### Key Features of DAGger

- It considers the **possible worlds semantics for uncertain data**
  - natural semantics for incomplete and probabilistic databases.
  - the input is a probability distribution over a set of possible worlds, where each world defines a set of input objects.
  - the output is equivalent to clustering within each world and defines probability distributions for objects belonging to clusters.
- It allows for **arbitrary correlations**, which are:
  - used in results to queries in probabilistic databases,
  - obtained by structuring text using Conditional Random Fields,
  - enforced by experts and learned from data in Bayesian Networks and Markov Logic Networks.
- If correlations are ignored, the output can be arbitrarily off from the true clustering result.
- It can **compute exact and approximate probabilities with error guarantees** for the clustering output.

### DAGger’s Approach

- The **uncertainty and correlations in the input data** are represented symbolically in a language of probabilistic events.
- **Clustering events** are captured within the same formalism.
- This formalism supports a wide range of tasks:
  - probability computation for clustering events,
  - sensitivity analysis and explanation of clustering output,
  - different clustering algorithms, e.g., \( k \)-medoids, Markov clustering.
- All clustering events are represented within one event network:
  - Common expressions are represented only once.
  - Yields a highly repetitive and interconnected structure due to the combinatorial nature of clustering.
  - For \( k \)-medoids and Markov clustering, the events have the same structure at each step, and at any iteration step are expressions over the events at the previous clustering iteration.
  - Compute the probability of all events by bulk-compiling an entire event network into one decision tree.
    - Only the current root-to-leaf path of this decision tree is kept at any one time, while exploring it depth-first.
    - **Anytime approximation with error guarantees** can be achieved by exploring small fragments of the decision tree.

### State-of-the-art techniques (e.g. UK-means, UKmedoids, MMVar):

- do not support the possible worlds semantics,
- lack support for correlations and assume probabilistic independence,
- use deterministic cluster medoids or expected means, and
- can only compute clustering based on expected distances.

In many cases, the output is a hard clustering that assigns each object to one cluster, like in deterministic \( k \)-medoids or \( k \)-means.

### \( k \)-Medoids Clustering of Certain Data

1. **(Initialization)** Initially choose an object as medoid for each cluster.
   - Given: objects \( o_1, \ldots, o_n \), and clusters \( C_1, \ldots, C_k \).
2. **(Assignment)** Assign object to the cluster of the closest medoid.
   - “closest” defined using any distance metric, e.g., Euclidean distance, Manhattan distance or Minkowski distance.
3. **(Update)** Choose new medoid for each cluster.
4. Repeat phases 2 and 3 for a number of iterations, or until fixpoint reached.

### Language of Probabilistic Events

- Propositional events over independent Boolean random variables.
- Construct that can succinctly express **real values conditioned on propositional formulas**:
  - \( \Phi \land \Psi \) expresses that the value \( v \in \mathbb{R} \) is conditioned by the formula \( \Phi \in \mathbb{B} \):
    - if \( \Phi \) then \( v \) else 0.
  - Sums of if-then-else expressions: \( \Phi_1 \land \Psi_1 + \ldots + \Phi_n \land \Psi_n \)
  - Comparisons of such sums: \( \Phi_1 \land \Psi_1 + \ldots + \Phi_n \land \Psi_n \leq \Phi_1 \land \Psi_1 + \ldots + \Psi_m \land \Psi_m \)

This language allows for succinct encoding – **independently of the number of possible variable assignments** – of sums of distances from an object to any other object in a cluster, conditioned on the uncertainty of these objects.

### \( k \)-Medoids Clustering of Uncertain Data

Our approach is a realisation of \( k \)-medoids clustering on uncertain data.

- It is equivalent to performing \( k \)-medoids clustering in each possible world of the input, yet avoids the explicit enumeration of possible worlds.
- The probability that an object belongs to a cluster is the sum of probabilities of those worlds in which this event occurs.
- Each object belongs to each cluster or is medoid with a certain probability.

Examples of clustering queries:

- **membership**: does a given object belong to a given cluster?
- **medoid**: is a given object the medoid of a given cluster?
- **co-occurrence**: are given objects clustered together?

**Membership event** \( m \Phi | o_i \in C_j \) for object \( o_i \) and cluster \( C_j \) at step \( t \geq 1 \):

\[
\Phi | o_i \in C_j = o_i \land | \bigcup_{1 \leq k \leq n, k \neq i} (1 - d_i(o_k, o_i) + \Phi | o_k \in C_j)
\]

**Medoid event** \( m \Phi | C_j = o_i \) for object \( o_i \) and cluster \( C_j \) at step \( t \geq 1 \):

\[
\Phi | o_i = c_j \land \bigwedge_{1 \leq k \leq n} (\Phi | o_k \in C_j \land (\Phi | o_k \in C_j) \leftarrow d_i(o_k, o_i) + \Phi | o_k \in C_j)
\]

**Legend:**

- \( \Phi o_i \) is the event that object \( o_i \) exists.
- \( d_i(o_k, o_i) \) is the distance function between objects.
- \( \Phi C_j \) is the total distance-sum of \( o_i \) to the objects in \( C_j \) at step \( t \).

### Exact and Approximate Probability Computation

Partial example of an event network with five layers encoding, highly interconnected events for clusters \( C_0 \) and \( C_1 \).

- **Compilation of event network into decision tree using Shannon expansion**: \( \Phi = \Phi_0 \land x_0 \land \ldots \land \Phi_4 \land x_4 \) This means that: \( P(\Phi) = P_0 \cdot P(\Phi_0) \cdot (1 - P_0) \cdot P(\Phi_4) \)
- If \( \Phi \) is the network, then the restrictions \( \Phi_0 \) and \( \Phi_4 \) are obtained by masking in \( \Phi \) those nodes that become true or false.
- Repeated application of Shannon expansion eventually masks nodes in the network and adds the probability of the variable assignments (x or -x) to the probability mass of these nodes.

**Approximate probability computation strategies** decide how to invest (eagerly, lazily, or hybrid) the error budget while exploring the decision tree.

### Experimental Evaluation with \( k \)-Medoids Clustering of Uncertain Data

- **naive** means \( k \)-medoids in each possible world.
- **types of correlations considered**: **positive**, **mutex** (block-independent disjoint); **conditional independence**.