Xplain Data
Discover Causality

WHITE PAPER



WIE XPLAIN DATA HILFT KAUSALITÄTEN ZU ERMITTELN

VON BEOBACHTUNGSDATEN ZU CAUSAL DISCOVERY

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

WAS SIE ERWARTET

ZUSAMMENFASSUNG 3

KORRELATION IMPLIZIERT NICHT KAUSALITÄT 4

BEOBACHTUNGSDATEN VS. EXPERIMENTELLE DATEN 5

CONFOUNDER ALS KRITISCHE FAKTOREN 6

SCHLUSSFOLGERUNGEN 6

CAUSAL DISCOVERY GEHT IHREN DATEN AUF DEN GRUND 7

WER PROFITIERT VOM XPLAIN DATA CAUSAL DISCOVERY PROZESS? 8

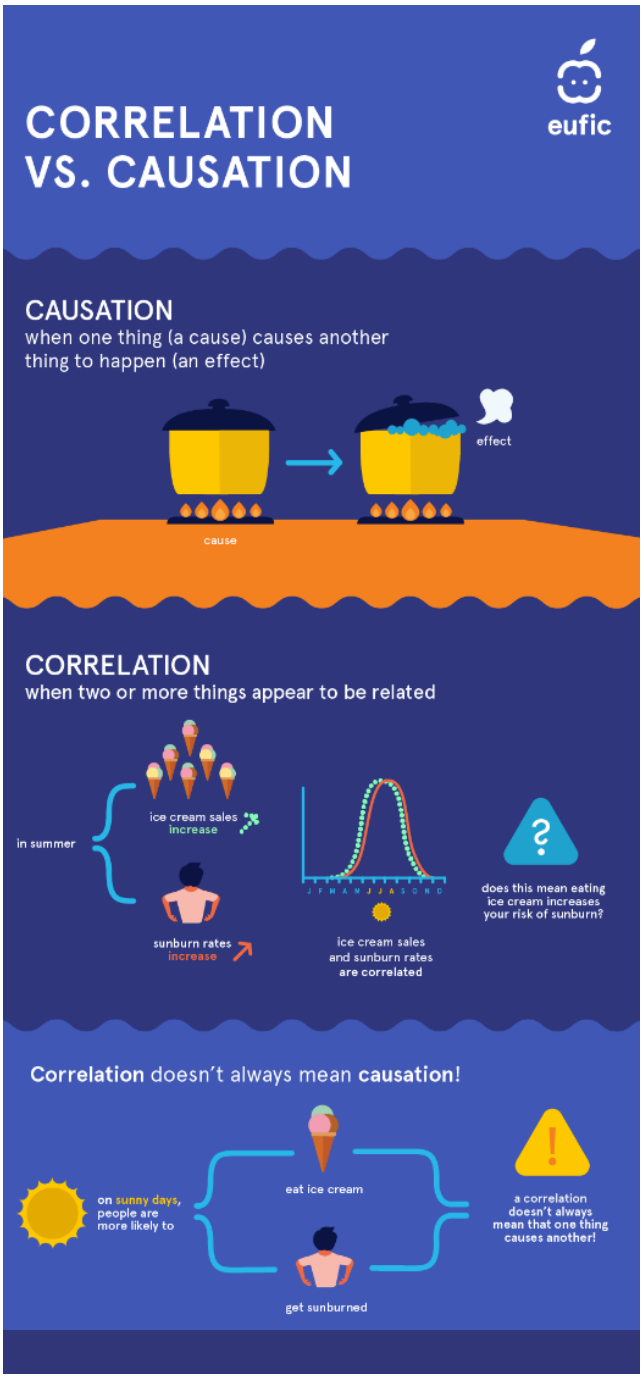
ZUSAMMENFASSUNG

„Big Data“ sind in der Regel „Beobachtungsdaten“ – also Daten, die nicht unter den kontrollierten Bedingungen eines Experimentes erhoben wurden, sondern rein beobachtend ohne Intervention. Die Analyse solcher Daten? Eine Herausforderung! So z. B. die Unterscheidung von Korrelationen und tatsächlich ursächlichem Zusammenhang zwischen zwei Größen. Eine maximale Transparenz über Ursache und Wirkung ist für die intelligente Steuerung von Prozessen, Systemen und Anlagen aber essenziell, um unerwünschte Zustände oder beobachtete Produktionsfehler künftig zu vermeiden. Was ist notwendig, um potenzielle Ursache-Wirkungs-Zusammenhänge aus Beobachtungsdaten zu ermitteln - und welchen Beitrag leistet Xplain Data zu deren sinnvoller und intelligenter Nutzung?

Doch wie kommen wir tatsächlich von „Big Data“ zu wertvollen Einsichten? Dieses White Paper erklärt beispielhaft den Unterschied zwischen Beobachtungsdaten und Daten, die unter experimentellen Bedingungen erhoben werden und wirft ein Auge auf die Herausforderungen, die reine Beobachtungsdaten mit sich bringen. Es wird erklären, was notwendig ist, um potenzielle Ursache-Wirkungs-Zusammenhänge aus solchen Daten zu ermitteln, und welchen Beitrag Xplain Data leistet, um große Mengen komplexer Beobachtungsdaten – sogenannte „Big Data“ – sinnvoll und intelligent zu nutzen.

“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



KORRELATION IMPLIZIERT NICHT KAUSALITÄT

Anhand von sogenannten „Beobachtungsdaten“ lassen sich Ursache-Wirkungs-Zusammenhänge nicht beweisen. Ein Beispiel: In den Daten einer Krankenversicherung zeigt sich, dass die Anzahl verschriebener Schmerzmittel mit der Häufigkeit einer Depression korreliert – je mehr Schmerzmittel einem Patienten verschrieben werden, desto häufiger wird auch eine Depression diagnostiziert.

Trotzdem gilt: Aus diesem **beobachteten** Zusammenhang lässt sich **nicht** auf eine Kausalität schließen – Schmerzmittel verursachen keine Depressionen.

Ein Beweis einer Kausalität bedarf eines Experiments. So kann beispielsweise der Zusammenhang zwischen der Betriebstemperatur eines Bauteils und dessen Ausfallhäufigkeit gemessen werden, um zu ermitteln, ab welchen Temperaturen Ausfälle zu erwarten sind. Experimente im Bereich der Gesundheit sind typischerweise randomisierte Studien.

Mit der Form der Erhebung der Daten wird sichergestellt, dass zwischen unterschiedlichen Testgruppen eine Vergleichbarkeit besteht.
So wird ein Ingenieur das exakt gleiche Bauteil bei unterschiedlichen Temperaturen vermessen. Im Bereich Gesundheit wird über die Randomisierung sichergestellt, dass die beiden Gruppen im statistischen Mittel vergleichbar sind.

BEOBACHTUNGSDATEN VS. EXPERIMENTELLE DATEN

„Big Data“ sind typischerweise keine „Labordaten“, die in einem Experiment unter kontrollierten Bedingungen erhoben wurden, sondern sogenannte Beobachtungsdaten. Bei Vergleichen von unterschiedlichen Gruppen in solchen Daten besteht die Gefahr, „Äpfel mit Birnen“ zu vergleichen.

Wie eingangs erwähnt, können Kausalitäten anhand von Beobachtungsdaten nicht bewiesen oder gar mathematisch definiert werden. Nichtsdestotrotz lassen sich damit Hinweise auf kausale Zusammenhänge gewinnen – und diese in ihrer Stärke vergleichen.

David A. Kenny (1979) fordert in seiner Definition von Kausalität[1] unter anderem:

- **Die Wirkung muss der Ursache immer zeitlich folgen.** Ursache und Wirkung haben mit „vorher und nachher“ bzw. einer beobachteten Veränderung nach dem Auftreten einer Ursache zu tun.

- **Es darf „keine andere Erklärung“ für den beobachteten Zusammenhang geben.** Im Falle „Schmerzmittel / Depression“ vermuten wir eine „andere Erklärung“ (als die direkte Kausalität) für den beobachteten Zusammenhang: das Alter. Denn je älter ein Patient ist, desto häufiger finden wir sowohl eine Schmerzmittel-Therapie als auch Depressionen. Das erklärt die häufige gemeinsame Beobachtung dieser beiden Größen, auch ohne einen kausalen Zusammenhang.

Um das Alter als einen sogenannten Confounder (gemeinsame Erklärung) zu erkennen, muss das Alter der Patienten natürlich in den Daten verfügbar sein. Fehlende Informationen zu wichtigen Confoundern führen zu falschen Schlüssen über Ursache und Wirkung. Wenn wir – ohne Berücksichtigung des Alters – Patienten, die Schmerzmittel bekommen, mit denen ohne Schmerzmittel vergleichen, vergleichen wir implizit ältere mit jüngeren Patienten (also „Äpfel mit Birnen“).

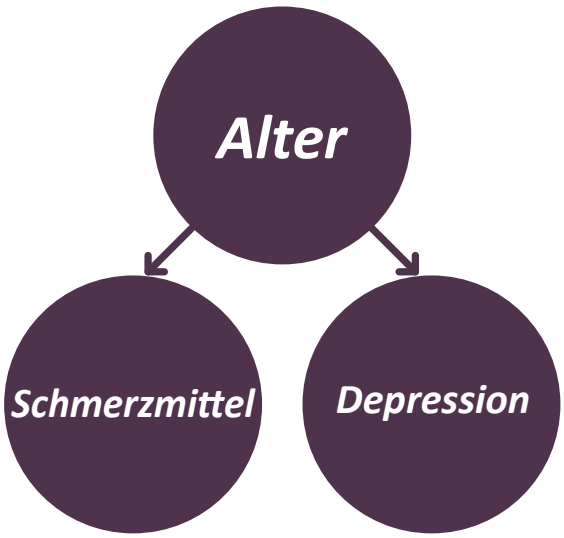


Abbildung 1 - Das „Alter“ erklärt indirekt den Zusammenhang zwischen Schmerzmittel und Depression - Man bezeichnet Alter als einen „Confounder“, eine gemeinsame Ursache für beides.

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

CONFOUNDER ALS KRITISCHE FAKTOREN

So kann es viele denkbare Faktoren geben, die als gemeinsame Erklärung in Frage kommen und sowohl die Einnahme eines Schmerzmittels als auch eine Depression nach sich ziehen, denn: viele schwere Erkrankungen gehen mit einer Schmerztherapie einher - und stimmen gleichzeitig den Patienten depressiv.

Alle diese „Confounder“ müssen demnach für eine verlässliche Aussage geprüft werden – nur wenn keiner davon (oder keine Kombination dieser Faktoren) einen beobachteten Zusammenhang erklären kann, entsteht für eine Variable der begründete Verdacht einer potenziellen, direkten Kausalität.

SCHLUSSFOLGERUNGEN

A. Die Basis zur Gewinnung nützlicher Hinweise auf potenziell kausale Zusammenhänge - und zur Unterscheidung von banalen Korrelationen - sind sehr umfangreiche Informationen, idealerweise ein 360°-Bild des analysierten Objektes inkl. aller historischen Daten. Der potenziell kausale Anteil einer beobachteten Korrelation ist der Teil, der sich nicht über andere Größen erklären lässt.

B. Umfangreiche Daten sind dennoch wertlos, wenn nicht gleichzeitig auch performante Algorithmen zur Verfügung stehen, die in der Lage sind, in solchen umfangreichen und komplexen Daten „alle alternativen Erklärungen“ zu suchen und zu bewerten.

Xplain Data hat sich diesen beiden Punkten verschrieben: Unsere ObjectAnalytics Datenbank gestattet analytischen Zugriff auf ein komplexes Objekt - übergreifend über alle Daten, die am Objekt hängen (rekursive Unterobjekte).

Im Bereich Gesundheit ist das sogenannte „Wurzel-Objekt“ in der Regel der Patient, mit Millionen gespeicherter einzelner Patienten-Beispiele (Objekt-Instanzen) und daran gebunden mehreren Milliarden Ereignissen wie verschriebene Arzneimittel, gestellte Diagnosen, durchgeführte Behandlungen und Operationen etc.

In der Praxis erfolgt die Zusammenführung verschiedener Datenquellen in ein Objekt-Modell entsprechend den individuellen Gegebenheiten. Unterschiedlichste Datenbanken, Tabellen und Bibliotheken – zum Teil über Jahre aufgebaut – werden damit einer übergreifenden Analyse zugänglich. Und oft wird erst so der Wert dieser Daten realisiert. (Zu ObjectAnalytics siehe auch entsprechendes White Paper. Der Ansatz ist durch unser US und EP Patent geschützt[2]).

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

CAUSAL DISCOVERY GEHT IHREN DATEN AUF DEN GRUND

Mithilfe eines patentierten Verfahrens, dem „Causal-Discovery“-Prozess, werden automatisch Millionen „alternativer Erklärungen“ aus den komplexen Daten gebildet, um einen beobachteten Zusammenhang indirekt (über Confounder) zu erklären. Nur wenn so keine indirekten Erklärungen gelingen, wird ein Faktor als potenzielle direkte Kausalität vorgeschlagen. Umfang der Daten und Stärke der Suchverfahren bestimmen dabei die Qualität des Ergebnisses.

„Potenziell“ kausal – denn auch mit sehr umfangreichen Daten liegt noch kein statistisch belegter Beweis vor.

Folglich müssen - vom Algorithmus vorgeschlagene - Hypothesen durch Anwendungsexperten bewertet und interpretiert werden. In der Beurteilung muss es Raum für Folgefragen geben, warum etwa ein erwarteter kausaler Faktor algorithmisch nicht als solcher identifiziert wurde – denn oftmals stellen Causal Discovery-Analysen die erstmalige Gelegenheit dar, subjektive Eindrücke und langjährige Erfahrungen mit eindeutigen Fakten in dieser Form zu vergleichen.

Hier schliesst sich die dritte Schlussfolgerung an:

C. Die algorithmische Intelligenz muss ergänzt werden durch eine Benutzeroberfläche, die Hypothesen für potenziell kausale Faktoren intuitiv verstehbar präsentiert und es Experten erlaubt, in einem Prozess von Diskussion und Folgefragen zu einem Verständnis zu gelangen.

Dieser Anforderung begegnet Xplain Data mit einer Benutzeroberfläche, welche die Beantwortung auch komplexer Fragestellungen in einem interaktiven Prozess ermöglicht. Umfang und Tiefe der Analysen werden dabei bedingt durch das individuelle Setting (Object Tree/ Fragestellungen) des Anwenders.

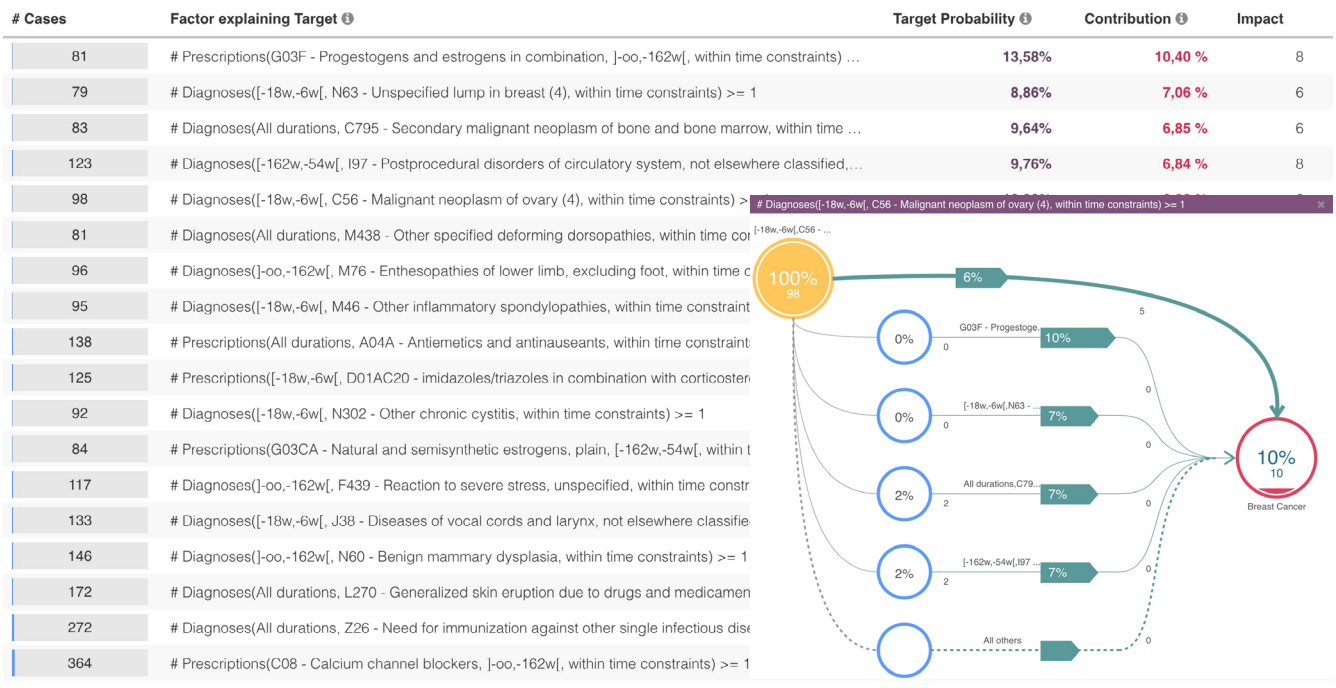


Abbildung 2 - Beispielergebnisse: Darstellung der gefundenen Faktoren in einer interaktiven Tabelle. Die Auswahl eines einzelnen Faktors zeigt, wie sich der Gesamteffekt aus einem direkten, potenziell kausalen Beitrag des Faktors und einem indirekten (über andere Faktoren) zusammensetzt.

WER PROFITIERT VOM XPLAIN DATA CAUSAL DISCOVERY PROZESS?

Typische Anwendungsbereiche sind Analysen im Pharma- und Gesundheitsumfeld sowie in der Fertigungsindustrie, z. B. um anhand sogenannter „Real World Data“ Zusammenhänge in der medizinischen Versorgung zu verstehen oder Ursachen für fehlerhaft produzierte Teile zu ergründen – und den Ausschuss (n.i.O.) zu reduzieren.

So wurde z. B. in einem Projekt mit der **Schwäbischen Werkzeugmaschinen GmbH** (SW) im Bereich der diskreten Fertigung bereits nach kurzer Zeit der Nutzen der Xplain Data Causal Discovery Algorithmen bei der kausalen Rückverfolgung von auffälligen Werkstücken nachgewiesen.

Xplain Data
Discover Causality

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