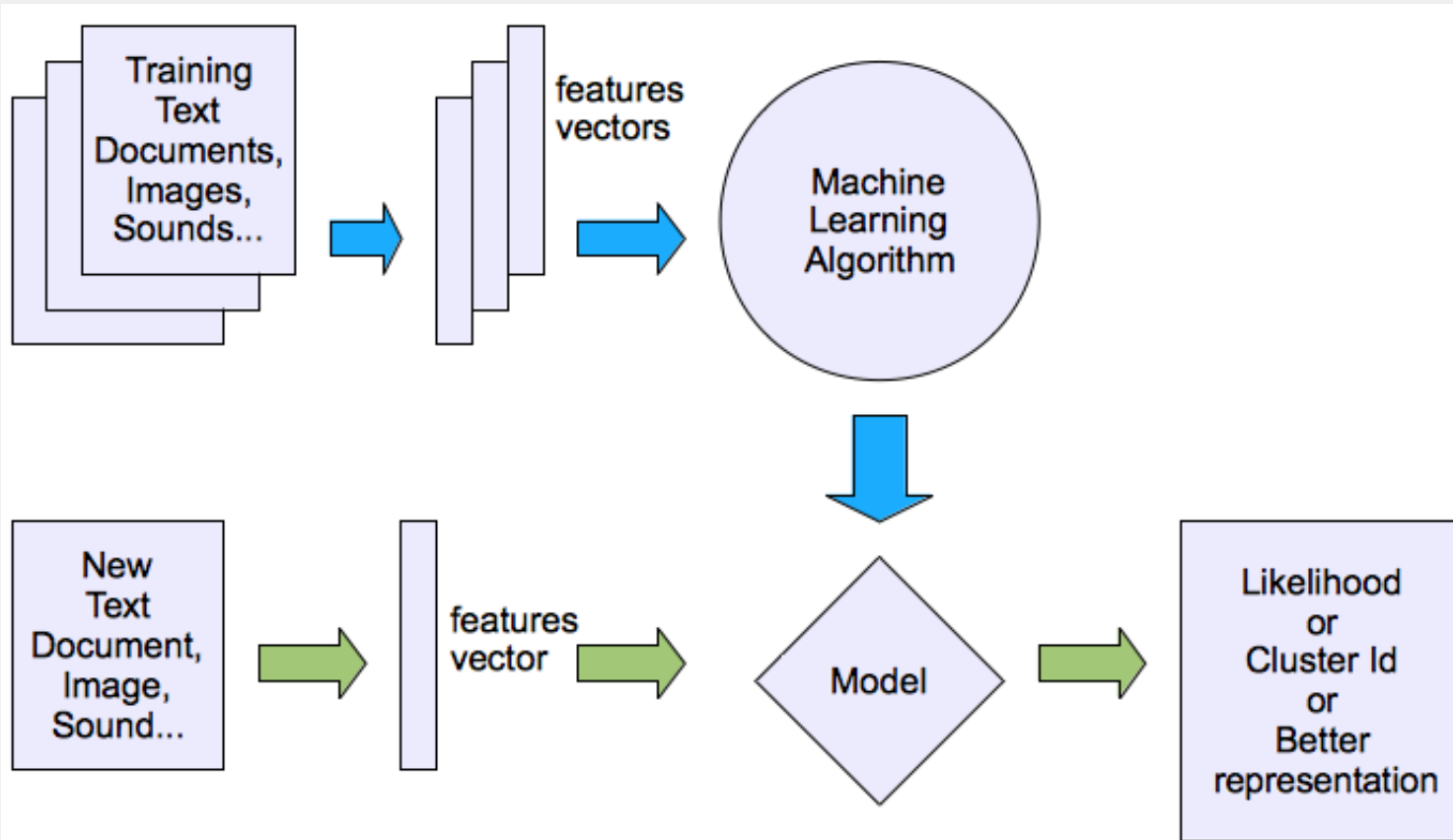
Unsupervised Learning

Supervised Learning



We seek low E-out or maximize probabilistic terms

Unsupervised Learning

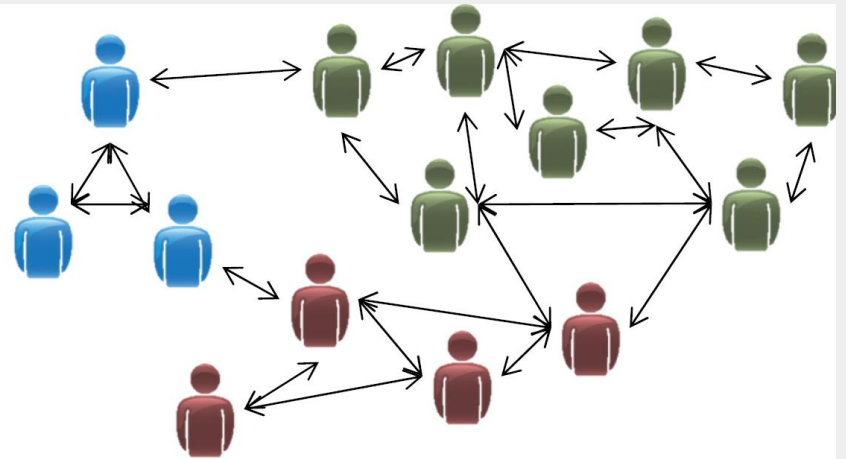


We seek minimum distance, MLE(maximum likelihood estimation) etc.

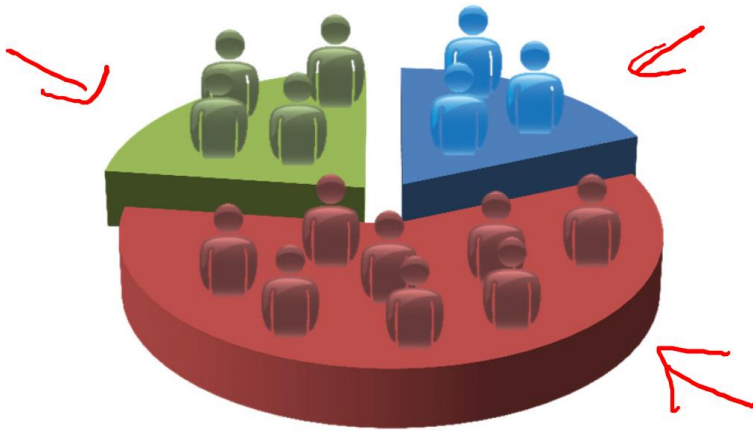
Sample Applications for Unsupervised Learning



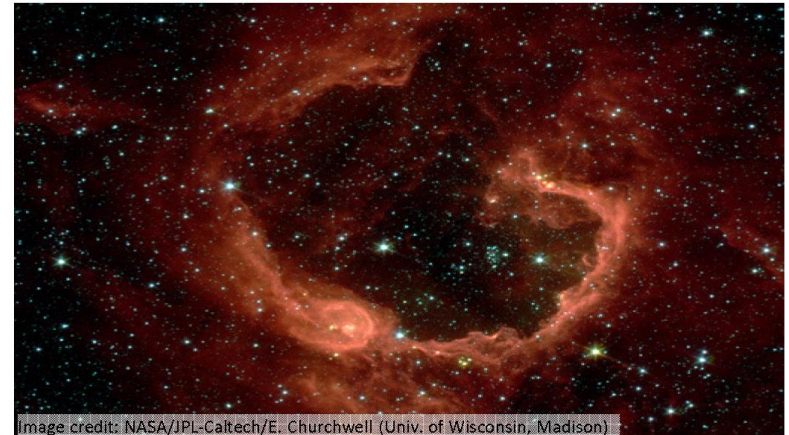
Organize computing clusters



Social network analysis



Market segmentation



Astronomical data analysis

#Checkpoint

Of the following examples, which would you address using an unsupervised learning algorithm?

- ☐ Given a database of customer data, automatically discover market segments and group customers into different market segments.
- ☐ Given email labeled as spam/not spam, learn a spam filter.
- ☐ Given a set of news articles found on the web, group them into set of articles about the same story.
- ☐ Given a dataset of patients diagnosed as either having diabetes or not, learn to classify new patients as having diabetes or not.