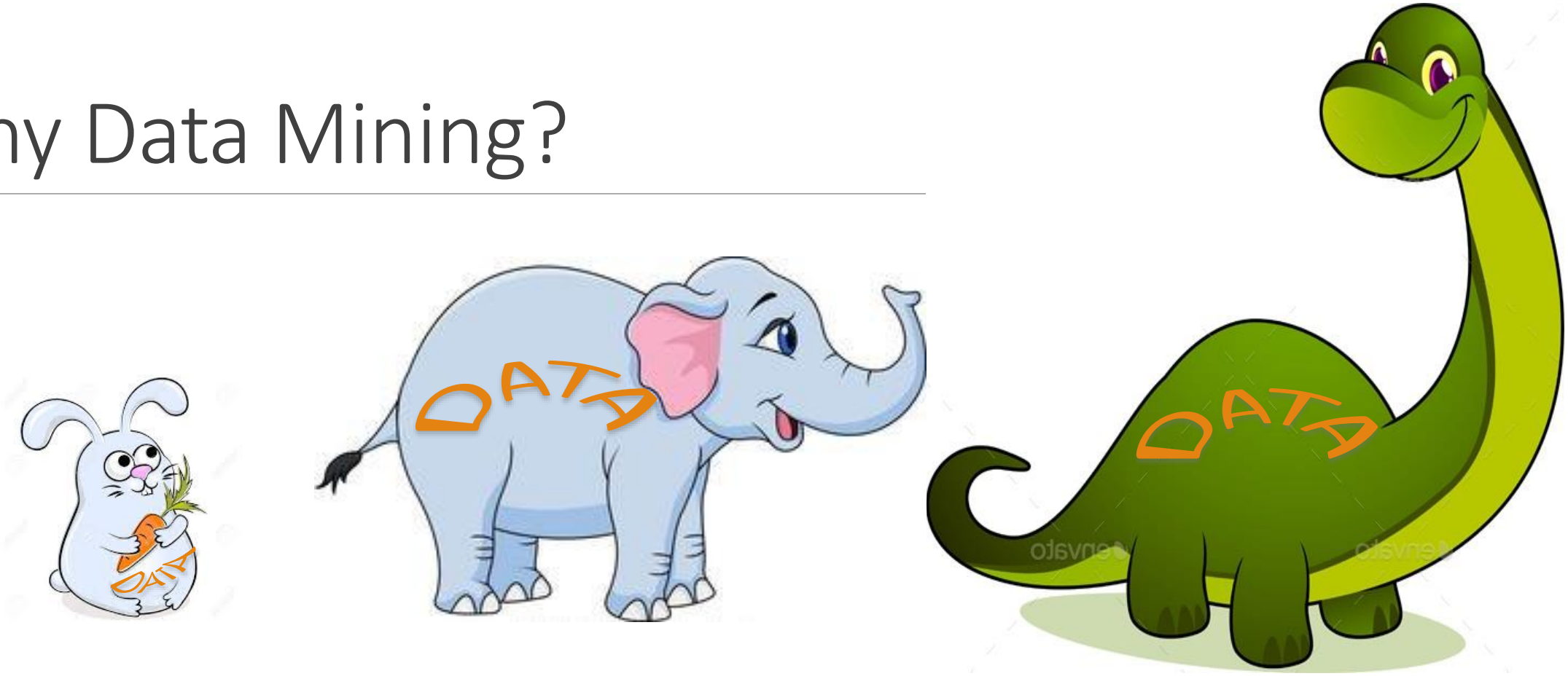
Introduction to machine learning

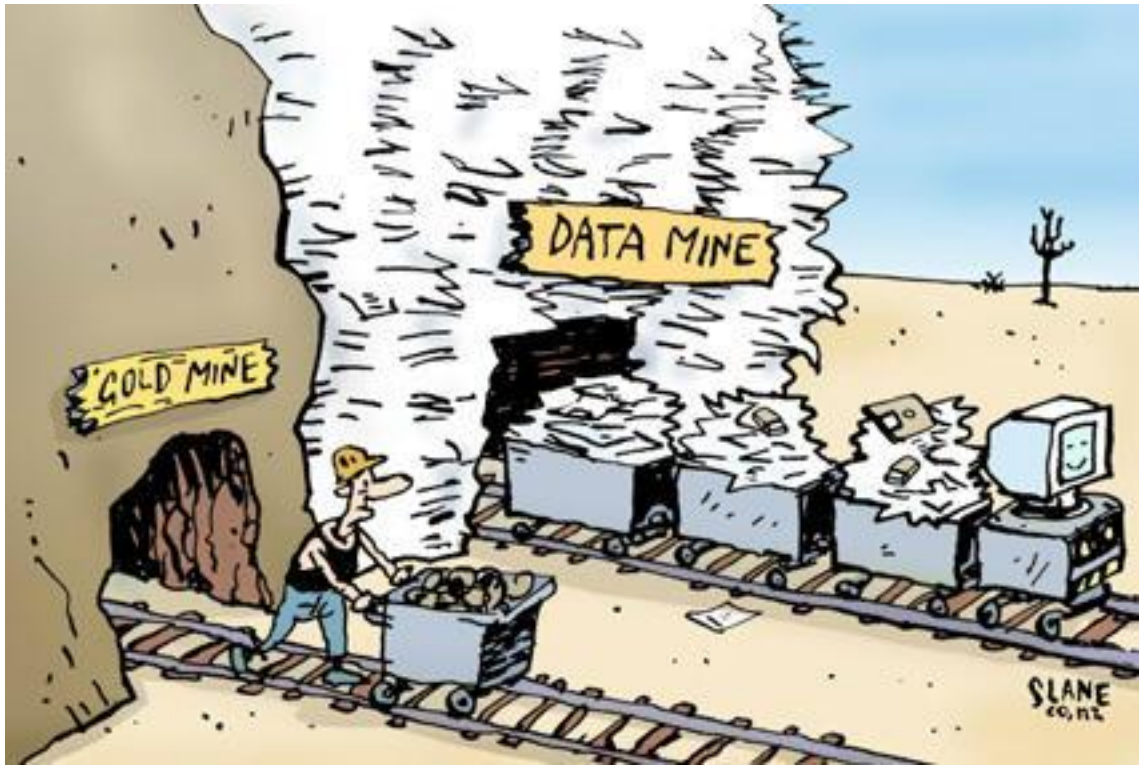
Why Data Mining?



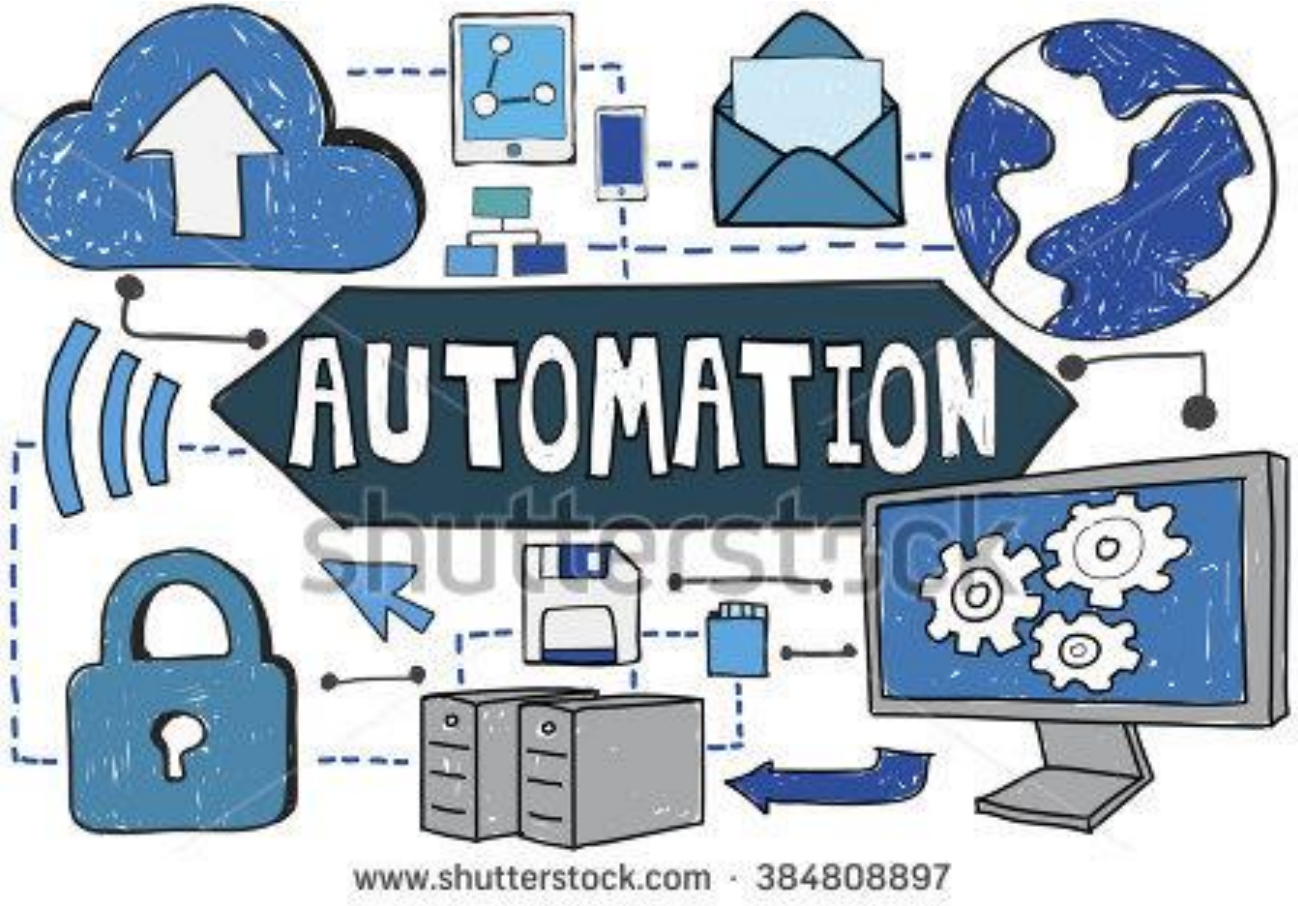
BIG AND BIGGER DATA

Why Data Mining?

Data is cheap and abundant (data warehouses, data marts);
knowledge is expensive and scarce.



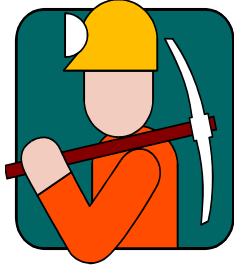
Why Data Mining?



Data mining—

Automated analysis of massive data sets

What Is Data Mining?



Data mining (knowledge discovery from data)

- Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
- Data mining: a misnomer?

Alternative names

- Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

Watch out: Is everything “data mining”?

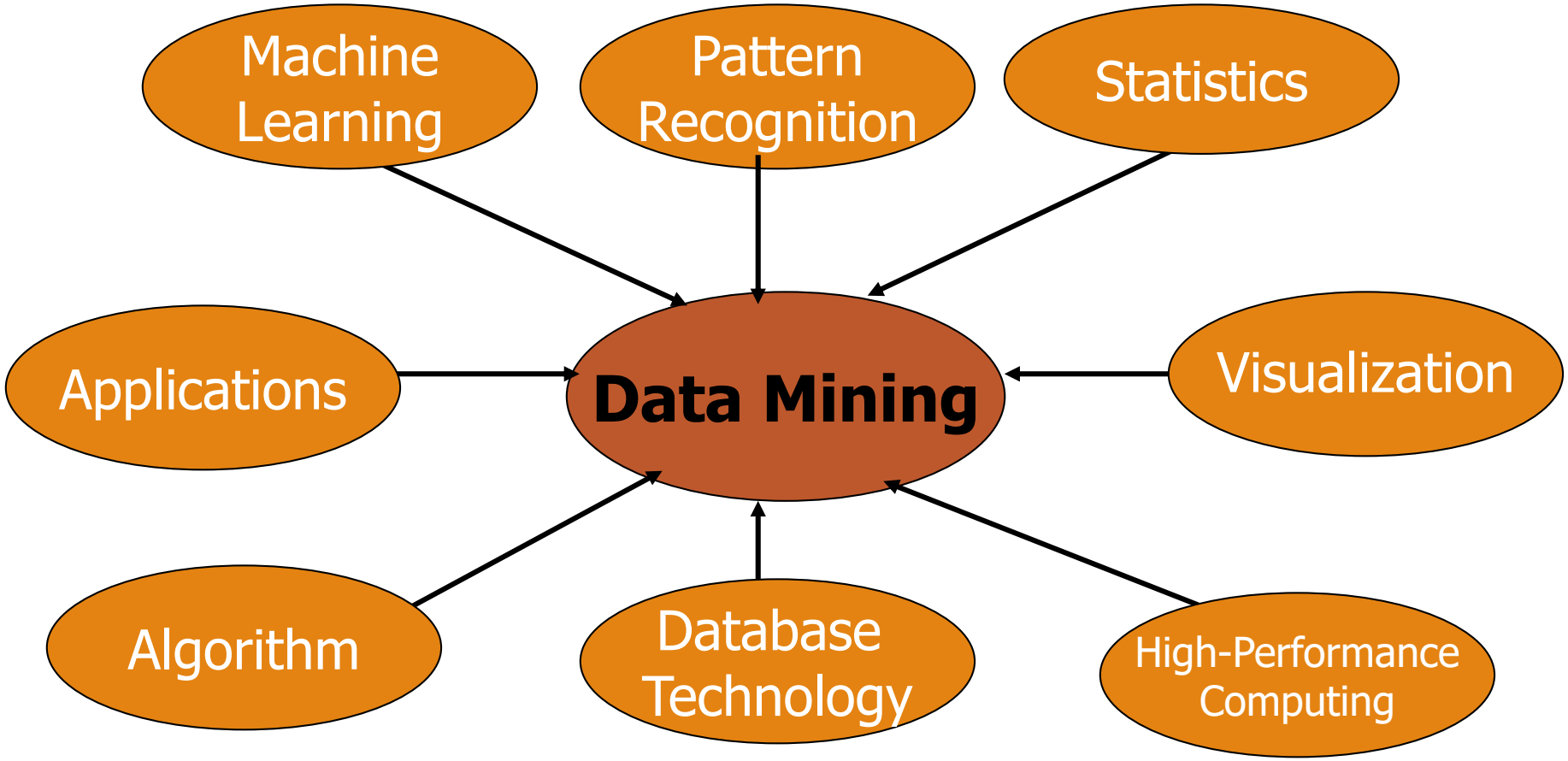
- Simple search and query processing
- (Deductive) expert systems

Examples: What is (not) Data Mining?

- **What is not Data Mining?**
  - Look up phone number in phone directory
  - Query a Web search engine for information about "Amazon"

- **What is Data Mining?**
  - Certain names are more prevalent in certain US locations (O'Brien, O'Rurke, O'Reilly… in Boston area)
  - Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,)

Data Mining: Confluence of Multiple Disciplines



Machine Learning in Data Mining Process

What is machine learning?

- Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.
- **Machine learning focuses on the development of computer programs** that can access data and use it learn for themselves.
- Typically, machine learning used to analyse data in data mining

What is Machine Learning?

Optimize a performance criterion using example data or past experience.

Role of Statistics: Inference from a sample

Role of Computer science: Efficient algorithms to

- Solve the optimization problem
- Representing and evaluating the model for inference

Machine learning and our focus

Like human learning from past experiences.

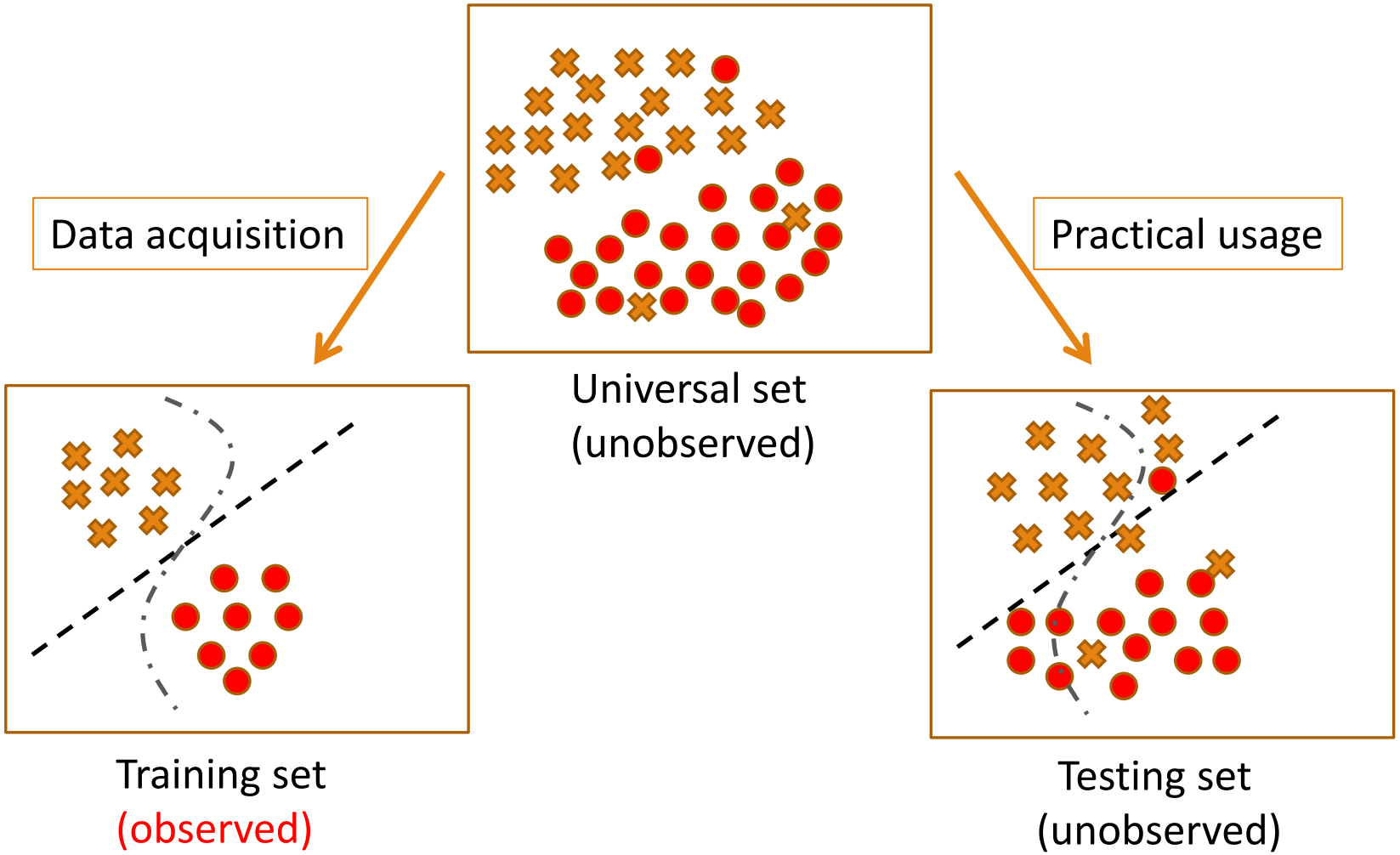
A computer does not have “experiences”.

A computer system learns from data, which represent some “past experiences” of an application domain.

Our focus: learn a target function that can be used to predict the values of a discrete class attribute, e.g., approve or not-approved, and high-risk or low risk.

The task is commonly called: Supervised learning, classification, or inductive learning.

Training and testing



Performance

There are several factors affecting the performance:

- **Types of training** provided
- The form and extent of any initial **background knowledge**
- The **type of feedback** provided
- The **learning algorithms** used

Two important factors:

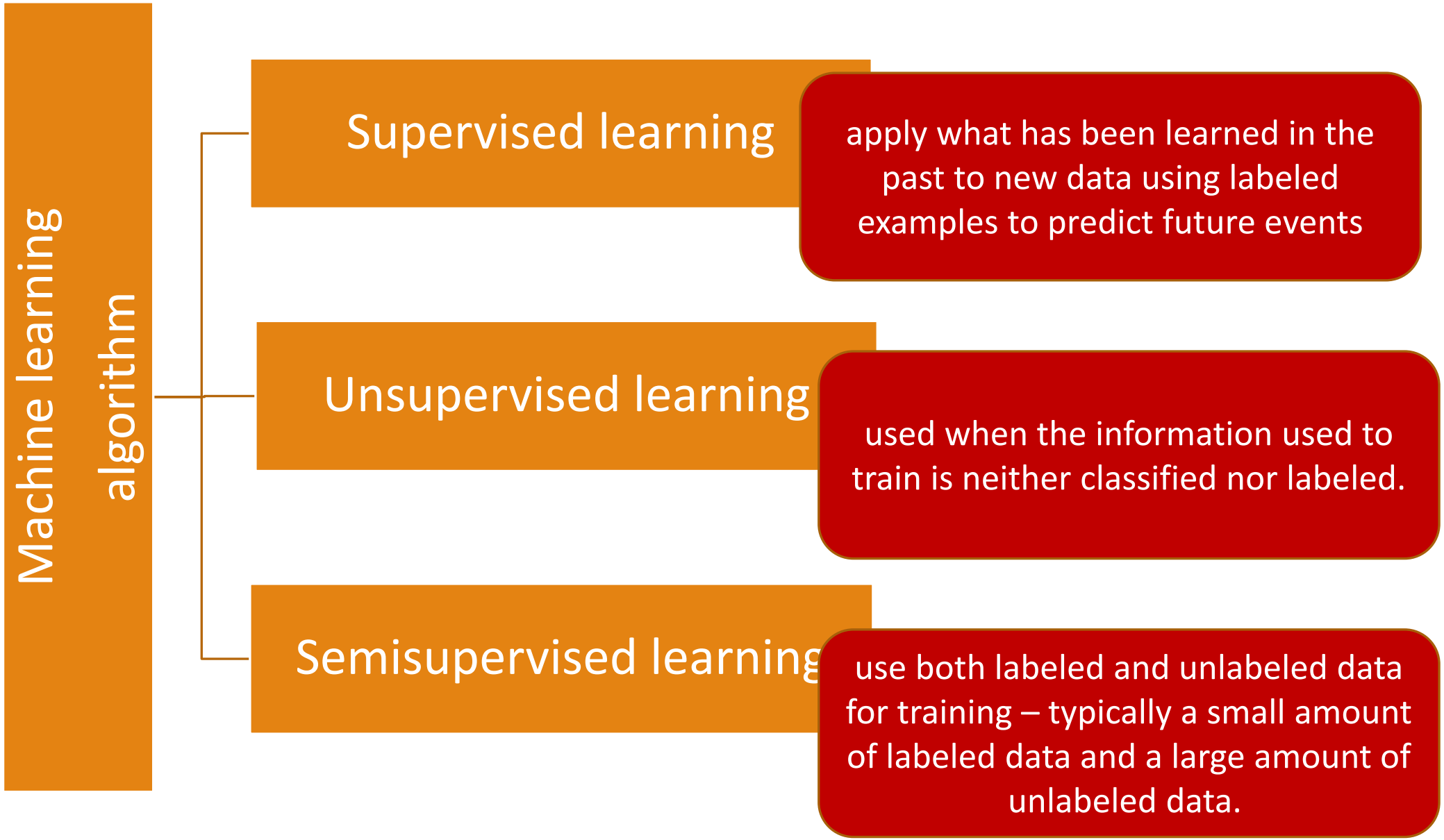
- Modeling
- Optimization

Algorithms

The success of machine learning system also depends on the algorithms.

The algorithms control the search to find and build the knowledge structures.

The learning algorithms should extract useful information from training examples.


Machine learning algorithm
Supervised learning
apply what has been learned in the past to new data using labeled examples to predict future events
Unsupervised learning
used when the information used to train is neither classified nor labeled.
Semisupervised learning
use both labeled and unlabeled data for training – typically a small amount of labeled data and a large amount of unlabeled data.

Machine Learning Algorithms *(sample)*

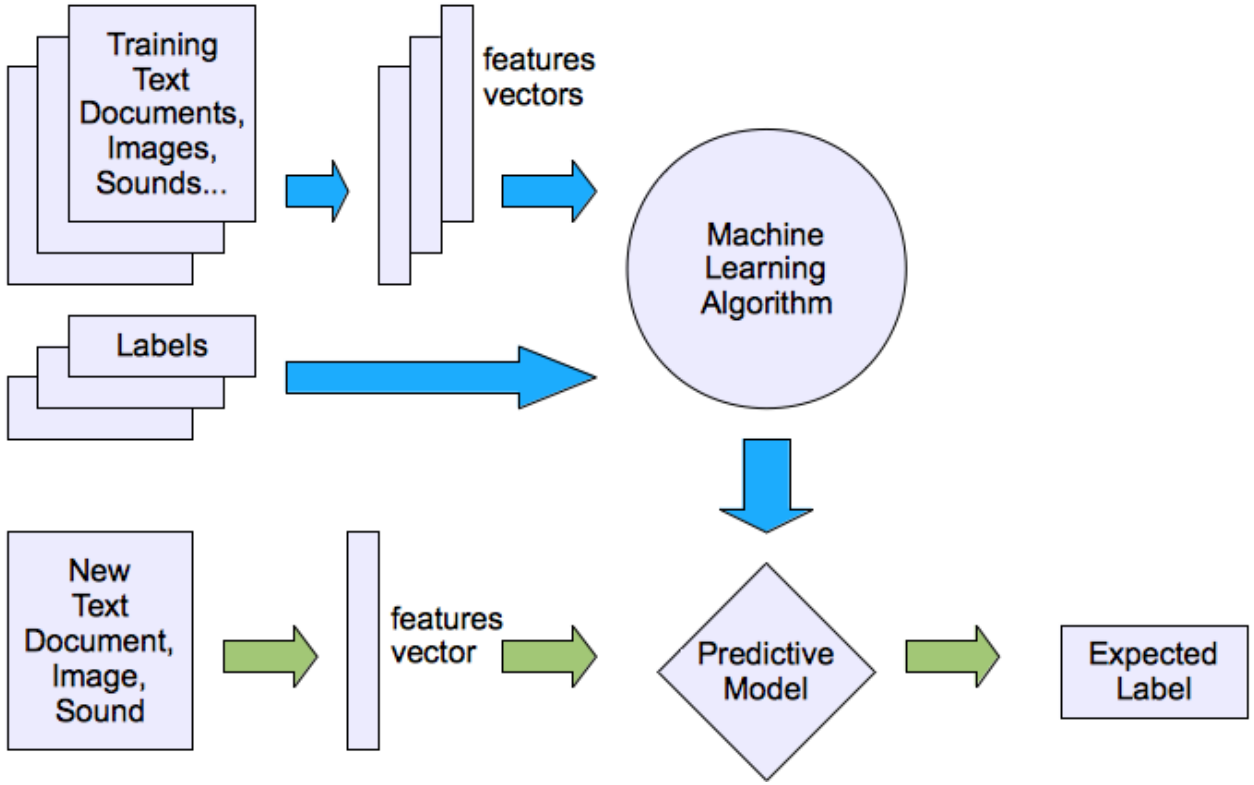
	Unsupervised	Supervised

Continuous	• Clustering & Dimensionality Reduction
	  ○ SVD
	  ○ PCA
	  ○ K-means
		• Regression
		  ○ Linear
		  ○ Polynomial
		• Decision Trees
		• Random Forests

Categorical	• Association Analysis
	  ○ Apriori
	  ○ FP-Growth
	• Hidden Markov Model
		• Classification
		  ○ KNN
		  ○ Trees
		  ○ Logistic Regression
		  ○ Naive-Bayes
		  ○ SVM

Machine learning structure

Supervised learning



What do we mean by learning?

Given

- a data set *D*,
- a task *T*, and
- a performance measure *M*,

a computer system is said to **learn** from *D* to perform the task *T* if after learning the system's performance on *T* improves as measured by *M*.

In other words, the learned model helps the system to perform *T* better as compared to no learning.

Fundamental assumption of learning

Assumption: The distribution of training examples is identical to the distribution of test examples (including future unseen examples).

In practice, this assumption is often violated to certain degree.

Strong violations will clearly result in poor classification accuracy.

To achieve good accuracy on the test data, training examples must be sufficiently representative of the test data.

An example: data (loan application)

Approved or not

ID	Age	Has_Job	Own_House	Credit_Rating	Class
1	young	false	false	fair	No
2	young	false	false	good	No
3	young	true	false	good	Yes
4	young	true	true	fair	Yes
5	young	false	false	fair	No
6	middle	false	false	fair	No
7	middle	false	false	good	No
8	middle	true	true	good	Yes
9	middle	false	true	excellent	Yes
10	middle	false	true	excellent	Yes
11	old	false	true	excellent	Yes
12	old	false	true	good	Yes
13	old	true	false	good	Yes
14	old	true	false	excellent	Yes
15	old	false	false	fair	No

An example

Data: Loan application data

Task: Predict whether a loan should be approved or not.

Performance measure: accuracy.

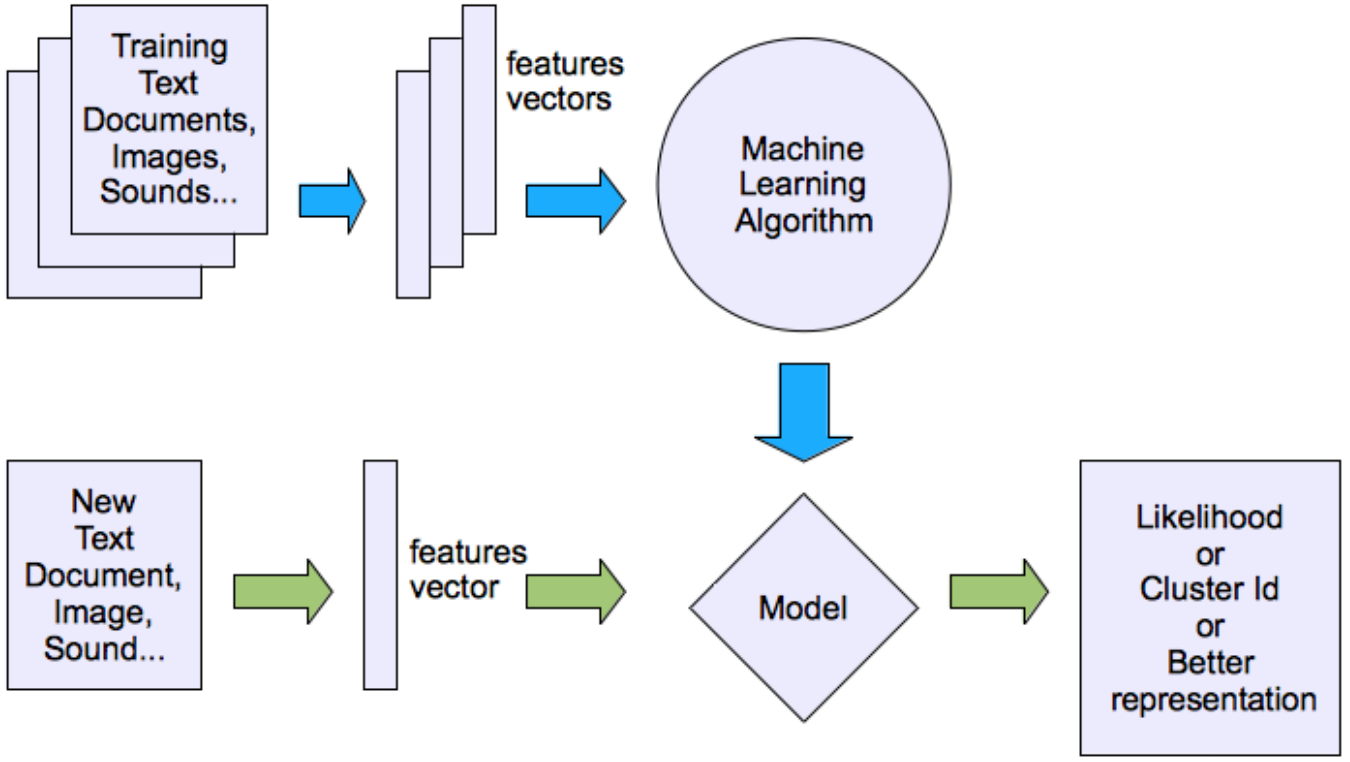
No learning: classify all future applications (test data) to the majority class (i.e., Yes):

Accuracy = 9/15 = 60%.

We can do better than 60% with learning.

Machine learning structure

Unsupervised learning



Examples: Applications of Machine Learning

Banking: loan/credit card approval

- predict good customers based on old customers

Customer relationship management:

- identify those who are likely to leave for a competitor.

Targeted marketing:

- identify likely responders to promotions

Fraud detection: telecommunications, financial transactions

- from an online stream of event identify fraudulent events

Manufacturing and production:

- automatically adjust knobs when process parameter changes

Examples: …(continued)

Medicine: disease outcome, effectiveness of treatments

- analyze patient disease history: find relationship between diseases

Molecular/Pharmaceutical: identify new drugs

Scientific data analysis:

- identify new galaxies by searching for sub clusters

Web site/store design and promotion:

- find affinity of visitor to pages and modify layout