Prediction-Driven Knowledge Discovery from Data and Prior Knowledge

Gheorghe Tecuci¹, Dorin Marcu¹, Steven Mirsky², Alison Robertson³
1Learning Agents Center, Department of Computer Science, School of Computing, George Mason University, Fairfax, VA 22030, USA
2Sustainable Agricultural Systems Lab, USDA, ARS, BARC-West, 10300 Baltimore Avenue, Beltsville, MD 20705, USA
3Department of Plant Pathology and Microbiology, Iowa State University, 2213 Pammel Drive, Ames, IA 50011, USA
tecuci@gmu.edu; dmarcu@gmu.edu; steven.mirsky@usda.gov; alisonr@iastate.edu

Abstract
This paper presents a novel approach to knowledge discovery. As opposed to the vast majority of existing approaches that use statistical machine learning to discover knowledge from data, we synergistically integrate ideas from artificial intelligence, machine learning, science of evidence, logic and probabilities to use both existing incomplete domain knowledge and imperfect data to discover new knowledge. We illustrate this approach in the precision agriculture domain considering the practices associated with cover crops, and learning how abiotic and biotic factors and cover crops management practices (planting and termination dates and methods) influence cereal cover crop biomass accumulation across corn-growing, soybean-growing, and cotton-growing regions of the U.S.

Introduction
Knowledge discovery from data (KDD) is a very active area of research that is concerned with the development and application of methods, such as classification, clustering, and association rule mining, to discover useful and interesting information from large collections of data (Tan et al., 2019).

We present research on the development of a novel approach that enables the discovery of new knowledge from existing data and prior knowledge, by synergistically integrating the scientific method of hypothesis generation and testing (Tecuci et al., 2016a), evidence-based reasoning (Schum et al., 2009; Tecuci et al., 2016b), multistrategy machine learning (Boicu et al., 2001; Tecuci, 1988; 1998), instructable agents (Tecuci and Hieb, 1996; Tecuci et al., 2000; 2002; 2005; 2007b; 2008; Huang et al., 2020), automated search, and mixed-initiative interaction (Tecuci et al., 2007a).

Section 2 presents the addressed discovery problem and method. Section 3 presents the domain used to illustrate them. Section 4 presents the method in detail. Finally Section 5 presents the current status and future work.
KDD approaches:

- Unlike the existing KDD approaches that rely on large amounts of data to draw conclusions, our approach can work with a few studies to formulate hypotheses that could explain the observed phenomenon and then test these hypotheses on the remaining studies. But this approach enables us to efficiently work with massive amounts of data as well.
- The individual study data do not need to be complete or uniform because they are treated as evidence on the considered hypotheses, evidence that can be incomplete, inconclusive, ambiguous, dissonant, or have various degrees of accuracy.
- We do not need to perform new experiments (that may be expensive and may require significant time and effort) because the formulated hypotheses can be tested on existing data.

**Biomass Accumulation of Cover Crops**

Cover crops (CC) are non-marketed crops that are planted between periods of cash crop production to provide a diverse array of ecosystem services including increased water and nutrient retention and availability, pest management, and greater soil health. As a result, CCs could also increase crop yield stability and resilience in a changing climate (Villamil et al., 2006; Kaspar et al., 2001; Meisinger et al., 1991; Ruffo et al., 2004; Shipley et al., 1992). Figure 2 summarizes our current partial knowledge on the main factors influencing CC biomass (Thapa et al., 2018). To reap various ecosystem services (reduced soil erosion, improved nutrient cycling, etc.), substantial CC biomass production is required. Extrinsic and intrinsic factors affect biomass production. We need to discover knowledge of how abiotic and biotic factors, and management practices (planting and termination dates, and methods) influence biomass accumulation across corn-, soybean-, and cotton- growing U.S. regions.

We start by selecting a reference farm case for which actual data exists. We then use our current knowledge on CC management to develop a predictive model in the form of a Wigmorean argumentation (Wigmore, 1913; 1937; Schum, 1987; 2001; Tecuci et al., 2016a) that explains how various intrinsic and extrinsic factors from the reference case support the production of high biomass. Next, we iteratively and automatically apply the predictive model to other farm cases for which individual data exist, to identify cases where the predicted results differed from the actual results. Explanation of the differences between the reference case and these farm cases leads to the discovery of new knowledge and the iterative improvement of the predictive model.

**Illustration of the Discovery Method**

**Step 1: Selection of a Reference Farm Case \( F_{ref} \)**

As shown in Figure 3, the first step of the investigation and discovery process is to select a reference farm case \( F_{ref} \) that will guide the discovery of knowledge applicable to the class of cases similar to it. In this illustration, \( F_{ref} \) is Maryland Farm1 during the 2018-2019 cover crop season, a past case in which the cover crops produced high biomass.

**Step 2: Argumentation that Explains the Phenomenon on \( F_{ref} \)**

In this step the current cover crops knowledge is used to develop a predictive model (in the form of a Wigmorean argumentation) that explains how the various intrinsic and extrinsic factors supported the production of high CC biomass in case \( F_{ref} \).

**Wigmorean Argumentation for Evidence-based Hypothesis Assessment**

Figure 4 shows an abstract example of a Wigmorean argumentation structure used to assess a hypothesis based on evidence.

The hypothesis \( H \) to be assessed is decomposed into simpler hypotheses by considering both favoring arguments (supporting the truthfulness of \( H \)), under the left (green) square, and disfavoring arguments (supporting the falsehood of \( H \)), under the right (pink) square. Each argument is an independent strategy of showing that \( H \) is true or false, and is characterized by a specific relevance or strength. The argument consists either of a single sub-hypothesis (e.g., \( H_3 \)) or a conjunction of sub-hypotheses (e.g., \( H_1 \ & \ H_2 \)). The sub-hypotheses from these arguments are further decomposed through other arguments, leading to simpler and simpler (sub-sub-)hypotheses that can be more accurately assessed based on evidence. Evidence is any observable sign, datum, or item of information that is relevant in deciding whether a statement or hypothesis (e.g., a scientific claim) is true or false (Schum, 2001).

Consider, for example, sub-sub-hypothesis \( H_{2b} \). There are two items of evidence relevant to this hypothesis, the favoring evidence item \( E_1 \), and the disfavoring evidence item \( E_2 \). Each item of evidence has three credentials that

![Figure 2. Cover crop knowledge.](image-url)
need to be assessed: accuracy, relevance, and inferential force. The accuracy of evidence answers the question: “What is the probability that the evidence is true?” The relevance of evidence to a hypothesis answers the question: “What would the probability of the hypothesis be if the evidence were true?” These two credentials are used to compute the inferential force or weight of the evidence on the hypothesis, which answers the question: “What is the probability of the hypothesis, based only on this evidence?” This is computed as the minimum between the accuracy and relevance. For example, the inferential force of $E_1$ is almost certain, that of $E_2$ is barely likely.

The probability of sub-sub-hypothesis $H_{2b}$ is determined by balancing the inferential force of the favoring evidence with that of the disfavoring evidence. Once the probabilities of the bottom-level hypotheses have been computed based on evidence, the probabilities of the upper level hypotheses are computed based on the logical structure of the Wigmorean argumentation (conjunctions and disjunctions of hypotheses), using min-max probability combination rules common to the Fuzzy probability view (Zadeh, 1983; Negoita and Ralescu, 1975; Schum 2001) and the Baconian probability view (Cohen, 1977; 1989; Schum, 2001). These rules are much simpler than the Bayes rule used in the Bayesian probability view (Schum, 2001), or the Dempster-Shafer rule in the Belief Functions probability view (Shafer, 1976).

Such Wigmorean arguments are easy to develop and understand, and can be learned by an intelligent software agent, such as the proposed KDA.

Thus, the specific Wigmorean argumentation for our example of biomass production of cover crops (Figure 5) shows how the factors in Figure 2 (e.g., climate, soil) supported the production of high biomass (top hypothesis) in the case of Maryland Farm1 during the 2018-2019 cover crop season ($F_{ref}$). There are two arguments favoring the top hypothesis. The left argument, based on the current cover crop knowledge, consists of three sub-hypotheses and states that favorable environmental, management, and genetic factors led to high biomass. Each of these factors has its own argument. For example, favorable environmental conditions were determined by favorable soil and climate conditions. Favorable soil conditions, in turn, were determined by high residual Nitrogen and excess drainage. These lower level conditions are supported by actual evidence (that is, data from the case study), $E_1$: Maryland Farm1 2018-19 has high residual N and $E_2$: Maryland Farm1 2018-19 has excess drainage. Now, following the inference steps from bottom-up, from these evidence items to the top hypothesis, one concludes high biomass on Maryland Farm1. The right argument of the top hypothesis is the direct evidence from the case data, that the biomass produced was high, $E_7$: Maryland Farm1 2018-19 has high biomass.

Thus, in this example, the argument based on knowledge correctly predicted the actual biomass produced. The question is: How can we determine whether this is true for
all the recorded cases? That is, how can we determine whether, in all the recorded cases, the result predicted using the current cover crop knowledge is consistent with what was actually produced? Any discovered inconsistency is an indication of an imperfect Wigmorean prediction model and thus of imperfect knowledge. We will now discuss how these inconsistencies can be used to correct or extend this knowledge.

Step 3: Knowledge-Based Generalization of the Argumentation

Ontology Development

Based on the argumentation from Figure 5 we develop the ontology (bottom right-hand Figure 5) to represent the entities from the argumentation, as well as similar ones. The ontology contains concepts from the application domain, such as farm, plot, plant and soil characteristic, which define the hierarchical types (taxonomies) for the entities in the argumentation. For example, Maryland Farm1 is a Maryland farm, while Plot1 is a plot. The ontology also represents the relationships between entities, for example, that Plot1 has high residual Nitrogen and excessive soil drainage.

Teaching the Discovery Assistant

We need to teach the KDA to automatically develop arguments like the one from Figure 5 based on farm data from other cases. As illustrated in Figure 6, the KDA learns a general hypothesis analysis rule from each specific...
argument that decomposes a hypothesis into sub-hypotheses. For example, from the top-left argument in Figure 6 the assistant learns the hypothesis analysis rule A1 shown at the bottom right of that figure.

The learned rule consists of the argument pattern obtained by replacing the entities from the top-left argument in Figure 6 (cereal rye P1, October 2018, Maryland Farm1, April 2019, …) with corresponding variables (i.e., ?O1, ?O2, etc.). The rule also has an applicability condition that indicates the possible values of these variables for which the reasoning pattern is likely to be correct, based on the hierarchy of concepts from the ontology in Figure 5. Notice however that, instead of a single applicability condition, the KDA learns a lower bound and an upper bound for this condition, using two complementary learning strategies:

The lower bound of the condition is generated by employing the strategy of a cautious learner that wants to minimize the chances of making mistakes when employing the learned pattern. This strategy increases the confidence of the KDA in the correctness of its reasoning. However, the KDA may fail to apply the reasoning pattern in situations where, in fact, it is applicable.

The upper bound of the condition is generated by employing the strategy of an aggressive learner that wants to maximize the opportunities of employing the learned pattern. This strategy increases the number of situations where the rule can be applied, although in some of these situations the reasoning may not be correct.

The two bounds may be refined, and may even become identical, based on additional example arguments encountered by the KDA.

The KDA also learns an evidence collection rule for each argument that reduces a hypothesis to an evidence item. A specialized collection agent can then search the data repository of recorded cases for the evidence item. The design and management of specialized collection agents are critical for the automatic extraction of evidence from existing farm data.

![Figure 6. Rules learned from the argumentation in Figure 5.](image-url)
The vast majority of the current machine learning approaches rely heavily on statistics and learn single functions from a large number of examples. Such approaches are not applicable for our learning problem because sets of examples to learn from (i.e., arguments) do not exist and would require a significant effort to create. Instead, an agricultural scientist explains to the KDA the individual arguments from Figure 5 by selecting the corresponding relations from the ontology or by defining them, and the agent learns rules as ontology-based generalizations of these arguments, as discussed above (Figure 6). The explanations provided by the agricultural scientist to the KDA point directly to the relevant features of the individual arguments, enabling rapid learning. Thus, these features do not need to be discovered through the statistical comparison of a large number of positive and negative argument examples (that are often not available), as current (statistics-based) inductive learning methods do (Witten et al. 2011; Flach, 2012; Alpaydın, 2020).

Step 4: Discovery of Favoring and Disfavoring Cases
This step involves a process of knowledge-based search and classification. The generalized argumentation is automatically applied to cases that are similar to the reference case $F_{ref}$. The cases are split into favoring cases (high biomass) and disfavoring cases (low or medium biomass).

Step 5: Selection of Most Similar Disfavoring Case
Since there are disfavoring cases, the argumentation from Figure 6 is incomplete and/or partially incorrect. We thus have to discover what factors were not taken into account, and improve this argumentation. To facilitate this complex knowledge discovery process, the KDA selects a new case $F_s$ from the set of disfavoring ones that is most similar to the reference farm case $F_{ref}$ (i.e., Maryland Farm1). Since there will be very fewer factors that are different, some may be responsible for the difference in cover crop biomass. This farm case $F_s$ might be, for example, Virginia Farm3 during the 2017-2018 cover crop season. The corresponding instantiation of the generalized argumentation from Figure 6 is shown in Figure 7. Notice that, in this case, the direct evidence $E_{37}$ disfavors the top hypothesis because the actual cover crop biomass produced was medium.

Step 6: Hypothesis-Driven Explanation Discovery
Now the agricultural scientist has to hypothesize an explanation for the difference in cover crop biomass between the two very similar cases $F_{ref}$ (Maryland Farm1) and $F_s$ (Virginia Farm3). After comparing data from these two farms, the agricultural scientist hypothesizes that the cause of medium biomass at Virginia Farm3 is the soil pH during the 2017-2018 season, which is too low. Since the soil pH at the Maryland Farm1 during the 2018-2019 season was neutral, the agricultural scientist hypothesizes that an additional relevant soil condition for high biomass (besides high residual Nitrogen and excessive drainage shown in Figure 5) is neutral pH.

Step 7: Explanation-Based Refinement of the Argumentation
As a result, the argumentation from Figure 5 inferring high biomass for the reference case $F_{ref}$ is extended (Figure 8) with the additional soil pH condition generalized as discussed in Step 3, validated on Virginia Farm3 during the 2017-2018 season, and automatically applied to all similar cases in Step 4. If the total number of favoring cases is not increased (or, equivalently, the total number of disfavoring cases not decreased), the hypothesis formulated in Step 6 is rejected and a new explanation has to be hypothesized. If, however, the total number of favoring cases is increased, the...
hypothesis is accepted and represents new discovered knowledge.

**Loop 5-6-7-3-4: Learning Argumentation-Based Explanation Models**

The 5-6-7-3-4 loop (Figure 3) is repeated, leading to new discovered knowledge, until all cases similar to $F_{ref}$ are correctly predicted, with the possible exception of a few anomalous cases. The result of this process is the discovery of new knowledge and a generalized argumentation that correctly infers and explains cover crop biomass for the class of cases similar to $F_{ref}$.

Then, the process restarts with Step 1, in which a new reference farm case is selected from the remaining data (if any). The entire process is repeated along the steps from Figure 3 as discussed so far, until all available farm cases are correctly predicted, with the possible exception of a few anomalous cases (including those for which the data may be incorrect or incomplete).

**Role of Cover Crops in Weed Suppression.** The role of CCs in weed suppression is complex and dependent on many factors, including geographical region, climate, and soil. Past research has shown that CC mulch levels are highly correlated with suppression of summer annual weeds. However, this relationship varies considerably with climate and soil type and other plant-soil interactions. To date, there has not been an integration of these factors to elucidate how climate, soil, and management intersect to drive weed suppression. Our incomplete understanding of the factors influencing weed growth stymies our ability to parameterize models that capture the complexity of cover crop-weed interactions. We plan to apply the KDA to improve our current understanding of how climate, soil, and weed seedbank size interact with CC biomass to drive weed suppression.

**Foliar Fungicide Applications to Field Crops.** In the past two decades, fungicide use on field crops has increased considerably (Mueller et al., 2021). Several factors have been suggested for this growth in fungicide use including increased commodity prices, more fungicide products registered for use on field crops, greater disease prevalence, and marketing (Wise and Mueller, 2011). Fungicides have traditionally been applied to reduce disease, however, in more recent years they have been applied for their physiological plant effects that may contribute to yield increases. Research in soybean and other crops, however, indicates these physiological effects are inconsistent, or do not always result in a measurable yield increase (Bertelsen et al., 2001; Khan and Carlson, 2009; Robertson et al., 2020). We need to understand the abiotic and biotic, management, and environmental factors that contribute to greater yields in field crops to identify situations where an application of a fungicide could result in greater yields and a return on investment.

The KDA will be freely available to scientists of any ilk and we will support its use.

**Current Status and Future Research**

We are currently developing an intelligent system, called the Knowledge Discovery Assistant (KDA) that implements the presented approach. Many of the modules of the KDA are customizations of modules of Cogent, a cognitive assistant for intelligence analysis (Tecuci et al., 2018) and of Disciple-EBR, a tool set for developing agents for evidence-based reasoning tasks (Tecuci et al. 2016b).

Once the prototype of the KDA is developed, it will be used by agricultural scientists to discover knowledge in three areas, one being the biomass accumulation of CC discussed above. The other two are briefly presented below.

**References**


