Discussion 11
Data Mining & Machine Learning I

Weijie Lyu

Data Warehouse:

Collects and organizes historical data from multiple sources

Data is periodically ETLed into the data warehouse:

- **Extract:** Extracting data from multiple remote sources
- **Transform:** Data cleaning and transforming them into a proper storage format for querying and analysis
- **Loaded**: Insertion of data into the final target database (Data Warehouse)

Data Lake

- Like data warehouse, but without ETL
  - Store unstructured data in raw format
  - Schema-on-Read: determine the best organization when **data is used**
- Beware of data swamp!
  - Save everything
  - Collect a rich history of **dirty data**

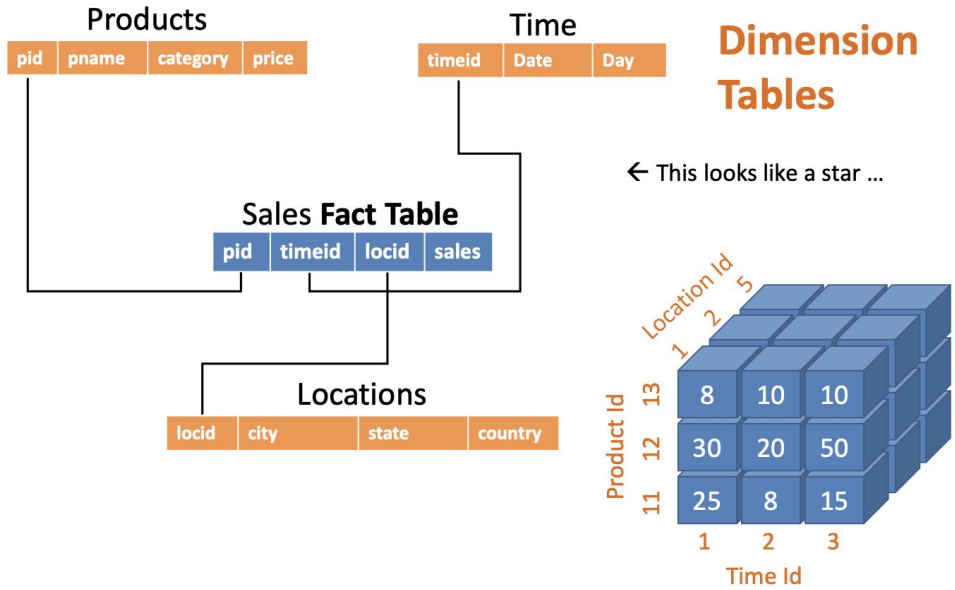
Example Sales Data:

pname	category	price	qty	date	day	city	state	country
Corn	Food	25	25	3/30/16	Wed.	Omaha	NE	USA
Corn	Food	25	8	3/31/16	Thu.	Omaha	NE	USA
Corn	Food	25	15	4/1/16	Fri.	Omaha	NE	USA
Galaxy 1	Phones	18	30	1/30/16	Wed.	Omaha	NE	USA
Galaxy 1	Phones	18	20	3/31/16	Thu.	Omaha	NE	USA
Galaxy 1	Phones	18	50	4/1/16	Fri.	Omaha	NE	USA
Galaxy 1	Phones	18	8	1/30/16	Wed.	Omaha	NE	USA
Peanuts	Food	2	45	3/31/16	Thu.	Seoul		Korea
Galaxy 1	Phones	18	100	4/1/16	Fri.	Seoul		Korea

- **Big** table: many *columns* and *rows*
  - Substantial redundancy → expensive to store and access
- Could we organize the data a little better?

Multidimensional Data Model

- Multidimensional “cube” of data
- StarSchema



OLAP Queries

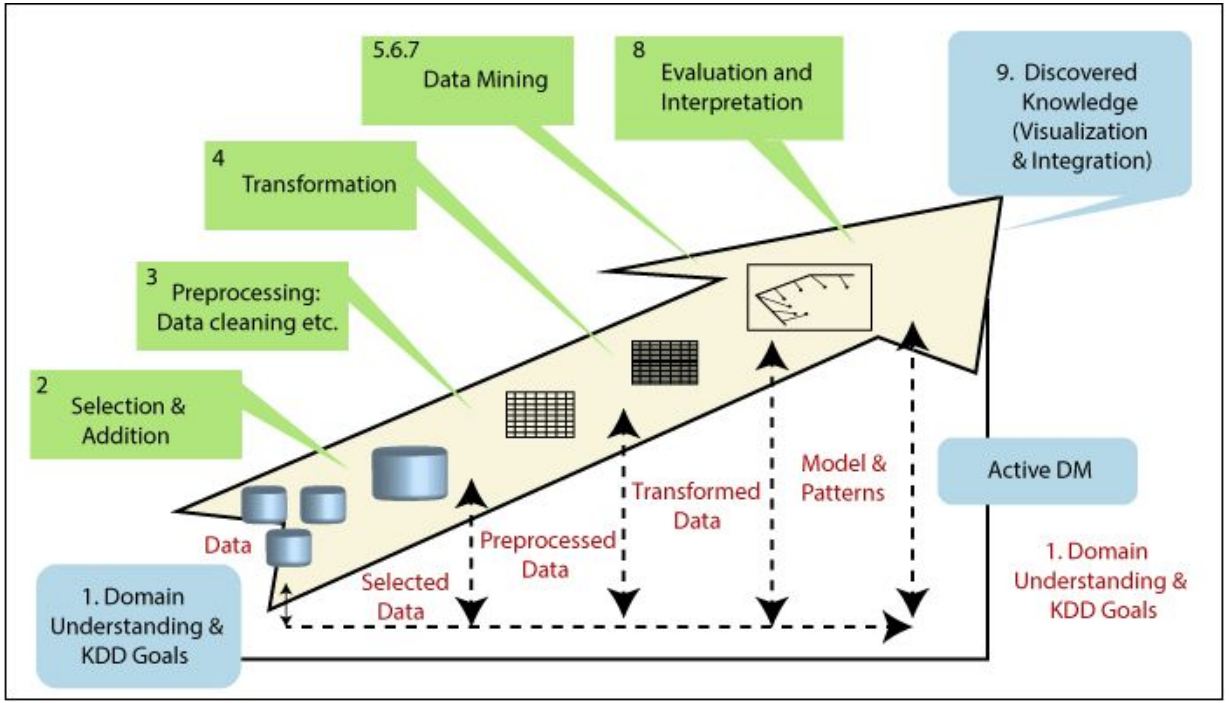
- Slicing
  - Select a value for a dimension
- Dicing
  - Select a range of values in multiple dimension
- Rollup
  - Aggregate along a dimension
- Drill-Down
  - De-aggregating along a dimension

Knowledge Discovery in Database (KDD)

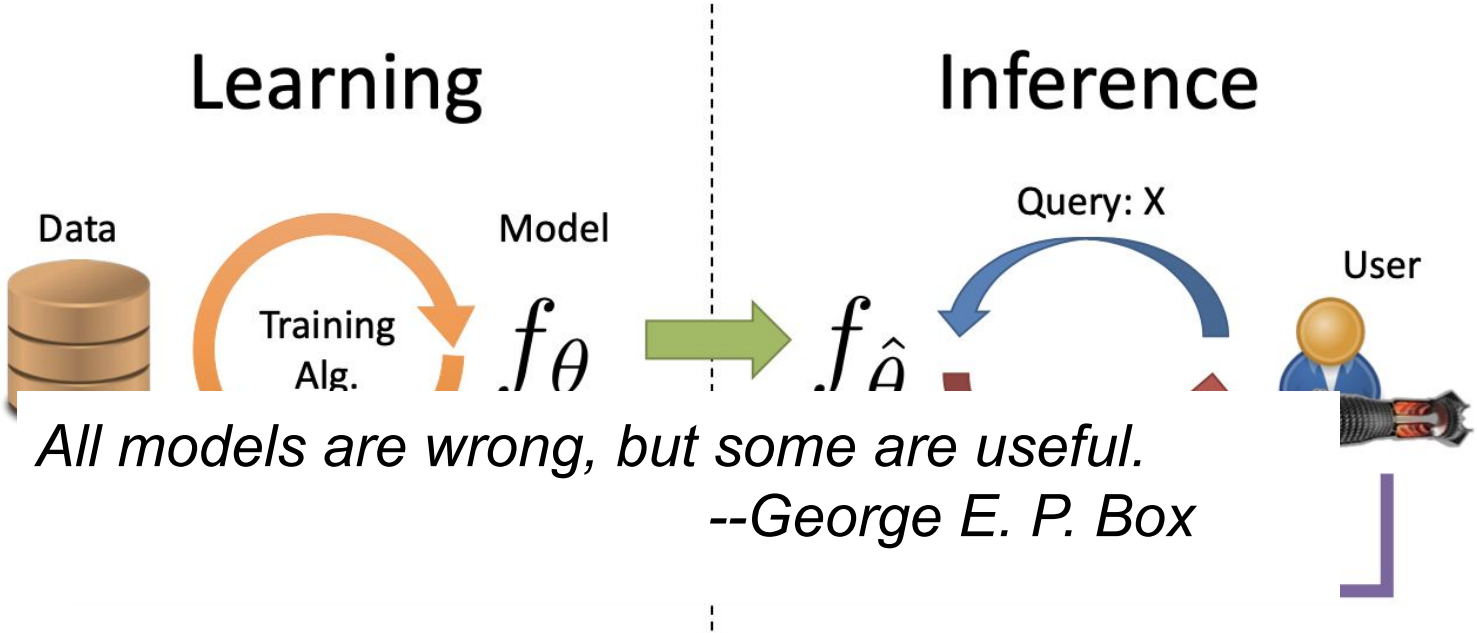
- Data Selection
- Data Cleaning
- Data Mining &

  Machine Learning

- Evaluation



Machine Learning

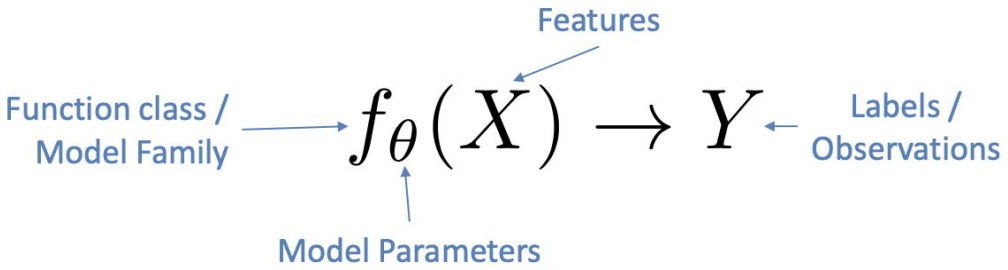


All models are wrong, but some are useful.

--George E. P. Box

Learning: Fitting the model

- Training Data:
  - X: Features
  - Y: Labels



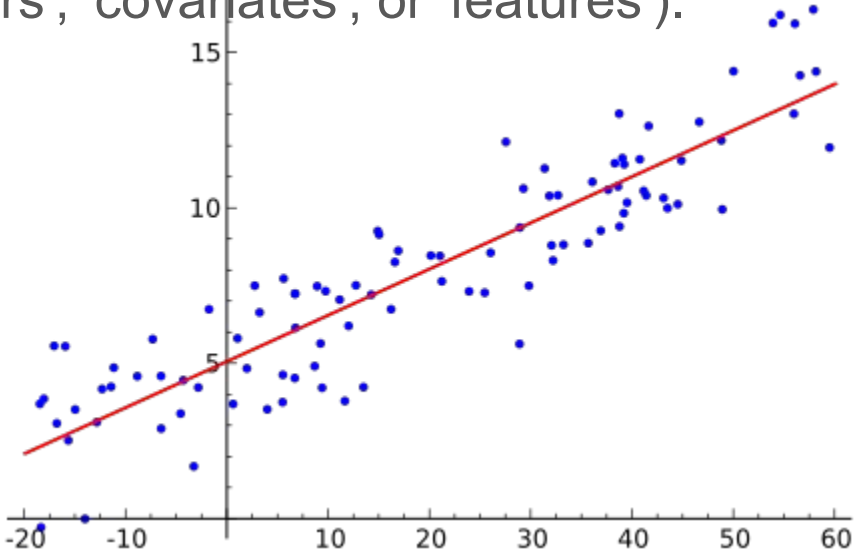
- Learn a function that generalizes the relationship between X and Y
  - Choosing a good function
  - Finding the best parameters

Taxonomy of Machine Learning

- Supervised Learning
  - Regression
  - Classification
- Reinforcement & Bandit Learning (Multi-armed bandit)
- Unsupervised Learning
  - Dimensionality Reduction
  - Clustering

Supervised Learning

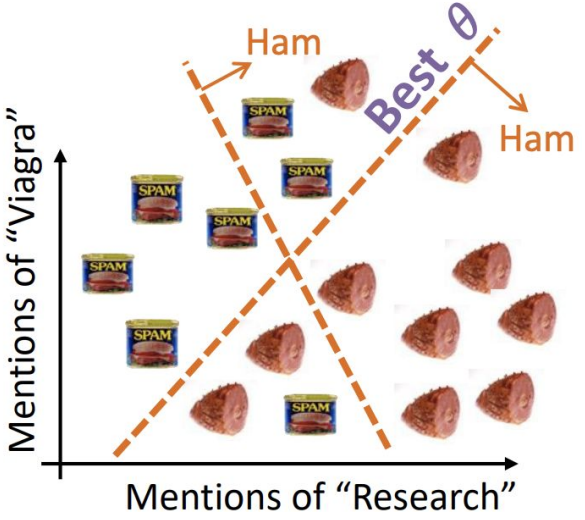
- Regression Analysis: A set of statistical processes for estimating the relationships between a **dependent variable** (often called the 'outcome variable') and one or more **independent variables** (often called 'predictors', 'covariates', or 'features').



Supervised Learning

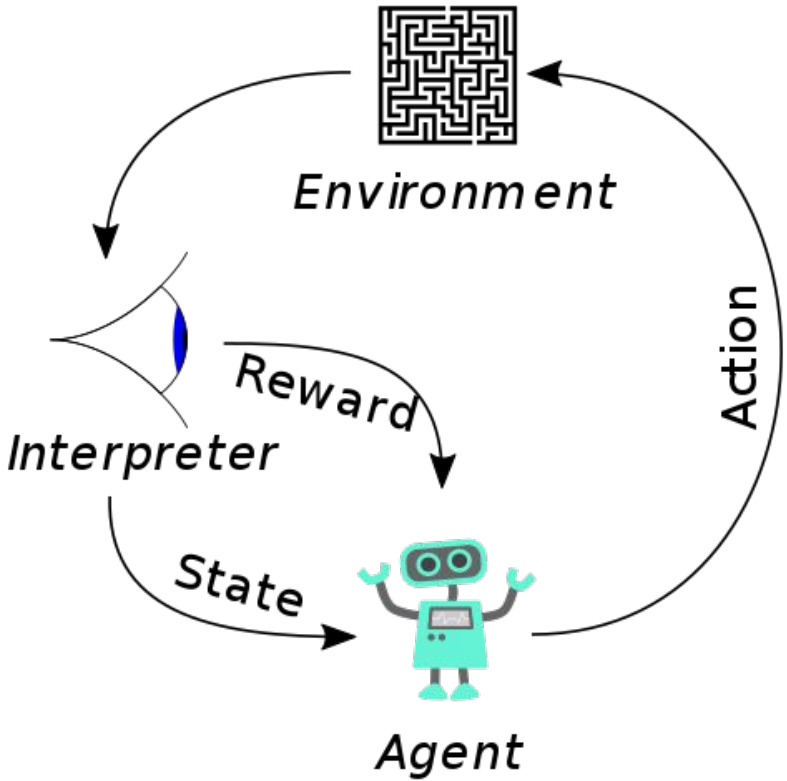
- Classification: The problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known.

- E.g.: Nearest Neighbor, Support Vector Machine, Logistic Regression …...



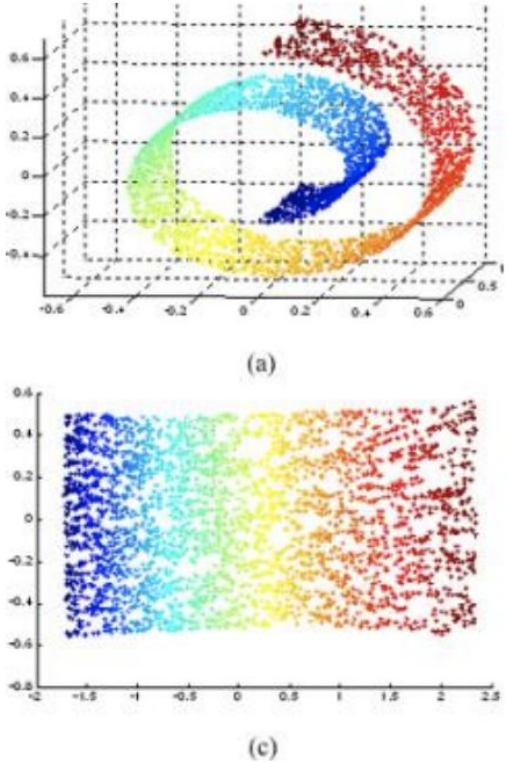
Reinforcement Learning

- **Reinforcement learning** (**RL**) is an area of machine learning concerned with how software agents ought to take actions in an environment in order to maximize the notion of cumulative reward.



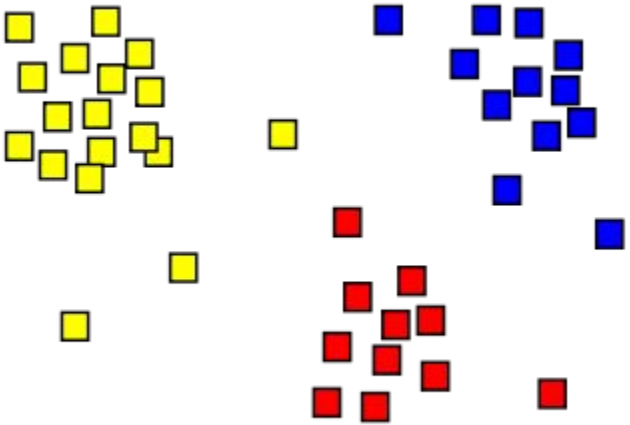
Unsupervised Learning

- Dimensionality Reduction is the transformation of data from a **high-dimensional space** into a **low-dimensional space** so that the low-dimensional representation retains some meaningful properties of the original data, ideally close to its intrinsic dimension.



Unsupervised Learning

- Clustering: is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense) to each other than to those in other groups (clusters).



K-means clustering

- Don’t confuse with **K-nearest neighbors (KNN)** algorithm!!!
- Main Steps:
  - Specify number of clusters K.
  - Randomly select K data points as initial centroids.
  - Keep iterating the following steps until converged (no changes to the centroids):
    - Assign each data point to the closest centroid (cluster).
    - Computer the new centroids for clusters by taking the average of all data points that belong to each cluster.

https://stanford.edu/class/engr108/visualizations/kmeans/kmeans.html