CS57300
PURDUE UNIVERSITY

MARCH 28, 2019

# DATA MINING

# ANNOUNCEMENTS

- Assignment 3 grade is out!
- Assignment 4 is due this Sunday (March 31), 11:59pm
  - If you are going to use any late days, please specify it clearly on your pdf report

# DESCRIPTIVE MODELING

# DATA MINING COMPONENTS

- Task specification: **Description**
- Knowledge representation
- Learning technique
- Evaluation and interpretation

# DESCRIPTIVE MODELS

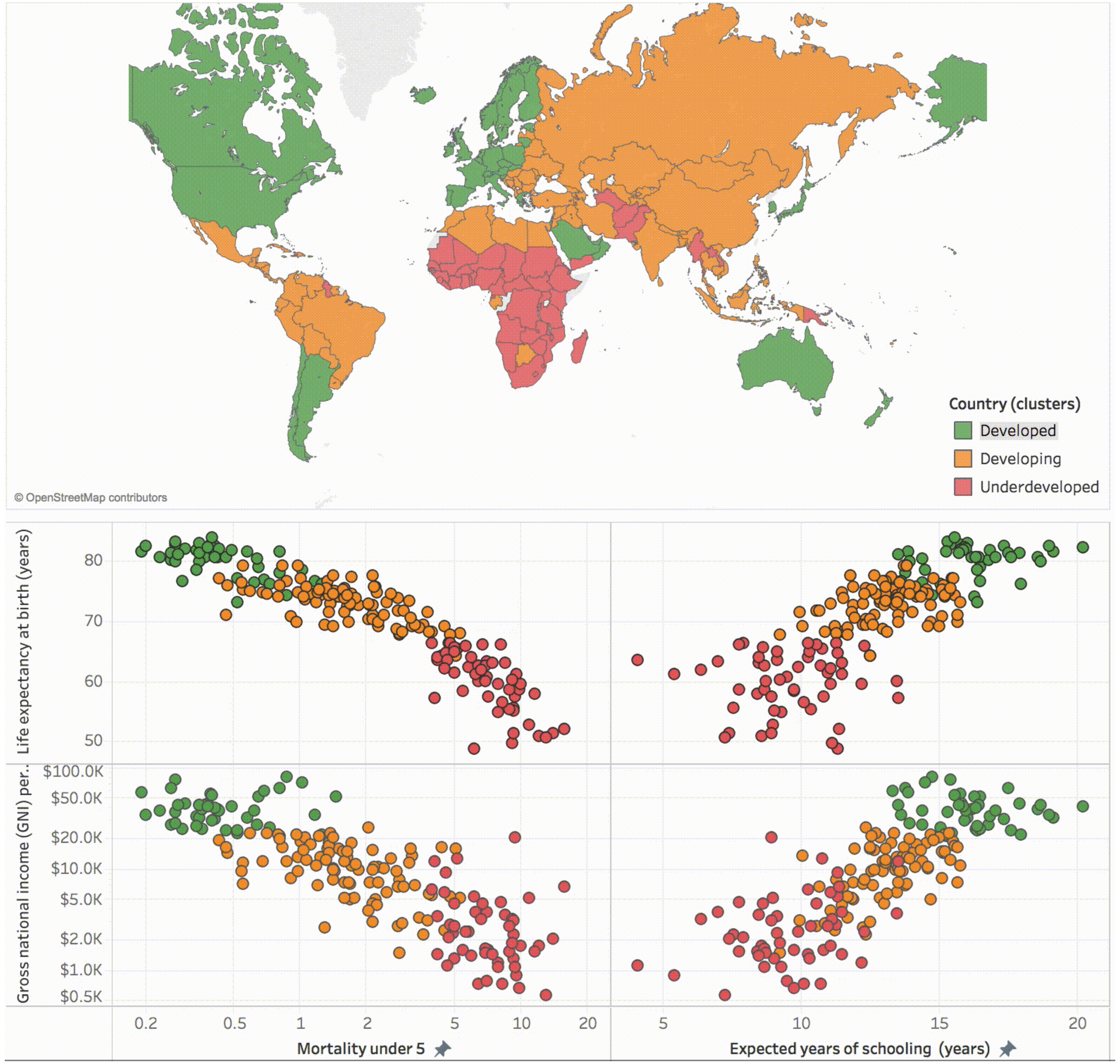
- Descriptive models **summarize** the data
  - Provide a global summary of the data which gives insights into the domain
  - May be used for prediction, but prediction is not the primary goal
- Also known as **unsupervised learning**
  - No predefined “class” labels for each data instance

# DESCRIPTIVE MODELING

- Data representation: data instances represented as attribute vectors ***x***$(i)$, often in the form of $n \times p$ tabular data (i.e., $p$ attributes)
- Task–depends on approach
  - Clustering: summarize the data by characterizing groups of similar instances
  - Structure learning and density estimation: determine a compact representation of the full joint distribution $P(\mathbf{X})=P(X_1,X_2,\ldots,X_p)$

# CLUSTER ANALYSIS

- Decompose or partition instances into groups s.t.:
  - **Intra**-group similarity is ***high***
  - **Inter**-group similarity is ***low***
- Measure of distance/similarity is crucial

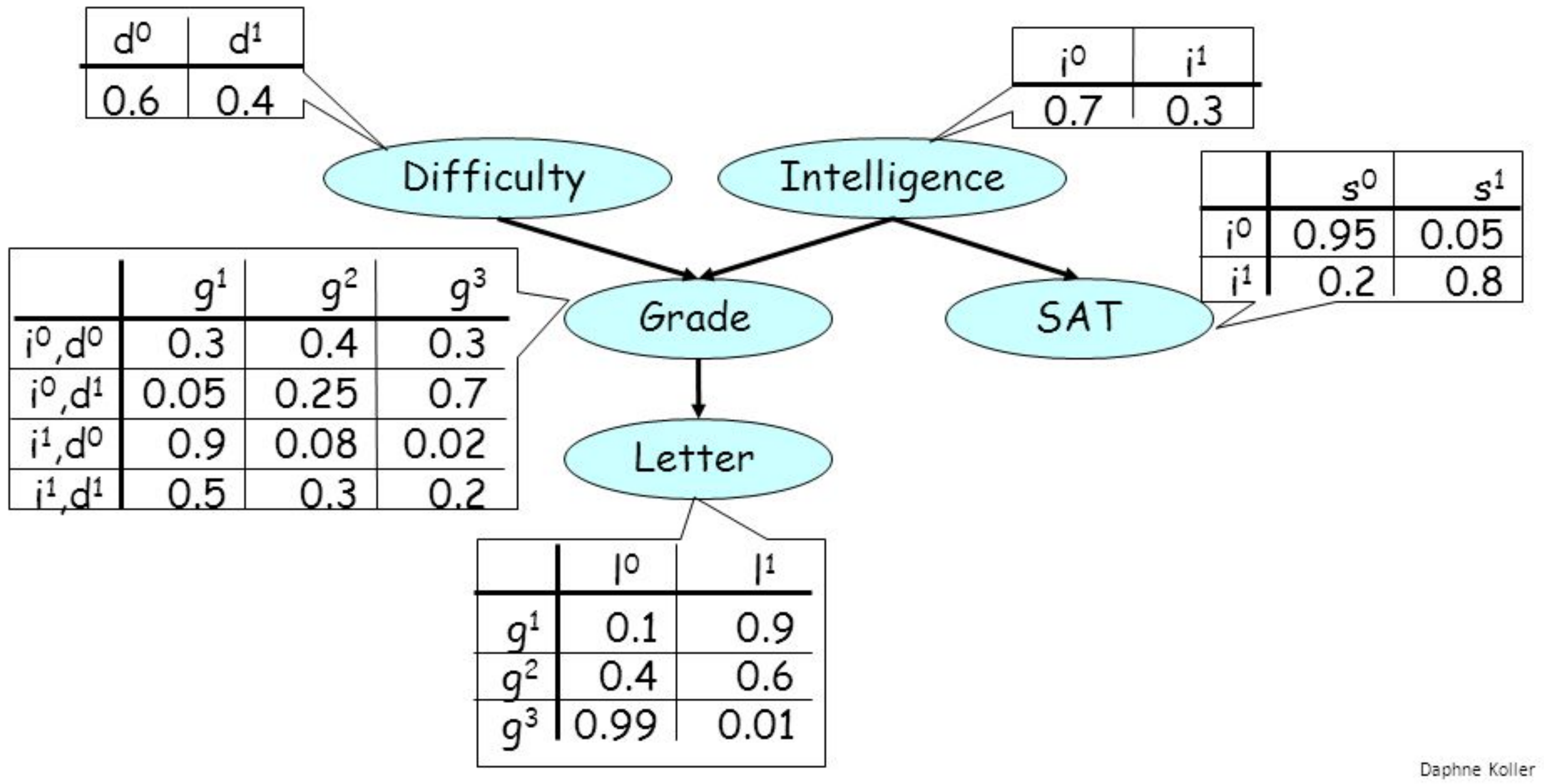


# APPLICATION EXAMPLES

- **Marketing**: discover distinct groups in customer base to develop targeted marketing programs
- **Land use**: identify areas of similar use in an earth observation database to understand geographic similarities
- **City-planning**: group houses according to house type, value, and location to identify "neighborhoods"
- **Earth-quake studies**: Group observed earthquakes to see if they cluster along continent faults

# STRUCTURE LEARNING AND DENSITY ESTIMATION

- Estimate the structure and parameters for the model that generates the observed data such that:
  - Likelihood of observing the data is high
  - Assumption: data is sampled independently from the same distribution (i.i.d)
- Example
  - Observe data: (student's IQ, student's SAT score, midterm exam difficulty, midterm exam grade, letter quality from the instructor)




| $d^0$ | $d^1$ |
|---|---|
| 0.6 | 0.4 |

| $i^0$ | $i^1$ |
|---|---|
| 0.7 | 0.3 |

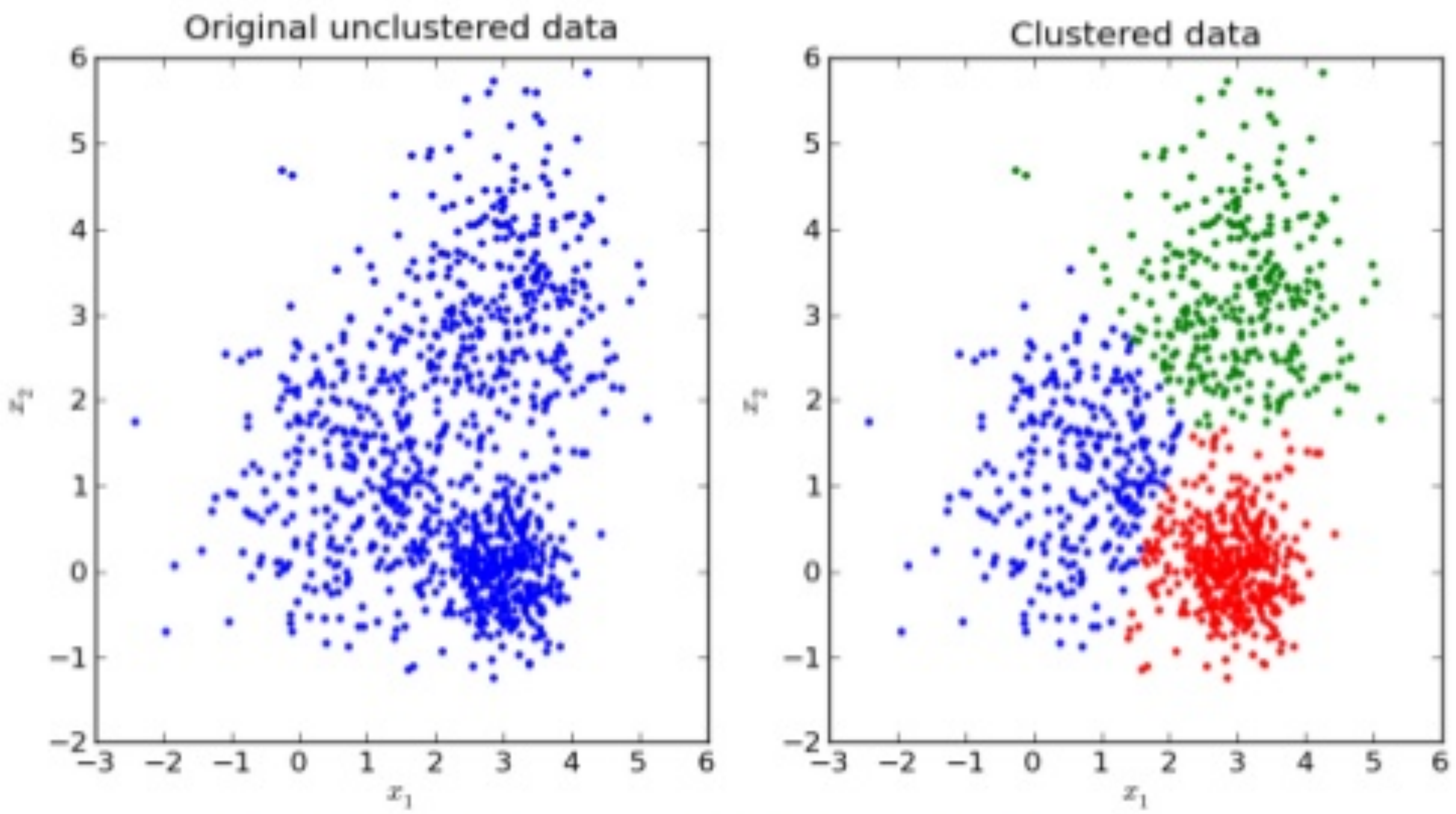
| | $s^0$ | $s^1$ |
|---|---|---|
| $i^0$ | 0.95 | 0.05 |
| $i^1$ | 0.2 | 0.8 |

| | $g^1$ | $g^2$ | $g^3$ |
|---|---|---|---|
| $i^0,d^0$ | 0.3 | 0.4 | 0.3 |
| $i^0,d^1$ | 0.05 | 0.25 | 0.7 |
| $i^1,d^0$ | 0.9 | 0.08 | 0.02 |
| $i^1,d^1$ | 0.5 | 0.3 | 0.2 |

| | $l^0$ | $l^1$ |
|---|---|---|
| $g^1$ | 0.1 | 0.9 |
| $g^2$ | 0.4 | 0.6 |
| $g^3$ | 0.99 | 0.01 |

# KNOWLEDGE REPRESENTATION

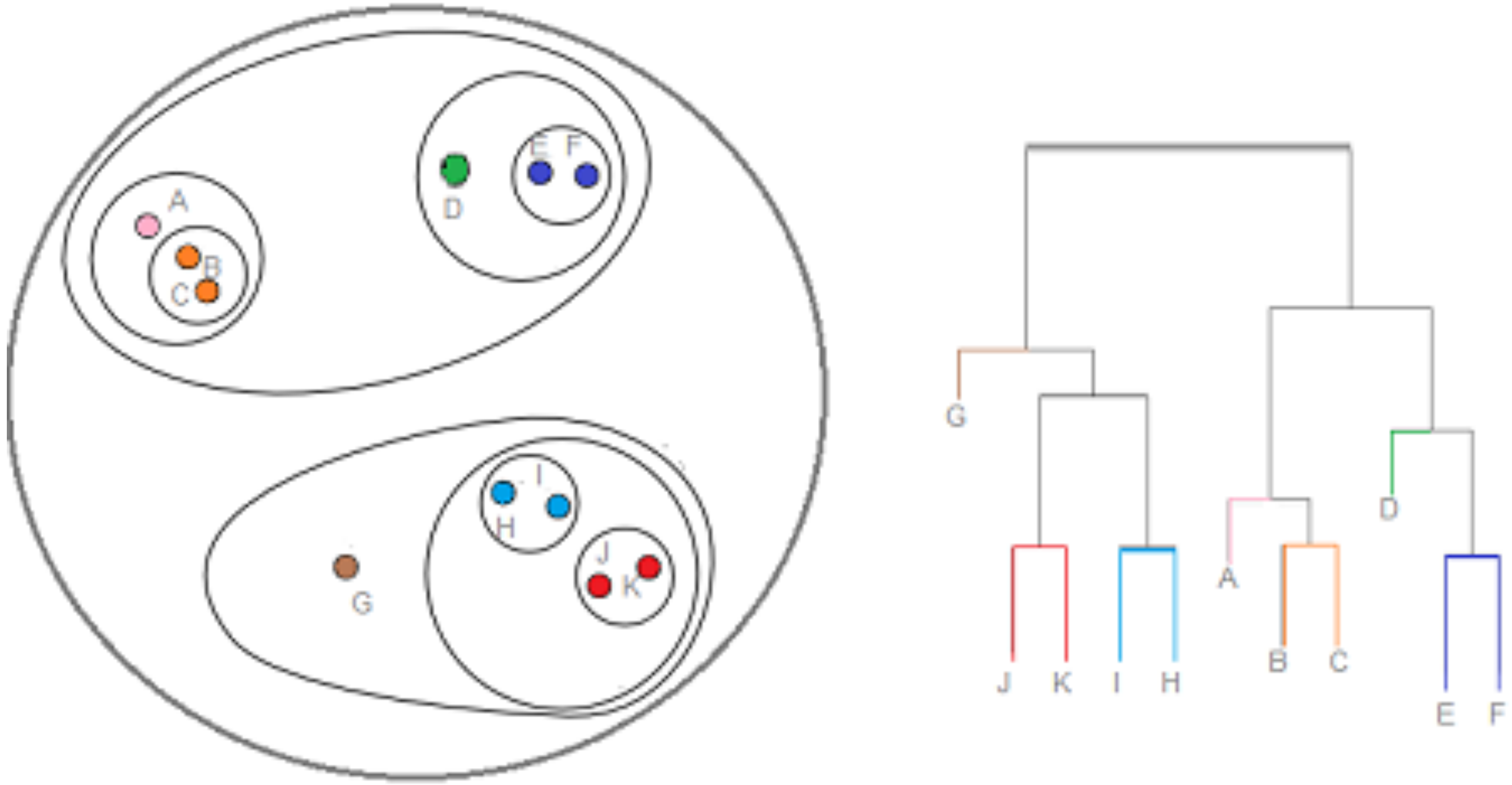
# PARTITION-BASED CLUSTERING




- Partition data instances into a fixed number of groups
- Representative algorithm: K-means

**Model space:**
all possible assignments of data instance to group

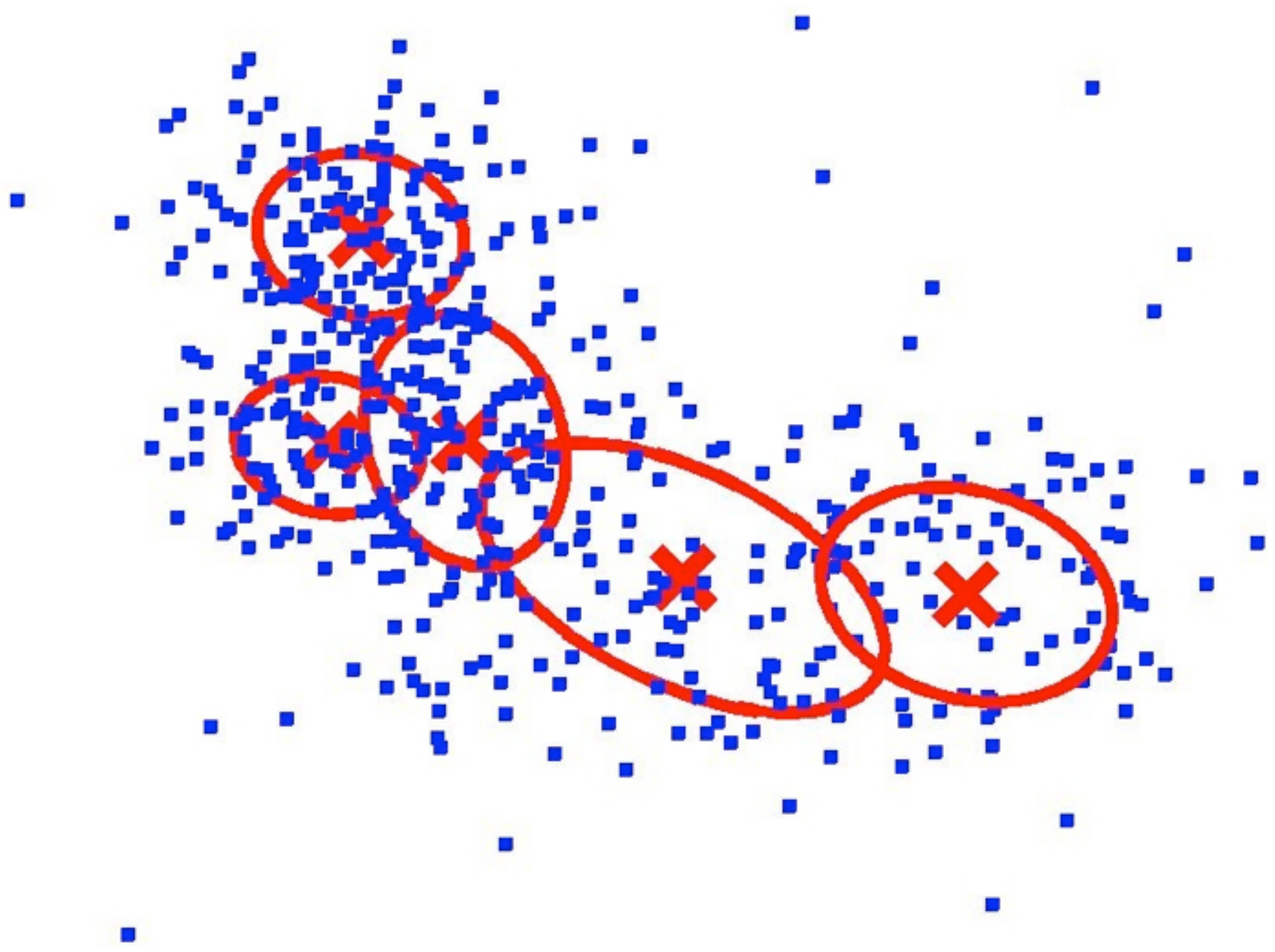
# HIERARCHICAL CLUSTERING




- Build a hierarchy of clusters given the data
- Can be agglomerative ("bottom-up") or divisive ("top-down")

**Model space:**
all possible hierarchies

# PROBABILISTIC MODEL-BASED CLUSTERING



$$f(x) = \sum_{k=1}^{K} w_k f_k(x; \theta)$$

**probability of observing x**

**likelihood of x being generated from cluster k**

**likelihood of point belonging to cluster k**

**Model space:**

$w_k$ and $f_k(x; \theta)$

# DESCRIPTIVE MODELING: LEARNING

# LEARNING DESCRIPTIVE MODELS

- Select a **knowledge representation** (a “model”)
  - Defines a **space** of possible models $M=\{M_1, M_2, \ldots, M_k\}$
- Define **scoring functions** to “score" different models
- Use **search** to identify “best” model(s)
  - Search the space of models
  - Evaluate possible models with **scoring function** to determine the model which best fits the data

# DESCRIPTIVE SCORING FUNCTIONS

- Clustering: What makes a good cluster?
  - High intra-group similarity, low inter-group similarity
  - Scoring function is often a function of within-cluster similarity and between-cluster similarity
- Example scoring functions

**cluster centroid:**

$$r_k = \frac{1}{n_k} \sum_{x(i) \in C_k} x(i)$$

**between-cluster distance:**

$$bc(C) = \sum_{1 \le j < k \le K} d(r_j, r_k)^2$$

**within-cluster distance:**

$$wc(C) = \sum_{k=1}^{K} wc(C_k) = \sum_{k=1}^{K} \sum_{x(i) \in C_k} d(x(i), r_k)^2$$

# DESCRIPTIVE SCORING FUNCTIONS

- Structure learning and density estimation: Does the model representation capture the observed data well?
  - Likelihood of the observed data is often used as the scoring function
  - Also applicable to probabilistic model-based clustering

## SEARCHING OVER MODELS

- Search over the model space to find the model structure / parameters that optimize the scoring function
- Discrete model space example: partition-based clustering
  - Find $k$ clusters among $n$ data instances: $k^n$ possible allocations
  - Exhaustive search is intractable
  - Most approaches use iterative improvement algorithms to search the model space heuristically

# SEARCHING OVER MODELS

- Continuous model space example: probabilistic model-based clustering
  - Searching for the cluster weight (i.e., $w_k$) and cluster parameters (i.e., $f_k(x, \theta)$) that gives the highest likelihood of observing the current data
  - Solution: **Expectation-maximization** to iteratively infer cluster member and estimate cluster parameters