# Probabilistic business analytics

The boom of Machine Learning and AI over the past decade has led to amazing achievements in many fields such as Computer Vision, Natural Language Processing, and Predictive models and fueled the advent of Data Science as a key method in Business consulting. Data Science projects typically focus on conventional Machine Learning methods that solve problems of prediction by finding patterns of association in large datasets, i.e., they answer questions such as: how can I predict the sales of my product based on data observed in the past?
Typical business problems, however, ask questions of interventions, i.e., they seek actions that lead to desired outcomes, e.g., in pricing (How should we set the price of my product to increase sales?), process optimization (How can we optimize our supply chain? ) or customer service (How can we prevent customer attrition?) to name just a few examples. To answer such questions, we need to employ Causal Inference, a method that goes beyond conventional Machine Learning by estimating causal effects between actions (interventions) and outcomes.

Moreover, the influence of uncertainty is often neglected or underestimated in Data Science applications. We think that considering uncertainty in data and business processes is essential, not only because it improves the reliability of analyses, but also because it enables nuanced decisions that calibrate the risk involved in a decision.
We here describe a principled workflow that ties together Causal Inference (causal effects), Probabilistic Programming (statistical modeling), and Bayesian Decision Making (optimal decision making) to tackle business problems with a Data Science approach (Figure 1). These technologies are increasingly gaining traction in the industry [1-5].

## A principled workflow

Our workflow consists of three phases: we start by using Causal Inference to define the problem and leverage domain expertise to make assumptions about causal relationships between variables in our problem.

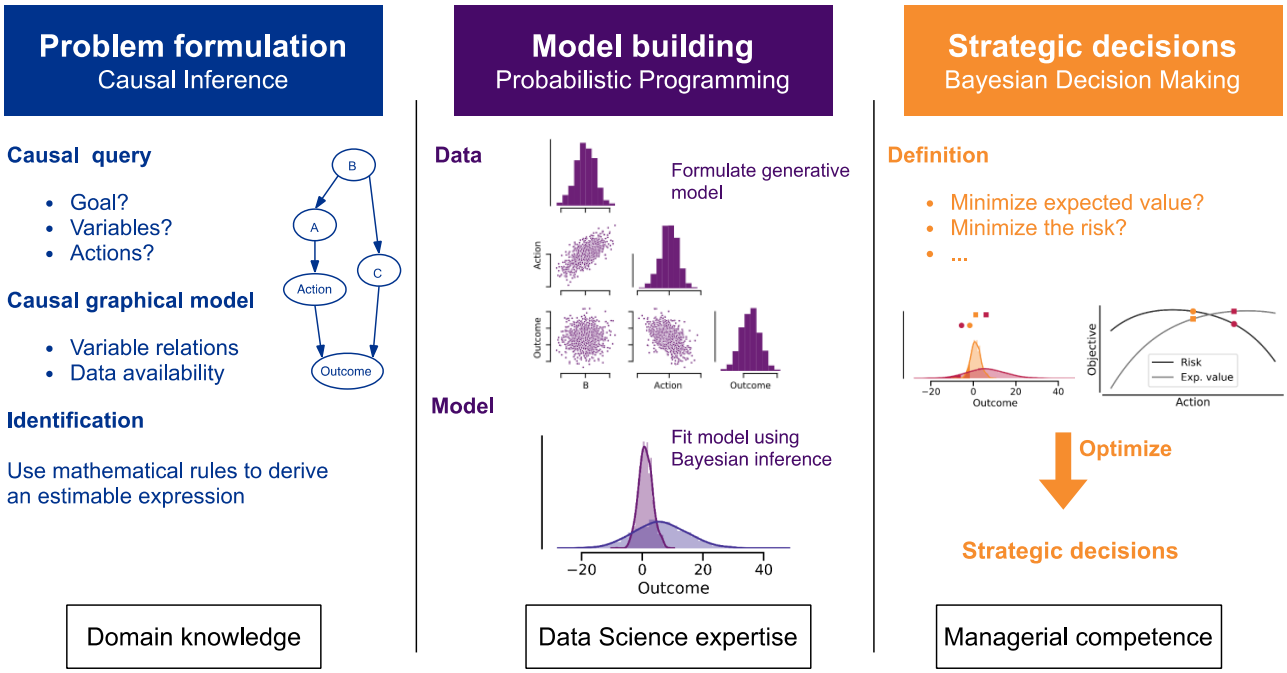


Figure 1: A principled workflow of probabilistic business analytics



© 2021 KPMG Ignition Tokyo, Inc., a company established under the Japan Companies Act and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

This leads to an estimable expression, a relationship that can be estimated with Machine Learning. In the second phase, we build a model to estimate this expression using Probabilistic Programming, a technique that allows us to incorporate uncertainties into the model. In the final step, we use this model in Bayesian Decision Making to find the optimal strategic decision under an objective defined with the client.

### Problem formulation with Causal Inference

We first define the problem that to be solved: what are the variables in the system? What are the possible actions we can take, i.e., which variables can be controlled by the client? What is our goal? We formulate this goal as the probability of an outcome given that a particular action is chosen, called an interventional distribution. Next, we leverage domain expertise of the client to make assumptions about the causal relations between the variables in the system: which variables influence which variables? These assumptions are summarized in a graphical representation, a causal graphical model (CGM).

To directly estimate the interventional distribution, we would need to run an experiment and test the impact of different actions. Since this is usually not possible in a business context, in the last step we convert the interventional distribution into an observational distribution, that can be estimated from observed (past) data alone. This step is called identification and is automated using the do-calculus, a set of mathematical rules that use the CGM to find out if the conversion is possible and, if possible, derive the observational distribution.

### Model building with Probabilistic Programming

In this stage, we build a model to estimate the expression identified in the first stage.

The thinking is generative: we imagine how the data could have been generated and formulate this as a probabilistic program; similar to a (deterministic) computer program, an algorithm computes output values. However, variables are stochastic: they are drawn from probability distributions that capture our uncertainty about the real process and the uncertainty in the data. The resulting model delivers an estimate of the question we defined in the first stage (what is the probability of an outcome given that we choose a particular action?).

### Deriving strategic decisions with Bayesian Decision Making

In business problems, we want to find an optimal decision for the action, i.e., an action that optimizes an objective function of the output variable(s). The objective function reflects a management decision: the client can choose to be more risk tolerant by e.g., maximizing the expected outcome or more conservative by minimizing the risk of a bad outcome. Bayesian decision making allows us to take such different routes because it incorporates the uncertainty in our problem and explores all possible outcomes triggered by a chosen action.

### Causal Inference is an essential tool in Business Analytics

Causal inference allows us to identify and quantify true causal effects between variables. While conventional statistics and Machine Learning focus on associative relations and ask questions like "What can I say about X given that I know the value of Y?", Causal Inference asks "How does Y cause X?" and "What happens to X if I set Y to a value?" In other words, Causal Inference estimates the effect of actions and is therefore important in Business Analytics where we seek decisions to influence outcomes of the business.

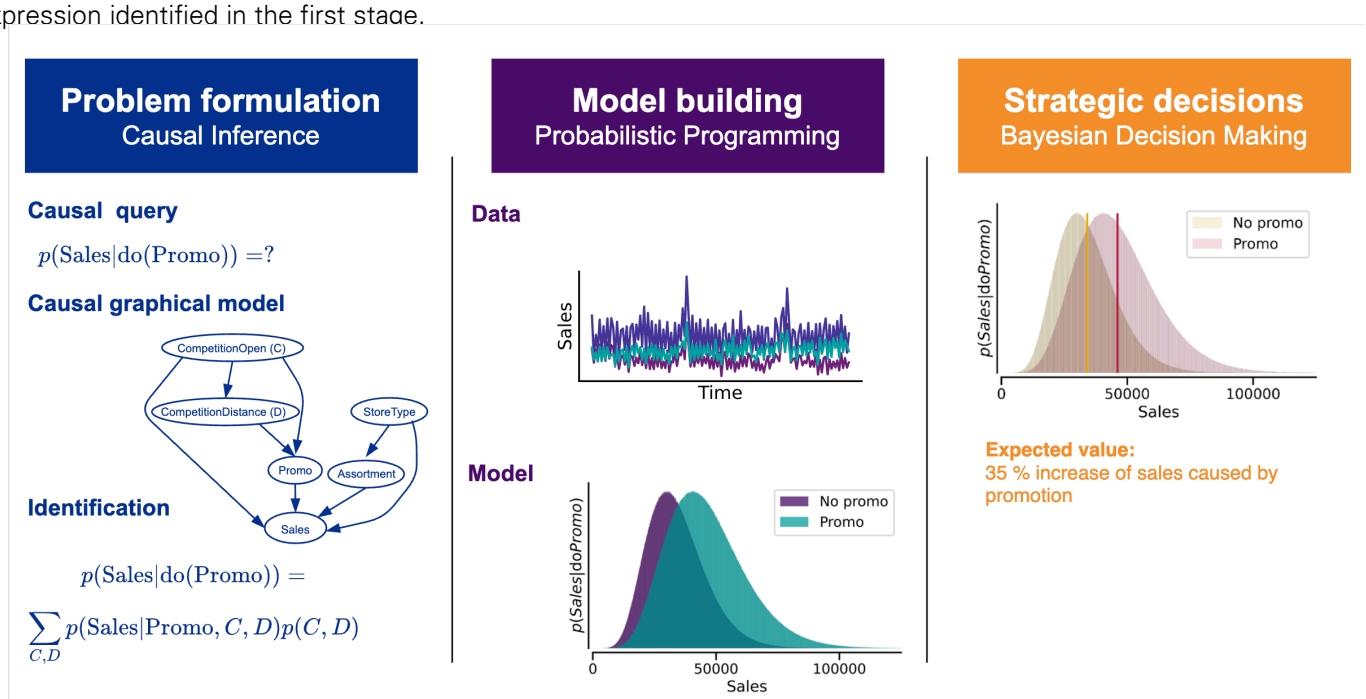




Figure 2: The principled workflow applied to find the causal effect of a promotion on sales

© 2021 KPMG Ignition Tokyo, Inc., a company established under the Japan Companies Act and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

It is a well-known proverb in statistics that Correlation is not Causation because spurious correlations in the observational data can distort the picture. One situation where this is true is when the interaction between the action and outcome of interest is overlaid by additional factors affecting both. Consider the case where a business has used a promotion campaign in its stores in the past and wants to know whether those promotions have increased sales (Figure 2).

In the first stage of our workflow, we find out that the presence of a competitor close to stores has influenced the decision to do a promotion in the past: if there is competitive pressure from a neighboring store, we are more likely to use a promotion to try to boost sales. At the same time, stores with such high competition probably have lower sales than unrivaled stores.

The presence of a competitor thus influences both the decision to run a promotion campaign and the sales. Therefore, to find the true causal effect of the promotion campaign on sales, we need to adjust for the presence of competitors in our analysis to get an unbiased view of the causal effect of the promotion on sales.

Defining the problem in a systematic way makes the potential biasing effect of competition on the analysis clear and simple to handle.

With Causal Inference we identify an expression that we can estimate using observed data, rather than having to run an experiment in the wild, which would be infeasible in this case. The workflow encourages explicit definition of variables and their relations, facilitating both the analysis as well as communication with clients.

## Probabilistic modelling allows us to make more informed decisions

Businesses operate in an uncertain world: demands of customers fluctuate over time, business relations pose risks (provided goods can be faulty, supply chains unstable), and global factors can have a large effect (overall economic situation, catastrophic events like natural disasters and pandemics). Consequently, we need to account for such uncertainty when modeling a business problem.

The framework of Probabilistic Programming allows us to do this in a simple and systematic manner. It yields probabilistic predictions for likely outcomes based on different choices for the actions in the problem. This enables us to define flexible objectives with the client: do they want to optimize the expected outcome? Or do they want to be more conservative and minimize the risk, i.e., minimize the possibility that undesirable outcomes happen?

Let's consider the problem of choosing the optimal number of vehicles to run a supply chain (Figure 3): on the one hand, the client wants to make sure to have sufficient capacity to satisfy the demands of our customers in the future. On the other hand, they strive to minimize the maintenance cost of our vehicle fleet. Balancing these two opposing interests requires a good forecast of customer demands to predict the probability of costs given a certain number of vehicles. After building a model and obtaining an estimate of the costs, the client has a choice between being more risk tolerant by minimizing the expected value of costs or more conservative by minimizing the risk of facing high costs. In our example, those choices lead to different decisions: choosing three vehicles leads to a lower expected value of costs by saving maintenance costs, while employing four vehicles leads to less risk because the client is more likely to satisfy customer demands with the increased capacity.

These are strategic choices that our approach enables the client to make: by considering all possible futures (with their associated probabilities) instead of focusing on one possible scenario for the future, we can distinguish those choices and offer the client a more informed view of the problem.

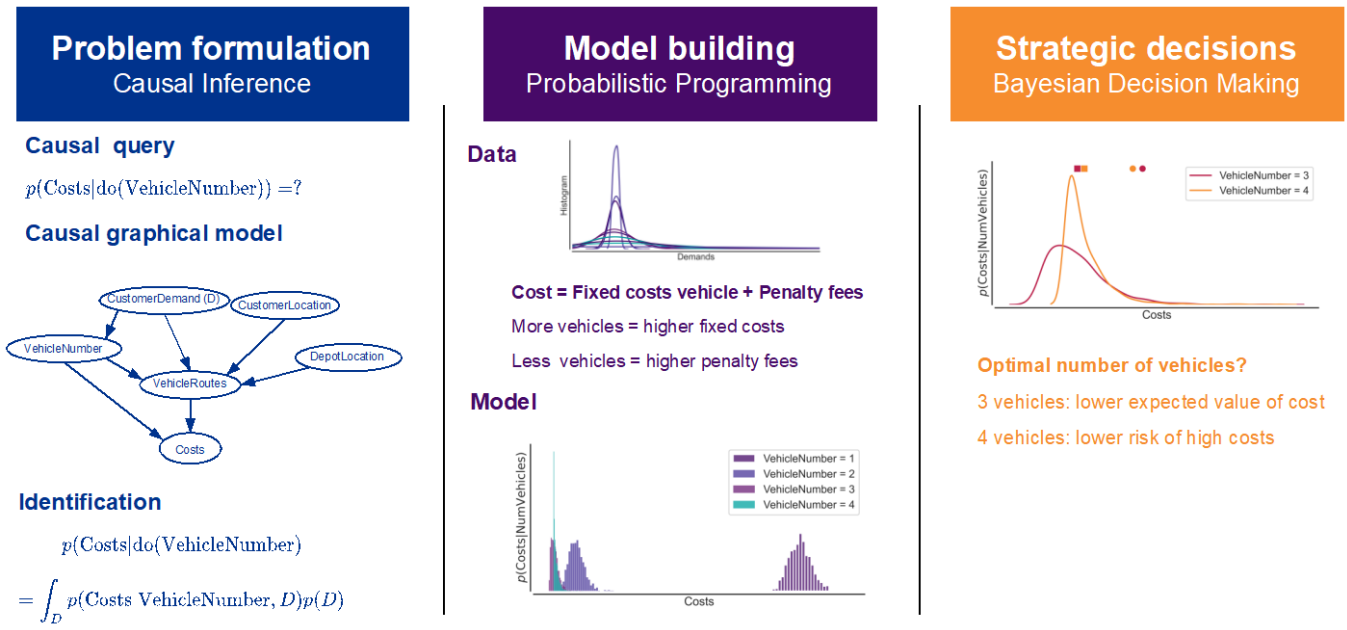


Figure 3: The principled workflow applied to find the optimal number of vehicles in a supply chain



© 2021 KPMG Ignition Tokyo, Inc., a company established under the Japan Companies Act and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

## Conclusion

We present a principled workflow that ties together the methods of Causal Inference, Probabilistic Programming, and Bayesian Decision Making to tackle complex business problems. With this workflow, we address the most important challenges in these problems: identifying causal effects of actions onto outcomes, producing flexible and accurate forecasts of future developments, and making strategic decisions based on those forecasts.

## Glossary

**Bayesian Decision Making**: The process of making a decision based on the likelihood of a successful outcome which is informed both by prior knowledge and data-driven models.

**Probabilistic Programming**: A method to build a generative model of data by formulating it as a computer program.

**Estimable expression**: A mathematical expression that can be estimated from observational data alone, i.e., does not require to run an active experiment.

**Objective function**: A mathematical formulation of a high-level objective such as e.g., "we want to maximize the expected value of sales".

**Observational distribution**: A probability distribution that quantifies the probability of an event given that the values of other variables are observed.

**Interventional distribution**: A probability distribution that quantifies the probability of an event given that we intervene on other variables, i.e., actively set them to some desired values.

**Causal graphical model**: A graphical formulation of the causal relationships between variables, where variables are nodes and their causal relationships are represented by arrows between them.

## References


[1] Silvia Chiappa and William Isaac. Causal bayesian networks: A flexible tool to enable fairer machine learning, 2019.
URL: https://deepmind.com/blog/article/Causal_Bayesian_Networks

[2] Totte Harinen and Li Bonnie. Using causal inference to improve the uber user experience.
URL: https://Eng.Uber.Com/Causal-Inference-At-Uber/

[3] Paul Hünermund, Jermain Kaminski, and Carla Schmitt. Causal machine learning and business decision making, 2021. URL: https://causalscience.org/blog/causal-data-science-in-practice

[4] Cloudera Fast Forward Labs. Causality for machine learning, 2020.
URL: https://ff13.fastforwardlabs.com/

[5] Jeffrey Wong and Colin McFarland. Computational causal inference at Netflix, 2020.
URL: https://netflixtechblog.com/computational-causal-inference-at-netflix-293591691c62


## Contacts


**KPMG Ignition Tokyo**

Otemachi Building 7F,
1-6-1 Otemachi, Chiyoda-ku,
Tokyo 100-0004 Japan

**home.kpmg/jp/KIT**

**Maximilian Schmidt**

Senior Engineer
DX Solution 2
+81 70 1043 7493
maximilian.schmidt@jp.kpmg.com

**Shungo Miyabe**

Senior Scientist
Smart Transaction



The information contained herein is of a general nature and is not intended to address the circumstances of any particular individual or entity. Although we endeavor to provide accurate and timely information, there can be no guarantee that such information is accurate as of the date it is received or that it will continue to be accurate in the future. No one should act on such information without appropriate professional advice after a thorough examination of the particular situation.

© 2021 KPMG Ignition Tokyo, Inc., a company established under the Japan Companies Act and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved. 21-5029

The KPMG name and logo are trademarks used under license by the independent member firms of the KPMG global organization.