Xplain Data
Discover Causality

WHITE PAPER

HOW XPLAIN DATA HELPS TO DETERMINE CAUSALITIES

FROM OBSERVATIONAL DATA TO CAUSAL DISCOVERY

Xplain Data GmbH
Grünlandstr. 27
85604 Zorneding
info@xplain-data.com
xplain-data.com

WHAT TO EXPECT

SUMMARY 3

CORRELATION DOES NOT IMPLY CAUSATION 4

OBSERVATIONAL DATA VS. EXPERIMENTAL DATA 5

CONFOUNDERS AS CRITICAL FACTORS 6

CONCLUSIONS 6

CAUSAL DISCOVERY GETS TO THE BOTTOM OF YOUR DATA 7

WHO BENEFITS FROM THE XPLAIN DATA CAUSAL DISCOVERY PROCESS? 8

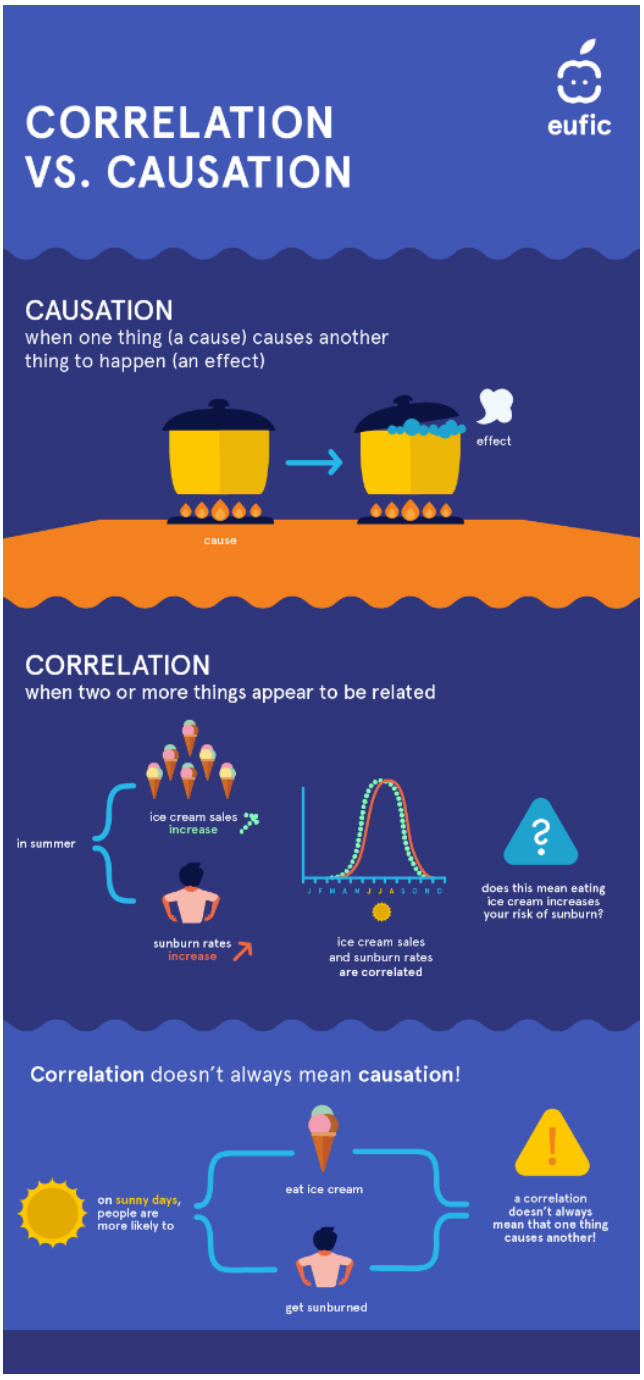
SUMMARY

"Big Data" in general means "observational data" - in other words, data has not been collected under controlled conditions of an experiment but purely observationally without intervention. The analysis of such data – it's a challenge! In particular, observed correlations in those data may not be misinterpreted as causal relationships. Knowledge about cause and effect, however, is key for targeted interventions into a system, e.g., for intelligent process control which aims at avoiding observed manufacturing defects in future operation. What is required to determine potential cause-and-effect relationships from observational data - and what does Xplain Data contribute to intelligently use causal knowledge?

That is how Gartner defines "Big Data". But how do we get from Big Data to valuable insights? This white paper explains, by way of example, the difference between observational data and data collected under experimental conditions. Thereby the challenges posed by observational data will become evident, especially with respect to understanding cause and effect. We will explain what is needed to identify potential cause-and-effect relationships from such data, and what solution Xplain Data offers to address this problem. A key to that is our Object Analytics approach, which facilitates the analysis of large volumes of complex observational data - so-called "Big Data".

"BIG DATA IS HIGH-VOLUME, -VELOCITY AND -VARIETY INFORMATION ASSETS THAT DEMAND COST-EFFECTIVE, INNOVATIVE FORMS OF INFORMATION PROCESSING FOR ENHANCED INSIGHT AND DECISION MAKING."

Gartner Group, Big Data Definition



CORRELATION DOES NOT IMPLY CAUSATION

Based on so-called "observational data" we cannot prove cause-and-effect relationships. As an example: Health insurance companies collect data about their patients; they simply collect what they get – observational data with no interventions. Those data may show that the number of painkillers prescribed correlates with the frequency of depression - the more painkillers a patient receives, the more frequently depression is diagnosed later. Does this mean that painkillers cause depression?

Proof of causality requires experimentation. An engineer, for example, might do an experiment and measure the failure rate of a part in dependence of the operating temperature. From that he may conclude at what temperature failures are to be expected. In healthcare, such experiments are typically randomized controlled trials (RCTs).

In both cases, the experimental setup ensures that there is comparability between different test groups. The engineer will evaluate exactly identical parts at different temperatures (usually brand-new parts). In healthcare, randomization ensures that the two groups are comparable on statistical average.

OBSERVATIONAL DATA VS. EXPERIMENTAL DATA

"Big Data" typically is not such "lab data" collected in an experiment with controlled comparison groups. Big Data in general means observational data. Comparing different groups in such data bears the risk of comparing "apples and oranges".

As mentioned above, causality cannot be proven based on observational data – indeed in that case the term "causation" cannot be defined in a rigorous mathematical way. Nevertheless, we may obtain more or less strong indications on potential causal relationships. For this, we rely on a definition of Kenny that includes the following two important elements:

DAVID A. KENNY[1] (1979):

- **The effect must always follow the cause in time.** Cause and effect has to do with "before and after" and an observed change after the occurrence of a cause.

- **No "other explanation" for the observed relationship may exist.** In our case of "painkillers and depression", we suspect an "other explanation" (other than a direct causal effect) for the observed relationship: the age. The older a patient is, the more frequently we will find an analgesic therapy, and at older patients also depression is more prevalent. This explains the frequent joint observation of these two variables, even without a causal relationship.

To identify age as a so-called confounder (common explanation), the age of the patient needs to be available in the data. Missing information on important confounders leads to flawed conclusions about cause and effect. If – without considering age – we compare patients receiving painkillers with those not receiving painkillers, we implicitly compare older patients with younger ones (i.e., we compare "apples and oranges").

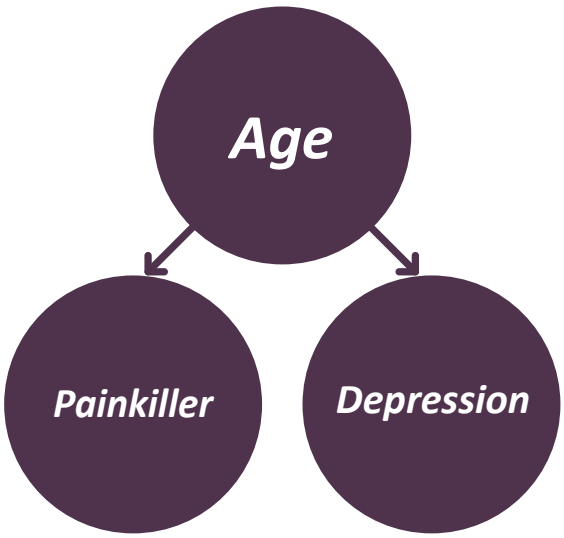


Figure 1 - „Age" indirectly explains the link between painkillers and depression - age is said to be a „confounder," a common cause of both.

[1]D. A. Kenny: Correlation and Causality, New York: John Wiley & Sons, 1979

CONFOUNDERS AS CRITICAL FACTORS

There may be many possible factors to consider as common explanation for taking painkillers and depression: many severe diseases go along with a pain therapy - and at the same time, serious illness often makes patients depressed. Drinking of alcoholic beverages may result in the need of a pain killer the next day, and alcohol abuse long-term causes depression… we may envision many factors causing both. All these potential "confounders" need to be evaluated – if none of them (or no combination of these factors) explains an observed relationship, only then a reasonable suspicion arises on a potential direct cause-and-effect relationship.

CONCLUSIONS

A. To obtain evidence on potential cause-and-effect relationships we need very comprehensive data, ideally a 360° view of the analyzed object including all historical data. Only then we will be able to discern trivial correlations from potential cause-and-effect relationships. The potentially causal part of an observed correlation is the part that cannot be explained via other available factors.

B. Extensive data is worthless unless there are high-performance algorithms that can search such data for confounders and evaluate "all alternative explanations". Note that comprehensive data typically means complex data, i.e., those algorithms must be able to process data far beyond a flat table.

Xplain Data is committed to these two issues: Our Object Analytics Database provides an analytical access to a complex object – across all data attached to the object (including recursive sub-objects).

In healthcare, the so-called "root object" is usually the patient, with millions of stored individual patient examples (object instances) and several billion events tied to them, such as prescribed drugs, diagnoses, performed procedures, measurements of clinical parameters, etc.

All these different data streams may be hooked to the root object into a joint object model – even data from different sources. With that, all the different data areas of a patient can be analyzed in relation to each other. Often only then the value of such data is realized. (For Object Analytics see also the corresponding White Paper. The approach is protected by our US and EP patent.)[2]

[2]US Patent Nr. 11,194,811 B2 / EP Patent EP3460680

CAUSAL DISCOVERY GETS TO THE BOTTOM OF YOUR DATA

Our “Causal Discovery” algorithms build upon Object Analytics. Based on that, millions of hypotheses for “alternative explanations” are automatically formed to explain an observed relationship indirectly (via confounders). Only if – in the wealth of data hooked into an object model – nothing can be found which explains an observed correlation between two variables, the corresponding factor is suggested as a potential direct causality. The scope of available data and the strength of the search algorithms determine the quality of the result.

We still term the detected factors as “potentially” causal - because even with very comprehensive data there is no proven evidence.

Consequently, hypotheses – proposed by the algorithm – need to be evaluated and interpreted by domain experts. In the assessment, there must be room for follow-up questions, for example, as to why an expected causal factor was not algorithmically identified. Vice versa, the expert needs to be able to reject an identified factor thereby requesting a second-best proposal, which may be more intuitive to understand.

This is where we arrive at the third conclusion:

C. Algorithmic intelligence must be complemented by a user interface that presents hypotheses for potentially causal factors in an intuitively understandable way. Experts need to be able to pose questions and follow-up questions, thereby getting their expectations and expert knowledge aligned with knowledge from data.

Xplain Data meets this requirement with a user interface that enables questions and follow-up questions to be answered in an interactive process.

# Cases	Factor explaining Target	Target Probability	Contribution	Impact
81	# Prescriptions(G03F - Progestogens and estrogens in combination,]-oo,-162w[, within time constraints) …	13,58%	10,40 %	8
79	# Diagnoses([-18w,-6w[, N63 - Unspecified lump in breast (4), within time constraints) >= 1	8,86%	7,06 %	6
83	# Diagnoses(All durations, C795 - Secondary malignant neoplasm of bone and bone marrow, within time …	9,64%	6,85 %	6
123	# Diagnoses([-162w,-54w[, I97 - Postprocedural disorders of circulatory system, not elsewhere classified,…	9,76%	6,84 %	8
98	# Diagnoses([-18w,-6w[, C56 - Malignant neoplasm of ovary (4), within time constraints) >= 1	10,20%	6,02 %	6
81	# Diagnoses(All durations, M438 - Other specified deforming dorsopathies, within time constraints) >= 1	7,41%	5,92 %	5
96	# Diagnoses(]-oo,-162w[, M76 - Enthesopathies of lower limb, excluding foot, within time c			
95	# Diagnoses([-18w,-6w[, M46 - Other inflammatory spondylopathies, within time constraint			
138	# Prescriptions(All durations, A04A - Antiemetics and antinauseants, within time constraint			
125	# Prescriptions([-18w,-6w[, D01AC20 - imidazoles/triazoles in combination with corticoster			
92	# Diagnoses([-18w,-6w[, N302 - Other chronic cystitis, within time constraints) >= 1			
84	# Prescriptions(G03CA - Natural and semisynthetic estrogens, plain, [-162w,-54w[, within t			
117	# Diagnoses(]-oo,-162w[, F439 - Reaction to severe stress, unspecified, within time constr			
133	# Diagnoses([-18w,-6w[, J38 - Diseases of vocal cords and larynx, not elsewhere classifie			
146	# Diagnoses(]-oo,-162w[, N60 - Benign mammary dysplasia, within time constraints) >= 1			
172	# Diagnoses(All durations, L270 - Generalized skin eruption due to drugs and medicamen			
272	# Diagnoses(All durations, Z26 - Need for immunization against other single infectious dise			
364	# Prescriptions(C08 - Calcium channel blockers,]-oo,-162w[, within time constraints) >= 1			

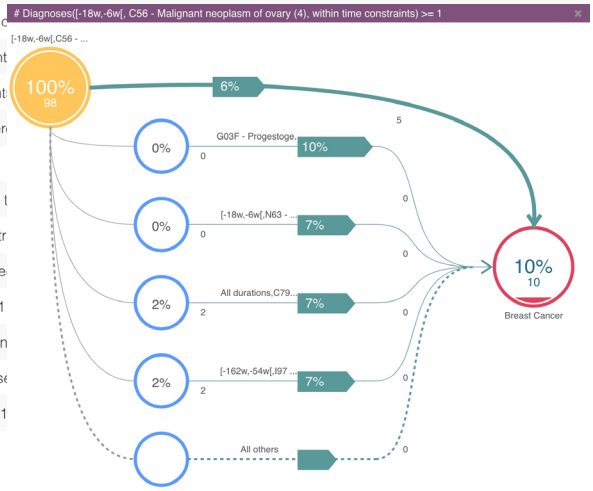


Figure 2 - example results: presentation of the factors found in an interactive table. Selection of a single factor shows how the overall effect is composed of a direct, potentially causal contribution of the factor and indirect (via other factors).

WHO BENEFITS FROM THE XPLAIN DATA CAUSAL DISCOVERY PROCESS?

Typical areas of application are analyses in the pharmaceutical and healthcare environment as well as in the manufacturing industry.
In healthcare, so-called “real-world data” plays an increasingly important role. While randomized controlled trials (RCTs) are expensive, real-world data emerges as a byproduct of electronically supported processes and becomes easily available on an increasing scale. Also, RCTs are one thing, but real-world environments might be something different. It is becoming increasingly important to understand effects of treatments under real-world conditions.

And it’s not much different in the manufacturing industry. Production devices are generating more and more data, and data is becoming available up- and downstream along the entire supply chain. As manufacturing processes evolve in complexity, it becomes increasingly important to understand causes for defectively produced parts, or why parts fail when they are later mounted in other devices and operated under real-world conditions.

In a project with **Schwäbische Werkzeugmaschinen GmbH (SW)** in the discrete manufacturing industry, the benefits of Xplain Data Causal Discovery algorithms immediately became evident. Failed parts could be traced back to quickly revealed causes for failures. New data is now constantly analyzed to detect upcoming failure causes as soon as possible.

If you would like to learn more about the benefits of Causal Discovery or you are interested in working with us, contact us at:

Xplain Data
Discover Causality

Xplain Data GmbH
Grünlandstr. 27
85604 Zorneding
Germany
info@xplain-data.com
xplain-data.com