JMLR Workshop and Conference Proceedings 6:1–38 NIPS 2008 workshop on causality

Causality: Objectives and Assessment

Isabelle Guyon ISABELLE@CLOPINET.COM
Clopinet, California, USA

Dominik Janzing DOMINIK.JANZING@TUEBINGEN.MPG.DE
Max Planck Institut für Biologische Kybernetik, Tübingen, Germany

Bernhard Schölkopf BERNHARD.SCHOELKOPF@TUEBINGEN.MPG.DE
Max Planck Institut für Biologische Kybernetik, Tübingen, Germany

Editor: Neil Lawrence

Abstract

The NIPS 2008 workshop on causality provided a forum for researchers from different horizons to share their view on causal modeling and address the difficult question of assessing causal models. There has been a vivid debate on properly separating the notion of causality from particular models such as graphical models, which have been dominating the field in the past few years. Part of the workshop was dedicated to discussing the results of a challenge, which offered a wide variety of applications of causal modeling. We have regrouped in these proceedings the best papers presented. Most lectures were videotaped or recorded. All information regarding the challenge and the lectures are found at http://www.clopinet.com/isabelle/Projects/NIPS2008/. This introduction provides a synthesis of the findings and a gentle introduction to causality topics, which are the object of active research.

Keywords: Causality, Bayesian Networks, Benchmark, Challenge, Competition, re-simulated data, probe method

1. Motivations

Machine learning has traditionally been focused on prediction: Given observations that have been generated by an unknown stochastic dependency, the goal is to infer a law that will be able to correctly predict future observations generated by the same dependency. Statistics, in contrast, has traditionally focused on "data modeling", *i.e.,* on the estimation of a probability law that has generated the data. During recent years, the boundaries between the two disciplines have become blurred and both communities have adopted methods from the other, however, it is probably fair to say that neither of them has yet fully embraced the field of causal modeling, *i.e.,* the detection of causal structure underlying the data. This has probably different reasons. Many statisticians would still shun away from developing and discussing formal methods for inferring causal structure, other than through experimentation, as they would traditionally think of such questions as being outside statistical science and internal to any science where statistics is applied. Researchers in machine learning, on the other hand, have too long focused on a limited set of problems, shying away from non i.i.d. data and problems of distribution shifts

between training and test set, neglecting the mechanisms underlying the generation of the data, including issues like stochastic dependence, and all too often neglecting statistical tools like hypothesis testing, which are crucial to current methods for causal discovery.

Since the Eighties there has been a community of researchers, mostly from statistics and philosophy, who, in spite of the pertaining views described above, have developed methods aiming at inferring causal relationships from observational data, building on the pioneering work of Glymour, Scheines, Spirtes, Pearl, and others. While this community has remained relatively small, it has recently been complemented by a number of researchers from machine learning. This introduces a new viewpoint to the issues at hand, as well as a new set of tools, including algorithms of causal feature selection, nonlinear methods for testing statistical dependencies using reproducing kernel Hilbert spaces, and methods derived from independent component analysis. Presently, there is a profusion of algorithms being proposed, mostly evaluated on toy problems or in application contexts where models cannot be falsified because of the lack of appropriate data. One of the main challenges in causal learning consists of developing strategies for an objective evaluation. This includes finding methods to acquire large representative data sets of both "observational" and "experimental" data. This, in turn, raises the question to what extent the regularities observed in these data sets provide sufficient evidence on unknown causal structures.

The two themes discussed at the NIPS 2008 workshop on causality reflect these concerns: (1) **Objectives:** Define **causal problems** *i.e.,* generic tasks involving causal modeling illustrated across various application domains. Formalize such tasks mathematically to clearly outline the objectives to be optimized. (2) **Assessment:** Devise reliable protocols of evaluation of solutions to causal problems. To address these objectives, we stated a program of data exchange and benchmarking: the "causality workbench" (Guyon et al., 2010). As part of the effort, we organized for NIPS 2008 a "pot-luck challenge" in which participants were invited to either contribute a solution to one of six proposed tasks or propose a new task.

This introduction is directed to researchers, students, and practitioners with no prior exposure to causality problems, but with some background in machine learning or data mining. It gently guides them through the maze of problems and techniques, without burdening them with mathematical notations and discusses the main outcomes of the workshop. A glossary is appended.

2. Contents overview

In these proceedings, we have gathered the contributions of researchers from a wide variety of horizons. Our collection of papers includes:

- a tutorial paper by one of the founders of the field, Judea Pearl, who revisits the problem of causal modeling with graphical models taking a counterfactual viewpoint,
- a thought provoking paper by Philip Dawid questioning the sanity of the "causal Bayesian network" methodology and proposing a different way of using graphical models for causal modeling, not necessarily interpreting arrows as causal relationships,
- an insightful paper by Lemeire and Steenhaut justifying some of the common model selection choices made in causal discovery using graphical models with the notion of Kolmogorov complexity,
- a novel vision of causal discovery as a game by Frederick Eberhardt,
- a machine learning approach by Haufe and collaborators to learning causal relationships from multivariate time series by enforcing model sparsity,

- two new algorithms by Mani and collaborators for discovering unconfounded causal relationships from observational data without assuming causal sufficiency (which precludes hidden common causes for the observed variables),
- a paper by Tillman and Spirtes analyzing under which conditions models using classical variable or feature selection methods may or may not outperform causal models, shedding light on the results of the causation and prediction challenge (WCCI 2008 (Guyon et al., 2008)).

The proceedings also include selected contributions to the NIPS 2008 "causality pot-luck challenge", proposing innovative solutions to:

- reverse engineering Boolean networks (the SIGNET task),
- finding local causal relationships around a target variable (the LOCANET task),
- finding all possible Markov boundaries, when there is a large number of possible solutions (the TIED task),
- learning a causal network from "heavy handed" manipulations affecting several variables simultaneously (the CYTO task),
- learning causal relationships among pairs of variables isolated from their context – therefore making impossible the use of conditional dependencies to unravel causal direction (the CauseEffectPairs task),
- quantifying the causal effect of promotions on sales (the PROMO task).

Kun Zhang and Aapo Hyvärinen received the best benchmark result award for their contribution to the CauseEffectPairs task (8/8 correct answers). The following authors received mentions: Ernest Mwebaze and John Quinn (for their work on the REGED dataset of the LOCANET task), and You Zhou, Changzhang Wang, Jianxin Yin, Zhi Geng (SIDO dataset, LOCANET task), Mehreen Saeed and the team of Cheng Zheng and Zhi Geng (SIGNET task), and Eugene Tuv (TIED task).

The tasks of the challenge and new proposed tasks contributed by the participants are summarized in Tables 1 and 2. The proceedings include papers describing these tasks, including the new contributions, which will be used in future challenges:

- learning causal relationships using time series when noise is corrupting data in a way that the classical Granger causality method may fail (the NOISE task),
- learning the structure of a fairly complex dynamic system that disobeys the equilibration-manipulation commutability, and predicting the effect of manipulations accurately when a manipulation does not cause an instability (the MIDS task),
- in a manufacturing process (wafer production), identifying measurements on the production line that allow engineers to detect early the pass/fail status at the end of the line (the SECOM task) or identifying faulty manufacturing steps affecting a performance metric (the SEFTY task).

The donor of the dataset NOISE (Guido Nolte) received the best dataset award. The reviewers appreciated that the task includes both real and artificial data and we want to encourage future data donors to move in this direction.

To facilitate the work of practitioners, we have also assembled a collection of "Fact Sheets" containing brief descriptions of the tasks of the challenge and their proposed solutions.

In the rest of this introduction, we develop the main problems addressed in the NIPS 2008 workshop on causality: "objectives" and "assessment". At the risk of missing important aspects, we focus on those concepts most related to machine learning. Section 3 reviews the various

Name (TP; NP; V)	Size	Description	Objective
CEP (Real; 5; 218)	P=8 pairs. N=2 variables.	**Cause Effect Pairs.** Pairs of real variables with known causal relationships.	Find the causal direction in all pairs.
CYTO (Real; 2; 394)	P$\simeq$800 samples per experimental condition $\times$ 9 conditions. N=11 proteins.	**Causal Protein-Signaling Networks in human T cells.** Protein activity monitored by flow cytometry. "Heavy-handed" manipulations are performed using chemical activators or inhibitors.	Learn the architecture of the protein signaling network.
LOCANET (Semi-artificial; 10; 558)	REGED & MARTI: P=500 patients; N=999 genes + target (disease). CINA: P=16033 persons; N=132 attributes + target (earnings). SIDO: P=12678 drugs; N=4932 descriptors + target (activity).	**LOcal CAusal NETwork.** Four datasets: REGED and MARTI (genomics), CINA (marketing), and SIDO (drug discovery). The datasets also include large test sets that were used in the "causation and prediction challenge" (Guyon et al., 2008).	Find the local causal structure around a given target variable (depth 3 network).
SECOM (Real; NA; 59)	P=1567 wafers. N=591 QC measurements + 1 binary target (pass/fail) and 1 date of processing	**Semiconductor manufacturing.** Production entities (wafers) are associated with quality control (QC) measurements on a fabrication line. The labels represent a pass/fail yield in line testing (classification problem).	Predict pass/fail in test data and identify predictive features.
TIED (Artificial; 1; 330)	P=750 training ex. N=1000 variables (including target).	**Target Information Equivalent Dataset.** A Bayesian network with 72 equivalent Markov blankets of the target variable.	Find all Markov blankets.

Table 1: **Atemporal datasets.** "TP" is the data type, "NP" the number of participants who returned results and "V" the number of views as of December 2008. The semi artificial datasets are generally "re-simulated" data, *i.e.,* data obtained from simulators of real tasks, usually trained with real data. Two datasets of LOCANET are made of real data augmented with artificial "probe" variables (SIDO and CINA). N is the number of variables and P is the number of examples (in training data; some datasets have test data too).

Name (TP; NP; V)	Size	Description	Objective
MIDS (Artificial; NA; 65)	T=12 sampled values in time (unevenly spaced); R=10000 simulations. N=9 variables.	**Mixed Dynamic Systems.** Simulated time-series based on linear Gaussian models with no latent common causes, but with multiple dynamic processes.	Use the training data to build a model able to predict the effects of manipulations on the system in test data.
NOISE (Real + artificial; NA; 43)	**Artificial:** T=6000 time points; R=1000 simulations; N=2 variables. **Real:** R=10 subjects. T$\simeq$200000 points sampled at 256Hz. N=19 channels.	**Real and simulated EEG data.** Learning causal relationships using time series when noise is corrupting data causing the classical Granger causality method to fail.	**Artificial task:** find the causal dir. in pairs of var. **Real task:** Find which region of the brain influences which other one.
PROMO (Semi-artificial; 3; 570)	T=365$\times$3 days; R=1 simulation; N=1000 promotions + 100 products.	**Simulated marketing task.** Daily values of 1000 promotions and 100 product sales for three years incorporating seasonal effects.	Predict a 1000$\times$100 boolean influence matrix, indicating for each (i,j) element whether the i^{th} promotion has a causal influence of the sales of the j^{th} product.
SEFTI (Semi-artificial; NA; 35)	R=4000 manufacturing lots; T=300 asynchronous operations (pair of values {one of N=25 tool IDs, date of processing}) + continuous target (circuit performance for each lot).	**Semiconductor manufacturing.** Each wafer undergoes 300 steps each involving one of 25 tools. A regression problem for quality control of end-of-line circuit performance.	Find the tools that are guilty of performance degradation and eventual interactions and influence of time.
SIGNET (Semi-artif.; 2; 415)	T=21 asynchronous state updates; R=300 pseudodynamic simulations; N=43 rules.	**Abscisic Acid Signaling Network.** Model inspired by a true biological signaling network.	Determine the set of 43 boolean rules that describe the network.

Table 2: **Time dependent datasets.** "TP" is the data type, "NP" the number of participants who returned results and "V" the number of views as of December 2008. The semi-artificial datasets are obtained from simulators of real tasks. N is the number of variables, T is the number of time samples (not necessarily evenly spaced) and R the number of simulations with different initial states or conditions.

settings of causal modeling. Section 4 identifies **objectives** for causal modeling and indicates the role that machine learning may play in pursuing such objectives. Section 5 gives a brief overview of **assessment** methods. Finally, in a discussion section (Section 6) we provide a perspective on challenges being faced, success stories, and open problems.

3. Causal systems *vs.* causal models

A proper definition for causality that regroups all the notions it encompasses in philosophy, psychology, history, law, religion, statistics, physics, and engineering has eluded scientists and philosophers for centuries. However, to avoid accusations of circularity, we give in this section tentative definitions, which, although not universally accepted, are useful to pursue machine learning objectives.

3.1 Causal systems

In the branch of causal studies closest to engineering, the notion of causality is intimately related to the idea that there exist self-contained systems, which have a number of input variables and output variables. Given values of the input variables (set by an external agent), there is a mechanism (a function), which determines the values of the output variables, eventually up to some uncontrollable "stochastic noise". In a certain sense, the values assumed by the input variables cause those of the output variables. There is an intrinsic asymmetry: inversely, if the external agent would force the output variables to assume given values, one would not expect the input variables to be influenced. Take the example of TV remote controllers: you can press a button and turn on or off the TV, but turning on or off the TV does not affect the buttons of the remote controller.

A wide variety of physical systems under equilibrium do not fall into that category. For instance, a perfect gas governed by the law $pV = nRT$, which states that the product of pressure p and volume V is proportional to the temperature T, would not constitute a "causal system" in the sense described above since any change in two of the variables $\{p, V, T\}$ results in a change in the third one. The role of the three variables p, V and T seems completely symmetrical. Even though there is much to say about the causal interpretation of particular systems subject to the law of perfect gases, we shy away from such controversial cases and limit ourselves to systems in which there is a consensus on their causal interpretation. For instance, there can hardly be any disagreement that if we record the altitude of given villages and their average yearly temperature, if there is a cause-effect relationship, it ought to be altitude that causes temperature and not the opposite.

In many applications, it is useful to broaden the notion of causality to a set of interrelated variables, not necessarily assuming either a role of input or output variable. It becomes then more difficult to define causality and determine to what extent we can say that a variable "causes" another variable. An "operational criterion of causality" (Glymour and Cooper, 1999) is sometimes adopted: consider a system characterized by a set of interdependent random variables (RV) generated by a "natural" stationary distribution, some of which corresponding to directly actionable variables (their values can be set by means of action or manipulation performed by an agent external to the system rather than drawn from the "natural" distribution). **A random variable *C* may be called a cause of another RV *E*, called its effect or consequence, if actions performed on *C* by an external agent result in changes in the distribution of *E*.** For instance, the variable *C*=*smoking* and *E*=*lung cancer* may have given "natural" distributions in a given population. Banning smoking (at least in some places) is an action that may be taken by an external agent (*e.g.,* the Surgeon General). Changes in lung cancer incidence as a

result of this action would indicate a causal link between smoking and lung cancer, according to this criterion. This operational criterion of causality provides a sufficient condition for C to be called a cause of E, but not a necessary condition, hence it cannot serve as a definition: An absence of change in the distribution of E under manipulation of C does not exclude that C is a cause of E. For instance, consider the outcome of tossing two fair coins C_1 and C_2 and the variable E that is positive if both coins fall on the same side and negative otherwise. Performing the action of forcing C_1 to be constantly on the "face" side does not change the distribution of E even though C_1 is a cause of E (in the sense that, in the unmanipulated system, E is determined both by C_1 and C_2). To broaden the notion of causality, we give a definition of causal relevance of a variable C to a target E, in the context of other variables (Guyon et al., 2007). For other definitions, see also (Glymour and Cooper, 1999; Pearl, 2000; Spirtes et al., 2000; Neapolitan, 2003; Koller and Friedman, 2009).

The notion of causality between RVs allows us to make simple connections to machine learning and to feature selection applications in which data are often represented as random vectors. It implicitly makes the assumption that similar events repeat themselves and statistics can be computed, hence it does not encompass single event causality (like legal responsibility in a crime). There are alternative ways of thinking of causality as relationships between objects, events or system states, which we do not cover in this introduction.

Our everyday-life concept of causality is very much linked to time dependencies (the causes precede their effects). However, many machine learning problem are concerned with "cross-sectional studies", which are studies where many samples are drawn at a given point in time. Thus, sometimes the reference to time is replaced by the notion of "causal ordering". Causal ordering can be understood as fixing a particular time scale and considering only causes happening at time t and effects happening at time $t + \Delta t$, where Δt can be made as small as we want. But, we will also consider applications in which time dependencies are critical (for instance to continuously monitor treatment in a changing environment) corresponding to problems encountered in so-called "longitudinal studies".

From the point of view described in this section, a "causal system" is characterized by a set of variables, including at least some observable and some directly actionable variables, and a set of permitted actions or manipulations, which may be performed by an external agent to evidence causal relationships between these variables. With some abuse of language we refer to such variables as "random variables" to indicate that they are governed by a "natural" probability distribution when the system is left to evolve according to its own dynamics, and that causal conclusions will be drawn from samples and have only a statistical validity (like "price" influences "sales" or "age" influences "health"). Throughout this introduction, we often use a population of patients under the care of a physician as an example of a causal system. Variables of interest include socio-economic factors, environmental factors, clinical variables, etc. and the physician plays the role of an external agent administering treatments (thought of as actions or manipulations). We put forward this setting for concreteness, but acknowledge that requiring a separation between an inside and an outside of the system and the notion of external agent and manipulations is the object of much debate. In particular, causality is sometimes defined in terms of **counterfactuals** (see glossary): "C causes E" means that "had C not occurred, E would not have taken place". However, because we cannot rewind history and replay events after making small controlled changes, causation can only be inferred, never exactly known. In that sense, it can be understood that the role of "external agents" performing scientific experiments and of statisticians analyzing observations is to approximate as well as possible counterfactuals.

3.2 Causal models

A long time debate in machine learning has been whether predictive models should or not model the data's generative process. Years of research and the results of recent benchmarks (Clopinet, 2009) seemed to have settled the question: there is no need to be concerned with the data's generative process; "agnostic" predictive models, in the vein of neural networks, decision trees and kernel methods, perform as well or better than generative models, at least for data-mining style tasks for which data are i.i.d. But one should be careful not to jump too quickly to conclusions: might the situation change when we switch from making predictions in a stationary environment (the i.i.d. case) to predicting the consequences of actions?

Assume that we have a system of only two random variables X and Y. In a stationary i.i.d. setting, all that is needed to make predictions is the joint distribution $P(X,Y)$, which does not inform us on whether X was generated from Y or vice versa. However, if actions are being performed, it is useful to know how data were generated. Assume that X is generated first according to $P(X)$ (say X is the atmospheric temperature) and then Y according to $P(Y|X)$ (say Y is the position of the needle of a thermometer). Then, if we force X to assume a given value (by a manipulation like by making a big bonfire), we expect a certain change in Y. Conversely, if we force the thermometer needle position, we do not expect this should have an impact on temperature. The effect of interventions on the joint distribution cannot be predicted by the Bayes formula $P(X,Y) = P(X|Y)P(Y) = P(Y|X)P(X)$. In particular, borrowing Pearl's notations (Pearl, 2000), $P(X|do(Y = y)))$ may be different from $P(X|Y = y)$, where $do(Y = y)$ means that Y has been forced to take the value y (by an external agent), while $Y = y$ means that Y has been observed to have the value y. In the case of the temperature example, we have $P(Y|do(X = x)) = P(Y|X = x)$ (observing a given temperature or forcing it artificially to attain the same value results in the same thermometer reading), but we have $P(X|do(Y = y)) \neq P(X|Y = y)$. In fact, $P(X|Y = y)$ obeys the Bayes formula $P(X|Y = y) = P(Y = y|X)P(X)/P(Y = y)$, but $P(X|do(Y = y))$ does not: $P(X|do(Y = y)) = P(X)$ (temperature does not change as a result of forcing the needle position).

From the above considerations, we can conclude that **some knowledge of how the data were generated should be useful** to build predictive models, **if predictions of the consequences of actions are to be made**. In our example, it is useful to choose between two alternative generative models: X generated first according to $P(X)$, then Y generated according to $P(Y|X)$; or, Y generated first according to $P(Y)$, then X generated according to $P(X|Y)$. Importantly, $P(X|Y)P(Y)$ is not the same as $P(Y|X)P(X)$, if the "do" operator is inserted. However, this does not mean that the data's generative process should be modeled faithfully to obtain best prediction performances. As always in machine learning, overfitting must be avoided when modeling data and **the best predictive model does not necessarily belong to the class of systems that generated the data**, owing to the celebrated bias-variance tradeoff (Geman et al., 1992). Therefore, what may appear at first sight to be over-simplifying assumptions (some of which are discussed in Section 6) may turn out to reduce the variance of the model class so effectively that, even though some bias is introduced, good performance is attained.

Before moving forward, we want for concreteness to give some examples of causal models. The use of graphical models in causality has a long history that can be traced back to "path analysis" (Wright, 1921), "structural equations" (Haavelmo, 1943), and modern graphical models that can have a causal interpretation (Spiegelhalter et al., 1993; Glymour and Cooper, 1999; Pearl, 2000; Spirtes et al., 2000; Neapolitan, 2003; Koller and Friedman, 2009). Many other types of models have been used to model causal relationships, including artificial neural networks, Boolean networks, and various types of Markov models, including hidden Markov models (HMM), partially observable Markov decision processes (POMDP). The type of causal

relationships under consideration have often been modeled as **Bayesian causal networks** or **structural equation models** (SEM) (Pearl, 2000; Spirtes et al., 2000; Neapolitan, 2003). In the graphical representation of such models, an arrow between two variables $A \rightarrow B$ indicates the direction of a causal relationship: A causes B. A node of the graph, labeled with a particular variable X, represents a mechanism to generate the value of X given the parent node variable values. For Bayesian networks, such evaluation is carried out by a conditional probability distribution $P(X|\texttt{Parents}(X))$ while for structural equation models it is carried out by a function of the parent variables, eventually distorted by stochastic noise (often but not necessarily additive noise). Learning a causal graph can be thought of as a model selection problem: Alternative graph architectures are considered and a selection is performed, either by ranking the architectures with a global score (*e.g.,* a marginal likelihood, or a penalty-based cost function), or by retaining only graphs that fulfill a number of constraints, such as dependencies or independencies between subsets of variables. Such graphical models usually make at least two simplifying assumption: the **causal Markov condition** (CMC) and the **causal faithfulness condition** (CFC), both of which are discussed in more details is Section 6.

The task of training and selecting causal models is significantly harder than that of training and selecting regular predictive models (classical machine learning from i.i.d. data). The main hurdle in classical machine learning is generally the lack of training data: In most practical applications, with a sufficient amount of training data, the true data distribution may be approached with arbitrary precision, then the problem is "solved". In the jargon of causal modeling, the data commonly used in machine learning are called **observational data**; those are data collected from systems, which are let to evolve according to their own dynamics, without external intervention. Cross-validation is highly effective to perform model selection in this setting.

In contrast, causal models can often not be effectively trained with only "observational data" and cross-validation is ineffective to perform causal model selection, because many models with entirely different causal architectures may perform equally well in an observational setting. It is still debated what the most effective causal model selection strategy should be, but many penalty-based cost functions privileging simple models or stable models have been proposed (Koller and Friedman, 2009). In addition, training and selecting causal models often require data collected after **external interventions** (also referred to as actions, manipulations, or experiments). Such **experimental data** can better distinguish between mere statistical dependence (due for instance to an unknown common cause, referred to as **confounding variable** or **confounder**) and true causation. A widely recognized methodology of unraveling causal relationships or validating causal assumptions is **randomized controlled trial** (RCT). RCTs are most often used for conducting planned experiments in healthcare, but are also employed in other areas of application including judicial, educational, and social research. RCTs involve the random allocation of different interventions (treatments or conditions) to subjects. As long as numbers of subjects are sufficient, this ensures that both known and unknown confounding factors are evenly distributed between treatment groups. Methods for learning cause-effect links without experimentation (learning from observational data) are attractive because observational data is often available in abundance and experimentation may be costly, unethical, impractical, or even plain impossible (London and Kadane, 2002). Still, many causal relationships cannot be ascertained without the recourse of experimentation and the use of a mix of observational and experimental data might be the most cost effective.

4. Objectives of causal modeling

One of the central topics of the NIPS 2008 workshop was to define objectives for causal modeling. In the previous section, we have tentatively defined causal systems and introduced causal models, not as data generative models, but as tools to predict the consequences of action. Predicting the consequences of actions is often considered to be the main charter of causal modeling. We now review a number of other related causal problems worth pursuing and then put them in the context of applications.

4.1 Causal problems

We collectively call "causal problems" problems requiring the notion of causality. We contrast such problems with machine learning applications using i.i.d. training and test data. In the i.i.d. setting, variables predictive of the target, regardless of causal relationships, may be useful. For instance, in medical diagnosis, the abundance of a protein in serum may be used as a predictor of disease. It is not relevant to know whether the protein is a cause of the disease (*e.g.,* resulting from a gene mutation), or a consequence (*e.g.,* an antibody responding to inflammation). If one is interested in a diagnosis, the abundance of this protein is enough and means disease. We differentiate the problem of making predictions in a stationary environment (diagnosis) with two other types of predictions: the prediction of the consequences of actions performed deliberately by and external agent and counterfactual predictions:

Prediction of the consequences of actions: More and more applications require the assessment of the results of given actions (also referred to as "manipulations" or "experiments"), performed by agents external to the system, thus disturbing the natural functioning of the system. Such assessment is essential in many domains, including epidemiology, medicine, ecology, economy, sociology and business, to assist the development of new treatments and new policies. Assessing the consequences of actions is radically different from making predictions in a stationary environment when the system is subject to its own dynamics. For instance, one might observe that both smoking and coughing are predictive of respiratory disease in a general population and use either predictor for diagnosis. One is a cause (smoking) and the other a symptom (coughing). Acting on the cause can change the disease state, but not acting on the symptom. Therefore if we are interested in treatment rather than in diagnosis it is extremely important to distinguish between causes and symptoms to predict the consequences of actions.

Counterfactual prediction: Another landmark of causal reasoning is **counterfactual** prediction. In fact, some philosophers and practitioners like defining causality via counterfactuals. A typical counterfactual question is: considering that a given patient who took a nicotine substitute stopped smoking, what would have happened if he had not take the medicine? Would he have stopped smoking anyway? More generally, considering a self-contained system of interdependent RVs, what would have been the values assumed by certain variables had some other variables taken values different from the ones observed?

We see that there are subtle differences between predicting the consequence of actions and counterfactuals. First, counterfactuals have to do with hypothetical events that could have taken place in the past whereas predicting the consequences of actions projects events into the future. Second, counterfactual predictions are usually point-wise predictions. For instance, we want to predict what would have happened to one particular patient. In contrast, we might want to optimize the consequences of future actions on a population of patients.

There are many other causal questions. Here we mention a few, which were raised at the NIPS 2006 workshop on causality:

- Determine what manipulations are needed to reach a desired system state with maximum probability (e.g., select variables and propose values to achieve a certain value of a response/target variable, with perhaps a cost per variable).
- Find a causal explanation for a certain observed state y of a target variable Y, *i.e.,* a set of variables having assumed given values, which lead with high probability to the given observation $Y = y$.
- Propose system queries to acquire training data, *i.e.,* design experiments, with perhaps an associated cost per variable and per sample and perhaps with constraints on variables, which cannot be controllable.
- Determine a local causal region around a response/target variable (causal adjacency).
- Determine the source cause(s) for a response/target variable.
- Predict the existence of unmeasured variables (not part of the set of variables provided in the data), which are potential confounders (are common causes of an observed variable and the target).
- Predict which variables called "relevant" by feature selection algorithms are potentially causally irrelevant because their statistical dependency to the target is the result of an experimental artifact (e.g. sampling bias or systematic error).
- Determine causal direction in time series data in which one variable is causing the other.

Defining causal problems supersedes the need for defining causal systems, if we think of causality as a means to an end (solving problems, attaining objectives). We do not need to ascertain that data are generated by a causal system to address a causal problem or answer a causal question. Let us go back to our example of the perfect gas for which the system of variables $\{p,V,T\}$ did not seem to be in any particular causal relationship. If we use a bicycle pump, the action of pumping has predictable consequences linking the reduction of volume of the gas in the pump to the increase in pressure. Hence, via action/manipulation/experimentation we can evidence a cause-effect relationship for this particular setup, without the prerequisite of solving the problem of whether the set of RVs involved form a "causal" system (this question might not even make sense). What matters to us, in this case, is that we can predict the consequence of actions, *i.e.,* we can use the bicycle pump for a purpose.

4.2 The role of machine learning

The main charter of Machine Learning is learning from data the structure and parameters of an optimal predictive model. We refer to this task as **model inference**. Once a model structure and its parameters are computed, another kind of inference can take place: the inference of variable statistics (point estimations, estimation of expectations, or distribution calculation) given the values of other variables. We refer to this other problem as **variable inference** to distinguish it from the first one. When authors refer to causal inference, they may either refer to variable inference, model inference, or both. We briefly review both aspects to contrast them.

VARIABLE INFERENCE

A lot of effort has been put into solving the problem of variable inference in Bayesian networks (BN) and Structural Equation Models (SEMs), independently of solving the problem of model inference. In many applications, the structure of a causal models is derived from prior knowledge. For instance, in the PROMO task of the challenge, the model structure is given by expect knowledge ("promotions" influence "sales"); only the parameters need to be estimated from data. In some applications, the parameters themselves cannot be subject to learning because of

lack of training data, but they can be derived from expert knowledge. For example, the methodology of the noisy-or model, which has been widely deployed for medical diagnosis (Russell and P., 2003) and fault diagnosis (Yongli et al., 2006), allows mapping expert knowledge to parameters. It makes simple independence assumptions between direct causes X_i, $i = 1,...n$ of a target Y. The influence of the X_i on Y is parameterized by only n parameters p_i, easy and intuitive to evaluate for experts. Using n intermediary influence variables Y_i such that Y is the simple logical OR of the Y_i, the parameters p_i represent the probabilities of successful influence: $P(Y_i = 1|X_i = 1) = p_i$ and $P(Y_i = 1|X_i = 0) = 0$. The models thus constructed are used for variable inference.

Variable inference makes use of the model to predict values of certain variables in various situations, including when values of some other variables are missing or imposed (manipulations or counterfactuals). Here is a typical example of variable inference in a simplified "alarm network" (Pearl, 1988): $Burglary \rightarrow Alarm \leftarrow Earthquake$. Assume that the alarm goes off and alerts the police by telephone. The question is: has there been a burglary. If there has just been an earthquake, the probability of a burglary goes down and it may be unnecessary to send a police officer. Calculation of conditional probabilities (such as $P(Burglary|Alarm, Earthquake)$) can be facilitated by causal networks in complex cases involving a large number of variables. Such uses of causal networks inherit directly from expert systems in artificial intelligence, adding the additional "uncertainty" dimension to the logical constructs. Bayesian networks or SEMs with designated architecture and parameters can be thought of as **motors of calculation of conditional probabilities**. Going one step beyond, in some cases, it is possible to predict conditional probabilities in a **post-manipulation distribution** given the pre-manipulation distribution (the so-called "natural" distribution) and some causal assumptions. For instance, one might want to compute $P(Burglary|do(Alarm), Earthquake)$, where $do(Alarm)$ means that the alarm is triggered by an external agent. The action of the agent disconnects the $Alarm$ variable from its original causes $Burglary$ and $Earthquake$, hence $P(Burglary|do(Alarm), Earthquake) = P(Burglary)$. A complete methodology to carry out such variable inference problems using causal networks (implemented with BNs or SEMs) has been developed by Pearl and his collaborators under the name of "do-calculus" (Pearl, 2000).

MODEL INFERENCE

While variable inference is an important aspect of causal inference with a well developed set of algorithms, model inference has recently become the focus of interest. In that realm, machine learning has various important roles to play:

- **The finite sample case.** Traditionally in the causal discovery community, algorithms for learning causal network structure have been developed with the assumption that there exists an "oracle" having perfect knowledge of the data distribution, and which is capable of answering without mistake questions about **conditional independence** between subsets of variables. This implicitly make the assumption that an infinite amount of training data are available. This raises the questions of developing robust and powerful statistical tests of conditional independence (Margaritis and Thrun, 2001). Kernel methods have moved in this direction (Gretton et al., 2005).
- **Feature and model selection.** Another tradition of the causal discovery community is to dismiss cross-validation for model selection and focus on penalty-based cost functions (most often using Bayesian priors) for reasons alluded to in Section 3.2. Yet, as demonstrated in the "causation and prediction" challenge, regular feature selection methods and cross-validation can take you very far to prune feature space (Guyon et al., 2008). Also, purely frequentist penalty-based model selection methods based on regularization, which

have been developed in machine learning, may provide effective means of causal model selection (Pellet and Elisseeff, 2008; Lozano et al., 2009). In the problem of cause-effect pairs for instance, where constraint-based methods using conditional independence tests are not applicable, such methods have proved to be effective (see the papers of Mooij and Janzing and that of Zhang and Hyvärinen in these proceedings).

- **Learning algorithms.** There is a wealth of algorithms developed in machine learning, which can find applications in learning causal models. We saw recently the application of the ICA algorithm to learning SEMs with non-Gaussian noise for linear models (Shimizu et al., 2006), extended in these proceedings to non-linear models by Zhang and Hyvärinen. Recent methods also use non-linear regression techniques to distinguish between cause and effect (Hoyer et al., 2008). Novel methods for identifying latent confounders use a combination of nonlinear dimensionality reduction and kernel dependence measures (Janzing et al., 2009).

4.3 Examples of applications

Recently, there has been a surge of interest in causal models in data mining, prompted by the need of assisting policy making and the availability of massive amounts of "observational data". Examples of applications of causal models include: biology, medicine and pharmacology (Oniśko et al., 1997; Herskovits and Dagher, 1997; Friedman et al., 2000; Kononenko, 2009), epidemiology (Aickin, 2002), climatology (Chu and Glymour, 2008), social and economic sciences (Kaplan, 2000; Demiralp and Hoover, 2003; Moneta, 2005), marketing (CFMDCY, 2006), neuroscience (Ding et al., 2006; Neves et al., 2008), psychology, law enforcement and crime prevention (Young, 2008), manufacturing,, quality control, and fault or security diagnosis (Qin and Lee, 2003; Kraaijeveld and Druzdzel, 2005). Among the most prominent applications, which have taken off in the past decade, uncovering regulatory networks of chemicals in living organisms and connecting those networks to disease, has been the object of much research. For a rather extensive bibliography, see (Markowetz, 2007). Epidemiology has long been one of the main areas of application of causal modeling (Rubin, 1974; Herskovits and Dagher, 1997; J.M. Robins, 2000). Epidemiologists have also embraced the new tools of genomics and proteomics to investigate gene-environment interactions (Vinei and Kriebel, 2006; Jenab et al., 2009).

5. Assessment of causal solutions

A second objective of the NIPS 2008 workshop was to find means of assessing the performances of solutions proposed to causal problems. We present in this section assessment methods, which have been used in our challenges, and point to other methods of interest.

5.1 Experimental verifications

The most established way of assessing causal theories is to carry out randomized controlled experiments to test hypothetical causal relationships. Fisher's book "The Design of Experiments" in 1935 laid the mathematical foundations for experimental design. The central idea is the systematic use of randomization to avoid confounding.

For example, in the medical domain, a causal relationships $C \rightarrow E$ between a treatment C and an effect E may be tested in a **Randomized Controlled Trial** (RCT). Variable C may be the choice of one of two available treatments for a patient with lung cancer and E may represent 5-year survival. If we randomly assign a large number of patients to the two treatments by flipping

a fair coin and observe that the probability distribution for 5-year survival differs between the two treatment groups, it may be concluded that the choice of treatment causally determines survival in patients with lung cancer. The double blind placebo-controlled Randomized Controlled Trial, where allocations are randomized and neither patient not doctor knows which treatment has been assigned, is now standard in clinical trials. In agriculture, complex experiments in which many factors are controlled simultaneously are commonly performed. Unfortunately, experimenting is a long and costly process, and, in many domains it is impractical or infeasible.

An ideal benchmark of causal discovery methods (uncovering causal relationships from observational data) would compare predictions obtained by applying algorithms to large observational databases with the outcome of well designed experimental studies. Because of the rarity of adequate observational data sets paired with appropriate randomized experiments, to our knowledge no such comparisons have been made.

The Causality Workbench project has started a program of benchmarks in which realistic simulated systems will be used for generating observational data and performing virtual experiments (Guyon et al., 2010). In the "causation and prediction challenge" (Guyon et al., 2008), we used matched sets of artificially generated data for various tasks: a training dataset drawn from a "natural" **unmanipulated distribution** and several test sets drawn from various types of **post-manipulation distributions**.

We present alternative evaluation methods in the following sections.

5.2 Established ground truth

Second best to pairing observational studies and the outcome of designed experiments is to compare causal relationships inferred from observational data to **ground truth** established from human expertise (see glossary). This method has been used for instance by Cooper and Spirtes, 1998 (Spirtes et al., 2000, page 369) to compare cause-effect relationships inferred from a database on hospitalized pneumonia patients to expert medical judgement. Here are a few examples of cause-effect pairs tested in this study: *Coronary artery disease* $\rightarrow$ *Myocardial infection*, *Employment status* $\rightarrow$ *Illegal drug abuse*, *Nausea* $\rightarrow$ *Vomiting*, and *Number of comorbid conditions* $\rightarrow$ *Dire outcome*. In the pot-luck challenge organized for NIPS 2008, one dataset used human judgement as ground truth: the CauseEffectPairs dataset. Examples include the pairs *Altitude* $\rightarrow$ *Temperature* and *Longitude* $\rightarrow$ *Precipitation* in German cities and *Age* $\rightarrow$ *Length* for the snail Abalone.

In biology, regulatory pathways obtained by curating thousands of peer reviewed papers constitute reference human knowledge for discovery studies performed with genomic and proteomic observational data (Kanehisa et al., 2008). In the pot-luck challenge, the CYTO dataset is a good example using this type of ground truth. Note, however that due to many inconsistencies in the biological literature there is a lot of uncertainty in the reference regulatory pathways.

Using artificially generated data is another way of having access to an established ground truth (*i.e.,* the structure of the data generative model). In the NIPS 2008 challenge, several datasets resorted to this means of assessment. The dataset TIED is purely artificial and was designed to illustrate a particular technical difficulty. The datasets REGED and MARTI were build from a simulator of a gene regulatory network influencing lung cancer, trained with real data. The dataset SIGNET was simulated from a set of Boolean rules representing knowledge of a plant regulatory pathway gathered from several published papers.

5.3 Statistical tests

We regroup in this section a variety of techniques making solely use of observational data to *validate causal structures* using some statistical argument. We think of such methods as the

weakest way of validating causal relationships, yet they are much useful because there are often no better alternatives.

1. **Validation of theoretical models by hypothesis testing.** Statistical hypothesis testing is used as "confirmatory analysis" (not for structure discovery via tests of conditional independence) in social sciences, psychology, and econometrics to validate theoretical models proposed by experts. The parameters of a causal model (typically a SEM) whose structure is determined from domain knowledge, are fitted to data. In ordinary least square regression (with several input features that represent alleged causes and a single target variable), the residuals of the model are compared to the residuals of a null model (*e.g.,* the expected value of the target, another previously proposed model, or, for time series, an auto-regressive model). Statistical tests used to perform such comparisons include the Chi-square test. The tested model is invalidated if its predictions cannot be found statistically significantly better than those of the null model. Individual parameters of the model can also be examined within the estimated model in order to see how well the proposed model fits the driving theory.

   For structural equation models (SEMs) assuming Gaussian noise models, the parameter calculations are based on the covariance matrix of the variables. Goodness-of-fit is based on comparing the observed covariance matrix with the covariance matrix estimated by the model. In the early literature on SEMs, analysts tested simply the null hypothesis that the specified model leads to an exact reproduction of the observed covariance matrix with a chi-square test, but this was later replaced by a comparison with the predictions of a null model (*e.g.,* a baseline model assuming that all variables are uncorrelated) (Bollen and Long, 1992). Recently, methods for testing structural parts of a model rather than the whole model have been proposed, providing a more detailed and insightful validation (Tsamardinos and Brown, 2008).

   Another type of test investigates whether the explanatory variables and the error terms are statistically independent, as recently used in (Shimizu et al., 2006; Hoyer et al., 2008), and by Kun Zhang and Aapo Hyvärinen in these proceedings. Since these dependencies are typically non-linear, tests must be able detect higher-order dependencies, not just simple correlations, . Kernel-based methods like HSIC (Gretton et al., 2005) seem to be useful for this task.

   It is important to remember that if such methods are to be used for structure validation, the structure of the tested model should not be obtained from the data used for testing (otherwise it is like testing on training data).[1] Also, a model passing such a test is not confirmed, but rather it is not rejected, because the evidence obtained from observational data is usually insufficient to confirm a causal model. The tested model should have falsifiable implications, which can be tested against the data.

2. **Instrumental variables.** In econometrics, epidemiology and related disciplines, the method of instrumental variables is used to estimate causal relationships when controlled experiments are not feasible. In attempting to estimate the causal effect of some variable C on another E, an instrument is a third variable I which affects E only through I's effect on C: $I \rightarrow C \rightarrow E$. The method can be thought of as a "natural" experiment in

1. This section focusses on model assessment or "validation" (testing), not on model selection, which we consider part of training. Statistical tests are also used sometimes for model selection. For instance, nested models with increasing numbers of variables may be created and pvalues may be computed. This can be understood as testing a model not only against a single model, but against all simpler models. Pvalues must be adjusted correctly to take into account the multiple testing problem.

which the instrument variables play the role of the "external agent". The success of the method hinges on the selection of suitable instruments. For instance, Cooper and Spirtes, 1998 (Spirtes et al., 2000, page 372) used *race*, *age*, and *gender* as an instruments in the determination of cause-effect pairs in the example of pneumonia covariates mentioned in the previous section. In Section 6.2, we give examples of Mendelian randomization in which naturally occurring *gene mutations* are used as instruments to manipulate the level of certain proteins in blood.

Other natural and quasi-natural experiments of various types are commonly exploited, for example (Miguel et al., 2004) use weather shocks to identify the effect of civil conflict on economic growth. Jared Diamond (Diamond, 1997) defends the thesis of the influence of climate and natural resources on societal development (including food production *vs.* hunting and gathering) using a natural controlled experiment: the scattering of populations of homogeneous ancestry over a relatively short period of time in the widely diverse Polynesian islands.

3. **Re-simulation and model architecture stability.** The consistency of the findings obtained by causal discovery algorithms on real data may also be tested by "re-simulation". The re-simulation method consists in: (1) Training a data generative model with real observational data; (2) Generating simulated datasets with the model under various noise conditions; (3) Training new models for every the simulated dataset; (4) Studying the model stability with respect to its architectures and its predictions made under manipulation. This methodology was used by Statnikov and collaborators (Aliferis et al., 2006) on the problem of lung cancer. The REGED dataset used in our challenges emerged from this study, but re-simulation was not used as an assessment method is the challenge.

   Re-simulation is a variant of an assessment methods often used for clustering algorithms in which the stability of the model under various perturbations of the data is studied (Ben-Hur et al., 2002). Perturbations may include resampling the training dataset or adding noise to the input variables. Clustering and other unsupervised learning methods including principal component analysis and factor analysis can be thought of as latent causal constructs (the latent variables or cluster centers being alleged hidden causes).

4. **Probe method.** Yet another type of method of assessment, very popular in the field of variable or feature selection, is to introduce in real data a number of artificial "distracter" variables called "contrasts" (Tuv et al., 2006) or "probes" (Stoppiglia et al., 2003; Guyon and Dreyfus, 2006), which are, by construction, not predictive of a target variable of interest. In the first causality challenge (Guyon et al., 2008; Guyon et al., 2008), we extended this method to the assessment of causal discovery algorithms.

   The use of probes is relatively straightforward for "regular" feature selection from i.i.d. data, with the goal of selecting predictive variables of a given target variable, regardless of causal relationships. In statistics, for algorithms providing a ranking of variables in order of relevance, it is standard to compare the index of ranked variables to the index of hypothetical variables (called probes) drawn from a null distribution representing irrelevant variables (Guyon and Dreyfus, 2006). In this way, one can test the null hypothesis that variables are irrelevant. For instance, assume that our target variable is binary (*e.g.,* the patient health status "cancer" or "healthy") and that we want to determine whether a given predictor variable of mean μ is individually predictive of the target (univariate association). A possible null hypothesis may be that variables are drawn from a Gaussian distribution of mean μ and the alternative hypothesis may be that it is drawn from a mixture model of two Gaussians with different means (but same variance). The

t-test may then be used to test the hypothesis of *equality of the means of the two classes* and determining *whether the predictor variable of interest significantly separates the two classes*. Choosing the right ranking criterion and a good null distribution has been the object of a lot of study and there is no one-size-fit all solution (see Guyon and Dreyfus, 2006, for a review). A completely non-parametric solution to the problem is to select a well suited ranking criterion, not corresponding to any known tabulated statistic (*e.g.*, the Relief criterion Kira and Rendell, 1992), then to generate random "probes" by permuting the values of randomly chosen real variables. In this way, the marginal distribution of the probes mimics that of the real variables, but the randomization of the order of the values make them independent of the target variable. This method bears resemblance with permutation tests (Pitman, 1937). It is widely applied in genomics.

Extending the idea of probes for the problem of "causal" feature selection is not as simple as it may seem. We move from the relatively simple question of separating "relevant" from "irrelevant" features to a multi-class problem including "causes" of the target, "effects" of the target, "confounded" variables and "unrelated" variables. Suppose for simplicity that we only want to determine *whether an algorithm correctly uncovers causes of a target variable*. "Irrelevant" variables include "unrelated" variables, "effects" and "confounded" variables. So, to test the efficacy of an algorithm to uncover causes of the target, we must introduce artificial distracter variables (probes) of several kinds. Specifically, we need to construct variables with a "null mechanism" (*e.g.,* a function plus some noise or a posterior distribution), taking as input subsets of the available real variables (including eventually the target) and previously constructed probes. This ensures that no probe will be a cause of the target, but that some will be predictive and some not.

One way of assessing the validity of a proposed set of causes of the target is to compute the fraction of probes (all non-causes of the target) in that subset. Large fractions of probes shed doubt to the validity of the proposed causes. The probability of getting a number of probes smaller than a certain threshold can serve as a basis for a statistical test.

In the causality challenges that we organized, we assessed "causal relevance" using the probe method. Algorithms were required to return an ordered list of variables, with, for instance, all causes coming first in order of preference or confidence. If the truth values of the causal relationships had been known, this ranking could simply have been evaluated with the Area Under the ROC curve (AUC, the area under the curve plotting the fraction of correctly detected causes *vs.* the fraction of false alarms, when a threshold on the number of top ranking causes is varied). Instead, we used the probe AUC (called *PAUC*) as a proxy (*correctly detecting causes* being replaced by *correctly excluding probes*). In (Guyon et al., 2008), we prove that, if the null distribution used to generate the probes is correct, in the limit of an infinite number of probes, we have $PAUC = (n_+/n_r)AUC + 0.5n_-/n_r$, where AUC is the true AUC (which cannot be computed) and n_+ and n_- are the unknown numbers of positive examples (causes) and negative examples (non-causes) for the $n_r = n_+ + n_-$ real variables. Hence, asymptotically $PAUC$ is monotonically related to the real AUC and therefore it can be used as a proxy to assess the relative performance of models.

The introduction of probes among the real variables induces a perturbation, which may distort the causal discovery problem (*e.g.,* by creating spurious conditional dependencies between the target and real variables). These perturbations may alter the real cause-effect relationships in unsuspected ways. Hence, for discovery, we recommend to re-run the algorithm on real data only, without the addition of probes.

6. Discussion: Failure breeds success

The old timers of machine learning and artificial neural networks will remember that the field has long been traumatized by the XOR problem. In the 1960's, Frank Rosenblatt, Bernard Widrow and others introduced various training algorithms for one layer neural networks. In 1969, Minsky and Papert in their book on Perceptrons (Minsky and Papert, 1969), inventoried problems, which were "non linearly separable", *i.e.,* could not be solved with one layer neural networks. The archetype or such problems is the XOR problem: the Boolean function XOR is not linearly separable. The book had a great impact and put the field of artificial neural networks in dormancy for nearly 20 years. During its revival in the 1980's when algorithms to train multi-layer Perceptrons emerged, no paper on artificial neural networks failed to address the XOR problem. It is worth noting though that linear discriminant functions are tremendously useful and failing to solve the XOR problem is not an indication that a learning machine is useless. For example, in the 1990's, the non-linear Support Vector Machine was invented (Boser et al., 1992), which brought attention to its linear version dating back from the 1960's. The linear SVM is now a very widely used method in text processing and bioinformatics.

The field of causal discovery has many problems similar to the XOR problem. However, neither solving them nor failing to solve them is necessarily an indication that the methods will not perform well in real world applications. While such problems should be used as tools to improve our methodology and we also should constantly remind ourselves that "failure breeds success" and that stumbling on any of these problems does not mean that unraveling causal relationships is a hopeless task and much less that causality is a useless concept. In this section, we first play devil's advocate and give 10 reasons why causal discovery might be a hopeless enterprize. Then, we tell 10 success stories proving the pessimists wrong. Finally, we list 10 open problems on which researchers are still stumbling.

6.1 Ten challenging problems

Several papers in these proceedings present cases in which common assumptions made are violated or cases in which common causal models either find spurious causal relationships or fail to uncover existing ones. Most of these problems are discussed thoroughly in causality textbooks (Pearl, 2000; Spirtes et al., 2000; Neapolitan, 2003). We present briefly ten of them.

1. **No formal definition of causality.** There is so far no formal mathematical definition of causality. Two approaches attempt to fill this vacuum: (1) Operational tests of causality (Glymour and Cooper, 1999) allow us to detect causality experimentally using controlled experiments, but they provide only sufficient criteria for causality, not necessary conditions (see Section 3.1). (2) Data generative models propose ways in which variables values may be generated from each other using defined mechanisms. Algorithms, which can reconstruct the architecture of a model using data generated by that model are called "causal discovery algorithms".

2. **Statistically dependent is not the same as correlated.** Two random variables X and Y are called independent if $P(X,Y) = P(X)P(Y)$ which is a stronger condition than absence of correlation, *i.e.,* $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$. This distinction is often overlooked by causal discovery algorithms, which use correlation as a symptom of causation instead of statistical dependency. Non-linear mechanisms can generate dependencies without any correlation (although, in typical cases, dependent variables are at least weakly correlated). Reciprocally, (partial) correlation can arise in the absence of (conditional) statistical dependence: Partial correlations are given by the correlations of the residuals after *linear* regression. If X and Y are non-linear functions of Z (up to noise terms independent of

each other and of Z), only non-linear regression would render them independent and uncorrelated. Hence X and Y can remain partially correlated, given Z, even though they are conditionally independent, given Z. Therefore, neglecting the possibility of non-linear mechanism and using statistical tests based only on correlation can lead both to false negative and false positive dependencies.

3. **Statistical dependence does not imply causation**[2]. According to the **principle of common cause** (PCC), every statistical dependency between two random variables X and Y has a causal explanation. Reichenbach (Reichenbach, 1956) formulated the following three (not necessarily exclusive) cases: (1) X causes Y, (2) Y causes X, or (3) there is a third variable Z (common cause or **confounder**) causing both X and Y. In this last case, conditioning on Z renders X and Y independent, if cases (1) and (2) do not hold. For instance, assume that "chocolate intake" (variable X) is found to positively correlate with "life expectancy" (variable Y). This does not necessarily imply that eating more chocolate will improve your chances of living longer. It is possible that in fact "gender" (variable Z) affects both "life expectancy" (females live longer) and "chocolate intake" (females eat more chocolate), but that in each "gender" sub-population (male or female) there is no dependence between "chocolate intake" and "life expectancy" (Simpson's paradox).

   The problem that confounders are often unobserved, unobservable or even unknown and that Z can be a high-dimensional vector of relevant factors, is one of the main obstacles of causal inference from **observational data**. Often it is even hard to quantify latent factors such as subject's personality and physical condition in a medical study. In other words, there is no way for reliably deciding whether the set of observed variables is **causally sufficient** (*i.e.,* does not exclude any common cause of any pair of variables)[3].

   For causally sufficient sets of variables, the postulate of the **causal Markov condition** (CMC) provides a practical principle for selecting candidate causal structures from observational data, by providing conditions under which statistical dependency may be linked to causality. Several equivalent versions of the CMC exist. The most commonly used version postulates **conditional independence between every variable and its non-effects, given its direct causes**. Pearl justified the CMC by a model of causality where every variable is a function of its direct causes and a noise variable that renders the causal mechanism probabilistic (**structural equation model** or SEM). Then the CMC follows, assuming joint statistical independence of the noise terms[4]. The most common violations of the CMC arise from violations of causal sufficiency or existence of correlated noise. In deterministic systems, violations of the CMC may result from the existence of constraints (such as conservation of mass, energy, or momentum); a classical example is that of the trajectories of two billiard balls hit by a third one (see the paper of Lemeire and Steenhaut in these proceedings).

4. **"Faithfulness" is not always justifiable by "stability".** This sentence is a shorthand to bag together a variety of related hypotheses commonly referred to as Causal Faithfulness Condition (CFC). While the CMC essentially states that dependency implies the existence

2. In light of the previous item, we avoid using the terser motto "correlation does not mean causation".
3. See section 6.2 in (Pearl, 2000), called "Why there is no statistical test for confounding, why many think there is, and why they are almost right".
4. There is a tight relation between CMC and PCC: The conditional independence of two effects, given their common cause, is just a special case of CMC with three variables. It can also be argued that the independence of noise terms in Pearl's model corresponds to an absence of common noise-generating mechanism, which follows from the PCC. Spirtes et al. (2000) proposed a weak causal Markov assumption, similar to the converse of the PCC, stating that if X and Y have no common cause (including each other), they are probabilistically independent. This weaker assumptions implies the CMC for SEMs

of a causal arrow, the CFC states the opposite, namely that independence implies no causal arrow. The CFC is more controversial than the CMC and it is the XOR problem of causality. Imagine two identical fair coins tossed simultaneously and let us call X_1 and X_2 the binary random variables corresponding to the outcome (heads or tail). Consider a outcome Y, which is whether or not both coins fell on different sides. Note that the logical relation $Y = X_1 \mathrm{XOR} X_2$ is fulfilled. Since both X_1 and X_2 are individually independent of Y (and independent of each other), according to the CFC one should not draw any causal arrow. However, clearly, there is a joint dependency between X_1, X_2 and Y. In the causal network framework one could represent the dependency with the unfaithful graph $X_1 \to Y \leftarrow X_2$, but the representation $[X_1, X_2] \to Y$ might be more suitable since the two variables jointly cause Y. Other classical examples of faithfulness violation for non-binary variables include cases in which two causal paths exactly cancel each other with a particular choice of parameters. In either case (XOR or canceled causal path) the "stability" argument in favor of the CFC is that if there is the smallest defect in the generative process (a coin not exactly fair or parameters not exactly tuned to cancel the causal paths), then the symmetry is broken and faithfulness is re-established. Critics of the CFC point out that, in practice, small asymmetries are difficult to detect from empirical data and that there are many systems in which there is an equilibrium leading to canceled causal paths (see for instance the paper of Voortman, Dash, and Druzdzel in these proceedings). Hence, many technical systems like systems of logical gates easily violate faithfulness.

5. **Markov equivalences.** Many causal graphs may generate identical probability distributions or at least entail the same set of conditional independencies between variables (Markov equivalent graphs). For instance $X \to Y \to Z$, $X \leftarrow Y \leftarrow Z$, and $X \leftarrow Y \to Z$ all have the same unique Markov property that X and Z are independent given Y. Most structure learning algorithms (from observational data) rely on the existence of so-called unshielded colliders of the form $X \to Y \leftarrow Z$, which do not have any other Markov equivalent graph. Such methods can unravel causal relationships in systems of at least three variables, up to Markov equivalent graphs. Hence, they are not applicable to the problem of cause-effect pairs. Recent methods have addressed this problem, such as the solutions proposed in these proceedings to the CauseEffetPairs task.

6. **Model selection.** When learning from observational data, classical cross-validation is not very useful to perform model selection since predictions are to be made on data from a different, post-manipulation, distribution. Hence, penalty-based methods like AIC (Akaike, 1973) or BIC (Schwarz, 1978) are sometimes used to drive model choices toward fewer parameters or minimal architectures. Yet, obviously, minimal models are not always the best. For an analysis, see the paper of Lemeire and Steenhaut in these proceedings.

7. **Measurement errors, quantization, and aggregation distort dependencies.** In his presentation at the NIPS 2008 workshop, Richard Scheines gave several examples in which measurement errors or data quantization limit causal discovery. For instance, a causal system of three variables X, Y, and Z may have the Markov property that X is independent of Z given Y (*i.e.,* one of these three graphs is valid: $X \to Y \to Z$, $X \leftarrow Y \leftarrow Z$, $X \leftarrow Y \to Z$), and yet, this Markov property may go undetected if Y is observed through a noisy or quantized version Y' (technically, Y' is a consequence of Y and therefore it does not d-separate X and Z). Similarly, variables X, Y, and Z may be the result of averaging over populations $X = \sum_i X_i$, $Y = \sum_i Y_i$, and $Z = \sum_i Z_i$. So, even though X_i might be independent of Z_i given Y_i for every i, it is possible that the property does not hold for the average.

8. **Sample bias and attrition bias plague experimental design.** The validity of randomized experiments relies on the quality of randomization. Spurious relationships may be found because of sampling. For instance, it may be found that there is a correlation between pregnancy and flu. If the patients were sampled only from an emergency room, this may simply indicate that patients with acute nausea or vomiting symptoms arising from multiple conditions are more likely to show up in the emergency room, not that the two conditions are causally related or have a common cause. Sample bias plagues retrospective studies, which analyze observational data collected without any particular design. Prospective longitudinal studies following patients over a period of time are usually less prone to sample bias because they are more carefully designed, but they are prone to attrition bias (some patients quit the study before the end, for instance when a treatment has undesirable side effects.)

9. **Markovian causal graphs do not represent suitably all data's generative processes.** Directed Acyclic Graphs (DAGs) cannot represent cyclic systems, by definition. This can be remedied by unfolding cycles in time, which, for discrete time systems amounts to using a classical Markov model. But, symmetric relationships (such as gravitational or electrical forces) or constraints (such as energy, mass and momentum conservation) are not suitably represented by arrows (which are usually interpreted as directional relationships). Accordingly, a given event Y may simultaneously generate multiple related consequences (a classical example is that of the billiard ball hitting two balls simultaneously). The notation $X \leftarrow Y \rightarrow Z$ suggests that X and Z are generated by Y from two independent mechanisms, rather that a single mechanism with underlying constraints. A new notation such as $Y \rightarrow [X,Z]$ may be more suitable. See the paper of Lemeire and Steenhaut in these proceedings for a discussion of this issue.

10. **Causality in time series is not necessarily an easier problem.** Causality is commonly thought of as a time-related concept (causes precede their effects). So how can causality in time series be harder to investigate than causality in time independent data? On one hand, the problem is indeed simpler because events that took place in the future may be pruned from the set of candidate causes of an event. Thus temporal causal models use only past values of variables to predict future values. On the other hand, modeling can be harder (i) if the time series are non-stationary (spurious correlations are easily found), (ii) if the variables are measured in presence of noise (see the NOISE dataset in these proceedings), (iii) if data are scarce (overfitting problems can be severe since the data points are not independent, therefore more data points are required than for i.i.d. data), (iv) if experiments are not properly designed (in particular, the non-commutativity of equilibration and manipulation might complicate matters, see the paper of Voortman, Dash, and Druzdzel in these proceedings).

This list is pretty scary, although non-exhaustive. On top of that, the availability of (quality) data, particularly experimental data, is usually limited. Causal models are perhaps even more prone to overfitting than regular predictive models, because in addition to estimating dependencies, one must estimate the direction of the causal relationships. When there is enough data, causal models suffer from a high computational complexity. Hence for a large number of variables, sub-problems must usually be solved (*e.g.,* focusing on the local neighborhood of a variable). And yet, there are success stories!

6.2 How causal conclusions changed our life - ten stories

Researchers working on causal inference are often confronted with three kind of objections:

(1) Philosophical concerns about **whether causality is a well-defined scientific concept**. In 1913, Bertrand Russel stated "the law of causality, I believe, like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm" (Russell, 1913). On one hand, this perspective seems to be supported by the way many physical laws are formulated, *e.g.,* as time-inversion symmetric differential equations in space time in Einstein's theory of special relativity, just discovered at Russel's time. On the other hand, physics can also be seen as predicting how outcomes of experiments depend on the experimental setup, which is an inherently causal formulation.

(2) Scepticism about **whether causal conclusions can be drawn from non-interventional observations**. The most radical version of this concern would be the belief that only randomized controlled studies yield valuable causal conclusions. A more moderate version, which probably many statisticians would agree to, states that causal inference from non-randomized studies relies essentially on background knowledge of the domain of the data, which makes it part of the respective field rather than being part of statistics.

(3) Scepticism about **whether causal discovery from observational data can be mathematically formalized** up to a degree that admits the implementation of reliable inference algorithms. There is, however, no clear boundary between (2) and (3) because the way the input of an algorithm is specified can contain an arbitrary amount of prior knowledge. For instance, how to formalize observations in terms of random variables already involves human judgements about which representation is natural for the respective problem – a decision that also occurs in other machine learning tasks.

The practical relevance of concern (1) is questionable since an essential part of scientific and technological progress consists in deriving *causal* statements as opposed to purely predictive ones because they are the only results that provide criteria for human actions. The examples of this section illustrate how causal insights from different scientific disciplines already influenced both private decisions and those in public health, economy and politics. Historical examples pre-dating the computer age involved more human reasoning than computerized data analysis, but we include them because of their exemplary nature. They also show that progress has been made by exploiting non-interventional data and not only by randomized control studies, which responds to concern (2). The more recent examples, including the practical impact of Granger causality, the use of instrumental variables in genetic studies (via Mendelian randomization) and successes of Bayesian network in biology and SEMs in social sciences respond to concern (3).

In selecting our success stories, we have applied very stringent criteria, which prevented us from including many promising on-going efforts mentioned in Section 4.3 (either because the conclusions have not yet been sufficiently validated or because their socio-economic impact has not been evaluated). Consequently, many recent algorithms are not yet illustrated in these success stories. However, in response to concern (3), it should be emphasized that causality research must not be reduced to developing algorithms (even though this is an important part). Human causal reasoning requires rationales to rely on. The goal to develop automatic causal discovery has already created a conceptual clarity (Pearl, 2000; Spirtes et al., 2000) that many previous discussions were lacking. Human causal inference also requires reliable criteria that state-of-the-art statistics do not provide. Pioneering work from (Janzing and Schölkopf, 2008) present a formal basis for causal inference that also work with single observations rather than relying on statistical ensembles. In deriving further principles, one should be encouraged by the following successes of causal thinking in science.

1. **Vitamin C and scurvy: A historical RCT.** Observational epidemiology and controlled experiments have revolutionized our understanding of causal risk factors predisposing to a variety of common diseases. While at sea in May 1747, a ship surgeon of the British Royal Navy, James Lind, provided some crew members affected by scurvy with two oranges and one lemon per day, in addition to normal rations, while others continued on their regular diet. In the history of science, this is considered to be the first occurrence of a controlled experiment comparing results of two populations where one factor is applied to one group only with all other factors the same. Following this discovery, in 1795 the Royal Navy provided a daily ration of fresh lime or lemon juice to the sailors and successfully fought scurvy. It is now established that citrus fruits contain Vitamin C, which is necessary for the treatment and prevention of scurvy. However, there is continuing debate within the scientific community over the best dose schedule of vitamin C for maintaining optimal health in humans and whether overdose may have adverse effects.

2. **Hygiene and infectious diseases: Can you believe what you can't see?** It is hard to believe that the use of basic hygiene precautions was at some point fought by the medical establishment. Yet, when the Hungarian physician Ignaz Philipp Semmelweis discovered in the 1840's that cases of puerperal fever (childbed fever) could be cut drastically if doctors washed their hands in a chlorine solution before gynaecological examinations, he was ridiculed and harassed. The validation of the germ theory by Pasteur's experiments in the 1860's was necessary before the cause-effect relationship between hygiene and infectious diseases was accepted. He exposed freshly boiled broth to air in vessels either directly exposed to air or protected by a filter stopping all particles. Nothing grew in the protected broths, therefore the living organisms that grew in unprotected broths came from outside (as spores on dust) rather than being generated within the broth. This initial work stimulated the development of techniques to kill germs in beverages (Pasteurization), protocols of antiseptic surgery, and immunization methods (vaccination). With the advent of more powerful microscopes and the progresses made in microbiology, a large body of work now supports that the underlying mechanisms of infectious diseases involve germs, which can be killed with anti-bacterial agents, thus providing an explanation for the causal link between hygiene and infectious diseases.

3. **Crop yield optimization in agriculture: Mathematical foundations of experimental design.** The first statistician to consider a formal mathematical methodology for designing experiments was Fisher, in his book "The Design of Experiments" (1935). He developed his methodology while working at the Rothamsted Experimental Station (England), one of the oldest agricultural research institutions, founded in 1843. Partly through these methods, researchers at Rothamsted have made significant contributions to agricultural science, including the discovery and development of systemic herbicides and pyrethroid insecticides, as well as pioneering contributions to the fields of virology, nematology, soil science and pesticide resistance. During World War II, aiming to increase crop yields for a nation at war, a team under the leadership of Judah Hirsch Quastel developed 2,4-D, still the most widely used weed-killer in the world. In medicine, the double blind Randomized Controlled Trial (RCT), where allocations are randomized and neither patient not doctor knows which treatment has been assigned, is now a standard experimental design in clinical trials.

4. **The smoking ban and lung cancer: Better err on the safe side.** Prior to World War I, lung cancer was considered to be a rare disease, which most physicians would never see during their career. With the postwar rise in popularity of cigarette smoking, however,

came an epidemic of lung cancer. In 1950, Richard Doll undertook with Austin Bradford Hill a study of lung cancer patients in 20 London hospitals, at first under the belief that it was due to the new material tarmac, or motor car fumes, but rapidly discovering that tobacco smoking was the only factor they had in common. Sir Ronald A. Fisher and other statisticians opposed the conclusions of Doll and Hill that smoking caused lung cancer on the ground that correlation does not imply causation. For instance, there may be an unknown genetic factor, which causes both lung cancer and craving for tobacco. Many studies followed (see Spirtes et al., 2000, page 239 for a detailed account), eventually leading to tobacco smoking bans in public places in several countries. Interestingly, the results of controlled studies on the effect of smoking on lung cancer are mixed, but there is a large consensus that the smoking ban reduced heart disease (Sources include, the US National Cancer Institute and the American Lung Association).

5. **NSAIDs, drug efficacy and drug toxicity.** Non-steroidal anti-inflammatory drugs (NSAIDs) include some of the most commercially successful drugs like Aspirin or Tylenol. They are used to treat pain, fever and inflammation. Most NSAIDs act as non-selective inhibitors of the enzyme cyclooxygenase, which catalyzes the formation of prostaglandins, messenger molecules in the process of inflammation causing pain and fever. This mechanism of action was elucidated by John Vane, who later received a Nobel Prize for his work in 1982. Medicines containing derivatives of salicylic acid, structurally similar to aspirin, have been in medical use since ancient times. A French chemist, Charles Frederic Gerhardt, was the first to prepare acetylsalicylic acid in 1853. In 1899, Bayer patented it for its use as a drug under the name Aspirin. Aspirin's popularity grew over the first half of the twentieth century, spurred by its effectiveness in the wake of the Spanish flu pandemic of 1918, and aspirin's profitability led to fierce competition and the proliferation of aspirin brands and products, especially after the American patent held by Bayer expired in 1917. Aspirin is no longer used in children and adolescents due to the risk of Reye's syndrome; paracetamol (the international non-proprietary name for the drug Tylenol) is now often used instead. In 1887 the clinical pharmacologist Joseph von Mering first tried paracetamol on patients. In 1893 he published his results comparing paracetamol with phenacetin, another aniline derivative, claiming that, unlike phenacetin, paracetamol had a slight tendency to produce methemoglobinemia (abnormal oxidation of hemoglobin to methemoglobin, reducing the oxygen transport capabilities of red blood cells). The toxicity of paracetamol was not challenged until the late 1940's when it was shown that phenacetin metabolizes to paracetamol (von Mering's results may have been due to some impurity). Paracetamol was first marketed in the United States in 1953 by Sterling-Winthrop Co., which promoted it as preferable to aspirin since it was safe to take for children. More recently, the commercially successful NSAID Vioxx, approved by the FDA in 1999, was voluntarily withdrawn from the market by Merck in 2004 because of concerns about increased risk of heart attack and stroke. This example illustrates the intricacy of determining positive and negative effects via a combination of observational, controlled studies, and understanding of mechanisms.

6. **Genetic epidemiology: Towards personalized medicine.** Genetic epidemiology is concerned with understanding heritable aspects of disease risk, individual susceptibility to disease, and ultimately with contributing to a comprehensive molecular understanding of pathogenesis and a medicine tailored to the individuals. It is also an area of intensive causal studies. According to Kraft and Hunter (Kraft and Hunter, 2009): "A major goal of the Human Genome Project was to facilitate the identification of inherited genetic variants that increase or decrease the risk of complex diseases. The completion of the

International HapMap Project and the development of new methods for genotyping individual DNA samples at 500,000 or more loci have led to a wave of discoveries through genome-wide association studies. These analyses have identified common genetic variants that are associated with the risk of more than 40 diseases and human phenotypes. Several companies have begun offering direct-to-consumer testing that uses the same single-nucleotide polymorphism chips that are used in genomewide studies." And, according to Goldstein (Goldstein, 2009): "More than 100 genomewide association studies have been conducted for scores of human diseases, identifying hundreds of polymorphisms that are widely seen to influence disease risk. After many years in which the study of complex human traits was mired in false claims and methodological inconsistencies, genomics has brought not only comprehensive representation of common variation but also welcome rigor in the interpretation of statistical evidence."

7. **Reverse causation and confounding resolved by Mendelian randomization.** Mendelian randomization makes a bridge between observational epidemiology studying environmental factors and genetic epidemiology. The problem of "reverse causality" occurs when the direction of a cause-effect relationship is inverted because the onset of the cause was not detectable. The problem was studied by Martijn Katan in 1986 (Katan, 2004; Keavney, 2004) for the association between low serum cholesterol levels and cancer. In this case, a pre-existing occult tumor might cause lower cholesterol levels, rather than lower cholesterol levels causing cancer (Garcia-Palmier et al., 1981). The association might also be explained by confounding factors (such as cigarette smoking) related both to future cancer risk and to lower circulating cholesterol (McMichael et al., 1984). Katan proposed a method using genetics to emulate a RCT without performing actual manipulations. His method was never tested but it was then generalized by Gray and Wheatley in 1991 (Gray and Wheatley, 1991; Wheatley and Gray, 2004; Smith, 2007) in a method called "Mendelian Randomization". The idea is to use a naturally occurring genetic polymorphism, with a well understood regulatory effect, as an instrument to manipulate a variable of interest (*e.g.,* raising blood cholesterol). Importantly, the genotype must only affect the disease status indirectly via its effect on the variable of interest (*e.g.,* blood cholesterol). Because genotypes are assigned randomly when passed from parents to offspring, the statistical dependence between the population genotype and the cancer cannot be confounded (as opposed to cholesterol, where confounding by social, behavioral or physiological factors is possible). The biggest success so far of Mendelian randomization studies were obtained using a mutation of *methylene tetrahydofolate reductase* as randomization instrument in studies of the implication of folate in coronary heart disease, fetus neural tube defects, and cancer (see Smith, 2007).

8. **System biology: Reverse engineering the cell.** One branch of system biology, which is an active area of causal studies, aims at modeling a whole cell. As part of that effort, Nir Friedman and his collaborators wrote several of the key papers using Bayes Networks for gene expression analysis and pathway modeling. This approach generalized the method of Boolean networks for pathway modeling traditionally used by chemical engineers to abstract metabolic and biochemical networks by modeling uncertainty and introducing hidden variables. For a review of Friedman's work see (Friedman, 2004). For the most part, published papers in this area propose networks based on analyzing empirical data and then compare the results with the existing literature. Few papers are followed by an experimental validation of new findings. Still, these results, which incorporate global simultaneous measurements, are a good complement to results coming from other sources

investigating in more details the interactions of few chemical species, including *e.g.,* via gene knockout experiments.

9. **College dropouts: Assisting policy-making in social sciences.** Using causal discovery algorithms to learn the structure of Structural Equation Models, Spirtes and collaborators have worked out a large number of problems previously published in the literature and found structures matching or closely resembling those built with expert knowledge (Spirtes et al., 2000). The examples include finding the causes of publishing probability, finding the influence of parent education on children education, and finding what influences abortion opinions. For illustration, we give an end-to-end story, which actually led to a change in policy: (Druzdzel and Glymour, 1999) performed a study at the request of the provost of Carnegie Mellon University (CMU) to investigate policies for lowering dropout rates. Using the US News and World Report database on American college and universities, they found that all variables in the database to be independent of college dropout given the results of test scores of the entering class (SAT test scores). Subsequent higher selection of students based on the SAT test results at CMU correlated with lower dropout rates (but may have been affected by other factors).

10. **Granger causality: Causality in time series.** Clive Granger and his collaborators published in 1970's and 1980's methods for determining whether some time series are useful in forecasting others. A time series $x(t)$ "Granger causes" another $y(t)$ if the bivariate model (using past values of x and y to predict y) is more predictive than the auto-regressive model (using only past values of y to predict y). This conclusion, however, is only correct if there are no instantaneous causal influences between $x(t)$ and $y(t)$ and if there is no common cause influencing both.

    Granger received the 2003 Nobel prize in economics for his work on co-integration and modeling of non-stationary time series. If both $x(t)$ and $y(t)$ are non-stationary, but some linear combination $ax(t)+by(t)$ is stationary, then $x(t)$ and $y(t)$ are said to be co-integrated. Granger proved that co-integrated time series must be in a Granger causal relationship. In spite of its limitations, Granger causality is a big leap forward as it eliminates many spurious correlation or spurious regression found by fitting models making stationarity assumptions using ordinary least squares. Granger's work has transformed the way economists deal with time-series data. Today, tests of stationarity and co-integration are carried out routinely as a stepping-stone to the specification of dynamic econometric models relating exchange rates and price levels, consumption and wealth, dividends and stock prices, and interest rates of different maturities (source: Nobel web site (Granger, 2003)).

6.3 Ten open problems

Much remains to be done in the domain of causal modeling. While successful causal studies have focused primarily on systems of just a few variables, more ambitious recent endeavors have ventured to unravel causal relationships in systems of thousands of variables, facing new challenges. We give ten research directions, which we think deserve attention.

1. **Optimizing directly defined objectives.** One of the two themes of the NIPS 2008 workshop was to define objectives for causal modeling. Assuming that we made a step in the right direction, the next step will be to develop methods to optimize such objectives. In pattern recognition, the old paradigm which consisted in developing separately the building blocks of recognition systems (preprocessing, classifier, and post-processing)

has made way to approaches, which globally optimize simultaneously all the parameters of the processing chain with respect to a global objective. Similarly, we anticipate that in causal modeling searching directly for optimal modes of action (policies) to attain given objectives may be easier and yield better solutions than attempting to faithfully unravel the data's generative process. The causal model would then just be a means to an end, not an end in itself. Such approaches may bridge between causal modeling, operations research and identification and control.

2. **Improving and comparing assessment methods.** The second theme of the workshop was the development and study of methods of assessment of causal models. As we pointed out in the course of the paper, the problems of model selection, model performance prediction, and model assessment are more difficult for causal models than for regular statistical models because data are not i.i.d. We briefly reviewed some assessment methods in Section 5. The next step will be to study and compare such methods (and others), eventually leading to best practice recommendations for data analysts.

3. **Understanding and modifying regularly made assumptions.** Assumptions like the CMC, causal sufficiency, the CFC, Gaussianity of the noise, linearity of the relationships, are often made out of convenience rather than out of an understanding of the data's generative process and of the possible consequences on the solution. Collecting pedagogical examples violating such assumptions should facilitate the work of data analysts and, in turn, inspire theoreticians to modify the assumptions. For instance, unfaithful distributions can arise from deterministic relations (Lemeire, 2007), which are not uncommon in nature. Finding appropriate meta-principles which imply faithfulness under specific conditions would be an option for future foundations of causal inference (see the paper of Lemeire and Steenhaut in these proceedings). In a Bayesian setting, this task would corresponds to finding good priors on the parameter space of a Bayesian network. One could ask for abstract properties such priors should have, following earlier work of (Meek, 1995).

4. **Developing versatile regularized models.** Bayesian networks based on directed acyclic graphs (DAGs) are praised for their simplicity, but have the limitations that we mentioned in Section 6.1. Many other models have been proposed to generalize them and/or address their limitations, including partial ancestral graphs (which model uncertainties about arrow directions), Markov random fields (for bi-directional connections), cyclic and dynamic models. Linear structural equation models (SEMs) with Gaussian noise variables have been generalized to non-linear and non-Gaussian noise models. Practitioners are at a loss to determine without domain knowledge which model may be best suited and avoid either underfitting or overfitting data. It may facilitate their work to move towards general-purpose versatile causal models, and use regularization methods to bias the search for optimal structures and parameters towards simpler solutions. Efforts in this direction have stated to emerge (see Lozano et al., 2009, and the paper of Zhang and Hyvärinen in these proceedings)

5. **Developing efficient and effective algorithms.** Much progress has been made recently towards scaling up algorithms to large numbers of variables and large numbers of examples. One approach has been to make use of regular feature selection methods developed in machine learning to prune the search for causes and effects (Aliferis et al., 2003). This and other efforts in the same direction need to be pursued.

6. **Developing a methodology for feature construction.** Variable definition and coding is not innocuous in causal modeling. We have seen in Section 6.1 that variable aggregation

can occlude some conditional independencies. Coding a categorical variable into several (dependent) variables using a complete disjunctive coding may result in similar problems. Hence a methodology for defining, constructing, and coding variables must be developed to guide practitioners. Steps in this direction have recently be made (Spirtes, 2008).

7. **Addressing imperfections in data.** Imperfection in data such as measurement errors, data quantization, missing values, sampling bias, attrition bias, and correlated noise may be responsible for modeling errors. While classical statistical models may degrade gracefully with such data imperfections, structural errors in causal models may yield entirely wrong conclusions as to which actions are susceptible to influence a desired outcome. Although it may not be possible to inventory all possible adverse situation, it is important to raise awareness among practitioners, find methods for diagnosing a number of classical problems, and eventually find remedies.

8. **Integrating heterogeneous information.** Merging data from a variety of sources is going to be one of the major challenge in some domains. In genomics and proteomics, for instance, understanding the role of specific genes and proteins in disease requires multidisciplinary approach. Relevant data come from sources as diverse as high-throughput tools (like DNA microarrays and mass-spectrometry), gene knock-out/knock-down techniques, protein characterization, metabolic profiling, high-content screening, phenotype, and clinical data. In medicine, it is generally admitted that the strongest evidence for therapeutic interventions is provided by systematic review of multiple Randomized Controlled Trials. The Cochrane Collaboration is a group of over 15,000 volunteers in more than 90 countries who review the effects of health care interventions tested in biomedical randomized controlled trials. There may be value in developing methods to integrate information from various sources, identify possible contradictions, and track them back to confounding factors or experimental errors.

9. **Designing studies combining observational and experimental data.** Observational studies and expert opinions are usually not considered reliable evidence, compared to controlled experiments. However, experiments being costly, time consuming and sometimes unethical or impractical, it seems that it could make sense to design studies in which both observational and experimental data would be collected, in an effort to maximize information for a given budget.

10. **Quantifying uncertainty.** Learning from a finite amount of observational and/or experimental data yields models and predictions tainted with uncertainty. Most causal discovery algorithms are justified in the infinite sample size limit. There is a need to quantify uncertainty *e.g.,* with bounds on the prediction error involving model complexity, data quality, and data quantity.

7. Conclusion

There is an intense activity in a nucleus of machine learning researchers interested in causality. We hope that this activity will result in improving techniques to unravel cause-effect relationships and expand the domain of application in areas where the number of features and variables is much larger than those usually considered in the past. At this stage, there seems to be an abundance of algorithms looking for good applications. Hence the most urgent questions are: How to get good problems? How to get good data? How to get conclusive results? For that reason, we are continuing our effort of data exchange and benchmark through the Causality Workbench project.

While we hope that our effort will lead to an improvement in methodology, we would like to borrow the wisdom of Petitti (Petitti, 2004), who makes the following four recommendations: (1) Do not turn a blind eye to contradiction. Do not ignore contradictory evidence but try to understand the reasons behind the contradictions. (2) Do not be seduced by mechanism. Even where a plausible mechanism exists, do not assume that we know everything about that mechanism and how it might interact with other factors. (3) Suspend belief. Do not be seduced by your desire to prove your case. (4) Maintain scepticism. Question whether the factors under investigation can really be that important; consider what other differences might characterize the case and control groups. Do not extrapolate results beyond the limits of reasonable certainty.

Acknowledgments

This project is an activity of the Causality Workbench supported by the Pascal network of excellence funded by the European Commission and by the U.S. National Science Foundation under Grant N0. ECCS-0725746. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We are very grateful to all the members of the causality workbench team for their contribution to organizing the pot-luck challenge: Constantin Aliferis, Greg Cooper, André Elisseeff, Jean-Philippe Pellet, Peter Spirtes, and Alexander Statnikov. We thank Hans Bitter, Jean-Philippe Pellet, Alexander Statnikov, and Ioannis Tsamardinos for commenting on the manuscript.

References

M. Aickin. *Causal Analysis in Biomedicine and Epidemiology: Based on Minimal Sufficient Causation*. Chapman and Hall/CRC, 2002.

H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. Petrov and F. Csaki, editors, *2nd International Symposium on Information Theory*, pages 267–281. Akademia Kiado, Budapest, 1973.

C. F. Aliferis, I. Tsamardinos, A. Statnikov, and L. Brown. Causal explorer: A probabilistic network learning toolkit for biomedical discovery. In *2003 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS)*, Las Vegas, Nevada, USA, June 23-26 2003. CSREA Press.

C. F. Aliferis, A. Statnikov, and P. P. Massion. Pathway induction and high-fidelity simulation for molecular signature and biomarker discovery in lung cancer using microarray gene expression data. In *APS Conference: Physiological Genomics and Proteomics of Lung Disease*, 2006.

A. Ben-Hur, A. Elisseeff, and I. Guyon. A stability based method for discovering structure in clustered data. In *Pacific Symposium on Biocomputing*, pages 6–17, 2002.

K. A. Bollen and J. S. Long. Tests for Structural Equation Models: Introduction. *Sociological Methods Research*, 21(2):123–131, 1992. doi: 10.1177/0049124192021002001. URL http://smr.sagepub.com.

B. E. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In *COLT*, pages 144–152, 1992.

CFMDCY. Committee on Food Marketing and the Diets of Children and Youth. *Food Marketing to Children and Youth: Threat or Opportunity*. The National Academies Press, Washington, D.C., 2006. URL http://www.nap.edu/catalog.php?record_id=11514#orgs.

T. Chu and C. Glymour. Search for additive nonlinear time series causal models. *JMLR*, 9: 967–991, 2008. ISSN 1533-7928.

Clopinet. Challenges in machine learning, 2009. URL http://clopinet.com/challenges.

S. Demiralp and K. D. Hoover. Searching for the causal structure of a vector autoregression. *Oxford Bulletin of Economics and Statistics*, 65(s1):745–767, December 2003. URL http://ideas.repec.org/a/bla/obuest/v65y2003is1p745-767.html.

J. Diamond. *Guns, Germs, and Steel: The Fates of Human Societies*. W.W. Norton and Company, 1997.

M. Ding, Y. Chen, and S. L. Bressler. Granger causality: Basic theory and application to neuroscience. *WILEY-VCH VERLAGE*, 2006:451, 2006. URL http://www.citebase.org/abstract?id=oai:arXiv.org:q-bio/0608035.

M. J. Druzdzel and C. Glymour. Causal inferences from databases: Why universities lose students. In C. Glymour and G. F. Cooper, editors, *Computation, Causation, and Discovery*, pages 521–539, Menlo Park, CA, 1999. AAAI Press.

N. Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303 (5659):799–805, February 2004. ISSN 1095-9203. doi: 10.1126/science.1094068. URL http://dx.doi.org/10.1126/science.1094068.

N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using Bayesian networks to analyze expression data. In *RECOMB*, pages 127–135, 2000. URL citeseer.ist.psu.edu/friedman99using.html.

M. R. Garcia-Palmier, P. D. Sorlie, J. Costas, R., and R. J. Havlik. An apparent inverse relationship between serum cholesterol and cancer mortality in Puerto Rico. *Am. J. Epidemiol.*, 114(1):29–40, 1981. URL http://aje.oxfordjournals.org/cgi/content/abstract/114/1/29.

S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Comput.*, 4(1):1–58, 1992. ISSN 0899-7667. doi: http://dx.doi.org/10.1162/neco.1992.4.1.1.

C. Glymour and G. Cooper, editors. *Computation, Causation, and Discovery*. AAAI Press/The MIT Press, Menlo Park, California, Cambridge, Massachusetts, London, England, 1999.

D. B. Goldstein. Common Genetic Variation and Human Traits. *N Engl J Med*, 360(17):1696–1698, 2009. doi: 10.1056/NEJMp0806284. URL http://content.nejm.org.

C. Granger. Statistical methods for economic time series, 2003. URL http://nobelprize.org/nobel_prizes/economics/laureates/2003/public.html.

R. Gray and K. Wheatley. How to avoid bias when comparing bone marrow transplantation with chemotherapy. *Bone Marrow Transplant*, 7(Suppl. 3):9–12, 1991.

A. Gretton, O. Bousquet, A. Smola, and B. Schoelkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *ALT 2005*, pages 63–78, 10/08/ 2005.

I. Guyon and G. Dreyfus. Chapter 2: Assessment methods. In I. G. e. Eds., editor, *Feature Extraction, Foundations and Applications*, Series Studies in Fuzziness and Soft Computing. Physica-Verlag, Springer, 2006.

I. Guyon, C. Aliferis, and A. Elisseeff. Causal feature selection. In H. Liu and H. Motoda, editors, *Computational Methods of Feature Selection*, pages 63–82. Chapman and Hall/CRC Press. Longer TR: http://clopinet.com/isabelle/Papers/causalFS.pdf, 2007.

I. Guyon, C. Aliferis, G. Cooper, A. Elisseeff, J.-P. Pellet, P. Spirtes, and A. Statnikov. Design and analysis of the causation and prediction challenge. In *JMLR W&CP*, volume 3, pages 1–33, WCCI2008 workshop on causality, Hong Kong, June 3-4 2008. URL http://jmlr.csail.mit.edu/papers/topic/causality.html.

I. Guyon, C. Aliferis, G. Cooper, A. Elisseeff, J.-P. Pellet, P. Spirtes, and A. Statnikov. Causality workbench. In P. M. Illaria, F. Russo, and J. Williamson, editors, *Causality in the Sciences*. Oxford University Press, (to appear), 2010.

I. Guyon et al. Datasets of the causation and prediction challenge. Technical Report, 2008. URL http://clopinet.com/isabelle/Projects/WCCI2008/Datasets.pdf.

T. Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica*, 11, 1943.

E. H. Herskovits and A. P. Dagher. Application of Bayesian networks to health care, 1997.

P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *NIPS*, pages 689–696. MIT Press, 2008. URL http://books.nips.cc/papers/files/nips21/NIPS2008_0266.pdf.

D. Janzing and B. Schölkopf. Causal inference using the algorithmic Markov condition. http://arxiv.org/abs/0804.3678, 2008.

D. Janzing, J. Peters, J. Mooij, and B. Schölkopf. Identifying confounders using additive noise models. In *25th Conference on Uncertainty in Artificial Intelligence*, pages 1–9, Corvallis, OR, USA, 06 2009. AUAI Press. URL http://www.cs.mcgill.ca/~uai2009/.

M. Jenab, N. Slimani, M. Bictash, P. Ferrari, and S. Bingham. Biomarkers in nutritional epidemiology: applications, needs and new horizons. *Human Genetics*, June 2009. URL http://dx.doi.org/10.1007/s00439-009-0662-5.

B. B. J.M. Robins, M.A. Hernan. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi. Kegg for linking genomes to life and the environment. *Nucleic Acids Res.*, 36:D480–D484, 2008.

D. Kaplan. *Structural Equation Modeling: Foundations and Extensions*, volume 10 of *Advanced Quantitative Techniques in the Social Sciences*. SAGE, 2000. ISBN 0-7619-1407-2.

M. B. Katan. Apolipoprotein E isoforms, serum cholesterol, and cancer. *Int. J. Epidemiol.*, 33 (1):9–, 2004. doi: 10.1093/ije/dyh312. URL http://ije.oxfordjournals.org.

B. Keavney. Commentary: Katan's remarkable foresight: genes and causality 18 years on. *Int. J. Epidemiol.*, 33(1):11–14, 2004. doi: 10.1093/ije/dyh056. URL http://ije.oxfordjournals.org.

K. Kira and L. A. Rendell. A practical approach to feature selection. In *ML92: Proceedings of the ninth international workshop on Machine learning*, pages 249–256, San Francisco, CA, USA, 1992. Morgan Kaufmann Publishers Inc. ISBN 15586247X. URL http://portal.acm.org/citation.cfm?id=142034.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

I. Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine*, 23(1):89–109, 2009.

P. C. Kraaijeveld and M. J. Druzdzel. Genierate: An interactive generator of diagnostic Bayesian network models. In *16th International Workshop on Principles of Diagnosis*, Monterey, California, USA, 2005. URL www.kbs.twi.tudelft.nl/Publications/MSc/2005-Kraaijeveld-MSc.html.

P. Kraft and D. J. Hunter. Genetic Risk Prediction – Are We There Yet? *N Engl J Med*, 360(17): 1701–1703, 2009. doi: 10.1056/NEJMp0810107. URL http://content.nejm.org.

J. Lemeire. Learning causal models of multivariate systems. PhD thesis, Brussels, 2007.

A. J. London and J. B. Kadane. Placebos that harm: sham surgery controls in clinical trials. *Statistical Methods in Medical Research*, 11:413–427, October 2002.

A. C. Lozano, N. Abe, Y. Liu, and S. Rosset. Grouped graphical Granger modeling methods for temporal causal modeling. In *15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 577–586, Paris, France, 2009.

D. Margaritis and S. Thrun. A Bayesian multiresolution independence test for continuous variables. In *17th Conference on Uncertainty in Artificial Intelligence (UAI)*, Seattle, Washington, August 2001.

F. Markowetz. A bibliography on learning causal networks of gene interactions, March 2007. URL http://genomics.princeton.edu/~florian/docs/network-bib.pdf.

A. J. McMichael, O. M. Jensen, D. M. Parkin, and D. G. Zaridze. Dietary and endogenous cholesterol and human cancer. *Epidemiol Rev*, 6(1):192–216, 1984. URL http://epirev.oxfordjournals.org.

C. Meek. Strong completeness and faithfulness in Bayesian networks. Proceedings of *11th Uncertainty in Artificial Intelligence (UAI), Montreal, Canada,* Morgan Kaufmann, pages 411–418, 1995.

E. Miguel, S. Satyanath, and E. Sergenti. Economic shocks and civil conflict: An instrumental variables approach. *Journal of Political Economy*, 112(4):725–753, 2004. doi: 10.1086/421174. URL http://www.journals.uchicago.edu/doi/abs/10.1086/421174.

M. L. Minsky and S. A. Papert. *Perceptrons*. MIT Press, Cambridge, MA, 1969.

A. Moneta. Causality in macroeconometrics: some considerations about reductionism and realism. *Journal of Economic Methodology*, 12(3):433–453, September 2005. URL http://ideas.repec.org/a/taf/jecmet/v12y2005i3p433-453.html.

R. E. Neapolitan. *Learning Bayesian Networks*. Prentice Hall series in Artificial Intelligence. Prentice Hall, 2003.

G. Neves, S. F. Cooke, and T. V. P. Bliss. Synaptic plasticity, memory and the hippocampus: a neural network approach to causality. *Nature Reviews Neuroscience*, 9:65–75, 2008. URL http://www.nature.com/nrn/journal/v9/n1/abs/nrn2303.html.

A. Oniśko, M. J. Druzdzel, and H. Wasyluk. Application of Bayesian belief networks to diagnosis of liver disorders. In *Proceedings of the Third Conference on Neural Networks and Their Applications*, pages 730–736, Kule, Poland, 14–18 October 1997.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufman, San Mateo, California, 1988.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

J.-P. Pellet and A. Elisseeff. Using Markov blankets for causal structure learning. *JMLR*, 9: 1295–1342, 2008. ISSN 1533-7928.

D. Petitti. Commentary: Hormone replacement therapy and coronary heart disease: four lessons. *Int. J. Epidemiol.*, 33(3):461–463, 2004. doi: 10.1093/ije/dyh192. URL http://ije.oxfordjournals.org.

E. J. G. Pitman. Significance tests which may be applied to samples from any population. *Royal Statistical Society Supplement*, 4, 1937.

X. Qin and W. Lee. Statistical causality analysis of INFOSEC alert data. In *RAID*, pages 73–93, 2003.

H. Reichenbach. *The Direction of Time*. University of Los Angeles Press, Berkeley, 1956.

D. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

B. Russell. On the notion of cause. *Proceedings of the Aristotelian Society*, 13:1–26, 1913.

S. J. Russell and N. P. *Artificial Intelligence: A Modern Approach (Second Edition)*. Prentice Hall, 2003.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978. doi: 10.2307/2958889. URL http://dx.doi.org/10.2307/2958889.

S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-Gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.*, 7:2003–2030, 2006. ISSN 1533-7928.

G. D. Smith. Capitalizing on Mendelian randomization to assess the effects of treatments. *J R Soc Med*, 100(9):432–435, 2007. doi: 10.1258/jrsm.100.9.432. URL http://jrsm.rsmjournals.com.

D. J. Spiegelhalter, A. P. Dawid, S. L. Lauritzen, and R. G. Cowell. Bayesian analysis in expert systems. *Statistical Science*, 8(3):219–247, August 1993.

P. Spirtes. Variable definition and causal inference. In *13th International Congress of Logic Methodology and Philosophy of Science*, 2008.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. The MIT Press, Cambridge, Massachusetts, London, England, 2000.

H. Stoppiglia, G. Dreyfus, R. Dubois, and Y. Oussar. Ranking a random feature for variable and feature selection. *JMLR*, 3:1399–1414, 2003. URL http://www.jmlr.org/papers/v3/stoppiglia03a.html.

I. Tsamardinos and L. E. Brown. Bounding the false discovery rate in local Bayesian network learning. In *AAAI*, pages 1100–1105, 2008.

E. Tuv, A. Borisov, and K. Torkkola. Feature selection using ensemble based ranking against artificial contrasts. In *IJCNN*, pages 2181–2186, 2006.

P. Vinei and D. Kriebel. Causal models in epidemiology: past inheritance and genetic future. *Environ Health*, 5(21), 2006.

K. Wheatley and R. Gray. Commentary: Mendelian randomization–an update on its use to evaluate allogeneic stem cell transplantation in leukaemia. *Int. J. Epidemiol.*, 33(1):15–17, 2004. doi: 10.1093/ije/dyg313. URL http://ije.oxfordjournals.org.

S. Wright. Correlation and causation. *Journal of Agricultural Research*, 20:557–585, 1921.

Z. Yongli, H. Limin, and L. Jinling. Bayesian networks-based approach for power systems fault diagnosis. *IEEE Transactions on Power Delivery*, 21(2):634–639, April 2006.

G. Young. Causality and causation in law, medicine, psychiatry, and psychology: Progression or regression? *Psychological Injury and Law*, 1(3):161–181, 2008.

Glossary

Action: An intervention performed by an external agent to disrupt the normal functioning of a system, which would otherwise be left to evolve according to its own dynamics.

Causal Bayesian Network: A model frequently used in causal discovery, using a directed acyclic graph (DAG) to model causal relationships between random variables. Using the network, it is possible to infer the probability distribution of some variable given measured values of others.

Causal Faithfulness Condition (CFC): The CFC is the faithfulness condition applied to a causal model (see "faithfulness"). The CFC essentially states that independence implies absence of a causal arrow. Complying with the CFC excludes modeling the XOR problem and cases in which multiple paths compensate each other.

Causal Markov Condition (CMC): The CMC is the Markov condition applied to a causal model (see "Markov property" or "Markov condition"). The CMC essentially states that statistical dependency implies the existence of a causal arrow. See also "principle of common cause". Systems with hidden confounders violate the CMC. See also "causal sufficiency".

Causal sufficiency: Causal sufficiency essentially states that there are no hidden variable that is a common cause of two variables considered, i.e., no hidden confounder. This commonly made assumption is very difficult to verify and the presence of a hidden confounder may invalidate completely a study. See "confounder".

Cause (as system state or event): Informally, a cause can be defined as a state C of a system of interest consistently followed by another state E (its effect) whenever the system is (actually or hypothetically) forced to assume the state C. The eventual existence of unobservable state variables makes it possible that correlated events succeeding each other are not in a causal relationship: both may be the consequence of an earlier common cause. For instance, lightning may trigger both thunder, followed by a fire alarm. "Thunder" and "fire alarm" are the consequence of the common cause "lightning", but are not causally related, even though "thunder" might happen consistently before "fire alarm". This ambiguity could be resolved if an external agent could perform an experiment and force "thunder" to happen with or without "lightning". See also manipulation or action.

Cause (as random variable): If a random variable is an indicator of presence/absence of an event, causal relationships between random variables are simple extensions of causal relationships between events. More generally, causal relationships between random variables can be defined via manipulations. For instance, given two random variables C and E and a manipulation $do(C)$, a univariate causal relationship between C (cause) and E (effect) is found if $P(E|do(C)) \neq P(E)$. For instance, in a randomized clinical trial, C can be the amount of medicine taken and E the health status of the patient. If the health status of patients having taken the medicine differs from that of patients in the control group, a causal effect is detected.

Conditional independence (CI): Two random variables X and Y are conditionally independent of a third one *iff* $P(X,Y|Z) = P(X|Z)P(Y|Z)$. This may be extended to subsets of variables. Regular statistical independence is equivalent to conditioning on the empty set.

Confounded variable: An alleged cause of a target variable whose dependency with the target can be explained by the presence of a confounder (see "confounder").

Confounding factor or confounder: A variable that shows statistical dependencies to a target variable and its alleged cause and that may be a common cause to both, hence potentially making us confuse statistical dependence and causation.

Consequence, effect: The effect can be defined as the manifestation of the cause, see cause.

Counterfactual: An event contrary to the fact. Causality and counterfactuals are intimately tied together. Some authors argue that all causal statements can be phrased in terms of counterfactuals: "the throw of the stone caused the window to break" may be replaced by "had the stone not been thrown, the window would not have broken". Causal models allow us to predict what would have happened under a situation that did not occur (*e.g.,* "would the patient have died had he not taken the treatment").

Cross-validation (CV): A method frequently used in machine learning to select models, *e.g.,* with different architectures or hyper-parameters. One selects the model with the best CV performance, obtained by splitting repeatedly the available (observational) training data into training and validation set and averaging the prediction results on the validation sets. If observational data are used, CV is not a good method for selecting among alternative causal architectures.

Do-calculus: A method for calculating conditional probabilities of certain variables in a post-manipulation distribution given only conditional probabilities from the pre-manipulation distribution and some causal assumptions. The method was originally developed by Judea Pearl.

D-separation (D-connection): A set $\mathbf{C}$ is said to d-separate A from B if $\mathbf{C}$ blocks every path between A and B. If A and B are not d-separated, then they are d-connected. A path Π between two variables A and B is blocked by a set of nodes $\mathbf{C}$ if (1) Π contains a chain $I \to C \to J$ or a fork $I \leftarrow C \to J$ such that C is in $\mathbf{C}$, or (2) Π does not contain a collider $I \to C \leftarrow J$ such that C or any of its descendants are in $\mathbf{C}$. D-separation is an algorithm to compute all the conditional independence relations entailed by a Bayesian network or a SEM.

Endogenous variable: A variable having explicit causes within a particular causal model. The characterization depends on the set of variables under consideration and the chosen causal model. Complementary concept: exogenous variable.

Exogenous variable: A variable having no explicit causes within a particular causal model. The characterization depends on the set of variables under consideration and the chosen causal model. Complementary concept: endogenous variable.

Experiment: Planned manipulations designed to determine causal relationships (see also "randomized controlled trial").

Experimental data: Data collected as a result of an experiment (see also "observational data").

Faithfulness: In the Bayesian network framework, a graph is faithful to a distribution if all the conditional independencies entailed by the distribution are reflected by Markov properties that can be read from the graph (see

"Markov property"). A distribution is faithful if there exists a faithful graph representing it.

Features: Variables potentially predictive of the target variable, also called *covariates*, *explanatory variables* , or *predictor variables* in statistics.

Ground truth: In the pattern recognition jargon, "ground truth" refers to verified information obtained by scouting the terrain *on the ground* as opposed to information collected from far away observations, like satellite images.

I.i.d: Independent and identically distributed. A common assumption about the data distribution in machine learning, which assumes a stationary data generating process. This assumption is violated when external agents perform manipulations on the system.

Inference: There are two types of inference: model inference and variable inference (see the corresponding definitions).

Instrumental variable: A variable I used to test an alleged causal relationship $C \to E$ by performing a "natural manipulation" of C. It must be known that I is exogenous and cannot influence E in any other way than though C.

Latent variable: An unobserved (hidden) variable, possibly unknown.

Manipulation: A set of actions performed by an external agent on a system under study to disrupt the normal functioning of a system. A manipulation of a random variable C denoted as $do(C)$ consists in making C assume values according to a distribution decided by the agent, distinct from the "natural" distribution of C conditioned on the other variables of the system.

Markov blanket and Markov boundary (MB): A Markov blanket of a target variable (called MB) is a sufficient set of variables such that all other variables are independent of the target, given MB. A minimal Markov blanket is called a Markov boundary. Under some conditions, the Markov boundary is unique. Under the faithfulness assumption (see "faithfulness") it coincides with the set of parents, children, and spouses of the target. Many people include the minimality restriction in the definition of Markov blankets, therefore identifying the Markov blanket and the Markov boundary.

Markov property and Markov condition: A stochastic process of random variables has the Markov property if its future states are independent of far away past states given the present and a finite number of near past states (*i.e.,* it is memoryless). All Markov processes have an equivalent first order Markov process in which future states are independent of past states given the present state. By extension, atemporal Bayesian networks and SEMs are (first order) Markov models in the sense that each node is independent of its non-descendants given its parents. This is also called the "Markov condition". For these models, a "Markov property" is a conditional independence property between a subset of variables. "Markov properties" read from the graph (see "d-separation") are all valid conditional independence properties.

Model inference, model fitting, training: In a learning problem, inference refers to choosing the model, its structure, hyper-parameters and parameters.

Model over-fitting: Training a model to make excellent predictions for training examples, but obtaining poor prediction performance on test examples.

Natural distribution: Synonym of "observational distribution" or "pre-manipulation" distribution.

Non-interventional observations: See "observational data".

Observational data: Data collected from the observation of a system let to evolve according to its own dynamics, without controlled intervention (see also "experimental data").

Observational distribution: The joint distribution of the variables of a system in the absence or any external perturbation. Also called "pre-manipulation distribution".

Pre-manipulation distribution: Same as "observational distribution".

Principle of Common Cause (PCC): The PCC states that if two variables are correlated but neither is the cause of the other, then there should be at least one common cause influencing both variables.

Post-manipulation distribution: The joint distribution of the variables of a system after an action was performed by an external agent.

Predictive model, predictor: A mathematical construct $y = f(x;a)$ parameterized by a parameter vector a, allowing to make predictions of an outcome y given an input datum x.

Randomized Controlled Trial (RCT): Planned experiments involving a random allocation of different interventions (treatments or conditions) to subjects. As long as the numbers of subjects are sufficient, this ensures that both known and unknown confounding factors are evenly distributed between treatment groups. There are many variants of RCTs including various blinding and randomizing techniques (see also "experiment").

Structural Equation Model (SEM): A model to represent causal relationships as a directed acyclic graph (DAG), similar to a Bayesian network, but in which variables are interconnected by functional relationships (eventually altered by stochastic noise) rather than conditional distributions. Noise variables are called "exogenous"; other (dependent) variables are called "endogenous". (See "exogenous variables" and "endogenous variables").

Target variable (or target): The outcome under study.

Variable inference: A trained model (e.g. a Bayesian network) can then be used to infer variable probability distributions from the partial knowledge of other variables.