Causal Feature Learning for Utility-Maximizing Agents

David Kinney DAVID.KINNEY@SANTAFE.EDU
Santa Fe Institute

David Watson DAVID.WATSON@OII.OX.AC.UK
Oxford Internet Institute

Abstract

Discovering high-level causal relations from low-level data is an important and challenging problem that comes up frequently in the natural and social sciences. In a series of papers, Chalupka et al. (2015, 2016a, 2016b, 2017) develop a procedure for *causal feature learning* (CFL) in an effort to automate this task. We argue that CFL does not recommend coarsening in cases where pragmatic considerations rule in favor of it, and recommends coarsening in cases where pragmatic considerations rule against it. We propose a new technique, *pragmatic causal feature learning* (PCFL), which extends the original CFL algorithm in useful and intuitive ways. We show that PCFL has the same attractive measure-theoretic properties as the original CFL algorithm. We compare the performance of both methods through theoretical analysis and experiments.

Keywords: Causal Feature Learning; Coarse-Graining; Bayesian Networks; Expected Utility.

1. Introduction

In many scientific contexts, one goal of inquiry is to discover types of fine-grained events that can be grouped together into a smaller set of more coarse-grained events. This is especially true in causal analysis, where the goal is often to find some hierarchical structure to improve both the tractability and the representational adequacy of models. Developing efficient and reliable computational methods for mapping low-level data to high-level phenomena is therefore an important step in the more general task of automating, at least partially, the process of scientific discovery.

Chalupka et al. (2015, 2016a, 2016b, 2017) propose a procedure called *causal feature learning* (CFL) to derive macrovariables from microvariables in datasets with some minimal causal structure. For example, Chalupka et al. (2016a) analyze wind speeds and sea surface temperatures in a particular region of the Western Pacific Ocean. Using CFL, they partition both sets of variables into coarse-grained clusters that reveal the causal association between these fine-grained observations and large-scale weather patterns including El Niño and La Niña.

Chalupka et al.'s approach is related to work by Hoel et al. (2013) and Hoel (2017). Although Hoel et al.'s formal approach (unlike Chalupka et al.'s) explicitly incorporates information theory, the approaches are similar in that they define an optimal coarse-graining for variables in a causal model using only the probabilistic relationships between variables. Additionally, work by Beckers et al. (2019) builds on work by Rubenstein et al. (2017) and Beckers and Halpern (2019) to define a scheme for coarse-graining causal variables such that the resulting causal graph is an approximation, to some degree of precision, of the underlying, fine-grained graph. While what we present here is in broad agreement with the spirit of this approach, our proposal is different in that it explicitly represents the extent to which the optimal level of approximation can be determined by the pragmatic interests of an agent.

The merits of CFL notwithstanding, we argue that the method is ill-equipped to handle the variable interests of real-world agents who may undertake causal analyses of a target system with different goals in mind. We demonstrate that by failing to incorporate pragmatic information, CFL is prone to errors in both directions – failing to cluster values of a causal variable that should be grouped together, and clustering values of an effect variable that should be kept separate. We present an algorithm for *pragmatic causal feature learning* (PCFL) that avoids these pitfalls without sacrificing any of the measure-theoretic advantages of the original CFL method.

The remainder of this paper is structured as follows. In Sect. 2, we outline the original CFL algorithm. We present two examples in Sect. 3 that demonstrate how CFL can generate errors even in relatively simple cases. We introduce PCFL in Sect. 4, outline an algorithm for implementing the procedure in Sect. 5, and present experimental results in Sect. 6. Sect. 7 concludes.

2. Causal Feature Learning

Let (C, E) be a pair of discretely-valued random variables in a set $\mathcal{V}$. Let $(\mathcal{V}, \mathcal{E}, p(\cdot))$ be a Bayesian network in which $\mathcal{E}$ is an acyclic set of ordered pairs (or directed edges) of variables in $\mathcal{V}$, and $p(\cdot)$ is a joint probability distribution over the cross-product of the ranges of the variables in $\mathcal{V}$. In the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, E is a descendant of C. In keeping with the theory of Bayesian networks developed by Pearl (1988), the graph $(\mathcal{V}, \mathcal{E})$ is Markov to the probability distribution $p(\cdot)$. Let $p(e_i|\hat{c}_j)$ be an *interventional conditional probability*, in which the notation $\hat{c}_j$ indicates that the value of C has been set via an exogenous intervention on the system represented by the Bayesian network $(\mathcal{V}, \mathcal{E}, p(\cdot))$. Let R_X be the range of a random variable X. Finally, a set A is a *quotient set* of a set B iff each element of A is an equivalence class of elements of B, according to some equivalence relation $\sim$.

Chalupka et al.'s proposal for coarsening variables in causal models is straightforward. First, they define equivalence relations over the ranges of cause and effect variables. These equivalence relations form the basis of a coarsening process that generates macrovariables whose values are just the equivalence classes of the microvariables. Chalupka et al. define the equivalence relation $\sim_c$ over the range R_C as follows:

Causal Equivalence: $c_j \sim_c c_k$ with respect to E iff $p(e_i|\hat{c}_j) = p(e_i|\hat{c}_k)\ \forall e_i \in R_E$.

Let $C^{\dagger[E]}$ be the *causal coarsening* of C with respect to E iff its range $R_{C^{\dagger[E]}}$ is the quotient set of R_C induced by the equivalence relation $\sim_c$ with respect to E. Chalupka et al.'s equivalence relation $\sim_e$ over the range R_E is defined as follows:

Effect Equivalence: $e_i \sim_e e_s$ with respect to C iff $p(e_i|\hat{c}_k) = p(e_s|\hat{c}_k)\ \forall c_k \in R_C$.

Let $E^{*[C]}$ be the *effect coarsening* of E with respect to C iff its range $R_{E^{*[C]}}$ is the quotient set of R_E induced by the equivalence relation $\sim_e$ with respect to C.

Chalupka et al.'s principal formal achievement is to prove that, except in a Lebesgue measure zero subset of cases, the causal coarsening and the effect coarsening of a given variable can be learned from observational rather than experimental data. To state their result in a perspicuous way, we first define additional equivalence relations and coarsenings over the variables in the cause-effect pair (C, E). Let us begin with the *observational causal equivalence* relation $\sim_{oc}$, which is defined as follows:

Observational Causal Equivalence: $c_j \sim_{oc} c_k$ with respect to E iff $p(e_i|c_j) = p(e_i|c_k)\ \forall e_i \in R_E$.

Note that the sole difference between observational causal equivalence and causal equivalence *simpliciter* is that the former is defined using observational conditional probabilities (i.e., conditional probabilities such that the conditioning event is *not* set via an intervention), whereas the latter is defined using causal conditional probabilities (i.e., conditional probabilities such that the conditioning event *is* set via an intervention). Let $C^{oc[E]}$ be the *observational causal coarsening* of C with respect to E iff the range $R_{C^{oc[E]}}$ is the quotient set of the range R_C induced by the equivalence relation $\sim_{oc}$ with respect to E. Similar definitions of an observational equivalence relation and accompanying coarsening can be given for the effect variable:

> **Observational Effect Equivalence**: $e_i \sim_{oe} e_s$ with respect to C iff $p(e_i|c_k) = p(e_s|c_k)$ $\forall c_k \in R_C$.

Let $E^{oe[C]}$ be the *observational effect coarsening* of E with respect to C iff the range of $E^{oe[C]}$ is a quotient set of the range of E according to the equivalence relation $\sim_{oe}$ with respect to C.

The distributions $p(E|c_j)$ and $p(E|\hat{c}_j)$ are not necessarily equivalent. For instance, the two may diverge due to some confounding variable Z that is a cause of both C and E. Alternatively, Z may be an effect of C and a cause of E. In both cases, there may be differences between observational and interventional distributions, such that the observational coarsenings will differ from causal coarsenings of both the cause and effect variables. Chalupka et al. argue that when we coarsen fine-grained cause and effect variables, we can almost always ignore the distorting influence of potential confounders.

Proposition 1 (Chalupka et al. 2017, p. 149)**.** *Let C, E, and Z be variables in a graph $\mathcal{G}$ such that C is an ancestor of E and Z is a possible confounder of the causal relationship between C and E. Consider the set of possible joint probability distributions over C, E, and Z. Let $C^{\dagger[E]}$ and $C^{oc[E]}$ be the causal coarsening and observational causal coarsening of C, respectively, in any such probability distribution. Let $E^{*[C]}$ and $E^{oe[C]}$ be the effect coarsening and observational effect coarsening of E, respectively, in any such probability distribution. The set of joint probability distributions over C, E, and Z such that the range of $C^{\dagger[E]}$ is not a quotient set of the range of $C^{oc[E]}$ and the range of $E^{*[C]}$ is not a quotient set of the range of $E^{oe[C]}$ is Lebesgue measure zero within the set of all possible joint distributions over C, E, and Z.*

This result has a putatively important practical consequence. Suppose that we want to find the causal coarsening of some fine-grained variable C that is a cause of some other fine-grained variable E. If we are guided solely by the coarsening strategies outlined above, then doing so would require a separate intervention for each value of C, since each causal conditional probability would be needed to determine which values of C and E are equivalent. However, equipped with Proposition 1, we can instead use a potentially more efficient procedure. Using just observational data, we can coarsen the microvariable C into the observational macrovariable $C^{oc[E]}$. We can then intervene to set $C^{oc[E]}$ to each of its values, and thereby determine which (if any) are equivalent according to the relation $\sim_c$ with respect to E, thereby generating the causal coarsening $C^{\dagger[E]}$. This procedure is potentially more efficient since, by construction, $|R_{C^{oc[E]}}| \leq |R_C|$. The procedure is justified just in case the joint probability distribution over C, E, and Z is such that $C^{\dagger[E]}$ is a coarsening of $C^{oc[E]}$, a condition that provably only fails for a measure zero set of probability distributions. An analogous argument applies to the effect variable E.

	[0,49]	[50,69]
Marlboro	.026	.25
Other	.024	.25
Nothing	.001	.05
	[70,90]	[90,Inf]
Marlboro	.698	.026
Other	.702	.024
Nothing	.948	.001

Table 1: Causal CPT

	[0,49]	[50,69]
Marlboro	-950	1100
Other	-990	1050
Nothing	-1000	1000
	[70,90]	[90,Inf]
Marlboro	2100	2150
Other	2050	2145
Nothing	2000	2050

Table 2: Utilities for Smoking Decision

3. Problems for Causal Feature Learning

Suppose that a person's mortality age in years is represented by an effect variable D such that $R_D = \{\texttt{[0,49]}, \texttt{[50,69]}, \texttt{[70,89]}, \texttt{[90,Inf]}\}$. The causal variable S describes the smoking habits of individuals, with $R_S = \{\texttt{Marlboro}, \texttt{Other}, \texttt{Nothing}\}$. All causal conditional probabilities are given in Table 1. In this example, the probability distribution over the effect variable given an intervention making someone a Marlboro smoker is very similar to the probability distribution over the effect variable given an intervention making someone a smoker of other brands. For many practical purposes, these two values of the causal variable are equivalent, and only trivially non-equivalent, such that the value space of the causal variable should be coarsened into the two-element set $\{\texttt{Smoker}, \texttt{Non-Smoker}\}$. However, CFL recommends against such a coarsening and therefore fails to satisfy an intuitive desideratum for any coarsening procedure – namely, that it groups together causal values that are only trivially non-equivalent in a given context.

CFL can also lead to counterintuitive results for effect variable coarsening. Consider again the conditional probability distribution in Table 1. On Chalupka et al.'s account, the values `[0,49]` and `[90,Inf]` should be coarsened into a single discontinuous value `[0,49]` $\vee$ `[90,Inf]`. However, this coarse-grained value does not seem to pick out any meaningful scientific category. After all, dying before age fifty and dying after age ninety are very different outcomes from both an ontological and a pragmatic standpoint. Yet these outcomes are equivalent according to Chalupka et al.'s definition. Thus, CFL fails to satisfy another desideratum of a causal coarsening procedure (viz., that it not group together effect values that are only trivially equivalent in a given context).

4. Pragmatic Causal Feature Learning

In this section, we present a novel approach to CFL wherein values of the cause and effect variables are coarsened together iff doing so does not decrease the maximum expected utility of intervening on the causal variable for an agent with a pragmatic interest in the target system. To formalize this idea, let $u(\cdot) : R_C \times R_E \to \mathbb{R}$ be a function that represents the utility some agent receives when a given pair of cause and effect values obtains.[1] For each value c_j of the causal variable, we define the following set of utilities $U_{E|c_j} = \{u(c_j, e_1), \ldots, u(c_j, e_n)\}$ and set of interventional conditional probabilities $P_{E|\hat{c}_j} = \{p(e_1|\hat{c}_j), \ldots, p(e_n|\hat{c}_j)\}$. The expected utility of setting causal variable C to value c_j via an intervention, for an agent with a utility function defined over the product space $R_C \times R_E$, is given by the inner product of these two sets. Let $\eta_u(C, E) = \max_{c_j \in R_C} \langle U_{E|c_j}, P_{E|\hat{c}_j} \rangle$ denote

1. This utility function can be obtained from data regarding an agent's preferences over cause-effect pairs; see von Neumann and Morgenstern (1944).

this agent's maximum expected utility over possible causal interventions. To illustrate, consider an agent who is deliberating whether to smoke cigarettes, and which brand to smoke. The agent's payoffs are given in Table 2. Using the interventional conditional probability table shown in Table 1, we calculate that if S is a variable such that the possible interventions on S are the smoker statuses listed in the rows of Table 2, and the four death age categories in the columns of Table 2 comprise the range of a variable D, then $\eta_u(S, D) \approx 1947.05$. In what follows, we define equivalence relations over fine-grained cause and effect variables such that the corresponding coarsenings maximize $\eta_u(\cdot)$.

We define a pragmatic equivalence relation $\sim_{pc}$ between two values of the causal variable:

> **Pragmatic Causal Equivalence**: $c_j \sim_{pc} c_k$ with respect to E iff either of the following holds: i) $\langle U_{E|c_j}, P_{E|\hat{c}_j} \rangle = \langle U_{E|c_k}, P_{E|\hat{c}_k} \rangle = \eta_u(C, E)$, ii) $\langle U_{E|c_j}, P_{E|\hat{c}_j} \rangle \neq \eta_u(C, E)$ and $\langle U_{E|c_k}, P_{E|\hat{c}_k} \rangle \neq \eta_u(C, E)$.

$C^{pc[E]}$ is the *pragmatic causal coarsening* of C with respect to E iff its range $R_{C^{pc[E]}}$ is the quotient set of R_C induced by the equivalence relation $\sim_{pc}$ with respect to E. By construction, $C^{pc[E]}$ is equivalent to a Boolean variable that takes a value of 1 for all and only those fine-grained c_j that maximize an agent's expected utility according to some function $u(\cdot)$. All other values of C correspond to sub-optimal interventions from the point of view of an agent who aims to maximize expected utility. Using Tables 1 and 2, one can verify that `Marlboro` and `Other` are pragmatic causal equivalents. Thus, PCFL successfully coarsens together these two values of C, whereas CFL fails to do so. We define a pragmatic equivalence relation $\sim_{pe}$ between two values of the effect variable:

> **Pragmatic Effect Equivalence**: $e_i \sim_{pe} e_s$ with respect to C iff $u(c_k, e_i) = u(c_k, e_s)$ $\forall c_k \in R_C$.

$E^{pe[C]}$ is the *pragmatic effect coarsening* of E with respect to C iff its range $R_{E^{pe[C]}}$ is the quotient set of R_E induced by the equivalence relation $\sim_{pe}$ with respect to C. Using Table 2, one can verify that `[0,49]` and `[90,Inf]` are *not* pragmatic effect equivalents. Thus, PCFL successfully refrains from coarsening together these two values of E, whereas CFL fails to do so.

We show that PCFL has the same measure-theoretic advantages as CFL. First, we define the following *observational* pragmatic equivalence relation over R_C:

> **Observational Pragmatic Causal Equivalence**: $c_j \sim_{opc} c_k$ with respect to E iff $\langle U_{E|c_j}, P_{E|c_j} \rangle = \langle U_{E|c_k}, P_{E|c_k} \rangle$.

There are two salient differences between observational pragmatic causal equivalence and pragmatic causal equivalence *simpliciter*. First, the observational pragmatic causal equivalence is calculated using conditional probability sets of the form $P_{E|c_j} = \{p(e_1|c_j), \ldots, p(e_n|c_j)\}$, which contains observational conditional probabilities rather than interventional conditional probabilities. Second, observational pragmatic causal equivalence is not assessed relative to any maximum expected utility; we only check whether expected utility is the same, given an observation of each value of C.

Let $C^{opc[E]}$ be the *observational pragmatic causal coarsening* of C with respect to E iff its range $R_{C^{opc[E]}}$ is the quotient set of R_C induced by the equivalence relation $\sim_{opc}$ with respect to E. We can now state the following proposition:

Proposition 2. *Let C, E, and Z be variables in a graph $\mathcal{G}$ such that C is an ancestor of E, and Z is a possible confounder of the causal relationship between C and E. Consider the set of possible*

Algorithm 1 CFL

input : $\mathcal{D} = \{(c_1, e_1), \ldots, (c_N, e_N)\}$
`Cluster` - a clustering algorithm
output : $W(c_i), T(e_i)$

1: Regress $f \leftarrow \text{argmin}_f \sum_1^N (f(c_i) - e_i)^2$;
2: Let $W(c_i) \leftarrow \texttt{Cluster}(f(c_1), \ldots, f(c_N))[c_i]$;
3: Let $\texttt{Range}(W) = \{c_1^{oc[E]}, \ldots, c_v^{oc[E]}\}$;
4: Let $\mathfrak{E}_\beta \leftarrow \{e_i | W(c_i) = c_\beta^{oc[E]} \text{ and } (c_i, e_i) \in \mathcal{D}\}$;
5: Let $g(e_i) \leftarrow [\texttt{kNN}(e_i, \mathfrak{E}_\beta), \ldots, \texttt{kNN}(e_i, \mathfrak{E}_v)]$;
6: Let $T(e_i) \leftarrow \texttt{Cluster}(g(e_1), \ldots, g(e_N))[e_i]$;

Algorithm 2 PCFL

input : $\mathcal{D} = \{(c_1, e_1), \ldots, (c_N, e_N)\}$
$\mathcal{U} = \{u(c_1, e_1), \ldots, u(c_N, e_N)\}$
`Cluster` - a clustering algorithm
output : $W_p(c_i), T_p(e_i)$

1: Regress $f \leftarrow \text{argmin}_f \sum_1^N (f(c_i) - u(c_i, e_i))^2$;
2: Let $W_p(c_i) \leftarrow \texttt{Cluster}(f(c_1), \ldots, f(c_N))[c_i]$;
3: Let $g(e_i) \leftarrow [u(c_1, e_1), \ldots, u(c_N, e_N)]$;
4: Let $T_p(e_i) \leftarrow \texttt{Cluster}(g(e_1), \ldots, g(e_N))[e_i]$;

joint probability distributions over C, E, and Z. Let $C^{pc[E]}$ and $C^{opc[E]}$ be the pragmatic causal coarsening and observational pragmatic causal coarsening of C, respectively, in any such probability distribution. The set of joint probability distributions over C, E, and Z such that the range of $C^{pc[E]}$ is not a quotient set of the range of $C^{opc[E]}$ is Lebesgue measure zero within the set of all possible joint distributions over C, E, and Z.[2]

There is no formal analog of the causal coarsening theorem for pragmatic effect coarsening. However, it is still the case that we can efficiently learn the pragmatic coarsening of an effect variable from just the observational conditional probability distribution over the causal variable, with failure only occurring in a Lebesgue measure zero subset of probability distributions. This is because the pragmatic effect equivalence relation holds or does not hold between two values of an effect variable independently of the probability distribution over the cause-effect pair (C, E). Thus, we can use the following procedure to learn the pragmatic coarsening of a pair (C, E) from observational data. First, obtain the observational pragmatic coarsening $C^{opc[E]}$. Next, intervene on each value of $c_j^{opc[E]} \in R_{C^{opc[E]}}$ and calculate $\langle U_{E|\hat{c}_j^{opc[E]}}, P_{E|\hat{c}_j^{opc[E]}} \rangle$ to obtain the pragmatic causal coarsening $C^{pc[E]}$. This procedure will only fail to obtain the true range of $C^{pc[E]}$ on a Lebesgue measure zero subset of probability distributions over (C, E). Finally, use the utility function over $R_C \times R_E$ to determine which values of the effect variable are pragmatic effect equivalents in order to obtain the pragmatically coarsened cause-effect pair $(C^{pc[E]}, E^{pe[C]})$.

5. An Algorithm for Pragmatic Causal Feature Learning

Alg. 1 presents Chalupka et al. (2016a)'s original CFL algorithm, rephrased to conform to the formalisms used in this paper. Note that this algorithm only finds the *observational* coarsenings of the fine-grained cause and effect variable. Thus, it serves mainly to compress the space of interventions needed to learn the full causal coarsenings $C^{\dagger[E]}$ and $E^{*[C]}$, under the assumption that the interventional conditional probability distribution over E, given each intervention on C, is not known and cannot be learned via inference from the true causal structure (i.e., the true causal structure is also not known). The algorithm assumes that the fine-grained variables C and E can take a finite number of continuous or categorical values, and takes as input N observations of values for C and E. Note

2. Note that both this result and Prop. 1 are only significant if one accepts that the distribution over the set of possible distributions over variables is given by the Lebesgue measure. Thus, we claim only that our approach is on equal footing with Chalupka et al.'s with respect to the putatively good-making features of an algorithm established by these geometric results.

that line 1 regresses E on C with L_2 loss, thereby learning the conditional expectation $\mathbb{E}[E|C]$. In general $\mathbb{E}[E|c_j] = \mathbb{E}[E|c_k]$ is a necessary but insufficient condition for $c_j \sim_{oc} c_k$ with respect to E. The mean value of a random variable can be identical under two distributions even if those distributions are not identical. However, Chalupka et al. use $\mathbb{E}[E|c_j] = \mathbb{E}[E|c_k]$ as a "heuristic indicator" for $c_j \sim_{oc} c_k$ with respect to E, while accepting that this heuristic may sometimes fail (2016a, p. 6). Note also that the algorithm `kNN` in line 5 returns the distance between e_i and its k-th nearest neighbor in the set constructed in Line 4, for some arbitrarily selected k.[3] If two values e_i and e_s of E are the same distance from their k-th nearest neighbor in each of the sets constructed in Line 4, such that they will be clustered together in line 6, then $p(e_i|c_\beta^{oc[E]}) = p(e_s|c_\beta^{oc[E]})$ for all $c_\beta^{oc[E]}$, as shown by Fukunaga and Hostetler (1973).

Alg. 2 provides pseudo-code for PCFL. Crucially, this algorithm takes as an additional input the utilities that the agent assigns to each observed cause-effect pair. Note that Line 1 regresses $\mathcal{U}$ on C with L_2 loss, thereby finding the conditional expected utility $\mathbb{E}[u(c_i, E)|c_i]$ for each c_i. Since $\mathbb{E}[u(c_i, E)|c_i] = \langle U_{E|c_i}, P_{E|c_i} \rangle$, the regression in line 1 allows us to define, in Line 2, a function $W_p(\cdot)$ whose range consists of all and only those values of the observational pragmatic causal coarsening $C^{opc[E]}$ that appear in the dataset $\mathcal{D}$. Unlike in CFL, the expectations $\mathbb{E}[u(c_i, E)|c_i]$ are not heuristic indicators of an equivalence relationship, the use of which may lead to inaccurate output. Rather, we can use these expectations to directly learn which values of C are observational pragmatic causal equivalents with respect to E.

6. Experimental Results

6.1 Simulated Data

We implement Alg. 1 and Alg. 2 on a simulated dataset to compare the performance of both methods.[4] Consider a Bayes net M (see Fig. 1) with two unobserved binary variables $Z_1 \sim \text{Bern}(0.5)$ and $Z_2 \sim \text{Bern}(0.5)$. Let $R_C = R_E = \{-2, -1, 1, 2\}$. Values of the causal variable C are fixed by the exogenous variables as follows: $C = -2$ if $Z_1 = 0$ and $Z_2 = 0$, $C = -1$ if $Z_1 = 0$ and $Z_2 = 1$, $C = 1$ and $Z_1 = 1$ and $Z_2 = 0$, and $C = 2$ if $Z_1 = 1$ and $Z_2 = 1$. Conditional probabilities for the effect variable E are given by the equation $P(E|C, Z_1) = \sigma(\boldsymbol{\alpha} + C\boldsymbol{\beta} + Z_1\boldsymbol{\gamma})$, where $\sigma(\cdot)$ denotes the softmax transformation and linear parameters are computed to induce the conditional probabilities listed in Table 3. Note that the exogenous variable Z_1 has directed edges into both C and E, thereby confounding our causal relationship of interest, $C \to E$. It is clear from Table 3 that $C = -1$ and $C = 1$ are observational causal equivalents, while $E = -2$ and $E = 2$ are observational effect equivalents, according to the definitions in Sect. 2. Thus, we expect CFL's coarsened conditional probability table to appear as it does in Table 4. This is empirically verified by a simulation experiment in which we run CFL on 10,000 samples drawn from $P(M)$. Resulting probabilities are reported in Table 5.

To test the performance of Alg. 2, we introduce a utility matrix over cause-effect pairs (see Table 6). Using these utilities, along with the conditional probabilities in Table 3, one can calculate the expected utility of each $c_j \in R_C$, and observe that $C = -2$ and $C = 2$ are observational pragmatic causal equivalents; both have an expected utility of 2.25. By contrast, the expected utilities of $C = -1$ and $C = 1$ are 4.50 and 7.50, respectively, and are therefore only observationally

3. Chalupka et al. use Euclidean distance, but other measures could of course be substituted here.
4. Code for all simulations and algorithms is available at `https://github.com/davidbkinney/pcfl`.

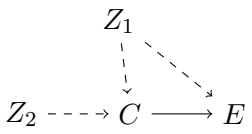

Figure 1: Causal diagram for the graph M.

	$E=-2$	$E=-1$	$E=1$	$E=2$
$C=-2$	.248	.189	.315	.248
$C=-1$	.252	.248	.248	.252
$C=1$	.252	.248	.248	.252
$C=2$	.248	.315	.189	.248

Table 3: Expected conditional probabilities, with expectation taken over the graph M.

	$E=-2\vee 2$	$E=-1$	$E=1$
$C=-2$	.496	.189	.315
$C=-1\vee 1$	.504	.248	.248
$C=2$	.496	.315	.189

Table 4: Expected output of CFL algorithm on data sampled from M.

	$E=-2\vee 2$	$E=-1$	$E=1$
$C=-2$	.491	.186	.323
$C=-1\vee 1$	.510	.255	.236
$C=2$	.514	.295	.191

Table 5: Observed output of CFL algorithm on data sampled from M.

pragmatically equivalent to themselves. In addition, we observe that $E=-1$ and $E=1$ are observational pragmatic effect equivalents, while $E=-2$ and $E=2$ are only observationally pragmatically equivalent to themselves. On this basis, and retaining the assumption that all values of C are equiprobable, we expect that PCFL will yield the coarsened conditional probabilities shown in Table 7. This is verified by our simulation test, where we obtain the correct pragmatic observational coarsening. Resulting probabilities are reported in Table 8.

We argue that, in the context of the utility function implied by Table 6, the coarsening in Table 8 is a more appropriate high-level representation of the target system than the coarsening in Table 5. This is because the former coarsens together those values of the cause and effect variables that represent equally desirable states of affairs, from the point of view of an agent whose preferences are represented by the utilities given in Table 6. The coarsening shown in Table 5 does not have a similar significance, and in some applications can lead to errors, as shown in Sect. 3. Note that the distinction is not reducible to continuity assumptions, since an agent may have valid reasons to coarsen together nonadjacent regions of the feature space. For instance, it is common in statistics to distinguish between significant and insignificant results of two-tailed hypothesis tests, a procedure which naturally groups together the largest and smallest values of a test statistic. Similarly, PCFL clusters $C=-2$ and $C=2$ together in this simulation, in accordance with the given utilities.

6.2 El Niño Data

We re-analyzed the dataset that Chalupka et al. (2016a) use in their study of El Niño using both CFL and PCFL. This dataset includes weekly average zonal wind speeds and sea surface temperatures

	$E=-2$	$E=-1$	$E=1$	$E=2$
$C=-2$	1	2	2	4
$C=-1$	8	5	5	0
$C=1$	5	8	8	9
$C=2$	4	2	2	1

Table 6: Utility matrix for PCFL simulation experiment.

	$E = -2$	$E = -1 \vee 1$	$E = 2$
$C = -2 \vee 2$	.248	.504	.248
$C = -1$	.252	.496	.252
$C = 1$	.252	.496	.252

Table 7: Expected output of PCFL algorithm on data sampled from M.

	$E = -2$	$E = -1 \vee 1$	$E = 2$
$C = -2 \vee 2$	.251	.498	.252
$C = -1$	.257	.494	.249
$C = 1$	.261	.487	.252

Table 8: Observed output of PCFL algorithm on data sampled from M.

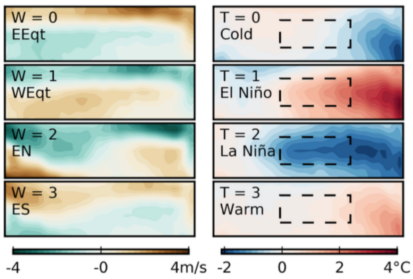

Figure 2: Alg. 1 output for the El Niño dataset (2016a). Zonal wind fields are visualized left, sea surface temperatures right.

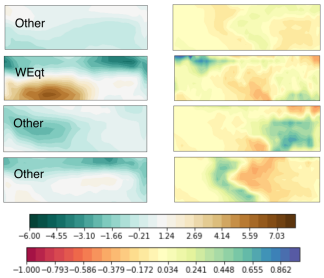

Figure 3: Alg. 2 output for the El Niño dataset. Zonal wind fields are visualized left, sea surface temperatures right.

from the same region of the Equatorial Pacific, each consisting of a 9×55 matrix of real numbers for each of 13,140 weeks. Fig. 2 shows the results of running CFL on this dataset; the algorithm picks out an Easterly Equatorial wind pattern, which causes slightly colder sea surface temperatures; a Westerly Equatorial wind pattern, which causes the warmer sea surface temperatures associated with the El Niño effect; an Easterly North of the Equator wind pattern, which causes the colder temperatures associated with the La Niña effect; and an Easterly South of the Equator wind pattern, which causes slightly warmer sea surface temperatures.

To implement PCFL, we let X and Y be the set of possible directional wind speeds and temperatures, respectively, in a given region of the Western Pacific. We then defined the following utility function over $R_X \times R_Y$, where y^* is the temperature rounded to the first decimal:

$$u(x, y) = -1 + \frac{1}{\sqrt{.02\pi}} e^{\frac{(y^* - 26)^2}{.02}} \tag{1}$$

Utilities for temperature observations are therefore determined by their distance from a mean of 26°C using a radial basis function kernel with small bandwidth and a scale shift of -1. This encodes a strong preference on the part of the relevant agent for temperatures near 26°C, rather than any observer-neutral value that this temperature might have. Note that in this case, the utility $u(x, y)$ is independent of the causal value x. As our earlier experiment on simulated data shows, our framework does not generally assume or require such independence.

Fig. 3 shows the output of PCFL when it is set to discover four clusters of both X and Y. The key difference between our results and those of Chalupka et al. is that whereas CFL picks out four trends in the data, the only distinct wind direction/speed pattern that PCFL identifies is an extreme version of the Westerly Equatorial wind pattern, shown in the second panel from the top. This causes mostly negative outcomes, as indicated in the pattern to the right of the Westerly Equatorial wind pattern; the darker orange regions represent highly negative outcomes, which are not offset by the

small patches of dark-blue, positive outcomes in the upper left corner. Recalling that the Westerly Equatorial wind pattern is a cause of the El Niño effect, our interpretation of this result is that an agent with the utility function defined by Eq. 1 is highly averse to the El Niño weather pattern and relatively unconcerned about other meteorological phenomena. Thus, this agent picks out two main classes of wind patterns: the Westerly Equatorial pattern and other patterns.

7. Discussion and Conclusion

The PCFL algorithm is a principled extension of Chalupka et al.'s CFL method. By incorporating pragmatic information about the inquiring agent's preferences over outcomes, we avoid certain counterintuitive and undesirable consequences of the original CFL algorithm without sacrificing its attractive measure theoretic properties. Our approach is also more computationally efficient if agents rely on utility functions less complex than the kNN algorithm. However, one could argue that this trivializes the problem of causal feature learning. If we are told what some agent cares about, then it seems that such an algorithm does not discover anything about the system under study, but instead regurgitates the interests of an arbitrary agent. In response, we argue that the dependence of our algorithm on an input utility function speaks to an important fact about the nature of scientific discovery. We maintain that the process of discovery, including the discovery of the salient possible macro-level states of a system from micro-level data, is a fundamentally goal-oriented process. A similar sentiment is echoed in work by Wellen and Danks (2016), who argue that the interaction between agential goals and the environment is essential to understanding feature learning. Another framework that uses a similar formal apparatus to discuss coarsening, but coarsens variables according to different equivalence relations, is the framework of "epsilon-machines for decisional states" put forward by Brodu (2011). It would be a potentially fruitful extension of this paper to explore the connections between our framework and these approaches. In addition, Beckers et al. (2019) propose a more complex formalization of the relationship between coarse-grained and fine-grained models than the quotient-set relation used by both Chalupka et al. and ourselves, but do not consider the pragmatic context in which coarsening occurs. It would be fruitful to consider how the approach proposed here contrasts with their model.

Appendix: Proof of Proposition 2

Proof. The proof closely follows the logic and methods of the first part of the proof of the causal coarsening theorem in Chalupka et al. (2017). We begin by introducing the following notation, for the purpose of concision: $i[i,l,j] = p(e_i|z_l,c_j)$, $\beta[j,l] = p(c_j|z_l)$, $\gamma[l] = p(z_l)$. Note that z_l is any value of a variable Z that is a potential confounder of the relationship between C and E, where R_Z has cardinality w. Next, we define the following three vectors: $\vec{i} = [i[1,1,1],\ldots,i[n,w,m]]$, $\vec{\beta} = [\beta[1,1],\ldots,i[m,w]]$, $\vec{\gamma} = [\gamma[1],\ldots,\gamma[w]]$. Each triple $(\vec{i},\vec{\beta},\vec{\gamma})$ is a point in a space $\mathbb{R}^d$. The set of all such points forms a simplex $\mathcal{S} \subseteq \mathbb{R}^d$. We want to show that the subset $\mathcal{S}' \subseteq \mathcal{S}$ containing all joint distributions such that there are two c_j and c_k such that $c_j \sim_{opc} c_k$ with respect to E but $c_j \not\sim_{pc} c_k$ with respect to E is Lebesgue measure zero in $P[\vec{i},\vec{\beta},\vec{\gamma}]$, where $P[\vec{i},\vec{\beta},\vec{\gamma}]$ is the space of all possible points $(\vec{i},\vec{\beta},\vec{\gamma})$. To do this, we first fix $\vec{i} = \vec{i}^*$ and $\vec{\beta} = \vec{\beta}^*$, so that $\vec{\gamma}$ is the only free parameter. Let $P[\vec{\gamma};\vec{i}^*,\vec{\beta}^*]$ be the set of all joint probability distributions consistent with this fixing of parameters. We proceed by showing that $\mathcal{S}' \cap P[\vec{\gamma};\vec{i}^*,\vec{\beta}^*]$ is Lebesgue measure zero in $P[\vec{\gamma};\vec{i}^*,\vec{\beta}^*]$, and then integrating over all possible $\vec{i}$ and $\vec{\beta}$. To show that $\mathcal{S}' \cap P[\vec{\gamma};\vec{i}^*,\vec{\beta}^*]$ is Lebesgue

measure zero in $P[\vec{\gamma}; \vec{i}^*, \vec{\beta}^*]$, pick any two values c_j and c_k such that $c_j \sim_{opc} c_k$ but $c_j \not\sim_{pc} c_k$. If no such pair exists, then we are done. If such a pair exists, then the fact that $c_j \sim_{opc} c_k$ means that the following constraint holds, for the fixed utility function $u(\cdot)$:

$$p(c_j)^{-1} \sum_{i=1}^{n} \sum_{l=1}^{w} u(c_j, e_i) i^*[i, l, j] \beta^*[j, l] \gamma[l] = p(c_k)^{-1} \sum_{i=1}^{n} \sum_{l=1}^{w} u(c_k, e_i) i^*[i, l, k] \beta^*[k, l] \gamma[l] \quad (2)$$

We note that $p(c_k) = \sum_{l=1}^{w} \beta^*[k, l] \gamma[l]$ and $p(c_j) = \sum_{l=1}^{w} \beta^*[j, l] \gamma[l]$, which means that, after some algebra, (2) implies:

$$\sum_{l'=1}^{w} \sum_{l=1}^{w} \gamma[l'] \gamma[l] \Big(\sum_{i=1}^{n} u(c_j, e_i) i^*[i, l, j] \beta^*[k, l'] \beta^*[j, l] - \sum_{i=1}^{n} u(c_k, e_i) i^*[i, l, k] \beta^*[j, l'] \beta^*[k, l] \Big) = 0 \quad (3)$$

Using an algebraic lemma from Okamoto (1973), we know that $\mathcal{S}' \cap P[\vec{\gamma}; \vec{i}^*, \vec{\beta}^*]$ is Lebesgue measure zero in $P[\vec{\gamma}; \vec{i}^*, \vec{\beta}^*]$ if this constraint is non-trivial, i.e. if there are $\vec{\gamma}$ such that (3) does not hold. To show that this is the case, suppose first that all entries in $\vec{\gamma}$ equal $1/w$. If (3) does not hold, then we are done. If (3) does hold under this condition, then there are at least two pairs of values (z_{q^+}, z_{v^+}) and (z_{q^-}, z_{v^-}) such that the following hold:

$$\sum_{i=1}^{n} u(c_j, e_i) i^*[i, q^+, j] \beta^*[k, v^+] \beta^*[j, q^+] - \sum_{i=1}^{n} u(c_k, e_i) i^*[i, v^+, k] \beta^*[j, q^+] \beta^*[k, v^+] > 0 \quad (4)$$

$$\sum_{i=1}^{n} u(c_j, e_i) i^*[i, q^-, j] \beta^*[k, v^-] \beta^*[j, q^-] - \sum_{i=1}^{n} u(c_k, e_i) i^*[i, v^-, k] \beta^*[j, q^-] \beta^*[k, v^-] < 0 \quad (5)$$

This assumes that there is a z_l such that (2) does not hold. Indeed, if (2) held, then it would not be the case that $c_j \sim_{opc} c_k$ with respect to E but $c_j \not\sim_{pc} c_k$ with respect to E (since the expected utility, given c_j or c_k, would be the same regardless of the value of the confounder, such that there would be no difference between the interventional and observational probability distribution over E, given either value) and the proof would already be complete. The two inequalities above imply that either $z_{q^+} \neq z_{q^-}$, $z_{v^+} \neq z_{v^-}$, or both. Assume $z_{q^+} \neq z_{q^-}$, and pick any positive $\epsilon < \min\{1/w, 1 - 1/w\}$. For any z_l such that $z_l \neq z_{q^+}$ and $z_l \neq z_{q^-}$, let $\gamma[l] = 1/w$, while $\gamma[q^+] = 1/w + \epsilon$ and $\gamma[q^-] = 1/w - \epsilon$. This way, $\sum_{l'=1}^{w} \sum_{l=1}^{w} \gamma[l'] \gamma[l]$ is unchanged from the case where each entry in $\vec{\gamma}$ is $1/w$, but it is nevertheless the case that (3) does not hold. We can repeat these same steps under the assumption that $z_{v^+} \neq z_{v^-}$, and under the assumption that $z_{q^+} \neq z_{q^-}$ and $z_{v^+} \neq z_{v^-}$, generating the same result in each case. Since (3) is non-trivial, we know that $\mathcal{S}' \cap P[\vec{\gamma}; \vec{i}^*, \vec{\beta}^*]$ is Lebesgue measure zero in $P[\vec{\gamma}; \vec{i}^*, \vec{\beta}^*]$. We integrate over $\vec{i}$ and $\vec{\beta}$ to show that $\mathcal{S}'$ is Lebesgue measure zero in $P[\vec{i}, \vec{\beta}, \vec{\gamma}]$. Let $\mathcal{S}' = \cup_{\vec{i}, \vec{\beta}} \tilde{P}[\vec{\gamma}; \vec{i}, \vec{\beta}] \subseteq P[\vec{i}, \vec{\beta}, \vec{\gamma}]$ be the Lebesgue measure zero set of all possible joint distributions $P[\vec{\gamma}; \vec{i}, \vec{\beta}]$ such that $c_j \sim_{opc} c_k$ with respect to E but $c_j \not\sim_{pc} c_k$ with respect to E. We define a characteristic function θ such that $\theta(\vec{i}, \vec{\beta}, \vec{\gamma}) = 1$ if $\gamma \in \tilde{P}[\vec{\gamma}; \vec{i}, \vec{\beta}]$ and $\theta(\vec{i}, \vec{\beta}, \vec{\gamma}) = 0$ otherwise. By the basic properties of positive measures, we have $\mu(S') = \int_{P[\vec{i}, \vec{\beta}, \vec{\gamma}]} \theta(\vec{i}, \vec{\beta}, \vec{\gamma}) \, \mathrm{d}\mu$. For concision, let $\mathcal{A} = \mathbb{R}^{w \times n}$, let $\mathcal{B} = \mathbb{R}^{n \times w}$, and let $\mathcal{G} = \mathbb{R}^w$. We calculate $\mu(S')$ as follows:

$$\begin{aligned} \mu(S') = \int_{\mathcal{A} \times \mathcal{B} \times \mathcal{G}} \theta(\vec{i}, \vec{\beta}, \vec{\gamma}) \, \mathrm{d}(\vec{i}, \vec{\beta}, \vec{\gamma}) &= \int_{\mathcal{A} \times \mathcal{B}} \int_{\mathcal{G}} \theta(\vec{i}, \vec{\beta}, \vec{\gamma}) \, \mathrm{d}(\vec{\gamma}) \, \mathrm{d}(\vec{i}, \vec{\beta}) \\ &= \int_{\mathcal{A} \times \mathcal{B}} \mu(\tilde{P}[\vec{\gamma}; \vec{i}, \vec{\beta}]) \, \mathrm{d}(\vec{i}, \vec{\beta}) = \int_{\mathcal{A} \times \mathcal{B}} 0 \, \mathrm{d}(\vec{i}, \vec{\beta}) = 0 \quad (6) \end{aligned}$$

Thus, the subset $\mathcal{S}' \subseteq \mathcal{S}$ containing all joint distributions such that $c_j \sim_{opc} c_k$ but $c_j \not\sim_{pc} c_k$ is Lebesgue measure zero in $P[\vec{i}, \vec{\beta}, \vec{\gamma}]$. □

References

S. Beckers and J. Y. Halpern. Abstracting causal models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2678–2685, 2019.

S. Beckers, F. Eberhardt, and J. Y. Halpern. Approximate causal abstraction. In *Poceedings of the 35th Conference on Uncertainty in Artificial Intelligence*, volume 2019, 2019.

N. Brodu. Reconstruction of epsilon-machines in predictive frameworks and decisional states. *Advances in Complex Systems*, 14(05):761–794, 2011.

K. Chalupka, P. Perona, and F. Eberhardt. Visual causal feature learning. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence*, 2015.

K. Chalupka, T. Bischoff, P. Perona, and F. Eberhardt. Unsupervised discovery of el nino using causal feature learning on microlevel climate data. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, pages 72–81, 2016a.

K. Chalupka, F. Eberhardt, and P. Perona. Multi-level cause-effect systems. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 361–369, 2016b.

K. Chalupka, F. Eberhardt, and P. Perona. Causal feature learning: an overview. *Behaviormetrika*, 44(1):137–164, 2017.

K. Fukunaga and L. Hostetler. Optimization of k nearest neighbor density estimates. *IEEE Transactions on Information Theory*, 19(3):320–326, 1973.

E. P. Hoel. When the map is better than the territory. *Entropy*, 19(5):188, 2017.

E. P. Hoel, L. Albantakis, and G. Tononi. Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences*, 110(49):19790–19795, 2013.

M. Okamoto. Distinctness of the eigenvalues of a quadratic form in a multivariate sample. *The Annals of Statistics*, pages 763–765, 1973.

J. Pearl. Probabilistic reasoning in intelligent systems: Networks of plausible inference. *Morgan Kauffmann Publishers Inc*, 1988.

P. K. Rubenstein, S. Weichwald, S. Bongers, J. M. Mooij, D. Janzing, M. Grosse-Wentrup, and B. Schölkopf. Causal consistency of structural equation models. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017.

J. von Neumann and O. Morgenstern. *Theory of games and economic behavior*. Princeton university press, 1944.

S. Wellen and D. Danks. Adaptively rational learning. *Minds and Machines*, 26(1-2):87–102, 2016.