Dealing with Uncertainty: A Survey of Theories and Practices

Yiping Li, Jianwen Chen, and Ling Feng, *Member, IEEE*

Abstract—Uncertainty accompanies our life processes and covers almost all fields of scientific studies. Two general categories of uncertainty, namely, *aleatory uncertainty* and *epistemic uncertainty*, exist in the world. While aleatory uncertainty refers to the inherent randomness in nature, derived from natural variability of the physical world (e.g., random show of a flipped coin), *epistemic uncertainty* origins from human's lack of knowledge of the physical world, as well as ability of measuring and modeling the physical world (e.g., computation of the distance between two cities). Different kinds of uncertainty call for different handling methods. Aggarwal, Yu, Sarma, and Zhang et al. have made good surveys on uncertain database management based on the probability theory. This paper reviews multidisciplinary uncertainty processing activities in diverse fields. Beyond the dominant probability theory and fuzzy theory, we also review information-gap theory and recently derived uncertainty theory. Practices of these uncertainty handling theories in the domains of economics, engineering, ecology, and information sciences are also described. It is our hope that this study could provide insights to the database community on how uncertainty is managed in other disciplines, and further challenge and inspire database researchers to develop more advanced data management techniques and tools to cope with a variety of uncertainty issues in the real world.

Index Terms—Uncertainty management, probability theory, Dempster-Shafer theory, fuzzy theory, info-gap theory, probabilistic database, fuzzy database

1 Introduction

Uncertainty is ubiquitous and happens in every single event we encounter in the real world. *Whether it rains or not tomorrow is uncertain; whether there is a train delay is uncertain...* Just as Socrates in ancient Greece said, "*as for me, all I know is I know nothing* [1]." Uncertainty distinguishes from certainty in the degree of belief or confidence. If certainty is referred to as a perception or belief that a certain system or phenomenon can experience or not, uncertainty indicates a lack of confidence or trust in an article of knowledge or decision [2]. According to the US National Research Council, "*uncertainty is a general concept that reflects our lack of sureness about something or someone, ranging from just short of complete sureness to an almost complete lack of conviction about an outcome* [3]."

1.1 Uncertainty Categorization

Uncertainty arises from different sources in various forms and is classified in different ways by different communities. According to the origin of uncertainty, we can categorize uncertainty into *aleatory* uncertainty or *epistemic* uncertainty [3], [4], [5], [6], [7], [8], [9]:

- *Aleatory uncertainty* derives from natural variability of the physical world. It reflects the inherent randomness in nature. It exists naturally regardless of human's knowledge. For example, in an event of flipping a coin, the coin comes up heads or tails with some randomness. Even if we do many experiments and know the probability of coming up heads, we still cannot predict the exact result in the next turn.

  Aleatory uncertainty cannot be eliminated or reduced by collecting more knowledge or information. No matter whether we know it, this uncertainty stays there all the time. Aleatory uncertainty is sometimes also referred to as *natural variability* [3], *objective uncertainty* [10], *external uncertainty* [11], *random uncertainty* [12], *stochastic uncertainty* [13], *inherent uncertainty*, *irreducible uncertainty*, *fundamental uncertainty*, *real-world uncertainty*, or *primary uncertainty* [14].
- *Epistemic uncertainty* origins from human's lack of knowledge of the physical world and lack of the ability of measuring and modeling the physical world. Unlike aleatory uncertainty, given more knowledge of the problem and proper methods, epistemic uncertainty can be reducible and sometimes can even be eliminated. For example, the estimation of the distance between Boston and Washington can be more precise if we have known the distance from Boston to New York.

  Epistemic uncertainty is sometimes also called *knowledge uncertainty* [3], *subjective uncertainty* [13], [10], *internal uncertainty* [11], *incompleteness* [15], *functional uncertainty*, *informative uncertainty* [16], or *secondary uncertainty* [14].

Taking the flood frequency analysis, for example, the probability distribution of the frequency curve is a representation of aleatory uncertainty, reflecting an inherent

- *The authors are with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China. E-mail: {liyp09, chen-jw08}@mails.tsinghua.edu.cn, fengling@mail.tsinghua.edu.cn.*

Manuscript received 3 Mar. 2012; revised 21 June 2012; accepted 29 Aug. 2012; published online 10 Sept. 2012.
Recommended for acceptance by N. Mamoulis.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-2012-03-0140.
Digital Object Identifier no. 10.1109/TKDE.2012.179.

randomness of the physical world. We cannot reduce this type of uncertainty. On the contrary, parameters of the frequency curve imply a kind of epistemic uncertainty, constrained by the existing knowledge and corresponding model. However, with the increase of information, we can always modify and refine the model to make it approach the realistic situation.

Although uncertainty is categorized into aleatory uncertainty or epistemic uncertainty, there is not a clear boundary between them, and they may even be dealt with in the same way. Nevertheless, this classification indeed reminds us what we should notice in representing and processing diverse uncertainty in our real world.

1.2 Uncertainty Management

Uncertainty complicates events and affects decision making in a number of unfavorable aspects. Even worse, some attempts that we take to manage and reduce uncertainty are accompanied with uncertainty themselves.

Though it is hard to completely eliminate uncertainty, it is worthwhile to recognize and cope with uncertainty to avoid unfavorable hazards for high-quality decisions. So far, uncertainty management covers almost all fields of scientific studies [17]. Berztiss outlined common methods of uncertainty management, including Bayesian inference, fuzzy sets, fuzzy logic, possibility theory, time Petri nets, evidence theory, and rough sets [18]. Walley [19] compared four measures of uncertainty in expert systems, including Bayesian probabilities, coherent lower and upper precisions, belief functions of evidence theory, and possibility measures in fuzzy theory. Klir [20] studied uncertainty and information as a foundation of generalized information theory.

In the early attempts of database community, *null* values were commonly used to manage uncertain information [21]. Imieliński and Lipski [22] represented different-leveled uncertainty information through different null values or variables satisfying certain conditions. Querying over the databases with null values was investigated by Abiteboul et al. [23]. So far, probabilistic, fuzzy, and possibilistic databases constitute major ways for uncertain data management.

Probabilistic databases. The framework of probabilistic databases was first presented in 1990 by Fuhr [24]. Query evaluation over probabilistic databases was extensively investigated by Dalvi and Suciu [25]. Pei et al. [26] surveyed various ways to answer probabilistic database queries. Sarma et al. [27], [28] presented a space of models for representing and processing probabilistic data based on a variety of uncertainty constructs and tuple-existence constraints. A recent good survey by Aggarwal and Yu [29] covered probabilistic data algorithms and applications, where traditional database management methods (join processing, query processing, selectivity estimation, OLAP queries, and indexing) and traditional mining problems (frequent pattern mining, outlier detection, classification, and clustering) are outlined.

Fuzzy and possibilistic databases. Fuzzy databases arose in 1990s, where tuple/attributewise fuzziness, similarity of fuzzy terms, and possibility distribution fuzzy data modeling and querying were researched. Good survey on fuzzy and possibilistic databases could be found in [30], [31], [32], [33], [34].

1.3 Our Study

This paper provides a cross-disciplinary view of uncertainty processing activities by different communities. We first examine existing uncertainty theories. Among them, probability theory [35], [36] and fuzzy theory [37], [38], [39] are the most common theories to model uncertainty. From the basic probability theory, three methods (i.e., Monte Carlo method, Bayesian method, and evidence theory) are derived. Beyond these, we also present information-gap (info-gap) theory originally developed for decision making [40], [41], as well as a recently derived uncertainty theory from probability and fuzzy theories, which intends to establish a mathematical model for general use [42], [43]. Based on these theories, different types of uncertainty are represented and handled. We overview some typical practices of the theories in different disciplines, spanning from economy, engineering, ecology, to information science. We hope the work reported here could advance uncertain database technology through cross-disciplinary research inspirations and challenges.

We list achievements made by the database community in uncertainty management through a running example of customers' interests to restaurants. Beyond classic probabilistic, fuzzy, and possibilistic databases, Monte Carlo and evidence-based database models and query evaluation are particularly described. We also discuss a few interesting issues for further data-oriented research.

The remainder of the paper is organized as follows: We outline four uncertainty handling theories in Section 2, followed by their applications in diverse domains in Section 3. We particularly review probabilistic and fuzzy database technologies developed in the database field in Section 4 and identify a few challenges ahead of the data-oriented research in Section 5. We conclude the paper in Section 6.

2 Uncertainty Handling Theories

2.1 Outline of Four Theories

Fig. 1 illustrates four uncertainty handling theories:

- *Probability theory* is the most widely used method in almost every field. It can deal with both natural aleatory uncertainty through random experiments and subjective aleatory uncertainty by statistics from questionnaires. Based on probability theory, Monte Carlo method, Bayesian method, and Dempster-Shafer evidence theory are developed.
    - Monte Carlo method can solve complicated situations where computing an exact result with a deterministic algorithm is hard. It approximates the exact value by repeated random sampling.
    - Bayesian method pursues an exact value based on a graphical model with prior and conditional probabilities. It is a good tool for inference.
    - Dempster-Shafer theory avoids the prior probability assumption. It computes the confidence interval, containing the exact probability, by evidences collected from different sources.
- *Fuzzy theory* is good at handling human ambiguity by modeling epistemic uncertainty through fuzzy sets with membership functions.

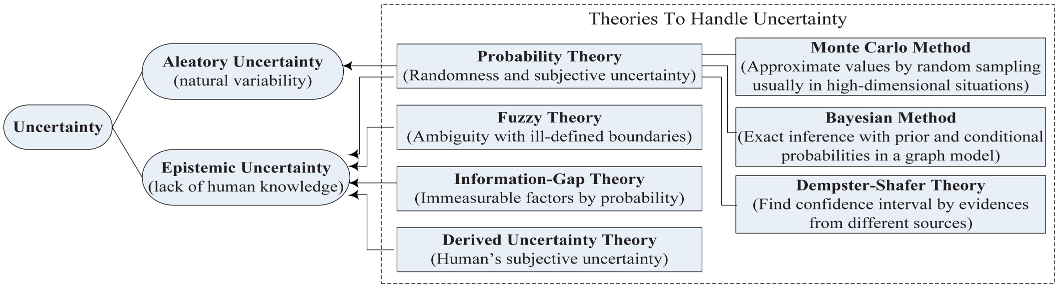

Theories To Handle Uncertainty
Uncertainty
Aleatory Uncertainty
(natural variability)
Epistemic Uncertainty
(lack of human knowledge)
Probability Theory
(Randomness and subjective uncertainty)
Fuzzy Theory
(Ambiguity with ill-defined boundaries)
Information-Gap Theory
(Immeasurable factors by probability)
Derived Uncertainty Theory
(Human's subjective uncertainty)
Monte Carlo Method
(Approximate values by random sampling usually in high-dimensional situations)
Bayesian Method
(Exact inference with prior and conditional probabilities in a graph model)
Dempster-Shafer Theory
(Find confidence interval by evidences from different sources)

Fig. 1. Uncertainty theories.

- *Info-gap theory* can address severe uncertainty whose probability cannot be easily measured or computed by using a range around the expected value to represent epistemic uncertainty.
- *Derived uncertainty theory* from probability and fuzzy theories aims at human's subjective epistemic uncertainty.

Details of the four uncertainty handling theories are reviewed in the following sections.

2.2 Probability Theory

The most well-established probability theory [35], [36] originally aims at random phenomena, such as flipping a coin. It states knowledge or belief that an event will occur or has occurred by means of *probability*, and the probability value is obtained based on statistics and random experiments through repeated trials and analysis. That is, in N times of independent random experiments, the occurrence times of an event approach a constant N_0. Then, N_0/N is the probability of the random event.

A function P is a way to represent the value of probability. Let Ω be a sample space. Each subset of Ω is called an event, denoted as $A_1, A_2, \ldots$. Assume A is an event, then P satisfies

1) $(Normality)$ $P(\Omega) = 1$.
2) $(Nonnegativity)$ $P(A) \geq 0$.
3) $(Additivity)$ For mutually disjoint events $A_1, A_2, \ldots,$ (1)
$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

Probability theory can deal with both aleatory and epistemic uncertainty. Random experiments usually deal with natural variability, which satisfies the definition of aleatory uncertainty. Through random experiments, one can calculate the frequency of a certain event, which is close to the real probability with the increase of running times. With the introduction of subjective probability, it is applied to subjective objects and situations that are not suitable for random experiments. Currently, in dealing with uncertainty, probability theory is at a dominant position. In most of today's applications, uncertainty problems are considered to be probabilistic ones.

2.2.1 Extensions of Probability Theory

Based on the classic probability theory, a few models such as the Monte Carlo method, Bayesian method, and Dempster-Shafer evidence theory are developed.

Monte Carlo methods. Monte Carlo methods origin from a famous experiment of dropping needles conducted by Buffon in 1777. It inspires researchers to simulate some values of interest by random sampling [44]. Now, the Monte Carlo method has become a well-known numerical calculation method based on the probability theory. It relies on repeated random sampling to compute the result (e.g., value of a parameter). In the random sampling, the Chernoff bound [45] can be used to determine the number of runs for a value by majority agreement. In n runs of random experiments flipping coins, p is the probability of heads coming up. For the assurance of $1 - \varepsilon$ accuracy that is the probability for majority agreement, the number of runs should satisfy

$$n \geq \frac{1}{\left(p - \frac{1}{2}\right)^2} \ln \frac{1}{\sqrt{\varepsilon}}. \tag{2}$$

Practically, Chernoff bound gives bounds on tail distributions of sums of independent random variables. Let $X_1, \ldots, X_n$ be independent random variables, $X = \sum_{i=1}^{n} X_i$, and μ is the expectation of X, then for any $\delta > 0$,

$$P(X > (1+\delta)\mu) < \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu}. \tag{3}$$

This bound measures how far the sum of random variables deviates from the expectation in n runs of random experiments.

Monte Carlo methods can generally solve complicated situations where computing an exact result with a deterministic algorithm is hard. It is especially good at simulating and modeling phenomena with significant uncertainty in inputs, such as fluids, disordered materials, strongly coupled solids, and cellular structures. It is widely used in mathematics, for instance, to evaluate multidimensional definite integrals with complicated boundary conditions. When Monte Carlo simulations are applied in space exploration and oil exploration, their prediction of failures, cost overruns, and schedule overruns are routinely better than human intuition or alternative *soft* methods.

A Markov chain, named for Andrey Markov, is a mathematical system that undergoes transitions from one state to another, between a finite or countable number of possible states. The popularly used first-order Markov chain is like:

$$\begin{aligned} &P(X_n = x_n \mid X_1 = x_1, X_2 = x_2, \ldots, X_{n-1} = x_{n-1}) \\ &\quad = P(X_n = x_n \mid X_{n-1} = x_{n-1}), \end{aligned} \tag{4}$$

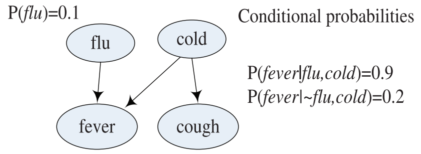

P(*flu*)=0.1
flu
cold
Conditional probabilities
P(*fever*|*flu*,*cold*)=0.9
P(*fever*|~*flu*,*cold*)=0.2
fever
cough

Fig. 2. A Bayesian network example.

where $X_i (1 \leq i \leq n)$ is a random variable of value x_i, stating that the next state depends only on the current state and not on the sequence of events that preceded it.

Furthermore, Markov chains join Monte Carlo simulations to have a Markov chain Monte Carlo (MCMC) method [46]. MCMC is a sampling approach for a desired probability distribution $\pi(x)$, where the sequence of samples satisfies a Markov property. Often, it is used as a mathematical model for some random physical process or complex stochastic systems. If the parameters of the chain are known, quantitative prediction can be made.

Bayesian methods. Bayes theorem, proposed by Bayes in 1763 [47], is based on the probability theory. It expresses relations between two or more events through conditional probabilities and makes inferences:

$$P(H \mid D) = \frac{P(D \mid H) P(H)}{P(D)}, \tag{5}$$

where H is a hypothesis with a prior probability $P(H)$, and $P(H \mid D)$ is H's posterior probability given observed data D. The value of $P(H \mid D)$ can be evaluated based on $P(D \mid H)$, $P(H)$, and $P(D)$.

Confronted with mutually exclusive hypotheses $H_1, H_2, \ldots, H_n$, we have $P(D) = \sum_{i=1}^{n} P(D \mid H_i) P(H_i)$. Therefore, the posterior probability of H_k is

$$P(H_k \mid D) = \frac{P(D \mid H_k) P(H_k)}{\sum_{i=1}^{n} P(D \mid H_i) P(H_i)}, \quad k = 1, 2, \ldots, n. \tag{6}$$

Bayes Theorem suits situations that are lack of direct information about an event. It involves logic reasoning, rather than random sampling as in the Monte Carlo method. Based on the Bayes theorem, a probabilistic graphical model, *Bayesian network*, is developed to represent a set of random variables and their conditional dependencies via a directed acyclic graph.

Example 1. Fig. 2 shows a Bayesian network representing the probabilistic relationships between *disease* and *symptom*. Each node represents a random variable with a prior probability. Edges represent the dependencies between nodes with conditional probabilities. Given the values of $P(flu), P(fever \mid flu, cold)$, and $P(fever \mid \sim flu, cold)$, according to the Bayes theorem, we have

$$\begin{aligned} P(fever \mid cold) &= P(fever, flu \mid cold) + P(fever, \sim flu \mid cold) \\ &= P(fever \mid flu, cold) * P(flu \mid cold) \\ &\quad + P(fever \mid \sim flu, cold) * P(\sim flu \mid cold) \\ &= P(fever \mid flu, cold) * P(flu) \\ &\quad + P(fever \mid \sim flu, cold) * P(\sim flu) \\ &= 0.9 * 0.1 + 0.2 * 0.9 = 0.27. \end{aligned}$$

TABLE 1
Evidence Theory-Based Probability Scope

Hypothesis	Mass	Belief	Plausibility	Prob. Scope
$\emptyset$	0	0	0	[0, 0]
$\{cold\}$	0.5	0.5	0.6	[0.5, 0.6]
$\{flu\}$	0.4	0.4	0.5	[0.4, 0.5]
$\{cold, flu\}$	0.1	1	1	[1, 1]

Dempster-Shafer theory (evidence theory). Dempster-Shafer theory (also called evidence theory), proposed by Dempster and Shafer [48], [49], combines evidence from different sources and arrives at a degree of belief by taking into account all the available evidence. It defines a space of *mass* and the belief mass as a function: $m : 2^X \rightarrow [0, 1]$, where X is the universal set including all possible states, and 2^X is the set of all the subsets of X. For a subset $S \in 2^X$, $m(S)$ is derived from the evidence that supports S:

$$\sum_{S \in 2^X} m(S) = 1. \tag{7}$$

In evidence theory, *belief* and *plausibility* are further defined as the low and upper boundary. *Belief* summarizes all the masses of the subsets of S, meaning all the evidence that fully supports S. *Plausibility* summarizes all the masses of the sets that have intersection with S, meaning all the evidence that partly or fully supports S:

$$belief(S) = \sum_{T \subseteq S} m(T), \quad plausibility(S) = \sum_{T \cap S \neq \emptyset} m(T). \tag{8}$$

The probability of a set $S \in 2^X$ falls into the range of *[belief(S), plausibility(S)]*.

Example 2. Reverting to the *disease* and *symptom* example in Fig. 2, a patient may be diagnosed to catch a cold or have a flu, i.e., $X = \{cold, flu\}$. The mass values $m(\{cold\})$, $m(\{flu\})$, $m(\emptyset)$, and $m(\{cold, flu\})$ (cold or flu) are determined according to the collected evidence from medical instruments or experiences of doctors, as shown in the second column of Table 1. Accordingly, the *belief*s and *plausibility*s of $m(\{cold\})$, $m(\{flu\})$, $m(\emptyset)$, and $m(\{cold, flu\})$ can be calculated. $belief(\{cold\}) = m(\emptyset) + m(\{cold\}) = 0 + 0.5 = 0.5$, $plausibility(\{cold\}) = m(\{cold\}) + m(\{cold, flu\}) = 0.5 + 0.1 = 0.6$. $m(\{cold, flu\})$ derives from the evidence that supports both *cold* and *flu*, such as the symptom of *fever*. However, $m(\{cold\})$ derives from the evidence that only supports *cold*.

The combination of two mass functions (e.g., m_1 and m_2) derived from different (possibly conflicting) sources evidence (e.g., different diagnoses from two doctors) is defined by Dempster as follows:

$$m_{1,2}(S) = \frac{1}{1 - K} \sum_{A \cap B = S \neq \emptyset} m_1(A) m_2(B), \tag{9}$$

where $K = \sum_{A \cap B = \emptyset} m_1(A) m_2(B)$.

Here, K is a normalization factor to ensure the total sum $m_{1,2}$ to be 1. It measures the amount of conflict between the two diagnoses.

Some researchers propose different rules for combining evidence, often with a view to handle conflict in evidence,

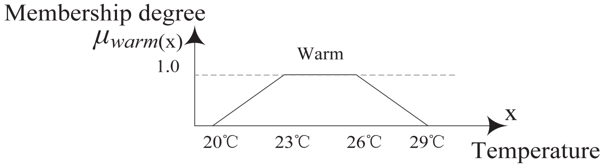


Fig. 3. A *warm* fuzzy set example.

like the transferable belief model [50] and coherent upper and lower previsions method [51].

Compared to the prior and error assumptions needed by the Bayesian method that are sensitive to the results, evidence theory does not enforce any applicable conditions and assumptions. It can thus deal with more uncertainty (including subjective uncertainty arising from experts) than the former Bayesian method. A comparison between evidence theory and Bayesian method in handling epistemic uncertainty has been given by Soundappan et al. [52] with some experiments.

2.3 Fuzzy Theory

Fuzzy theory, proposed by Zadeh in 1965 [37], is another good way to deal with vagueness uncertainty arising from human linguistic labels. It provides a framework for modeling the interface between human conceptual categories and data, thus reducing cognitive dissonance in problem modeling so that the way that humans think about the decision process is much closer to the way it is represented in the machine. The concept of *fuzzy set* extends the notion of a regular crisp set and expresses classes with ill-defined boundaries such as *young*, *good*, *important*, and so on. Within this framework, there is a gradual rather than sharp transition between nonmembership and full membership. A degree of membership in the interval [0, 1] is associated with every element in the universal set X. Such a membership assigning function ($\mu_A : X \rightarrow$ [0, 1]) is called a *membership function* and the set (A) defined by it is called a *fuzzy set*.

Example 3. When we are not sure about the exact centigrade degree of the day, we usually estimate the weather to be *warm*, *cool*, *cold*, and *hot*, and put on more or less clothes accordingly. The concept *warm* can be described through a fuzzy set and its membership function $\mu_{warm}(x)$. We think that 26 °C is the most appropriate temperature for the set *warm*. Then, the grade of membership function of $x = 26^\circ$ is 1, denoted as $\mu_{warm}(26^\circ) = 1$. Similarly, $\mu_{warm}(20^\circ) = 0$, $\mu_{warm}(23^\circ) = 1$, $\mu_{warm}(26^\circ) = 1$, $\mu_{warm}(29^\circ) = 0$. The fuzzy set *warm* can thus be represented as $\{0/20^\circ, 0.33/21^\circ, 1/23^\circ, 1/26^\circ, 0/29^\circ\}$. It can also be expressed in a function, as shown in Fig. 3.

Let A and B be two fuzzy sets. $\mu_{A\cup B}(x) = \max(\mu_A(x), \mu_B(x))$, $\mu_{A\cap B}(x) = \min(\mu_A(x), \mu_B(x))$.

Based on the fuzzy set theory, *fuzzy logic* is developed [39]. A fuzzy proposition is defined on the basis of the universal set X, and a fuzzy set F, representing a fuzzy predicate, such as "*tall.*" Then, the fuzzy proposition "*Tom is tall*" can be written as $\mu_F(Tom)$, representing the membership degree. Different from the classic two-value logic (either *true* or *false*), a fuzzy proposition can take values in the interval [0, 1]. $\mu_F(Tom) = 0.6$ means the truth degree of "*Tom is tall*" is 0.6.

Possibility theory extends fuzzy set and fuzzy logic [38] as a counterpart of probability theory. Let Ω be the universe of discourse. $A_1, A_2, \ldots$ are events, which are subsets of Ω. The possibility distribution Pos is a function from Ω to [0, 1], satisfying the following axioms:

$$\begin{aligned} &1)\ (Normality)\ Pos(\Omega) = 1.\\ &2)\ (Nonnegativity)\ Pos(\emptyset) = 0.\\ &3)\ (Maximality)\ \text{For disjoint events} A_1, A_2, \ldots,\\ &\qquad Pos\left(\bigcup_{i=1}^{\infty} A_i\right) = \max_{i=1}^{\infty} Pos(A_i). \end{aligned} \tag{10}$$

The above axioms show that the possibility distribution satisfies the maximality, which is distinct from the additivity property of probability theory.

2.4 Info-Gap Theory

Info-gap theory is proposed by Ben-Haim in the 1980s [40], [41]. It models uncertainty mainly for decision making and comes up with model-based decisions involving severe uncertainty independent of probabilities. This severe uncertainty belongs to the epistemic uncertainty category and is usually immeasurable or uncalculated with probability distributions and is considered to be an incomplete understanding of the system being managed, thus reflecting the information gap between what one does know and what one needs to know.

The info-gap decision theory consists of three components (i.e., *performance requirements, uncertainty model*, and *system model*) and two functions (i.e., *robustness function* and *opportuneness function*).

The *performance requirements* state the expectations of the decision makers, such as the minimally acceptable values, the loss limitations, and the profit requirements. These requirements form the basis of decision making. Different from probability theory that models uncertainty with probability distributions, the *uncertainty model* of info-gap theory models uncertainty in the form of nested *subsets*: $U(\alpha, \tilde{u})$, where $\tilde{u}$ is a point estimate of an uncertain parameter, and $\alpha(\geq 0)$ is the deviation around $\tilde{u}$. The *robustness* and *opportuneness functions* determine the settings of α. An example uncertainty model is

$$U(\alpha, \tilde{u}) = \{u : |u - \tilde{u}| \leq \alpha\tilde{u}\}, \tag{11}$$

which satisfies two basic axioms:

$$\begin{aligned} &1)\ (Nesting)\ \text{For}\ (\alpha < \alpha'), U(\alpha, \tilde{u}) \subseteq U(\alpha', \tilde{u}).\\ &2)\ (Contraction)\ U(0, \tilde{u}) = \{\tilde{u}\}. \end{aligned} \tag{12}$$

The *robustness function* represents the greatest level of uncertainty, at which minimal performance requirements are satisfied and failure cannot happen, addressing the pernicious aspect of uncertainty. The *opportuneness function* exploits the favorable uncertainty leading to better outcomes, focusing on the propitious aspect of uncertainty. Through the two functions, uncertainty can be modeled, and the information gap can be quantified and further be reduced with some actions [40], [41]. The decision-making process actually involves the construction, calculation, and optimization of the two functions.

The third component of info-gap theory is an overall *model of system* considering all factors and requirements.

Example 4. A worker faces a choice of cities to live: City A or City B. The salary s he could earn is uncertain. In City A, he might earn 80$ as an estimate every week, i.e., $\widetilde{s} = 80\$$. If he earns less than 60$, he cannot afford the lodging and is in danger of sleeping in the street. But if he earns more than 95$, he can afford a night's entertainment as a windfall. In City B, he might earn 100$ as an estimate. The lodging costs 80$, and the entertainment costs 150$.

Based on the system requirements (avoiding sleeping in the street, or affording a night's entertainment), for City A, the uncertain salary of a worker can be represented as a subset: $U(\alpha, 80\$) = \{s : |s - 80\$| \leq 80\$ * \alpha, \alpha \geq 0\}$. That is, the worker's income s falls into the interval $[80\$^*(1-\alpha), 80\$^*(1+\alpha)]$. Then, the robustness/opportuneness functions determining α are

$$\begin{aligned} Robust(s, 60\$) &= \max\left\{\alpha : \min_{s\in U(\alpha,80\$)} s \geq 60\$\right\}\\ &= \max\{\alpha : 80\$ * (1-\alpha) \geq 60\$\} = \max\{\alpha : \alpha \leq 0.25\}\\ &= 0.25.\\ Opportune(s, 95\$) &= \min\left\{\alpha : \max_{s\in U(\alpha,80\$)} s \geq 95\$\right\}\\ &= \min\{\alpha : 80\$ * (1+\alpha) \geq 95\$\} = \min\{\alpha : \alpha \geq 0.1875\}\\ &= 0.1875. \end{aligned} \tag{13}$$

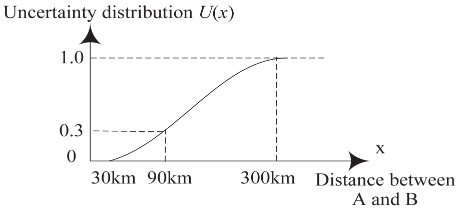
In a similar fashion, for City B, the α values returned from the robustness and opportuneness functions are 0.2 and 0.5, respectively. As α represents the deviation from the estimate value, the bigger α takes, the less danger the salary is below the hazard threshold 60$ (or 100$); and the smaller α takes, the higher chance to enjoy a night's entertainment. Therefore, moving to City A appears to be better than to City B.

Info-gap theory applies to the situations of limited information, especially when there is not enough data for other uncertainty handling techniques such as probability theory. Ben-Haim [40] once argues that probability theory is too sensitive to the assumptions on the probabilities of events. In comparison, info-gap theory stands upon an uncertain range rather than a probability and is thus more robust.

2.5 Derived Uncertainty Theory

A derived uncertainty theory from probability and fuzzy theories is later presented by Liu [42], [43] in 2007 to handle human ambiguity uncertainty. Its three key concepts, i.e., *uncertain measure, uncertain variable,* and *uncertain distribution* are defined as follows:

Let Γ be a nonempty set. $\mathcal{L}$ is a σ-algebra of Γ. A σ-algebra over a set A is defined as a nonempty collection of all subsets of Γ (including Γ itself). Each element in $\mathcal{L}$ is an event, expressed as $\Lambda_1, \Lambda_2, \ldots$. The *uncertain measure* $M(\Lambda)$ represents the occurrence level of an event. $(\Gamma, \mathcal{L}, M)$ is called an uncertainty space with the following axioms:



99-Table depicting uncertain distance between A and B

$U(x)$	0.01	0.02	...	**0.30**	0.31	...	0.99
x	30	32	...	**90**	91	...	300

Fig. 4. An uncertainty distribution example in a 99-Table.

$$\begin{aligned} &1)\ (Normality)\ M(\Gamma) = 1.\\ &2)\ (Self\text{-}duality)\ M(\Lambda) + M(\Lambda^C) = 1.\\ &3)\ (Subadditivity)\ M\left(\bigcup_{i=1}^{\infty}\Lambda_i\right) \leq \sum_{i=1}^{\infty} M(\Lambda_i).\\ &4)\ (Product\ measure)\ M\left(\prod_{k=1}^{n}\Lambda_k\right) = \min_{1\leq k\leq n} M_k(\Lambda_k). \end{aligned} \tag{14}$$

An *uncertain variable* is a function $\xi(\Gamma, \mathcal{L}, M) = 2^R$, where $(\Gamma, \mathcal{L}, M)$ is an uncertain space and 2^R is a set of real numbers. For any real number x, the *uncertain distribution* of an uncertain variable ξ is an increasing function defined as: $U(x) = M(\xi \leq x)$.

In the derived uncertainty theory, instead of uncertainty distribution functions, a discrete 99-Table (Fig. 4) is used to state the uncertainty distribution. A 99-Table usually accommodates 99 points in the curve of uncertain distribution and is helpful when uncertainty functions are unknown. Considering comprehensive requirements of storage capacity and precision, 99 points are usually taken from the uncertain distribution for calculation. However, this is not strictly restricted. One can also take 80 or 150 points according to different precision and storage requirements.

Example 5. Suppose an application is interested in the city distances below a threshold. We can view the distance between Cities A and B as an uncertain variable. The uncertain distribution $U(x)$ with its discrete 99-Table expression is plotted in Fig. 4. The second row of 99-Table represents the values that the uncertain variable can take, and the first row means the corresponding uncertainty. $U(90\text{ km}) = M(x \leq 90\text{ km}) = 0.3$ states that the distance between A and B is lower than 90 kilometers with the uncertainty degree 0.3.

Derived uncertainty theory also possesses *contradiction* and *excluded-middle* properties. Let H denote a proposition (e.g., "*restaurant ABC has a good reputation*") with a truth value $T(H)$. $T(H \vee \neg H) = 1$ and $T(H \wedge \neg H) = 0$. Besides, it conforms to the monotonicity. It does well in describing interval-based uncertainty measures and is suitable to handle subjective epistemic uncertainty in risk analysis, reliability analysis, and finance [43].

Table 2 summarizes the above four uncertainty theories in managing diverse uncertainty in the real world.

3 Uncertainty Handling Practices

Table 3 lists some prototypical practices in the fields of economics, engineering, ecology, and information science.

TABLE 2
Summary of the Four Uncertainty Handling Theories

	Probability Theory	Fuzzy Theory	Derived Uncertainty Theory	Info-Gap Theory
Managed Uncertainty	Randomness	Ambiguity with ill-defined boundaries	Human's subjectiveness uncertainty	Immeasurable factors of incomplete understanding
Uncertainty Measure	Probability measure $P(x)$	Membership function $\mu_A(x)$	Uncertain measure $M(x)$	A nested set $U(\alpha, \widetilde{u})$
Uncertainty Distribution	Probability density function (pdf) and cumulative distribution function (cdf)	Possibility distribution $Pos(x)$	99-Table (Uncertain distribution $U(x)$)	-
Property	Additivity $P(\overset{\infty}{\underset{i=1}{\cup}} A_i) = \sum_{i=1}^{\infty} P(A_i)$	Non-additivity $Pos(\overset{\infty}{\underset{i=1}{\cup}} A_i) = \max_{i=1}^{\infty} Pos(A_i)$	Sub-additivity $M(\overset{\infty}{\underset{i=1}{\cup}} A_i) \leq \sum_{i=1}^{\infty} M(A_i)$	Nesting, for $\alpha < \alpha'$ $U(\alpha, \widetilde{u}) \subseteq U(\alpha', \widetilde{u})$
	Self-duality $P(A) + P(A^c) = 1$	No self-duality	Self-duality $M(A) + M(A^c) = 1$	Contraction $U(0, \widetilde{u}) = \{\widetilde{u}\}$
Operation	Add when mutually exclusive $P(A \cup B) = P(A) + P(B)$	Maximum $Pos(A \cup B) = \max(Pos(A), Pos(B))$	Maximum when independent $M(A \cup B) = \max(M(A), M(B))$	-
	Multiply when independent $P(A \cap B) = P(A) * P(B)$	Minimum $Pos(A \cap B) = \min(Pos(A), Pos(B))$	Minimum when independent $M(A \cap B) = \min(M(A), M(B))$	-

TABLE 3
Uncertainty Handling Techniques in Practice

Field	Probability Theory	Fuzzy Theory	Derived Uncertainty Theory	Info-Gap Theory
Economics	Decision Tree in budget making	-	-	Credit risk analysis
Engineering	Risk analysis: Event Tree Analysis, Fault Tree Analysis, Probabilistic life cycle assessment	Fuzzy life cycle assessment	Reliability analysis, Risk analysis	-
Ecology	Population forecasting	-	-	Conservation management
Information Science	Social networking	Social networking,	Project scheduling	-

3.1 Uncertainty Handling in Economics

Economics is a classical field for uncertainty and risk analysis [53], [54], [55]. While uncertainty is generally considered to be lack of certainty with possible states and multiple outcomes, risk refers to a potential loss or undesired effect, which may cause something bad or unexpected [56]. In economic risk analysis, risk is usually expressed as a quantity, measured with the use of probabilities. In comparison, severe uncertainty is restricted to the nonquantitative case, caused by the lack of information or knowledge [53], where the underlying statistical distribution is unknown.

3.1.1 Probability-Based Economic Risk Analysis

Economic budget planning is a common activity involving risk analysis, where a probabilistic *decision tree* is usually exploited in the process. All possible situations with respective probabilities are outlined in the decision tree, based on which possible target benefits are computed.

Example 6. Suppose a factory wants to build a new workshop to produce products for 10 years. The construction cost for a large workshop is 90K dollars and 40K for a small workshop. The factory expects to get some revenues from the products according to the sales situation. Fig. 5 shows the decision tree that illustrates all possible financial solutions. Based on it, the expected 10-year revenues are computed as follows:

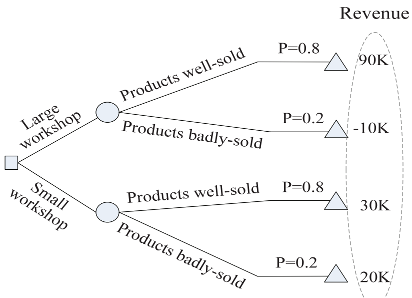
$$R_{large} = 0.8 \times 90 \times 10 + 0.2 \times (-10) \times 10 - 90 = 610\text{K},$$
$$R_{small} = 0.8 \times 30 \times 10 + 0.2 \times 20 \times 10 - 40 = 240\text{K}.$$

The result signifies that building a large workshop is better than building a small one.

3.1.2 Info-Gap-Based Credit Risk Analysis

Besides probability theory, info-gap theory also initiates a new way to assist economic risk evaluation during decision making [57].

Example 7. In offering loans, financial institutions try to make optimal decisions to manage credit risk by reducing potential losses from arrears on loans. In credit risk analysis [58], they can classify customers into categories $1, 2, \ldots, N$ of different credit rates. The corresponding



Yearly revenue after workshop construction

	Large Workshop	Small Workshop
Products Well-Sold	90K	30K
Products Badly-Sold	-10K	20K

Fig. 5. A probabilistic decision tree in an example plan.

probability of default (arrears) in category i $(1 \leq i \leq N)$ is denoted as d_i, and $d = (d_1, d_2, \ldots, d_N)$. Let $\tilde{d}$ be the best estimate of d, and α constrain the scope of uncertainty. In info-gap theory, the *uncertainty model* of d can be represented as a nested subset:

$$U(\alpha, \tilde{d}) = \left\{d : d = \tilde{d} + \varepsilon, \varepsilon^T C^{-1} \varepsilon \leq \alpha^2 \right\}, \quad \alpha \geq 0, \tag{15}$$

where C is an $N \times N$ positive symmetric matrix, whose entry C_{ij} is a model parameter that could be taken as an element of a correlation matrix.

Let l_i and r_i denote the loan amount and interest rate to customers in category i. Let $l = (l_1, l_2, \ldots, l_N)$ and $r = (r_1, r_2, \ldots, r_N)$. The total loan amount $L = l_1 + l_2 + \cdots + l_N$. Further, let δ and ρ denote the fractions of loss and profit in terms of the percentage of L. To ensure the normal business, two *performance requirements* must be satisfied. That is, the loss is not greater than $\delta * L$, and the profit is not less than $\rho * L$.

Based on the above uncertainty model, as well as the loss and profit requirements, one can compute the relations among the arrear probabilities, loan amounts, and interest rates with the *system model*:

$$\begin{aligned} Robust_{Loss} &= \max\left\{\alpha : \left(\max_{d \in U(\alpha, \tilde{d})} l^T d\right) \leq \delta * L\right\}. \\ Robust_{Profit} &= \max\left\{\alpha : \left(\min_{d \in U(\alpha, \tilde{d})} (l^T r - l'^T d)\right) \geq \rho L\right\}. \end{aligned} \tag{16}$$

l' is a vector, and $l'_i = (1 + r_i) * l_i (i = 1, 2, \ldots, N)$. $l^T r$ expresses the income from loan interests, and $l'^T d$ expresses the loss due to arrears.

With the analysis of the credit risk system model, the financial institution can make the following decision: The higher the probability of arrear a customer has, the higher interest rate and less amount of loan the bank should assign to him/her. The robust-optimal decisions can be obtained by maximizing $Robust_{Loss}$ and $Robust_{Profit}$ with respect to decision variables such as l and r. For detailed numerical analysis, refer to [58].

Besides credit risk analysis, info-gap decision theory is also applied to investment risk analysis, policy formulation, microeconomics of demands, and the equity premium puzzle [57].

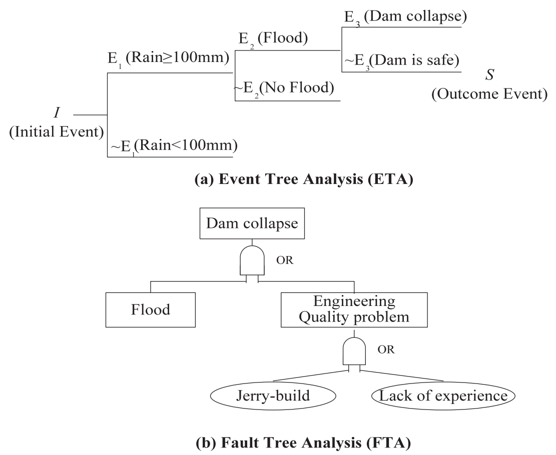
3.2 Uncertainty Handling in Engineering

Uncertainty exists in engineering risk management and life-cycle assessment (LCA).

3.2.1 Probabilistic Engineering Risk Analysis

Back to the 1970s, probabilistic risk analysis has been used to evaluate the risk of operations of nuclear power plants [59]. In probabilistic risk analysis, two data structures (i.e., *event tree analysis* (ETA) and *fault tree analysis* (FTA)) are usually adopted [60]. They both model the problem or process in the form of *tree* and can make both quantitative and qualitative analysis.

ETA illustrates an inductive process from reasons to results. It starts from an initial event, constructs the tree following casual relations, and comes up to some outcome events. ETA can aid to make predictions. In comparison, FTA plots a deductive process from results to reasons to facilitate discovery of some potential risks or failure causes. ETA and FTA can be combined for synthetic research.



(a) Event Tree Analysis (ETA)

(b) Fault Tree Analysis (FTA)

Fig. 6. *Dam* risk analysis example.

Example 8. Fig. 6a shows an ETA tree example in measuring a *dam*'s reliability, where I is an initiating event, E_1, E_2, and E_3 are different events, and S is an outcome event. The possibility of outcome S can be gained by $P(S) = P(I)P(E_1)P(E_2)P(\sim E_3)$. Fig. 6b shows a *top-down* analysis by listing factors causing *dam collapse*.

Greenland [61] describes the use of Monte Carlo and Bayesian risk uncertainty assessment in analyzing skin cancer risks from coal tar containing products.

3.2.2 Derived Uncertainty Theory-Based Reliability Analysis

Liu [43] analyzes system reliability with respect to the factors of lifetime, production rate, cost, and profit based on derived uncertainty theory. The term *reliability index* is used to measure the degree of hazard, representing the uncertainty of system reliability.

Assume a cascading system is composed of n components. It fails if any of its components does not work. Let $\xi_1, \xi_2, \ldots, \xi_n$ denote the lifetimes of the n components. The system lifetime is $min(\xi_1, \xi_2, \ldots, \xi_n)$.

If the system lifetime is expected to be longer than T, the system reliability function can be defined as

$$f(\xi_1, \xi_2, \ldots, \xi_n) = min(\xi_1, \xi_2, \ldots, \xi_n) - T.$$

Then, the system can work reliably, if and only if

$$f(\xi_1, \xi_2, \ldots, \xi_n) = min(\xi_1, \xi_2, \ldots, \xi_n) - T \geq 0.$$

The reliability index of the system is defined as

$$Reliability = M(f(\xi_1, \xi_2, \ldots, \xi_n) \geq 0) = \alpha.$$

Due to $M(\xi \leq x) = U(x)$, we can get

$$\begin{aligned} &M(f(\xi_1, \xi_2, \ldots, \xi_n) \geq 0) \\ &\quad = 1 - M(f(\xi_1, \xi_2, \ldots, \xi_n) < 0) = 1 - U(0) = \alpha. \end{aligned} \tag{17}$$

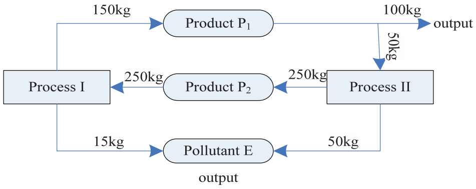


Fig. 7. A life-cycle system example [67].

According to the 99-Table method, if $\xi_i (i = 1, 2, \ldots, n)$ can be represented as the following 99-Table:

$U_i(x)$	0. 01	0. 02	$\cdots$	0. 99
x	x_1^i	x_2^i	$\cdots$	x_{99}^i

then the uncertainty distribution of $f(\xi_1, \xi_2, \cdots, \xi_n) = \min(\xi_1, \xi_2, \ldots, \xi_n) - T$ can be represented as the 99-Table:

$U(x)$	0.01	0.02	$\cdots$	0.99
x	$\min\limits_{1\leq i\leq n} x_1^i - T$	$\min\limits_{1\leq i\leq n} x_2^i - T$	$\cdots$	$\min\limits_{1\leq i\leq n} x_{99}^i - T$

Then, according to (17), we can use the 99-Table of $f(\xi_1, \xi_2, \ldots, \xi_n)$ to compute the value of α. In this table, we find the value $x = 0$, then the corresponding $U(x = 0) = 1 - \alpha$. Thus, the risk index α of the cascading system can be calculated.

With the risk index, decision makers can evaluate systems, make predictions, and take prevention actions. Similarly, the reliability index can also be presented based on a Boolean system.

3.2.3 *Probabilistic and Fuzzy LCA*

LCA refers to the evaluation of environmental and social impacts relating to the life cycle of products, going through all the stages of the products from raw materials, semi-finished products, finished products, to products' waste and recovery. Various uncertainties [62], [63], [64] arising from data sources, models, measurement errors, preferences of analysts, and physical systems need to be coped with for comparison and decision-making purposes [65].

Monte Carlo simulation, Bayesian statistics, and fuzzy theory have been applied to LCA. Although some researchers think that fuzzy sets are more suitable than random sampling methods, not only due to the ambiguity in LCA, but also due to the less computing time with fuzzy theory [66], [67], Monte Carlo is the most widely used approach in the literature [63]. Next, we illustrate a fuzzy LCA approach and a probabilistic LCA approach.

Fuzzy LCA approach. Tan [67] presents a matrix-based fuzzy model to deal with data variability during the evaluation of pollutant emission in LCA. Assume in a two-process life-cycle system, two commodities P_1 and P_2 are involved, and pollutant E is released. As illustrated in Fig. 7, Process I consumes P_2 and produces P_1 and pollutant E, and Process II consumes P_1 and produces P_2 and pollutant E. Through the LCA, one can compute the total emissions of pollutant E.

The amount ratios of commodities P_1, P_2, and pollutant E are uncertain, which can be represented as a fuzzy number. As shown in Fig. 8, the amount ratio of commodity 1 to commodity 2 in Process I is represented as a fuzzy number, written as $(0.5, 0.6, 1)_T$, where 0.5 and 1 are the least plausible values as the lower and upper boundaries of the ratio, 0.6 is the most plausible value, and the subscript T expresses that the fuzzy number is in the shape of triangle distribution. At any given degree of possibility α, it is possible to find a corresponding interval (L, U). In an *α-cut* where $\alpha = 0.4$, the resulting interval is $(0.54, 0.84)$. Other amount ratios of the life-cycle system are shown in Fig. 9.

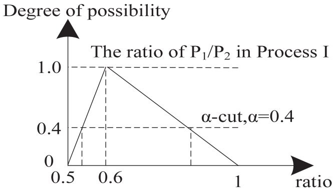


Fig. 8. A triangular distribution of fuzzy numbers in LCA.

According to the uncertain ratios in the second and third columns, and the output amount of the final column, the emission x of pollutant E can be calculated based on the following matrix-based fuzzy model:

$$g_{U,\alpha} = B_{U,\alpha} A_{L,\alpha}^{-1} f, \quad g_{L,\alpha} = B_{L,\alpha} A_{U,\alpha}^{-1} f. \tag{18}$$

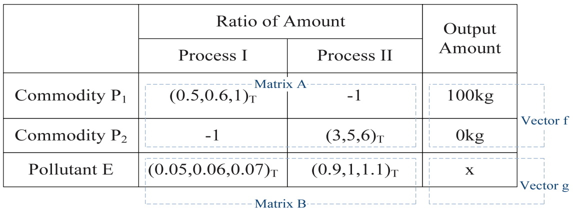
In this model, at the degree of possibility α, $g_{U,\alpha}$ and $g_{L,\alpha}$ are the upper and lower bounds of the emission inventory vectors, $B_{U,\alpha}$ and $B_{L,\alpha}$ are the upper and lower bounds of the emission intervention matrix, $A_{U,\alpha}$ and $A_{L,\alpha}$ are the upper and lower bounds of commodity-relevant technology matrix, and f is the functional unit vector of output. With this equation, the lower and upper bound of emission can be estimated at the degree of possibility α.

According to Fig. 9, at $\alpha = 1$,

$$g_{U,\alpha} = (x), \quad B_{U,\alpha} = (0.06 \quad 1)$$
$$A_{L,\alpha} = \begin{pmatrix} 0.6 & -1 \\ -1 & 5 \end{pmatrix}, \quad \text{f} = \begin{pmatrix} 100 \\ 0 \end{pmatrix}. \tag{19}$$

Using the fuzzy model, we can get $g_{U,\alpha} = 65$. That is, the most plausible value of pollutant E is 65 kg, as shown in Fig. 7. Varying the value of α, we can gain different emissions, forming a fuzzy distribution of pollutant E. The environmental impacts can thus be derived.

Probabilistic LCA approach. In a Monte Carlo-based incineration system [68], wastes are collected in different cities, and transported to incineration plants. After incineration, electricity is recovered and ash is disposed. Assume the probability distribution of CO_2 emission satisfies the normal distribution $N(1\,\text{kg}, 0.05\,\text{kg})$ in the use of diesel. Suppose one chooses $0.9\,\text{kg}$ as one CO_2 emission value. A concrete total CO_2 emission from the transportation process and the incineration process can then be computed according to a predefined formula. To repeat this process



	Ratio of Amount		Output Amount
	Process I	Process II	
Commodity P₁	$(0.5,0.6,1)_T$	-1	100kg
Commodity P₂	-1	$(3,5,6)_T$	0kg
Pollutant E	$(0.05,0.06,0.07)_T$	$(0.9,1,1.1)_T$	x

Matrix A
Vector f
Vector g
Matrix B

Fig. 9. A matrix-based fuzzy model for LCA [67].

for hundreds of times with different parameter settings by the Monte Carlo method, one can obtain a statistic final result of CO_2 emission [68].

3.3 Uncertainty Handling in Ecology

Good forecasting of ecological phenomena can help policy making and social/environmental planning.

3.3.1 Probabilistic Forecasting of Population

Confronted with the threat of global population increase, population forecasting constitutes another focal research point in ecology [69], [70], [71]. Basically, three factors (i.e., fertility, mortality, and migration) affect demographic change. Population forecasting needs to analyze each individual factor and their combined effects. Throughout analysis procedures, much uncertainty lying in population growth has to be coped with. According to the statistics made against demographic forecasting from 1985 to 2005, errors in forecasting come from four sources, which are *model misspecification, parameter estimation, random variation,* and *informed judgment* [72], [73].

To address them, probabilistic forecasting methods are introduced to deal with some component uncertainty in the manner of stochastic population renewal. Three complementary approaches are particularly developed to estimate forecast uncertainty: *model-based ex-ante error estimation, expert-based ex-ante error estimation,* and *ex-post error estimation,* relying on the extrapolative techniques, expert knowledge, and past forecast, respectively.

Besides, the Bayesian method is also applied to population forecasting with some flexibility [74]. The relation between the uncertain parameter θ and observed data $y_{\{T\}} = \{y_1, y_2, \ldots, y_T\}$ is formulated as

$$f(\theta \mid y_{\{T\}}) = \frac{f(y_{\{T\}} \mid \theta) f(\theta)}{f(y_{\{T\}})}, \tag{20}$$

where $f(\theta)$ is the prior distribution.

According to Bayes theorem, precedent values of y_T can be predicted:

$$\begin{aligned} &f(y_{T+1}, \ldots, y_{T+K} \mid y_{\{T\}}) \\ &= \int f(\theta \mid y_{\{T\}}) \prod_{k=1}^{K} f(y_{T+k} \mid y_{\{T+k-1\}}, \theta) d\theta. \end{aligned}$$

In this way, the next K values can be calculated with the joint predictive and posterior distribution based on the observed data $y_{\{T\}}$.

3.3.2 Info-Gap-Based Conservation Management

Conservation management intends to avoid potential risk of population decline or extinction and maximize opportunities of population persistence. Conservation biologists usually need to decide appropriate actions taken for endangered species. In the following example, a *utility* measurement is used to represent the environmental outcome of a conservation action.

Assume there are three options of conservation management that are translocation, new reserve, and captive breeding, denoted as $a_j (j = 1, 2, 3)$. The three options may lead to four possible outcomes that are poaching, loss of habitat, demographic accidents, and disease. Let $p_i (i = 1, 2, 3, 4)$ denote the probability of each outcome. Further, let v_{ij} represent the utility association between the jth option and the ith outcome. Then, the expected utility of the jth option is $EV(a_j) = \sum_{i=1}^{4} p_i v_{ij} (j = 1, 2, 3)$.

In the info-gap decision theory, the uncertainty vector p and v can both be represented with a subset, $U_p(\alpha, \tilde{p})$ and $U_v(\alpha, \tilde{v})$, satisfying

$$\frac{p_i - \tilde{p}_i}{\tilde{p}_i} \leq \alpha, \quad \frac{v_{ij} - \tilde{v}_{ij}}{\tilde{v}_{ij}} \leq \alpha (\alpha \geq 0,\ i = 1, 2, 3, 4,\ j = 1, 2, 3). \tag{21}$$

There is a critical value EV_c below which the utility is unacceptable. Then, the robustness function for option a_j $(j = 1, 2, 3)$ can be formulated by

$$\begin{aligned} &Robust(a_j, EV_c) \\ &\quad = \max\left\{ \alpha : \min_{p \in U_p(\alpha, \tilde{p}), v \in U_v(\alpha, \tilde{v})} \sum_{i=1}^{4} p_i v_{ij} \geq EV_c \right\}. \end{aligned}$$

Robust is the robustness function, representing the greatest horizon of uncertainty up to which all probabilities and utilities result in an expected utility no worse than EV_c. Through experiments, when $EV_c = 0.07$, the option "new reverse" holds the larger robustness ($Robust = 0.34$) than the other two options. Hence, it is a good choice.

3.4 Uncertainty Handling in Information Science

In the information domain, there exist unreliable information sources, system errors, imprecise information gathering methods, and/or model restrictions [75]. Unreliable information sources may be due to fault-reading instruments, incorrect input forms, and so on. System errors lie in transmission noises or delays in processing updated transactions, and so on. Information gathering may be affected by constantly varying phenomena. In the modeling process, some approximation techniques may be required and used, which result in uncertainty as well.

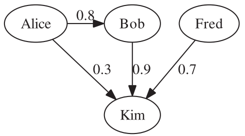
3.4.1 Probabilistic and Fuzzy Social Networking

A social network accommodates various uncertainty relationships of people such as the belief degree of two friends to be handled, where probability theory and fuzzy theory are generally brought in to deal with the uncertain situations. We use a music recommendation example to show how the two approaches are applied.

Probability approach. People may easily be influenced by friends in choosing music [76]. Fig. 10 gives a probabilistic social networking example, where directed edges indicate the influence upon song selection by each other. For example, *Alice* influences *Kim* with a probability of 0.3. Three probabilistic tables, *Preference* (*Name, Genre, Prob.*), *Song* (*Name, Genre, Prob.*), and *MusicInfluence* (*Name*1, *Name*2, *Prob.*), record users' song preferences, songs' genres, and music influence with certain probabilities, respectively.

To find out users who probably like song A, the following SQL can be issued:

```
SELECT Preference.Name
FROM Preference P, Song S
WHERE P.Genre=S.Genre AND S.Name='A'
```



(a) $MusicInfluence$

$Name1$	$Name2$	$Prob.$
Alice	Bob	0.8
Alice	Kim	0.3
Bob	Kim	0.9
Fred	Kim	0.7

(b) $Preference$

$Name$	$Genre$	$Prob.$
Alice	Classical	0.8
Alice	Pop	0.8
Bob	Rap	0.5
Bob	Pop	0.5
Kim	Country	0.75
Fred	Pop	0.2

(c) $Song$

$Name$	$Genre$	$Prob.$
A	Country	0.1
A	Pop	0.5
B	Rap	0.5
C	Classical	0.8
D	Country	0.6

Fig. 10. A probabilistic social networking example [76].

The probability that *Kim likes song A* is thus $0.75 \times 0.1 = 0.075$. Furthermore, considering *MusicInfluence* in Fig. 10a and song A is of genre *Country* with probability 0.1 and *Pop* with probability 0.5, *Alice, Bob,* and $Fred$ may all recommend song A to *Kim*. The final result probability that *Kim* accepts song A is $0.075 \times (1-(1-0.3)\times(1-0.9)\times(1-0.7)) = 0.073$.

Fuzzy approach. Considering the ambiguous properties of human thinking, fuzzy sets are employed to represent and evaluate users' social relationships. Some intuitive linguistic concepts (like *strong*, *weak*, or *medium*), rather than a raw probability value, can be used to express the users' influence factor upon song selection, as shown in Table 4a.

A similarity relation of influence can be defined in a matrix (Table 4b to replace classical membership functions of fuzzy sets—*strong, weak*, and *medium*). Here, 0.7 means that "*medium*" satisfies the concept "*strong*" with a degree of 0.7. Users who strongly influence *Kim* include (*Bob*, 1), (*Fred*, 0.7), and (*Alice*, 0.2).

3.4.2 Uncertain Decision Making in Project Management

Fuzzy approach. Efficient project management needs to be aware of project goal and various constraints toward a high-quality project output. Three factors (*project cost* (PC), *project duration*, and *project quality* (PQ)) can generally quantify the internal efficiency of project management in terms of how well the project is managed and executed [77], [78].

Dweiri and Kablan [79] conduct a fuzzy decision making to evaluate the internal efficiency of project management. It describes the PC, *project duration* (PT), and PQ through fuzzy sets. For instance, the PC is fuzzified into *very low, low, medium, high,* and *very high*. Different factors have different priorities to the project management, which are defined as project cost weighting factor (PCWF), project time weighting factor (PTWF), and project quality weighting factor (PQWF). It satisfies $PCWF + PTWF + PQWF = 1$.

The project management internal efficiency can be measured based on some fuzzy rules like "IF *project cost is low* AND *project cost weighting factor is high,* THEN *project management internal efficiency is very high.*"

Derived uncertainty theory approach. The derived uncertainty theory is used to solve the project scheduling problem in the form of uncertain programming. This process involves allocating resources to reduce total cost and complete the project in time. In the project scheduling model [43], these two factors are considered:

TABLE 4
Fuzzy *MusicInfluence* Relation and Similarity Matrix

(a) $MusicInfluence$

$Name1$	$Name2$	$InfluenceDegree$
Alice	Kim	Weak
Alice	Bob	Strong
Bob	Kim	Strong
Fred	Kim	Medium

(b) $Similarity$

$Influence$	**Strong**	**Medium**	**Weak**
Strong	1	0.7	0.2
Medium	0.7	1	0.4
Weak	0.2	0.4	1

$$\begin{cases} \min & E(C(x,\xi)) \\ s.t. & M(T(x,\xi) \le T^0) \ge \alpha,\ x \ge 0. \end{cases} \tag{22}$$

Here, x is the allocating time vector needed for all activities. ξ is the uncertain duration time vector of all activities. $E(C(x,\xi))$ is the expectation of the total cost $C(x,\xi)$ to be minimized, T is the complete time of the project that should be earlier than the due date T^0, with an occurrence level not less than α.

If the uncertain distribution of the total cost $C(x,\xi)$ can be represented with a 99-Table:

$U(C)$	0. 01	0. 02	$\cdots$	0. 99
C	c_1	c_2	$\cdots$	c_{99}

then the expectation $E(C(x,\xi)) = (c_1 + c_2 + \cdots + c_{99})/99$.

4 Uncertainty Processing in Databases

In the data management field, uncertain information is typically managed by a *probabilistic* or *fuzzy* database whose theoretic foundation is probability theory or fuzzy theory. We illustrate various uncertainty handling efforts by the database community, with an emphasis on uncertainty database model and query processing.

4.1 Probabilistic Data Management

Databases based on the classic probability theory, Monte Carlo methods, and evidence theory are described.

4.1.1 Classic Probabilistic Data Management

In a tuplewise probabilistic database relation, each tuple can be regarded as a description of a *basic* probabilistic event and associated with an explicitly given event identifier [80], [81], as shown in Table 5a. The column *Prob.* means the probability that a tuple belongs to the relation (e.g., tuple $r1$ belongs to the relation *Restaurant* with a probability of 0.7, or $r1$ does not appear in *Restaurant* with a probability of 0.3). Tuples within a probabilistic relation as well as among relations are assumed to be independent. The probabilistic database is based on the *possible world semantics* that is a probability distribution on all database instances [81]. It can be regarded as a finite set of database instances with the same schema. Each tuple in the database may or may not appear in a database instance, and the database instance is associated with a probability. Table 5b shows all the possible worlds of relation *Restaurant*. The probability of each possible world can be computed by multiplying its tuples' probabilities.

TABLE 5
Probabilistic *Restaurant* Relation and Its Possible Worlds

(a) Probabilistic $Restaurant$ Relation

ID	$RName$	$Type$	$Discount$	$Prob.$
r1	Starbucks	Dinning	8	0.7
r2	PizzaHut	Dinning	7	0.8
r3	KFC	Dinning	9.5	0.55

(b) Possible Worlds of Relation $Restaurant$

Possible world	Probability	Tuple list sorted by $Discount$
$W_1 = \{r_1, r_2, r_3\}$	$0.7 * 0.8 * 0.55 = 0.308$	r_3, r_1, r_2
$W_2 = \{r_1, r_3\}$	$0.07 * (1 - 0.8) * 0.55 = 0.077$	r_3, r_1
$W_3 = \{r_2, r_3\}$	$(1 - 0.7) * 0.8 * 0.55 = 0.132$	r_3, r_2
$W_4 = \{r_3\}$	$(1 - 0.7) * (1 - 0.8) * 0.55 = 0.033$	r_3
$W_5 = \{r_1, r_2\}$	$0.7 * 0.8 * (1 - 0.55) = 0.252$	r_1, r_2
$W_6 = \{r_1\}$	$0.7 * (1 - 0.8) * (1 - 0.55) = 0.063$	r_1
$W_7 = \{r_2\}$	$(1 - 0.7) * 0.8 * (1 - 0.55) = 0.108$	r_2
$W_8 = \varnothing$	$(1 - 0.7) * (1 - 0.8) * (1 - 0.55) = 0.027$	-

Besides tuplewise uncertainty, attribute-level uncertainty (e.g., *Location* attribute in the *CustomerLoc* relation in Table 6b) considers an attribute value as a set of discrete possible values or continuous values modeled through a probability density function [82], [83].

Querying over a probabilistic database will deliver a probabilistic result relation where each result tuple is associated with a *complex* event expression that is a Boolean combination of the events corresponding to the base tuples from which it is derived. Evaluation of typical *aggregate, join, range, top-k,* as well as *lineage* and *correlation*-based queries is done based on the possible world semantics.

(**Aggregate Query**) *"Find the average discount of restaurants."*

```
SELECT AVG (Discount) AS avgDiscount
FROM Restaurant
```

Each possible world in Table 5c leads to a partial result, and the final outcome is $(avgDiscount, Prob.) = \{(8.17, 0.308), (8.75, 0.077), (8.25, 0.132), (9.5, 0.033), (7.5, 0.252), (8, 0.063), (7, 0.108), (NULL, 0.027)\}$. As the world Ø does not include any tuple, *NULL* is returned as a result alternative according to [84].

Instead of listing all of the alternatives in exhaustive aggregation, some variants of aggregation just consider a low bound, a high bound, or a mean aggregate value of non-*NULL* alternatives with the confidence setting as 1.0 (i.e., (7, 1), (9.5, 1), or (7.68,1)) for query efficiency [84]. Aggregation on data streams can also be computed efficiently given a specific accuracy [85].

(**Join Query**) *"Find restaurants of interest to customers."*

```
SELECT RName
FROM Restaurant, Customer
WHERE Restaurant.Type=Customer.Interest
```

TABLE 6
Probabilistic *Customer* and *CustomerLoc* Relations

(a) Probabilistic $Customer$ Relation (Tuple-Level Uncertainty)

ID	$CName$	$Interest$	$Prob.$
c1	Lily	Dinning	0.9
c2	Tom	Photo	0.6
c3	Jessica	Entertainments	0.8

(b) Probabilistic $CustomerLoc$ Relation (Attribute-Level Uncertainty)

ID	$CName$	$Location$
l1	Lily	$N(2, 1)$
l2	Tom	$N(5, 1)$
l3	Jessica	$N(3, 1)$

$N(\mu, \sigma^2)$*: normal distribution with probabilistic density function:* $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$*, where* μ *is mean and* σ^2 *is variance.*

The above join query involves operations over multiple probabilistic relations, and the result is shown in Table 7. The approach of evaluating join queries on each possible world, called the *intentional semantics*, is precise yet not efficient, and the complexity increases exponentially with the number of tuples. Dalvi and Suciu [81] present an efficient evaluation method without listing all possible worlds based on the *extensional semantics*, where SQL queries are represented in an algebra and operators are modified to compute the probabilities of output tuples. To avoid dependencies in the querying process that may lead to incorrect results, they give a safe-plan algorithm that can compute most queries efficiently and correctly, and prove that the complexity of evaluating a query with a safe query plan is in PTIME. However, there exist some queries with a #P-complete data complexity, implying that these queries do not admit any efficient evaluation method. 2 out of the 10 TPC/H queries fall in this category, and only when all their predicates are uncertain. For these queries, a few techniques have been developed [81], [86], [87], including aggressively using previously computed query results (materialized views) to rewrite a query in terms of views; using heuristics to choose a plan that avoids large errors; using a Monte Carlo simulation algorithm (to be discussed later in Section 4.1.2), which is more expensive but can guarantee arbitrarily small errors.

(**Range Query**) *"Find customers whose locations are between 1km and 3km."*

SELECT CName **FROM** CustomerLoc
WHERE Location $\geq$ 1km **AND** Location $\leq$ 3km

In sensing and moving objects' applications, data recordings usually change continuously, making the query of an interval more meaningful than a single value. In Table 6b, each customer's location is considered to be a continuous random variable, satisfying the normal distribution $N(\mu, \sigma^2)$, where μ is the mean, and σ^2 is the variance [82]. For $c1(Lily)$, her location within [1 km, 3 km] has a probability $\int_1^3 \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-2)^2}{2}} dx = 0.68$. Combined with the overall tuple probability 0.9, one can get $Lily : 0.68 * 0.9 = 0.61$.

TABLE 7
Join Query Result (Relation *RestInt*)

ID	$RName$	$Prob.$
t1	Starbucks	0.7*0.9=0.63
t2	PizzaHut	0.8*0.9=0.72
t3	KFC	0.55*0.9=0.495

Similarly, one can compute other customers' probabilities, with the final result as $\{(Lily, 0.61), (Tom, 0.01), (Jessica, 0.38)\}$. As the value distribution of an attribute is regarded as a probabilistic density function, this approach incurs a high complexity due to the expensive cost of integration computation. Frequently, one must resort to approximation.

(**Top-k Query**) *"Find top 2 restaurants with the highest discounts."*

```
SELECT * FROM Restaurant
ORDER BY Discount DESC LIMIT 2
```

Several factors influence result ranking, including *Discount* and *Prob.* of each tuple, probability of each possible world, and the ranked tuple position in the possible world. They make top-k ranking an intriguing issue, leading to multiple query semantics [88]:

1. Ré et al. [89] rank query results based on tuples' probabilities and return $<r_2, r_1>$ (Table 5), because their probabilities are among the top 2.
2. *Uncertain top-K* (U-TopK) query [90] returns a tuple vector with the maximum aggregated probability of being top-k across all possible worlds. It returns $<r_3, r_1>$, contributed by world W_1 and W_2 as the result.
3. *Uncertain rank-k* (U-kRanks) query [90] returns a list of k tuples, where the ith tuple appears at rank i with the highest probability in all possible worlds. The top-2 query results are thus $<r_3, r_1>$, since r_3 has the highest probability to be ranked first in all possible worlds (r_3 is ranked first in possible world W_1, W_2, W_3, W_4 whose corresponding probabilities sum up to be 0.55), and r_1 has the highest probability to be ranked second in all possible worlds.
4. *Probabilistic threshold top-k* query [91] returns all tuples whose probabilities in the top-k list are larger than a prespecified threshold. When threshold $=$ 0.25, the returned results are $<r_1, r_3>$ (r_2 is ignored with a probability of 0.132).
5. *Expected rank* query [92] returns a list of k tuples that have the highest expected ranks, computed by summarizing the product of ranked tuple position and probability in each possible world. The top-2 results are $<r_3, r_1>$.
6. *Expected score* query [92] returns a list of k tuples that have the highest expected *discounts*, computed by multiplying tuple's *Discount* and *Prob.* The top-2 results are $<r_1, r_2>$.
7. *Parameterized ranking function*-based query [93]. A parameterized ranking function is first defined $\Upsilon_\omega(t) = \sum_{i>0} \omega(t, i) Pr(r(t) = i)$, where $r(t)$ is a random variable denoting the rank of t in all possible worlds, and $\omega(t, i)$ is a weight function: $T \times N \to C$ (T is the set of all tuples, N is the set of all ranking positions, and C is the set of complex numbers). The top-k query returns k tuples whose $|\Upsilon_\omega(t)|$ values are among the top k.

In Table 5, when the weight function is set to $\omega(t, i) = 4 - i$, meaning the weight function is independent of t, the top-2 results are $<r_1, r_3>$. By setting appropriate weights, the parameterized ranking function can approximate many of the previously proposed ranking semantics, except for the U-TopK query.

k-nearest neighbor queries over uncertain data can be classified to two types: One is probabilistic nearest neighbor, which ranks uncertain objects based on their probabilities of being the nearest neighbor of a query point [94], [95], [96]; the other approach is based on a distance metric, where similar query semantics like U-TopK query, uncertain rank-k query, expected score query, and expected rank query can be applied [97], [98], [99], [100].

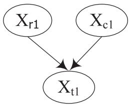
In evaluating top-k queries based on ranking functions, two efficient techniques are mainly involved:

1. *Generating function technique.* Li et al. [93] presented a polynomial algorithm based on generating functions avoiding listing all possible worlds in exponential time complexity. Let a list of tuples $T_i = \{t_1, \ldots, t_i\}$ in a nonincreasing order by their score, a generating function can be constructed in the form of $\mathcal{F}^i(x) = (\prod_{t \in T_{i-1}} (1 - \Pr(t) + \Pr(t) \cdot x))(\Pr(t_i) \cdot x) = \sum_{j \geq 0} c_j x^j$ to compute the probability that the tuple t_i is at rank j (i.e., $Pr(r(t_i) = j)$). In this formula, the coefficient c_j of x^j in $\mathcal{F}^i$ is exactly the value of $Pr(r(t_i) = j)$. Besides, $\mathcal{F}^i$ can be obtained from $\mathcal{F}^{i-1}$, which further simplifies the calculation. This approach is also applied in [101], [102].
2. *Poisson binomial recurrence technique.* The probability of a tuple t_i to be ranked in the top-k list can be computed as $Pr(t_i, j) = Pr(t_i) Pr(S_{t_i}, j - 1)$, where $Pr(t_i)$ is the probability that t_i appears, S_{t_i} is the set of all the tuples that satisfy the query and are ranked higher than t_i, and $Pr(S_{t_i}, j - 1)$ is the probability that $j - 1$ tuples in S_{t_i} appear in possible worlds. Based on the Poisson binomial recurrence $Pr(S_{t_i}, j) = Pr(S_{t_{i-1}}, j - 1) Pr(t_i) + Pr(S_{t_{i-1}}, j)(1 - Pr(t_i))$ [103], the value of $Pr(S_{t_i}, j)$ can be computed recursively. Poisson binomial recurrence and variants have been applied in [91] to efficiently answer probabilistic threshold top-k queries, [104] to do pruning in inverse ranking query (for a user-specified query point q, it computes all the possible ranks of q with probability greater than a predefined threshold), [100] to compute probabilistic similarity ranking on uncertain vector data, and other works [105], [106], [107], [108].

(**Lineage Query**) *"Find customers contributing to 'Starbucks' in RestInt."*

```
SELECT Customer.CName FROM RestInt, Customer
WHERE lineage(RestInt, Customer)
AND RestInt.RName='Starbucks'
```

The Trio system [109] accommodates tuples' lineage/provence information. For the relation *RestInt* (Table 7), $lineage(t1) = \{r1, c1\}$, $lineage(t2) = \{r2, c1\}$, $lineage(t3) = \{r3, c1\}$. $lineage(RestInt, Customer)$ in the *WHERE* clause holds true if the lineage of *RestInt* tuples includes a *Customer* tuple. As $t1.RName =$ "*Starbucks*" and $lineage(t1) = \{r1, c1\}$, the above query result is *"Lily."*



(i)f_{r1}

X_{r1}	$f_{X_{r1}}$
0	0.3
1	0.7

(ii)f_{c1}

X_{c1}	$f_{X_{c1}}$
0	0.1
1	0.9

(iii)$f^{and}_{t1,r1,c1}$

X_{t1}	X_{r1}	X_{c1}	$f^{and}_{t1,r1,c1}$
0	0	0	1
0	0	1	1
0	1	0	1
0	1	1	0
1	0	0	0
1	0	1	0
1	1	0	0
1	1	1	1

Fig. 11. A graphical model with factors representation.

(**Correlation-based Query**) *"Find restaurants of interest to Lily."*

```
SELECT RName FROM Restaurant, Customer
WHERE Restaurant.Type=Customer.Interest
AND CName='Lily'
```

Computation of result probability of $t1$ (Table 7) after joining $r1$ with $c1$ (Table 5) involves a logical-*and* relationship. Sen and Deshpande [110] present the correlation among tuples in probabilistic graphical model, as shown in Fig. 11.

In the model, each tuple is associated with a Boolean random variable, where 1 represents tuple existence, and 0 otherwise (i.e., X_{r1}, X_{c1}, and X_{t1}). The correlation in the graphical model can then be represented with random variables and factor functions in Fig. 11, and a 3-argument factor $f^{and}_{t1,r1,c1}$ is defined to express this *and* correlation. Then, query evaluation can be casted as inference over the probabilistic graphical model by computing the marginal probability of result tuples with all the tuples (regarded as random variables) and factors (representing tuple correlations) involved:

$$P(X_{t1}=1) = \sum_{X_{r1},X_{c1}} f_{r1}(X_{r1}) f_{c1}(X_{c1}) f^{and}_{t1,r1,c1}(X_{t1},X_{r1},X_{c1})$$
$$= 0.7 * 0.9 = 0.63.$$

Based on the model, various forms of correlation/dependency such as implication, mutually exclusivity, xor relationship can be represented. To reduce the storage and computing complexity, the *random variable elimination graph* (rv-elim graph) is built where nodes are marked with labels allowing the recognition of shared correlation factors [111]. The compression of the rv-elim graph enables to speed up query processing significantly [112]. A PrDB model for managing and exploiting rich correlation for query evaluation is further presented [112], [113]. Antova et al. [114], [115] handle complex correlations by vertical partitioning, where a databases relation is partitioned into several U-relations, so that only the involved partition is used in the query process. Correlations in probabilistic streams are also examined in [116], [117].

4.1.2 Monte-Carlo-Based Data Management

Data imprecision usually results in a lot of possible query answers. To address queries of a #P-complete data complexity, Dalvi and Suciu [81] propose the use of the pseudorandom Monte Carlo method [118] to approximately compute the result probability. The simulation algorithm can run in polynomial time and approximate the probabilities to an arbitrary precision. Basically, given a DNF formula with N clauses and any ϵ and δ, the algorithm runs in time $O(\frac{N}{\epsilon^2}\ln\frac{1}{\delta})$, guaranteeing that the probability of the error being greater that ϵ is less than δ.

Monte Carlo simulations are also exploited to evaluate top-k queries. Considering the importance of correct ranking rather than exact probabilities, Ré et al. [86] develop a Monte Carlo approximation algorithm, which calculates the top-k answers for many steps. After N steps, an approximation interval $[a^N, b^N]$ is returned for a result tuple's probability p, whose width shrinks as N increases. The ranking among all the result tuples is conducted to find the top-k probabilities according to the approximation intervals.

Jampani et al. [119], [120] present a Monte Carlo-based uncertain data management approach called *MCDB*. It does not encode uncertainty within the data model itself, and all its query processing is over the classical relational data model. MCDB allows a user to define arbitrary *variable generation* (VG) functions that embody the database uncertainty. It then uses these functions to pseudorandomly generate realized values for the uncertain attributes and runs queries over the realized values. Moreover, these VG functions can be parameterized on the results of SQL queries over *parameter table* that are stored in the database. By storing parameters rather than probabilities, it is easy to change the exact form of the uncertainty dynamically, according to the global state of the database [119], [120]. In MCDB, each query is evaluated only once, regardless of value N (number of Monte Carlo iterations) supplied by the user. Each "database tuple" that is processed by MCDB is actually an array or a "bundle" of tuples, where $t[i]$ for tuple bundle t denotes the value of t in the ith Monte Carlo database instance. The performance benefit of such a "tuple bundles" approach is that relational operations can efficiently operate in batch across all N Monte Carlo iterations that are encoded in a single tuple bundle. Most of the classic relational operations can be modified slightly to handle the fact that tuple bundles move through the query plan. Some additional operators are also defined to facilitate uncertain database querying, such as seed operators, instantiate operator, split operator, and inference operator.

Arumugam et al. [121] further extend MCDB by randomly generating database samples and exploring the tails of the query-result distribution. By adapting the Gibbs sampling (a special case of an MCMC) and cloning techniques developed in the simulation field, they present a statistical method for both estimating a user-specified quantile on a query-result distribution and deriving a set of samples from the tail. The approach finds a good place in risk analysis [121].

Example 9. MCDB approximately simulates the quantile of financial loss in enterprise risk analysis. In the relation *Loss(CustomerID, Val),Val* is an uncertain attribute whose values are generated through a predefined VG function in the query process. The quantile is defined as *"the value c such that there is a probability p of seeing a total-loss of c or more,"* $Prob.(SUM(Val) \geq c) = p$.

Assume that there are r customers. The computation of quantile c is as follows:

1. Generate four DB instances $S = \{D^{(1)}, D^{(2)}, D^{(3)}, D^{(4)}\}$, each including r tuples.

TABLE 8
An Evidence Database Example

(a) Evidence-based $Customer$ Relation

CID	$CName$	$Interest$	$Confidence$
c1	Lily	$unknown$	[1,1]
c2	Tom	$\{i1, i2, i3\}$	[1,1]
c3	John	$< \{i1\}, 0.7 >, < \{i1, i2\}, 0.3 >$	[1,1]
c4	Jessica	$< \{i1, i3\}, 0.7 >, < \{i2\}, 0.3 >$	[1,1]

$i1$:Dinning, $i2$:Entertainment, $i3$:Photo

(b) Result of Evidence-based Query

CID	$CName$	$Interest$	$Confidence$
c1	Lily	$unknown$	[0,1]
c2	Tom	$\{i1, i2, i3\}$	[0,1]
c3	John	$< i1, 0.7 >, < \{i1, i2\}, 0.3 >$	[0.7,1]
c4	Jessica	$< \{i1, i3\}, 0.7 >, < i2, 0.3 >$	[0.7,0.7]

2. Implement *SUM* query Q on each DB instance, and discard the DB instances whose *SUM* values are in the lowest $100(1 - p^{1/m})$ percent percentile. The remaining DB instances are cloned to ensure there are still four DB instances in S.
3. Implement Gibbs update in the newly cloned DB instances to eliminate duplicate instances for a new version of S, where tuples in a DB instance are updated one by one in a conditional way.
4. Repeat m times from Step 2.

Finally, quantile c is obtained through the *SUM* query on the final version of S.

4.1.3 Evidence-Based Data Management

Based on evidence theory, evidence-oriented database was proposed in the 1990s [122] to support data ignorance. In an evidence-oriented database, the value of an uncertain attribute in a relation is represented as a probability distribution on the power set of its domain. Each tuple has an additional *confidence* attribute in the form of *[belief, plausibility]*, stating the confidence level of its belonging to the relation. For an uncertain attribute A, its belief mass function is defined as $m : 2^{dom(A)} \rightarrow [0, 1]$.

Table 8a shows an evidence-based *Customer* relation. John's $Interest$ attribute is expressed as $<\{i1\}, 0.7>$, $<\{i1, i2\}, 0.3>$, meaning that John is interested in $i1$ with probability 0.7, or interested in $i2$ or $i3$ with probability 0.3. That is, $m(\{i1\}) = 0.7$ and $m(\{i1, i2\}) = 0.3$. Due to $\sum_{S \subseteq 2^{dom(A)}} m(S) = 1$, $m(\emptyset) = 0$, $m(\{i2\}) = 0$, $m(\{i3\}) = 0$, $m(\{i2, i3\}) = 0$, $m(\{i1, i3\}) = 0$, $m(\{i1, i2, i3\}) = 0$.

(**Evidence-based Query**) *"Find customers who may be interested in i1 or i3".*

```
SELECT CName,Interest FROM Customer
WHERE (Interest=i1) OR (Interest=i3)
```

The resulting table is shown in Table 8b. Taking $John$ tuple, for example,

$$belief(\{i1, i3\}) = \sum_{B \subseteq \{i1, i3\}} m(B) = m(\{i1\}) = 0.7,$$

$$plausibility(\{i1, i3\}) = \sum_{B \cap \{i1, i3\} \neq \emptyset} m(B) = m(\{i1\}) + m(\{i1, i2\}) = 1.$$

Then, the confidence level of *John* tuple in the resulting table is [0.7, 1].

TABLE 9
Tuple-Attributewise Fuzzy *Restaurant* Relation

RID	$RName$	$Discount$	$Reputation$	μ
r1	Starbucks	{0.5/7, 1/8, 0.6/9}	High	0.9
r2	PizzaHut	{0.6/9, 1/9.5}	VeryHigh	0.8
r3	KFC	{1.0/8}	Medium	0.5

Evidence-based compound query has more query conditions connected by logical connectives (conjunction, disjunction, or negation). The conjunction of two independent events A and B can be computed as

$$belief(A \wedge B) = belief(A) * belief(B),$$
$$plausibility(A \wedge B) = plausibility(A) * plausibility(B).$$

Similarly, due to $A \vee B = \neg(\neg A \wedge \neg B)$, the disjunction of two events A and B can also be computed according to the following equations in independent situations:

$$belief(A \vee B) = 1 - (1 - belief(A)) * (1 - belief(B)),$$
$$plausibility(A \vee B) = 1 - (1 - plausibility(A)) * (1 - plausibility(B)).$$

4.2 Fuzzy Data Management

Fuzzy data management deals with imprecisely defined vaguely bounded linguistic terms and statements like the discount is *high* (or *around* 4) rather than the discount is 4. Fuzzy approximate queries like *"the discount is around 4"* are explained according to the definition of fuzzy sets with vague boundaries.

Developed fuzzy data models can represent tuple-level, attribute-level, or both-leveled uncertainty. To show different representation and query mechanisms between probabilistic and fuzzy databases, we use the same example about customers' interests in restaurants with discounts. The tuple-level fuzzy data model considers a relation as a fuzzy set that includes each tuple as an element with a membership degree [123]. As shown in Table 9, attribute μ gives the fuzzy measure of the association among *RName*, *Discount*, and *Reputation*. The attribute-level possibility data model represents an uncertain attribute value that is represented with a possibility distribution [124], [125], [126] (e.g., *Discount* in Table 9), or a linguistic term (e.g., *Reputation* in Table 9). The possibility-distribution-fuzzy data model combines the above two methods [127], [128].

Based on fuzzy theory, the fuzzy membership degree of a conjunctive operation on fuzzy sets A and B is $\mu_{A \cup B}(x) = \max(\mu_A(x), \mu_B(x))$, and the fuzzy membership degree of a disjunctive operation on A and B is $\mu_{A \cap B}(x) = \min(\mu_A(x), \mu_B(x))$ [129].

We illustrate three typical fuzzy queries over Table 9.

(Fuzzy Query over Possibility Distributions) *"Find restaurants which offer a High discount.".*

```
SELECT RName FROM Restaurant
WHERE Discount=High
```

Assume term *HighDiscount* is represented as a possibility distribution $\mu_{HighDiscount}(x) = \{0.2/8, 0.7/9, 1/9.5\}$, and *Starbucks* tuple has the discount possibility distribution $\mu_{Starbucks.Discount}(x) = \{0.5/7, 1/8, 0.6/9\}$. We use the notations A and B to denote the fuzzy sets *HighDiscount* and

TABLE 10
Similarity Representation for *Reputation*

Reputation	**VeryLow**	**Low**	**Medium**	**High**	**VeryHigh**
VeryLow	1.0	0	0	0	0
Low	0	1.0	0.6	0.6	0.6
Medium	0	0.6	1.0	0.8	0.8
High	0	0.6	0.8	1.0	0.9
VeryHigh	0	0.6	0.8	0.9	1.0

Starbucks.Discount separately. The degree that A matches B can be computed as follows:

Ma [130] defined the semantic inclusion degree (SID) for two fuzzy sets μ_A and μ_B as follows:

Definition 1. *Let μ_A and μ_B be fuzzy sets. In the universe X, the degree that μ_A semantically includes μ_B is*

$$SID(\mu_A(x), \mu_B(x)) = \sum_{x_i \in X} \min(\mu_A(x_i), \mu_B(x_i)) \Big/ \sum_{x_i \in X} \mu_B(x_i). \tag{23}$$

For the above query, in the universe of discount $X = \{7, 8, 9, 9.5\}$,

$$\begin{aligned} SID(\mu_A(x), \mu_B(x)) &= (0.2 + 0.6)/(0.5 + 1 + 0.6) = 0.38, \\ SID(\mu_B(x), \mu_A(x)) &= (0.2 + 0.6)/(0.2 + 0.7 + 1) = 0.42. \end{aligned} \tag{24}$$

Then, the similarity between the possibility distribution $\mu_B(x)$ and possibility distribution $\mu_A(x)$ is $\min(SID(\mu_A(x), \mu_B(x)), SID(\mu_B(x), \mu_A(x))) = 0.38$.

Considering the overall *Starbucks* tuple uncertainty 0.9, the result *Starbucks* tuple satisfies high discount with the degree of $min(0.38, 0.9) = 0.38$. In the same way, we can obtain the final result as $(Starbucks, 0.38)$, $(PizzaHut, 0.8)$, and $(KFC, 0.13)$. Thus, we find that $PizzaHut$ is the most likely to offer *High* discount.

(Fuzzy Query over Fuzzy Terms) *"Find restaurants that have a high reputation."*.

```
SELECT RName FROM Restaurant
WHERE Reputation=High
```

A similarity measurement between two fuzzy linguistic terms can be expressed in Table 10.

As *Starbucks* has reputation *High*, its satisfying degree of *HighReputation* is 1.0 according to the above table. Thus, its membership degree in the final answer is $min(1.0,\ 0.9) = 0.9$. The query result is $\{(Starbucks, 0.9)$, $(PizzaHut, 0.8)$, $(KFC, 0.5)\}$.

(Fuzzy Query with Vague Condition) *"Find restaurants whose discounts are around 8"*.

```
SELECT RName FROM Restaurant
WHERE Discount≈8
```

The fuzzy condition *about* 8 can be defined a membership function $\mu_{around\,8}(Discount) = \{0.2/6, 0.7/7, 1.0/8,\ \ 0.7/9, 0.2/9.5\}$, which is similar to possibility distribution in the previous fuzzy query example. Result computation can follow exactly the same way through the SID.

We summarize tuplewise and attributewise uncertainty representation in uncertain data management in Table 11. In tuplewise uncertainty, two ways are used to express tuple existence uncertainty: a single value (probability in probabilistic databases, or membership degree in fuzzy databases) or a range of confidence degree (evidence databases). In attributewise uncertainty, various representations are used: a set of discrete attribute values and corresponding probabilities (e.g., (a_m, p_{a_m})), a number of value sets with probabilities (e.g., $(A_i, m(A_i))$, a probabilistic density function $f(A)$ denoting the distribution, several characteristic values for a distribution function (Monte Carlo-based databases), and a possibility distribution (fuzzy databases).

5 Challenges to Uncertain Data Management

Great achievements have been made on uncertain data management. On the other hand, when we look at different origins of uncertainty as well as different handling practices in diverse fields, we may find some interesting issues for further data-oriented research, particularly from user and domain perspectives.

5.1 Leveled Uncertainty Representation

Suciu et al. [81], [133], [87], [134] have made great efforts to reduce the complexity of uncertain database query processing. Inspired by uncertainty handling activities in diverse fields, domain-specific knowledge could be a help to further reduce the complexity based on some specific uncertain data representation.

For example, in a group-purchase application scenario, customers may be interested in such restaurants that offer discounts *less than 4, between 3 and 5,* and so on. On the other hand, the possible discounts available from restaurants themselves may fall into a scope rather than as a specific value. A tabular representation of uncertain *discount* values

TABLE 11
Uncertainty Representation in Data Management

Tuple-level (Tuple X)	**Attribute-level (Attribute A)**
Prob. DB: 1) (X, p_X) where $p_X \in [0, 1]$ is the probability [80], [81]. 2) $(X, [belief_X, plausibility_X])$ where $[belief_X, plausibility_X] \subseteq [0, 1]$ [122].	***Prob. DB:*** 1) $\{(a_1, p_{a_1}), \cdots, (a_m, p_{a_m})\}$, where $p_{a_i} \in [0, 1], a_i \in dom(A)$ $(1 \leq i \leq m)$ [84]. 2) $< A_1, m(A_1) >, \cdots, < A_m, m(A_m) >$, where $m(A_i) \in [0, 1]$ is the mass (degree of belief), $A_i \subseteq dom(A)$ $(1 \leq i \leq m)$ [122] 3) Probabilistic density function $f(A)$ [131], [82], [132]. 4) Characteristic values for pseudo-random generation [119], [120].
Fuzzy DB: (X, μ_X) where $\mu_X \in [0, 1]$ is the possibility [123].	***Fuzzy DB:*** $\{a_1/\mu_A(a_1), \cdots, a_m/\mu_A(a_m)\}$ where $\mu_A(a_i) \in [0, 1]$ $(1 \leq i \leq m)$ [124], [125], [126].

TABLE 12
Tabular Form of *Restaurant*'s Uncertain *Discount*

$Uncertainty$	0. 1	0.15	0.4	$\cdots$	1.0
$Discount$	≤ 1	≤ 2	≤ 5	$\cdots$	≤ 9

for a *restaurant* is illustrated in Table 12, whose pairwise uncertainty values are increasingly ordered. For instance, the first pair $\langle 0.1, \leq 1\rangle$ represents that the probability of $(Discount \leq 1)$ is 0.1. Apparently, the probability of $(Discount \leq 2)$, including $(Discount \leq 1)$, is greater than 0.1, as shown in Table 12. Such a property could be exploited to query optimization and is particularly good at range queries. It is a compromise between discrete and continuous probability distributions and offers a complementary way when a continuous probability distribution is unavailable or impractical due to the high integration computation complexity.

The size (or granularity) of the uncertainty table reflects the precision level of uncertainty handling behavior itself, forming a hierarchy of uncertainty tables $(\mathcal{T}_u, \prec_u)$, where $\mathcal{T}_u = (T_1, T_2, \ldots, T_s)$ of s levels, and $\prec_u$ is a partial order among the levels of $\mathcal{T}_u$, such that $(T_1 \prec_u T_i \prec_u T_s)$ where $(1 < i < s)$. Given a query or inference task, dynamically selecting the right uncertainty table from the hierarchy based on certain measurements is needed for different applications.

5.2 Domain-Driven Uncertainty Management

Different applications have different requirements on uncertainty management. Bringing application logics to uncertainty management is important. Taking sensing and monitoring domains, for example, real-time sensing and response with low latency are very much desirable. However, due to the inherent uncertainty in the real world, the execution time and waiting time of a query may vary. It is thus hard to precisely predicate and ensure real-time query response performance.

For instance, for a query Q issued at time T_s and expected to finish at T_d, let T_e denote its uncertain waiting and execution time, represented in the form of a value scope around the estimated value:

$$U(\alpha, \tilde{T_e}) = \{T_e : |T_e - \tilde{T_e}| \leq \alpha\tilde{T_e}\}, \tag{25}$$

where $\widetilde{T_e}$ is the estimated time for waiting and execution and $\alpha \in [0,1]$ is the derivation. A robust function can be defined to compute the largest allowable scope of query waiting and execution time to meet the query deadline:

$$Robust = \max\Big\{\alpha : \max_{T_e \in U(\alpha, \tilde{T_e})} (T_s + T_e) \leq T_d\Big\}. \tag{26}$$

Based on the *robust* value, the variation range of query waiting and execution time, $[\tilde{T_e}(1-\alpha), \tilde{T_e}(1+\alpha)]$, can be derived, which could then be used to guide time-sensitive query scheduling and resource management of the system.

5.3 Leveraging User Knowledge

Human users are good at and highly successful in coping with uncertainty throughout their daily lives, as most human knowledge in the real world is uncertain. While working with probabilistic database query and inference mechanisms to infer sensible and actionable information from underlying uncertain data, we could involve users in the loop of query evaluation for feedback.

For example, based on the observation that a user is usually more likely to recognize mistakes in basic uncertain tuples leading to the final ranked answer, than mistakes in the answer itself, the query engine could consider to display those influential underlying probabilistic tuples to the ranked query result, and then leverage user's personal knowledge to clarify the uncertainty degrees and precisions of the basic tuples. After that, the query engine can recompute the query and tailor its uncertain query result toward a better quality from the perspective of the specific user. More important, by opening the black box of the query engine and showing to the user how it comes up with the answer and which uncertain tuples it is based on, the user with his/her knowledge can decide how much confidence to be placed on the system, thus enhancing both the intelligibility of system behavior and accountability of human users.

Here, a few critical questions need to be answered, like how can we interact with the user for result explanation and uncertainty clarification without bringing much burden on the user? How can we correct the query/inference result after users uncertainty clarification without incurring much computing overhead on the query engine? and how can we reconcile different users' uncertainty clarification upon the uncertain database? Solutions to the above questions determine the effectiveness of the approach.

5.4 Crowdsourcing for Uncertain Data Management

As queries may require information from human knowledge that is missing in the databases, such as the recognition of misspelling words, efforts to deeply involve crowd on the internet in query processing have been made [135], [136], [137], [138], [139], [140]. For instance, *CrowdDB* [135] leverages human capability by crowdsourcing missing values of tuples, crowdsourcing new tuples from the inner relation that matches the tuple of the outer relation in join operations, and crowdsourcing comparison work. *Qurk* system [138] addresses the workflow management of crowd-powered querying tasks by balancing monetary cost, spending time, and result accuracy.

Currently, crowdsourcing in data management is still in the initial stage, leaving some challenges to be solved, such as quality assessment and improvement, latency, scheduling, cost optimization, privacy, and social issues [136]. The high ambiguity and multiple sources from human inputs need to be tackled for a smart interface to machine processing.

6 CONCLUSION

Uncertainty is unavoidable. It penetrates our lives. A good knowledge of uncertainty and uncertainty processing techniques is helpful to know the physical world and make better decisions. Until now, decades of efforts have been made to tackle uncertainty. In this survey, we review the research of uncertainty in diverse fields. To represent and model uncertainty, some mathematical tools are needed.

We overview four uncertainty handling theories and give some comparisons. Their applications to the fields of economics, engineering, ecology, and information science are described. We particularly describe uncertainty management achievements made by the database community and list some potential problems for future work from the data modeling and querying perspectives. We hope that uncertain data management methods and uncertainty handling practices could inspire each other. Therefore, practical problems can be handled more efficiently with the development of uncertainty management technologies.

Acknowledgments

The authors thank the referees, associate editor, and editor in chief for insightful comments that greatly help improve the paper. The work was supported by Chinese Major State Basic Research Development 973 Program (2011CB302203-2), Important National Science & Technology Specific Program (2011ZX01042-001-002-2), National Natural Science Foundation of China (60773156 and 61073004), and the Tsinghua-Tencent Joint Laboratory for Internet Innovation Technology.

References

[1] M. Stokes, *Apology of Socrates.* Aris & Phillips, 1997.

[2] H. Zimmermann, "Uncertainty Modelling and Fuzzy Sets," *Proc. Int'l Workshop Uncertainty: Models and Measures,* pp. 84-100, 1997.

[3] Nat'l Research Council, *Risk Analysis and Uncertainty in Flood Damage Reduction Studies.* Nat'l Academy Press, 2000.

[4] F. Hoffman and J. Hammonds, "Propagation of Uncertainty in Risk Assessment: The Need to Distinguish between Uncertainty Due to Lack of Knowledge and Uncertainty Due to Variability," *J. Risk Analysis,* vol. 14, no. 5, pp. 707-712, 1994.

[5] M.B.A. van Asselt and J. Rotmans, "Uncertainty in Integrated Assessment Modelling—From Positivism to Pluralism," *J. Climatic Change,* vol. 54, pp. 75-105, 2002.

[6] Environment Agency, "Climate Adaptation Risk and Uncertainty: Draft Decision Framework," Environment Agency Report, no. 21, June 2000.

[7] MAFF, "Flood and Coastal Defence Project Appraisal Guidance Notes: Approaches to Risk," FCDPAG4, Feb. 2000.

[8] V. Gelder, "Statistical Methods for the Risk Based Design of Civil Structures," PhD dissertation, Delft Univ., 1999.

[9] M. Paté-Cornell, "Uncertainties in Risk Analysis: Six Levels of Treatment," *J. Reliability Eng. and System Safety,* vol. 54, pp. 95-111, 1996.

[10] H. Natke and Y. Ben-Haim, "Uncertainty: A Discussion from Various Points of View," *Proc. Int'l Workshop Uncertainty: Models and Measures,* 1996.

[11] D. Kahneman and A. Tversky, "Variants of Uncertainty," *Variants of Uncertainty,* D. Kahneman, P. Slvic, and A. Tversky, eds., Cambridge Univ. Press, 1982.

[12] M. Henrion and B. Fischhoff, "Assessing Uncertainty in Physical Constants," *Ann. J. Physics,* vol. 54, no. 9, pp. 791-797, 1986.

[13] J. Helton, "Treatment of Uncertainty in Performance Assessments for Complex Systems," *J. Risk Analysis,* vol. 14, no. 4, pp. 483-511, 1994.

[14] T. Koopmans, *Three Essays on the State of Economic Science.* Martino Publishing, 1957.

[15] R. von Schomberg, "Controversies and Political Decision Making," *Controversies and Political Decision Making,* R. von Schomberg, ed., Kluwer Academic Publishers, 1993.

[16] G. Klir, "Uncertainty Theories, Measures and Principles," *Uncertainty Theories, Measures and Principles,* H.G. Natke and Y. Ben-Haim, eds., Akademie Verlag, 1996.

[17] G. Bammer and M. Smithson, *Uncertainty and Risk: Multidisciplinary Perspectives.* Earthscan Publications, 2008.

[18] A. Berztiss, "Uncertainty Management," Univ. of Pittsburgh, Dept. of Computer Science, Pittsburgh, PA, USA, 2002.

[19] P. Walley, "Measures of Uncertainty in Expert Systems," *Artificial Intelligence,* vol. 83, pp. 1-58, 1996.

[20] G. Klir, *Uncertainty and Information, Foundations of Generalized Information Theory.* Wiley, 2006.

[21] E. Codd, "A Relational Model of Data for Large Shared Data Banks," *Comm. ACM,* vol. 13, pp. 377-387, 1970.

[22] T. Imieliński and W. Lipski, "Incomplete Information in Relational Databases," *J. ACM,* vol. 31, no. 4, pp. 761-791, 1984.

[23] A. Abiteboul, P. Kanellakis, and G. Grahne, "On the Representation and Querying of Sets of Possible Worlds," *SIGMOD Record,* vol. 16, no. 3, pp. 34-48, 1987.

[24] N. Fuhr, "A Probabilistic Framework for Vague Queries and Imprecise Information in Databases," *Proc. 16th Int'l Conf. Very Large Data Bases (VLDB),* 1990.

[25] N. Dalvi and D. Suciu, "Management of Probabilistic Data Foundations and Challenges," *Proc. ACM SIGMOD-SIGACT-SIGART Symp. Principles of Database Systems,* 2007.

[26] J. Pei, M. Hua, Y. Tao, and X. Lin, "Query Answering Techniques on Uncertain and Probabilistic Data: Tutorial Summary," *Proc. ACM SIGMOD Int'l Conf. Management of Data,* 2008.

[27] A. Sarma, O. Benjelloun, A. Halevy, and J. Widom, "Working Models for Uncertain Data," *Proc. Int'l Conf. Data Eng. (ICDE),* 2006.

[28] A. Sarma, O. Benjelloun, A. Halevy, S. Nabar, and J. Widom, "Representing Uncertain Data: Models, Properties, and Algorithms," *J. VLDB,* vol. 18, no. 5, pp. 989-1019, 2009.

[29] C. Aggarwal and P. Yu, "A Survey of Uncertain Data Algorithms and Applications," *IEEE Trans. Knowledge Data Eng.,* vol. 21, no. 5, pp. 609-623, May 2009.

[30] M. Kamel, B. Hadfield, and M. Ismail, "Fuzzy Query Processing Using Clustering Techniques," *Information Processing and Management,* vol. 26, pp. 279-293, 1990.

[31] A. Yazici, B. Buckles, and F. Petry, "A Survey of Conceptual and Logical Data Models for Uncertainty Management," *Fuzzy Logic for Management of Uncertainty,* pp. 607-644, John Wiley and Sons Inc., 1992.

[32] E. Kerre and G. Chen, "An Overview of Fuzzy Data Modeling," *Fuzziness in Database Management Systems,* pp. 23-41, Physica-Verlag, 1995.

[33] Z. Ma and L. Yan, "A Literature Overview of Fuzzy Database Models," *J. Information Science and Eng.,* vol. 24, pp. 189-202, 2008.

[34] Z. Ma and L. Yan, "A Literature Overview of Fuzzy Conceptual Data Modeling," *J. Information Science and Eng.,* vol. 26, pp. 427-441, 2010.

[35] P.S. de Laplace, "Analytical Theory of Probability," 1812.

[36] A. Kolmogorov, *Foundations of the Theory of Probability.* Chelsea Pub., 1950.

[37] L.A. Zadeh, "Fuzzy Sets," *J. Information and Control,* vol. 8, pp. 338-353, 1965.

[38] L.A. Zadeh, "Fuzzy Sets as a Basis for a Theory of Possibility," *J. Fuzzy Sets and Systems,* vol. 1, pp. 3-28, 1978.

[39] L.A. Zadeh, "Fuzzy Logic," *IEEE Computer Magazine,* vol. 21, no. 4, pp. 83-93, Apr. 1988.

[40] Y. Ben-Haim, *Information-Gap Theory: Decisions under Severe Uncertainty.* Academic Press, 2001.

[41] Y. Ben-Haim, *Info-Gap Theory: Decisions under Severe Uncertainty,* second ed. Academic Press, 2006.

[42] B. Liu, *Uncertainty Theory: An Introduction to Its Axiomatic Foundations,* second ed. Springer-Verlag, 2007.

[43] B. Liu, *Uncertainty Theory: A Branch of Mathematics for Modeling Human Uncertainty,* fourth ed. Springer-Verlag, 2010.

[44] N. Metropolis and S. Ulam, "The Monte Carlo Method," *J. Am. Statistical Assoc.,* vol. 44, no. 247, pp. 335-341, 1949.

[45] H. Chernoff, "A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the Sum of Observations," *Annals Math. Statistics,* vol. 23, no. 4, pp. 493-507, 1952.

[46] W. Gilks, S. Richardson, and D. Spiegelhalter, *Markov Chain Monte Carlo in Practice.* Chapman & Hall, 1996.

[47] T. Bayes and R. Price, "An Essay towards Solving a Problem in the Doctrine of Chance. By the Late Rev. Mr. Bayes, Communicated by Mr. Price, in a Letter to John Canton, M. A. and F. R. S.," *Philosophical Trans. Royal Soc. of London,* vol. 53, pp. 370-418, 1763.

[48] A. Dempster, "Upper and Lower Probabilities Induced by a Multivalued Mapping," *Annals Math. Statistics,* vol. 38, no. 2, pp. 325-339, 1967.

[49] G. Shafer, *A Mathematical Theory of Evidence.* Princeton Univ. Press, 1976.

[50] P. Smets, "Belief Functions and the Transferable Belief Model," The Imprecise Probabilities Project, http://ippserv.rug.ac.be, 2013.

[51] P. Walley, "Coherent Upper and Lower Previsions," The Imprecise Probabilities Project, http://ippserv.rug.ac.be, 1998.

[52] P. Soundappan, E. Nikolaidis, R. Haftka, R. Grandhi, and R. Canfield, "Comparison of Evidence Theory and Bayesian Theory for Uncertainty Modelling," *J. Reliability Eng. System Safety,* vol. 85, nos. 1-3, pp. 295-311, 2004.

[53] F. Knight, *Risk, Uncertainty, and Profit.* Houghton Mifflin Company, 1921.

[54] C. Schinckus, "Economic Uncertainty and Econophysics," *Physica A,* vol. 388, pp. 4415-4423, 2009.

[55] S. Markridakis, R. Hogarth, and A. Gaba, "Forecasting and Uncertainty in the Economics and Business World," *J. Forecasting,* vol. 25, pp. 794-812, 2009.

[56] D. Hubbard, *How to Measure Anything: Finding the Value of Intangibles in Business.* John Wiley & Sons, 2007.

[57] Y. Ben-Haim, *Info-Gap Economics: An Operational Introduction.* Palgrave Macmillan, 2010.

[58] B. Beresford-Smith and C. Thompson, "Managing Credit Risk with Info-Gap Uncertainty," *J. Risk Finance,* vol. 8, no. 1, pp. 24-34, 2007.

[59] Reactor Safety Study, "An Assessment of Accident Risks in Us Commercial Nuclear Power Plants," Technical Report WASH-1400 (NUREG-75/OI4), United States Nuclear Regulatory Commission (USNRC), Washington, DC, 1975.

[60] V. Bier and L. Cox, "Probabilistic Risk Analysis for Engineered Systems," *Advances in Decision Analysis,* pp. 279-301, Cambridge Univ. Press, 2007.

[61] S. Greenland, "Sensitivity Analysis, Monte Carlo Risk Analysis, and Bayesian Uncertainty Assessment," *J. Risk Analysis,* vol. 21, pp. 579-583, 2001.

[62] M. Huijbregts, W. Gilijamse, A. Ragas, and L. Reijnders, "Evaluating Uncertainty in Environmental Life-Cycle Assessment," *J. Environmental Science and Technology,* vol. 37, pp. 2600-2608, 2003.

[63] S. Lloyd and R. Ries, "Characterizing, Propagating, and Analyzing Uncertainty in Life-Cycle Assessment: A Survey of Quantitative Approaches," *J. Industrial Ecology,* vol. 11, no. 1, pp. 161-179, 2007.

[64] Y. Goh, L. Newnes, A. Mileham, C. McMahon, and M. Saravi, "Uncertainty in Through-Life Costing-Review and Perspectives," *IEEE Trans. Eng. Management,* vol. 57, no. 4, pp. 689-701, Nov. 2010.

[65] J. Baker and M. Lepech, "Treatment of Uncertainties in Life Cycle Assessment," *Proc. Int'l Congress Structural Safety and Reliability,* 2009.

[66] R. Tan, A. Culaba, and M. Purvis, "Application of Possibility Theory in the Life-Cycle Inventory Assessment of Biofuels," *J. Energy Research,* vol. 26, pp. 737-745, 2002.

[67] R. Tan, "Using Fuzzy Numbers to Propagate Uncertainty in Matrix-Based LCI," *J. Life Cycle Assessment,* vol. 13, pp. 585-592, 2008.

[68] M. Hung and H. wen Ma, "Quantifying System Uncertainty of Life Cycle Assessment Based on Monte Carlo Simulation," *J. Life Cycle Assessment,* vol. 14, no. 1, pp. 19-27, 2009.

[69] W. Lutz, J. Vaupel, and D. Ahlburg, *Frontiers of Population Forecasting.* Population Council, 1999.

[70] J. Bongaarts and R. Bulatao, *Beyond Six Billion: Forecasting the World's Population.* Nat'l Academy Press, 2000.

[71] W. Lutz and J. Goldstein, "Introduction: How to Deal with Uncertainty in Population Forecasting?" *Int'l Statistical Rev.,* vol. 72, pp. 1-4, 2004.

[72] J. Alho, "Stochastic Methods in Population Forecasting," *J. Forecasting,* vol. 6, pp. 521-530, 1990.

[73] H. Booth, "Demographic Forecasting: 1980 to 2005 in Review," *J. Forecasting,* vol. 22, no. 3, pp. 547-581, 2006.

[74] G. Abel, J. Bijak, J. Forster, J. Raymer, and P.W.F. Smith, "What Do Bayesian Methods Offer Population Forecasters?" Working Paper 6/2010, ESRC Research Centre for Population Change, Univ. of Southampton, 2010.

[75] A. Motro, "Management of Uncertainty in Database Systems," *Modern Database Systems: The Object Model, Interoperability, and Beyond,* W. Kim, ed., pp. 457-476, ACM Press, 1994.

[76] E. Adar and C. Ré, "Managing Uncertainty in Social Networks," *IEEE Data Eng. Bull.,* vol. 30, no. 2, pp. 23-31, 2007.

[77] A. Shenhar, O. Levy, and D. Dvir, "Mapping the Dimensions of Project Success," *Project Management J.,* vol. 28, pp. 5-13, 1997.

[78] D. Baccarini, "The Logical Framework Method for Defining Project Success," *Project Management J.,* vol. 30, no. 4, pp. 25-32, 1999.

[79] F. Dweiri and M. Kablan, "Using Fuzzy Decision Making for the Evaluation of the Project Management Internal Efficiency," *Decision Support Systems,* vol. 42, pp. 712-726, 2006.

[80] N. Fuhr and T. Rolleke, "A Probabilistic Relational Algebra for the Integration of Information Retrieval and Database Systems," *ACM Trans. Information Systems,* vol. 15, pp. 32-66, 1997.

[81] N. Dalvi and D. Suciu, "Efficient Query Evaluation on Probabilistic Databases," *Proc. 30th Int'l Conf. Very large Data Bases,* 2004.

[82] R. Cheng, D. Kalashnikov, and S. Prabhakar, "Evaluating Probabilistic Queries over Imprecise Data," *Proc. ACM SIGMOD Int'l Conf. Management of Data,* 2003.

[83] R. Cheng, S. Singh, and S. Prabhakar, "Efficient Join Processing over Uncertain Data," *Proc. ACM Int'l Conf. Information and Knowledge Management,* 2006.

[84] R. Murthy and J. Widom, "Making Aggregation Work in Uncertain and Probabilistic Databases," *Proc. Int'l VLDB Workshop Management of Uncertain Data,* pp. 76-90, 2007.

[85] T. Jayram, S. Kale, and E. Vee, "Efficient Aggregation Algorithms for Probabilstic Data," *Proc. Ann. ACM-SIAM Symp. Discrete Algorithms (SODA),* 2007.

[86] C. Ré, N. Dalvi, and D. Suciu, "Efficient Top-k Query Evaluation on Probabilistic Data," *Proc. Int'l Conf. Data Eng. (ICDE),* 2007.

[87] N. Dalvi, C. Ré, and D. Suciu, "Queries and Materialized Views on Probabilistic Databases," *J. Computer System Science,* vol. 77, no. 3, pp. 473-490, 2011.

[88] I. Ilyas, G. Beskales, and M. Solimam, "A Survey of Top-K Query Processing Techniques in Relational Database Systems," *ACM Computing Surveys,* vol. 40, article 11, 2008.

[89] C. Ré, N. Dalvi, and D. Suciu, "Efficient Top-K Query Evaluation on Probabilistic Data," *Proc. Int'l Conf. Data Eng. (ICDE),* 2007.

[90] M. Soliman, I. Ilyas, and K.C.-C. Chang, "Top-k Query Processing in Uncertain Databases," *Proc. Int'l Conf. Data Eng. (ICDE),* 2007.

[91] M. Hua, J. Pei, W. Zhang, and X. Lin, "Ranking Queries on Uncertain Data: A Probabilistic Threshold Approach," *Proc. ACM SIGMOD Int'l Conf. Management of Data,* 2008.

[92] G. Cormode, F. Li, and K. Yi, "Semantics of Ranking Queries for Probabilistic Data and Expected Ranks," *Proc. Int'l Conf. Data Eng. (ICDE),* 2009.

[93] J. Li, B. Saha, and A. Deshpande, "A Unified Approach to Ranking in Probabilistic Databases," *Proc. VLDB Endowment,* vol. 2, pp. 502-513, 2009.

[94] G. Beskales, M. Soliman, and I. Ilyas, "Efficient Search for the Top-k Probable Nearest Neighbors in Uncertain Databases," *Proc. VLDB Endowment,* vol. 1, pp. 326-339, 2008.

[95] R. Cheng, J. Chen, M. Mokbel, and C.-Y. Chow, "Probabilistic Verifiers: Evaluating Constrained Nearest-Neighbor Queries over Uncertain Data," *Proc. Int'l Conf. Data Eng. (ICDE),* 2008.

[96] P. Agarwal, A. Efrat, Swaminathan, and W. Zhang, "Nearest-Neighbor Searching under Uncertainty," *Proc. 31st Symp. Principles of Database Systems,* 2012.

[97] V. Ljosa and A. Singh, "APLA: Indexing Arbitrary Probability Distributions," *Proc. Int'l Conf. Data Eng. (ICDE),* 2007.

[98] R. Cheng, L. Chen, J. Chen, and X. Xie, "Evaluating Probability Threshold K-Nearest-Neighbor Queries over Uncertain Data," *Proc. Int'l Conf. Extending Database Technology: Advances in Database Technology,* 2009.

[99] Y. Zhang, X. Lin, G. Zhu, W. Zhang, and Q. Lian, "Efficient Rank Based KNN Query Processing over Uncertain Data," *Proc. Int'l Conf. Data Eng. (ICDE),* 2010.

[100] T. Bernecker, H.-P. Kriegel, N. Mamoulis, M. Renz, and A. Zuefle, "Scalable Probabilistic Similarity Rankling in Uncertain Databases," *IEEE Trans. Knowledge and Data Eng.,* vol. 22, no. 9, pp. 1234-1246, Sept. 2010.

[101] J. Li and A. Deshpande, "Ranking Continuous Probabilistic Datasets," *Proc. VLDB Endowment,* vol. 3, pp. 638-649, 2010.

[102] T. Bernecker, T. Emrich, H.-P. Kriegel, N. Mamoulis, M. Renz, and A. Züfle, "A Novel Probabilistic Pruning Approach to Speed up Similarity Queries in Uncertain Databases," *Proc. IEEE 27th Int'l Conf. Data Eng.,* 2011.

[103] K. Lange, "Numerical Analysis for Statisticians," *Statistics and Computing,* Springer Verlag, 1999.

[104] X. Lian and L. Chen, "Probabilistic Inverse Ranking Queries over Uncertain Data," *Proc. 14th Int'l Conf. Database Systems for Advanced Applications,* 2009.

[105] K. Yi, F. Li, G. Kollios, and D. Srivastava, "Efficient Processing of Top-k Queries in Uncertain Databases," *Proc. Int'l Conf. Data Eng. (ICDE),* 2008.

[106] K. Yi, F. Li, and G. Kollios, "Efficient Processing of Top-k Queries in Uncertain Databases with x-Relations," *IEEE Trans. Knowledge and Data Eng.,* vol. 20, no. 12, pp. 1669-1682, Dec. 2008.

[107] M. Hua and J. Pei, "Continuously Monitoring Top-k Uncertain Data Streams: A Probabilistic Threshold Method," *Distributed and Parallel Databases,* vol. 26, no. 1, pp. 29-65, 2009.

[108] T. Bernecker, H.-P. Kriegel, M. Renz, F. Verhein, and A. Züfle, "Probabilistic Frequent Itemset Mining in Uncertain Databases," *Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining,* 2009.

[109] P. Agrawal, O. Benjelloun, A. Sarma, C. Hayworth, S. Nabar, T. Sugihara, and J. Widom, "Trio: A System for Data, Uncertainty, and Lineage," *Proc. 32nd Int'l Conf. Very Large Data Bases,* 2006.

[110] P. Sen and A. Deshpande, "Representing and Querying Correlated Tuples in Probabilistic Databases," *Proc. Int'l Conf. Data Eng. (ICDE),* 2007.

[111] P. Sen, A. Deshpande, and L. Getoor, "Exploiting Shared Correlations in Probabilistic Databases," *Proc. VLDB Endowment,* vol. 1, pp. 809-820, 2008.

[112] P. Sen, A. Deshpande, and L. Getoor, "PrDB: Managing and Exploiting Rich Correlations in Probabilistic Databases," *VLDB J.,* vol. 18, pp. 1065-1090, 2009.

[113] B. Kanagal and A. Deshpande, "Lineage Processing over Correlated Probabilistic Databases," *Proc. ACM SIGMOD Int'l Conf. Management of Data,* 2010.

[114] L. Antova, T. Jansen, C. Koch, and D. Olteanu, "Fast and Simple Relational Processing of Uncertain Data," *Proc. Int'l Conf. Data Eng. (ICDE),* 2008.

[115] J. Huang, L. Antova, C. Koch, and D. Olteanu, "MayBMS: A Pobabilistic Database Management System," *Proc. ACM SIGMOD Int'l Conf. Management of Data,* 2009.

[116] B. Kanagal and A. Deshpande, "Efficient Query Evaluation over Temporally Correlated Probabilistic Streams," *Proc. Int'l Conf. Data Eng. (ICDE),* 2009.

[117] C. Ré, J. Letchner, M. Balazinska, and D. Suciu, "Event Queries on Correlated Probabilstic Streams," *Proc. ACM SIGMOD Int'l Conf. Management of Data,* 2008.

[118] R. Karp and M. Luby, "Monte-Carlo Algorithms for Enumeration and Reliability Problems," *Proc. Ann. ACM Symp. Theory of Computing,* 1983.

[119] R. Jampani, F. Xu, and M. Wu, "MCDB: A Monte Carlo Approach to Managing Uncertain Data," *Proc. Ann. ACM Symp. Theory of Computing,* 2008.

[120] L. Perez, S. Arumugam, and C. Jermaine, "Evaluation of Probabilistic Threshold Queries in MCDB," *Proc. Ann. ACM Symp. Theory of Computing,* 2010.

[121] S. Arumugam, R. Jampani, L. Perez, F. Xu, C. Jermaine, and P. Haas, "MCDB-R: Risk Analysis in the Database," *Proc. VLDB Endowment,* vol. 3, pp. 782-793, 2010.

[122] S. Lee, "Imprecise and Uncertain Information in Databases: An Evidential Approach," *Proc. Int'l Conf. Data Eng. (ICDE),* 1992.

[123] J. Baldwin, "A Fuzzy Relational Inference Language for Expert Systems," *Proc. Int'l Symp. Multiple-Valued Logic,* pp. 416-423, 1983.

[124] H. Prade and C. Testemale, "Generalizing Database Relational Algebra for the Treatment of Incomplete or Uncertain Information and Vague Queries," *Information Sciences,* vol. 34, pp. 115-143, 1984.

[125] H. Prade and C. Testemale, "Fuzzy Relational Databases: Representational Issues and Reduction Using Similarity Measures," *J. Am. Soc. for Information Science,* vol. 38, pp. 118-126, 1988.

[126] D. Dubois and H. Prade, *Possibility Theory: An Approach to Computerized Processing of Uncertainty.* Plenum Press, 1988.

[127] M. Umano and S. Fukami, "Fuzzy Relational Algebra for Possibility-Distribution-Fuzzy-Relational Model of Fuzzy Data," *J. Intelligent Information Systems,* vol. 3, pp. 7-27, 1994.

[128] P. Bosc and O. Pivert, "Imprecise Data Management and Flexible Querying in Databases," *Fuzzy Sets, Neural Networks and Soft Computing,* Ch. 19, pp. 368-395, Van Nostrand Reinhold, 1994.

[129] P. Bosc and O. Pivert, "SQLf: A Relational Database Language for Fuzzy Querying," *IEEE Trans. Fuzzy Systems,* vol. 3, no. 1, pp. 1-17, Feb. 1995.

[130] Z. Ma, *Fuzzy Databases Modeling with XML,* Ch. 7, pp. 97-148. Kluwer Academic Publishing, 2005.

[131] Y. Tao, R. Cheng, X. Xiao, W.K. Ngai, B. Kao, and S. Prabhakai, "Indexing Multi-Dimensional Uncertain Data with Arbitrary Probability Density Functions," *Proc. 31st Int'l Conf. Very Large Data Bases,* 2005.

[132] C. Böhm, M. Gruber, P. Kunath, A. Pryakhin, and M. Schubert, "ProVeR: Probabilistic Video Retrieval Using the Gauss-Tree," *Proc. Int'l Conf. Data Eng. (ICDE),* 2007.

[133] N. Dalvi, K. Schnaitter, and D. Suciu, "Computing Query Probability with Incidence Algebras," *Proc. ACM SIGMOD-SIGACT-SIGART Symp. Principles of Database Systems,* 2010.

[134] C. Ré and D. Suciu, "Understanding Cardinality Estimation Using Entropy Maximization," *Proc. ACM SIGMOD-SIGACT-SIGART Symp. Principles of Database Systems,* 2010.

[135] M. Franklin, D. Kossmann, and T. Kraska, "CrowdDB: Answering Queries with Crowdsourcing," *Proc. ACM SIGMOD Int'l Conf. Management of Data,* 2011.

[136] A. Doan, M. Franklin, D. Kossmann, and T. Kraska, "Crowdsourcing Applications and Platforms: A Data Management Perspective," *Proc. Int'l Conf. Very Large Data Bases (VLDB),* 2011.

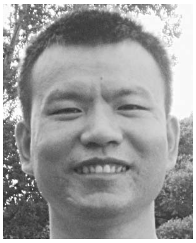
[137] S. Amer-Yahia, A. Doan, J. Kleinberg, N. Koudas, and M. Franklin, "Crowds, Clouds, and Algorithms: Exploring the Human Side of 'Big Data' Applications," *Proc. ACM SIGMOD Int'l Conf. Management of Data,* 2010.

[138] A. Marcus, E. Wu, D. Karger, S. Madden, and R. Miller, "Crowdsourced Databases: Query Processing with People," *Proc. Biennial Conf. Innovative Data Systems Research (CIDR),* 2011.

[139] A. Parameswaran and N. Polyzotis, "Answering Queries Using Humans, Algorithms and Databases," *Proc. Conf. Innovative Data Systems Research (CIDR),* 2011.

[140] A. Parameswaran, A. Sarma, and H. Garcia-Molina, "Humanassisted Graph Search: It's Okay to Ask Questions," *Proc. VLDB Endowment,* vol. 4, pp. 267-278, 2011.

Yiping Li received the bachelor's degree from the Beijing University of Technology in China in 2009. She is currently working toward the PhD degree in computer science and technology at Tsinghua University, China. Her research interest includes uncertain data management.



Jianwen Chen received the bachelor's degree from Nanjing University, China, in 2000. He is currently working the PhD degree in computer science and technology at Tsinghua University, China. His research interest includes uncertain data management.

Ling Feng is a professor of computer science and technology at Tsinghua University, China. Her research interests include context-aware data management toward ambient intelligence, knowledge-based information systems, data mining and warehousing, and so on. She received the 2004 innovational VIDI Award by the Netherlands Organization for Scientific Research and 2006 Chinese ChangJiang professorship Award by the Ministry of Education. She is a member of the IEEE.