Supervised Learning

Data Mining

Prof. Sujee Lee

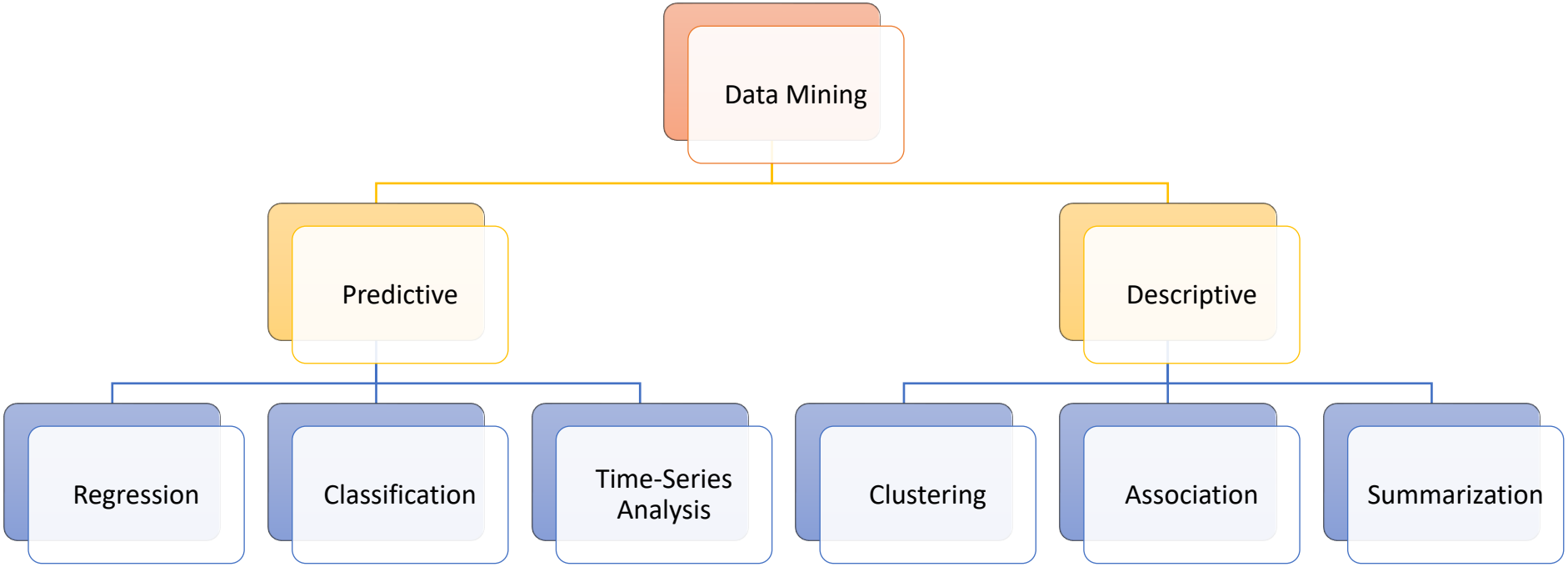
Department of Systems Management Engineering

Sungkyunkwan University

Learning Tasks

Data Mining Tasks

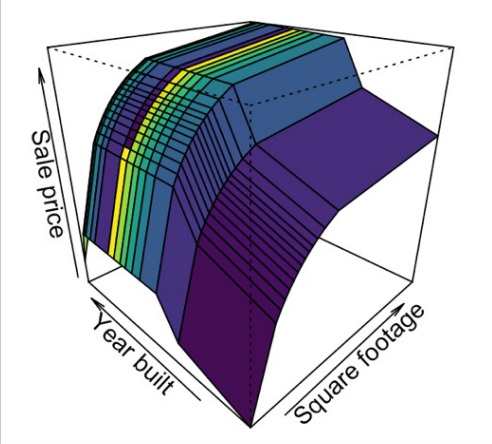
- **Data Mining Tasks**



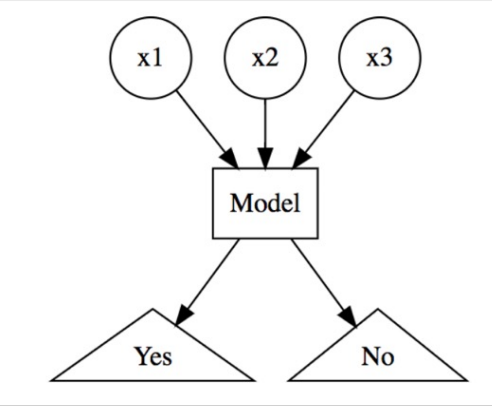
* We often use the following terms interchangeably:
  Data mining, Machine learning
  Predictive analysis, Supervised learning
  Descriptive analysis, Unsupervised learning

Supervised vs. Unsupervised Learning

- **Supervised Learning**
  - A **predictive model** is used for tasks that involve the prediction of a given output (or target) using other variables (or features) in the data set.
  - Given a set of data, the learning algorithm attempts to optimize a function to find the combination of feature values that results in a predicted value that is as close to the actual target output as possible.
  - e.g. **classification** : X -> y (class) / **regression** : X -> y (numeric)



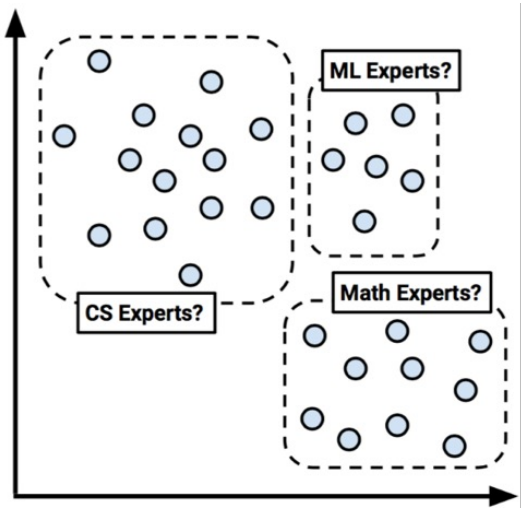
Prediction (numeric target)



Classification (categorical target)

Supervised vs. Unsupervised Learning

- **Unsupervised Learning**
  - There is no target variable to predict or classify. Unsupervised learning includes a set of techniques to better understand and describe the data.
  - **Clustering**: Segment observations into smaller groups based on the observed variables.
  - **Association Rules**, **Collaborative Filtering**, **Recommendation Systems**: Rule-based pattern detection in large databases or users' preferences given their purchase, rating, browsing, etc.



transaction ID	items purchased
1	{flowers, get well card, soda}
2	{plush toy bear, flowers, balloons, candy bar}
3	{get well card, candy bar, flowers}
4	{plush toy bear, balloons, soda}
5	{flowers, get well card, soda}

Learning Algorithms

Classification Algorithms

- Logistic Regression
- Naive Bayes
- Discriminant Analysis
- K-NN
- Classification Tree
- Support Vector Machine
- Neural Network

Regression Algorithms

- Linear Regression
- K-NN
- Regression Tree
  - & Ensembles
- Support Vector Regression
- Neural Network

Clustering Algorithms

- K-means
- Hierarchical Clustering
- Gaussian Mixture Model
- Spectral Clustering

Dimension Reduction

Association Rule Mining

Recommender Systems

Supervised Learning

Supervised Learning

- **Labeled Dataset**

Column: **variable**, attribute, **feature/target**, predictor/response, …

Row: **data point**, **instance**, example, record, pattern, object, …

	Input				Output
id	X_1	X_2	X_3	…	X_d	Y
1	x_{11}	x_{12}	x_{13}	…	x_{1d}	y_1
2	x_{21}	x_{22}	x_{23}	…	x_{2d}	y_2
3	x_{31}	x_{32}	x_{33}	…	x_{3d}	y_3
4	x_{41}	x_{42}	x_{43}	…	x_{4d}	y_4
5	x_{51}	x_{52}	x_{53}	…	x_{5d}	y_5
6	x_{61}	x_{62}	x_{63}	…	x_{6d}	y_6
7	x_{71}	x_{72}	x_{73}	…	x_{7d}	y_7
…	…	…	…	…	…	…

Supervised Learning

- **Supervised Learning**
  - *Labeled dataset* $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots, (\boldsymbol{x}_n, y_n)\} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$, where each $\boldsymbol{x}_i = (x_{i,1}, x_{i,2}, \dots, x_{i,d})$ is associated with a label y_i, is given
  - *Input variables*: $X_1, X_2, \dots, X_d$ (d **features**), *output variable*: y (**target**)
  - Find the functional relationship between input and output variables $\hat{y} = f(\boldsymbol{x})$ (**model**) from the dataset
  - Use input variables to predict unknown or future values of output variable

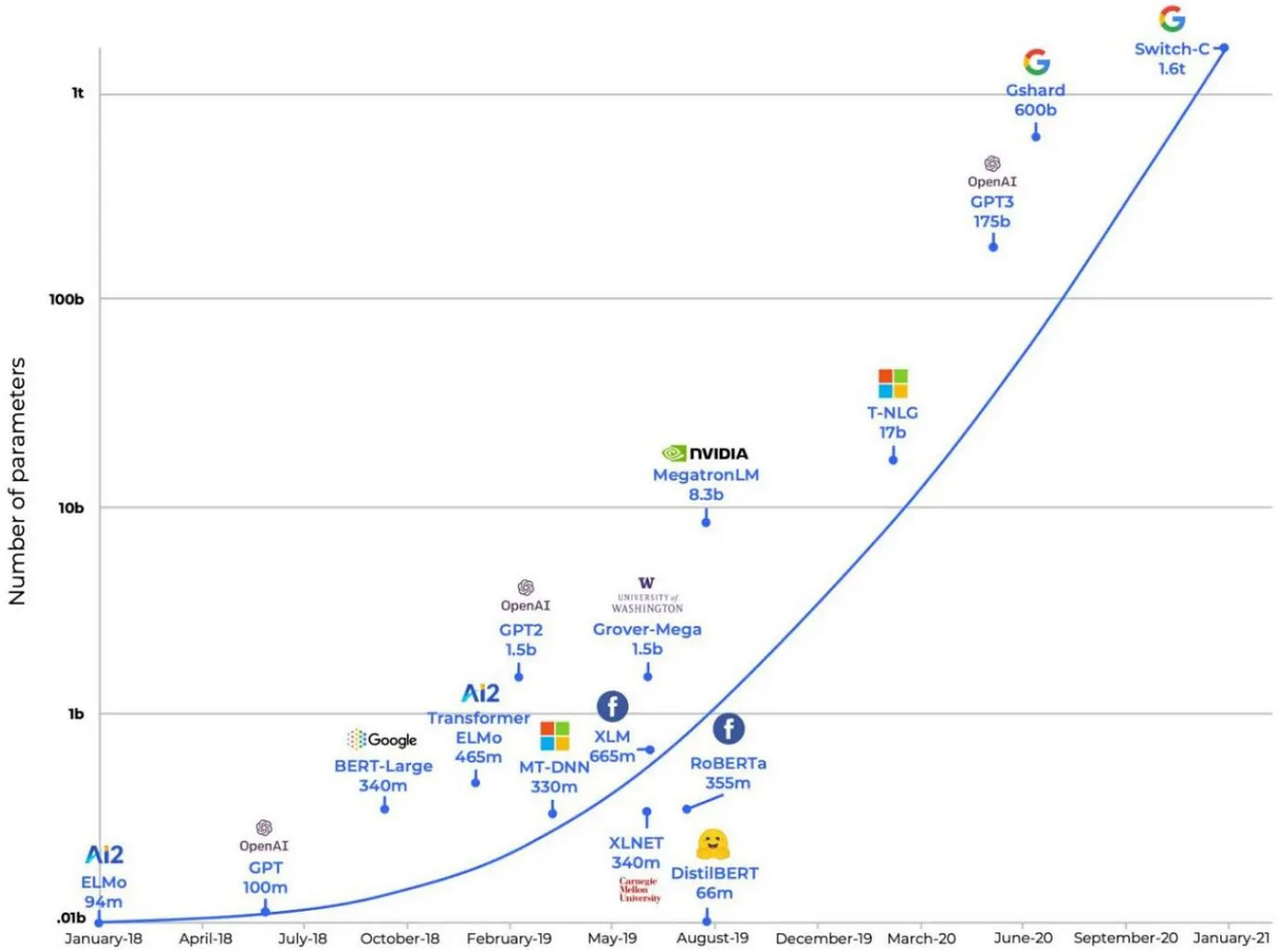
- **Examples**
  - **Linear Regression**

$$\hat{y} = f(\boldsymbol{x}) = \boldsymbol{w}\boldsymbol{x} + b = w_1 x_1 + \cdots + w_d x_d + b$$

  - **Neural Network**

$$\hat{y} = f(\boldsymbol{x}) = f_1\left(\boldsymbol{w_{f1}}\left(f_2\left(\boldsymbol{w_{f2}} f_3\left(\dots f_n\left(\boldsymbol{w_{fn}} x + b_{f_n}\right)\right) + \cdots\right)\right)\right)$$

Supervised Learning



Performance Evaluation

- **Evaluation Metrics for Regression**
  - Given a data set $\boldsymbol{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$
  - **Mean Squared Error (RMSE):** $\mathrm{MSE} = \frac{1}{n}\sum_i (y_i - \widehat{y_i})^2$
  - **Root Mean Squared Error (RMSE):** $\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_i (y_i - \widehat{y_i})^2}$
  - **Mean Absolute Error (MAE):** $\mathrm{MAE} = \frac{1}{n}\sum_i |y_i - \widehat{y_i}|$
  - **Coefficient of Determination (R^2):** $R^2 = 1 - \frac{\sum_i (y_i - \widehat{y_i})^2}{\sum_i (y_i - \bar{y})^2}$, $\bar{y} = \frac{1}{n}\sum_i y_i$

Performance Evaluation

- **Evaluation Metrics for Classification**
  - Given a test set $\boldsymbol{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$
  - **Accuracy**: the fraction of correctly classified data points

$$\text{Accuracy} = \frac{1}{n}\sum_{i} \mathbb{I}(y_i = \hat{y}_i) \times 100\%$$

  - **Confusion Matrix**
    - **Precision**: The proportion of true positives (TP) among the predicted positives (FP + TP).
      - assess the performance of positive predictions.
    - **Recall**: The proportion of true positives (TP) among the actual positives (FN + TP).
      - evaluate how well the model predicts actual positive cases (also called sensitivity or True Positive Rate, TPR).
    - **F1-score**: A metric that combines precision and recall.