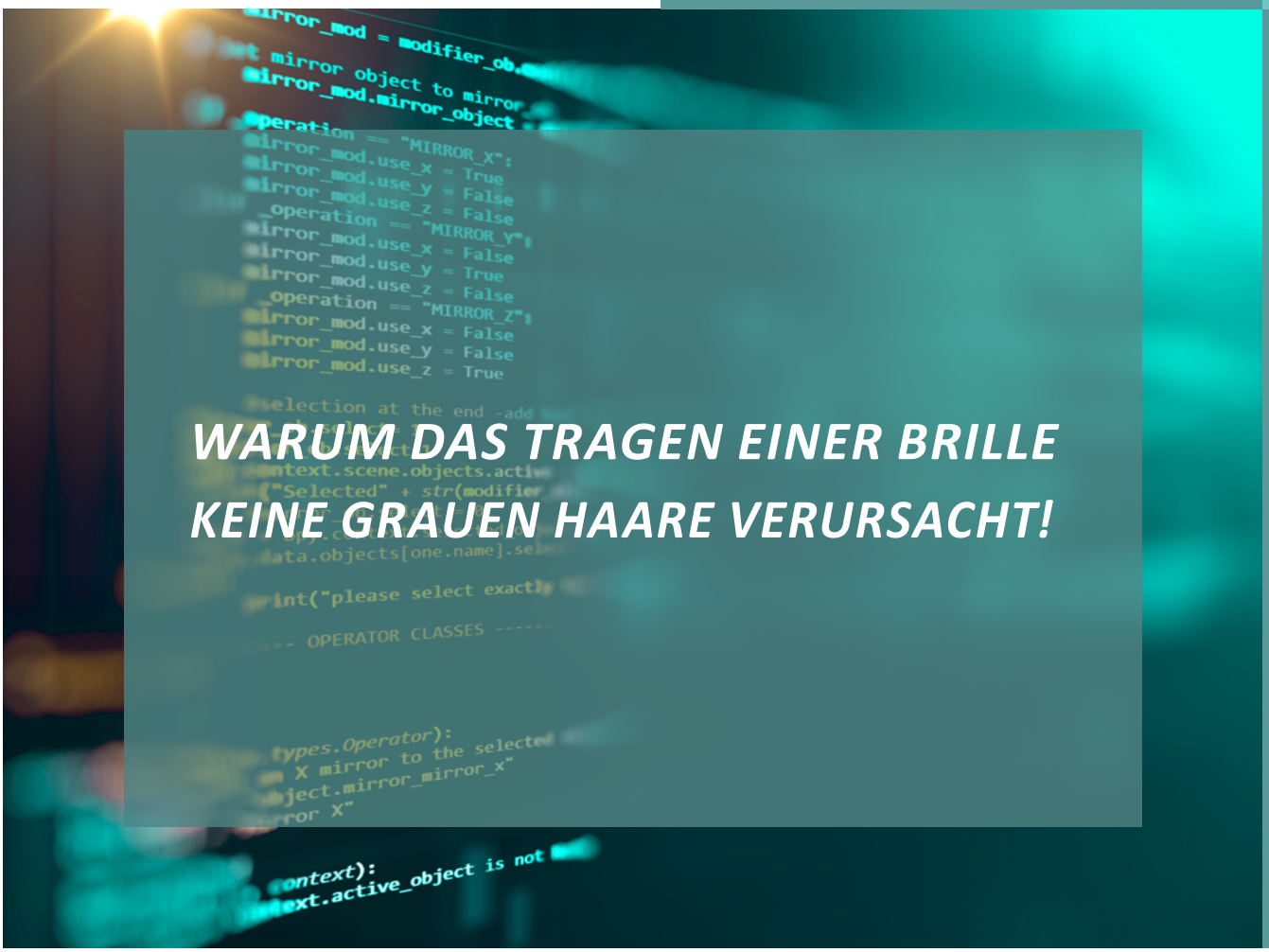
Xplain Data
Discover Causality

WHITE PAPER



WARUM DAS TRAGEN EINER BRILLE KEINE GRAUEN HAARE VERURSACHT!

VON KORRELATION ÜBER KAUSALITÄT ZU KÜNSTLICHER INTELLIGENZ …

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

WAS SIE ERWARTET

ZUSAMMENFASSUNG 3

BRILLEN & GRAUES HAAR 4

KORRELATION VS KAUSALITÄT 4

WAS IST KAUSALITÄT? 4

DIE SCHLECHTE NACHRICHT KAUSALITÄT – "KEINE ALTERNATIVE ERKLÄRUNG" 5

EIN PRAKTISCHER ANSATZ FÜR KAUSALE SCHLÜSSE 6

GANZHEITLICHE SICHTWEISE – EIN MUSS 6

EINBINDUNG VON EXPERTEN 7

EIN BEISPIEL 8

ANWENDUNG DER KAUSALITÄT: INTELLIGENTE MESSUNGEN, INTELLIGENTE INTERVENTIONEN 9

ZUSAMMENFASSUNG

Korrelation bedeutet nicht gleich Kausalität. Dennoch treffen wir oft Entscheidungen auf Basis von Korrelationen statt auf Basis von “Ursache und Wirkung”.

Kausalitätsbeziehungen lassen sich leider nicht anhand von Beobachtungsdaten nachweisen. Wir können jedoch einige Hinweise auf die Kausalität erhalten: Direkte kausale Faktoren lassen sich nicht “anderweitig erklären”, und eine intensive Suche nach alternativen Erklärungen bringt eine kleine Gruppe direkter und potenziell kausaler Faktoren ans Licht.



Ganzheitliche Daten sind daher für den Kausalschluss wichtig. Interventionen, die darauf abzielen, ein System in Richtung eines gewünschten Ziels zu lenken, erfordern Kenntnisse über Ursache und Wirkung. Kausalität wird daher eine wichtige Säule für zukünftige Systeme der Künstlichen Intelligenz sein. Präzisionsmedizin und individualisierte Behandlungen sind ohne sie kaum denkbar.

Wir zeigen ein Beispiel, bei dem wir depressive Episoden vorhersagen und dabei Wirkungen und Nebenwirkungen bestimmter Medikamentenkategorien aufzeigen - und wie diese auf verschiedene Patientengruppen wirken.

BRILLEN & GRAUES HAAR

Setzen Sie sich in ein Café, betrachten Sie die vorbeigehenden Menschen und zählen Sie, wie viele von ihnen eine Brille tragen und wie viele graue Haare haben ...

KORRELATION VS KAUSALITÄT

Sie werden feststellen, dass es einen Zusammenhang zwischen "Brille tragen" und "grauen Haaren" gibt: Von den Passanten, die eine Brille tragen, hat ein erhöhter Anteil auch graue Haare.

Sind Sie Brillenträger?

Keine Sorge! Sie müssen Ihre Brille nicht ablegen, um zu verhindern, dass Ihr Haar grau wird.
Es gibt zwar einen Zusammenhang zwischen "Brillenträgern" und "grauen Haaren", aber die Brille ist nicht die Ursache für graue Haare.

Das Muster "graue Haare und Brille" ist ein sehr bekanntes. Unser Gehirn ist sehr gut in der Mustererkennung, und wir ziehen - intuitiv und richtig - nicht den Schluss "Brille verursacht graue Haare". Dafür gibt es einen einfachen Grund, wie wir später sehen werden (Sie kennen ihn wahrscheinlich schon).

Bei der Datenanalyse werden jedoch oft falsche Schlüsse über die Kausalität von Korrelationen gezogen. Einige Beispiele:

- Zielgruppenorientierte Marketingmaßnahmen korrelieren möglicherweise positiv mit den Verkaufszahlen. Bedeutet das, dass Ihre Marketingmaßnahmen zu einer Umsatzsteigerung geführt haben? Marketingmanager neigen dazu, die Effizienz einer Kampagne auf der Grundlage einer solchen Korrelation zu beurteilen.

- Maschinenausfälle können mit bestimmten Betriebsparametern vor dem Ausfall korrelieren - bedeutet dies, dass diese Betriebsbedingungen Ausfälle verursachen?

- Genexpressionen können mit Krebs in Verbindung gebracht werden - bedeutet das, dass diese Expressionen Krebs verursachen?

Offenbar sind "Korrelation" und "Kausalität" zwei verschiedene Dinge.
Das wirft die Frage auf ...

WAS IST KAUSALITÄT?

Es gibt eine genaue mathematische Definition für Korrelation, z. B. in Form des Pearson-Korrelationskoeffizienten. Gibt es eine ähnlich genaue Definition für "Kausalität" auf der Grundlage von Beobachtungsdaten? Wie können wir sie messen?

In der Tat können wir Kausalität auf eine ähnlich strenge Art und Weise messen - allerdings nur, wenn wir Experimente machen dürfen.
"Randomisierte kontrollierte Studien (RCTs)" sind der Goldstandard für die Bewertung kausaler Auswirkungen. RCTs sind weit verbreitet, z. B. in der Arzneimittelentwicklung, wo Patienten nach dem Zufallsprinzip einer Kontroll- und einer Interventionsgruppe zugewiesen werden.

Wenn wir randomisierte Experimente durchführen können, ist alles in Ordnung. Um z. B. zu beweisen, dass Brillen keine grauen Haare verursachen, müssten wir vorbeilaufende Passanten nach dem Zufallsprinzip anweisen, von nun an keine oder eine Brille zu tragen, und später die Auswirkungen in den beiden verschiedenen Gruppen beobachten. In der Praxis ist das nicht machbar.

Experimente sind in der realen Welt selten möglich, insbesondere keine randomisierten Experimente. In der heutigen digitalen Welt häufen sich die Daten, diese sind meist Beobachtungsdaten (wobei Beobachtung bedeutet, dass

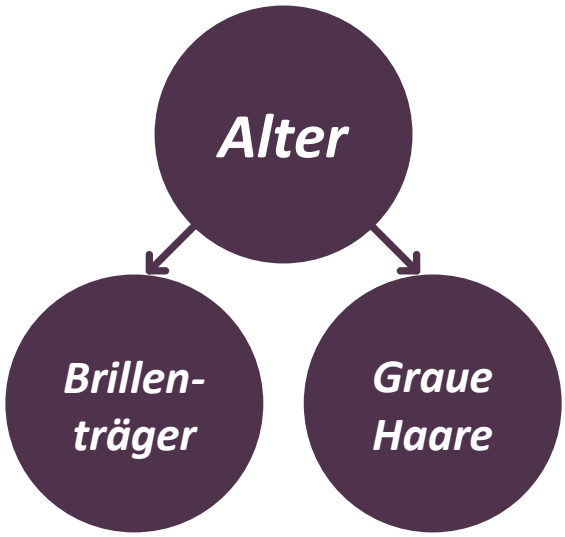
keine kontrollierten/randomisierten Interventionen/Experimente durchgeführt wurden). Die entscheidende Frage lautet daher: Wie können wir Kausalität auf der Grundlage von Beobachtungsdaten definieren oder messen?

DIE SCHLECHTE NACHRICHT

Wir können Kausalität nicht anhand von reinen Beobachtungsdaten messen – und könnten nicht einmal den Begriff "Verursachung" auf eine strenge mathematische Weise[1] definieren.

Das ist etwas enttäuschend. Im Zeitalter von Big Data gibt es keine Möglichkeit, diese Daten zu nutzen, um "Ursache und Wirkung" zu verstehen? Können wir nicht wenigstens eine Schätzung oder Grenzen für die Kausalität erhalten? Oder bestimmte Korrelationen als nicht kausal ausschließen, so dass wir am Ende eine Liste mit sehr "heißen" Kandidaten für potenziell kausale Beziehungen haben?

Wenn wir nur mehr oder weniger gute Beweise für die Kausalität erhalten können, lautet die nächste wichtige Frage: Wie müssen wir Daten sammeln, um bestmögliche Erkenntnisse über die Kausalität zu erhalten? Oft werden Unmengen von Daten gesammelt, nur um später festzustellen, dass ein entscheidender Teil fehlt.



KAUSALITÄT – "KEINE ALTERNATIVE ERKLÄRUNG"

Nach Kenny[2] müssen drei Bedingungen erfüllt sein, um von einem kausalen Einfluss der Variable (X) auf die Variable (Y) auszugehen:

1. X muss Y zeitlich vorausgehen.

2. X muss zuverlässig mit Y korreliert sein (über den Zufall hinaus).

3. Alternative Erklärungen für die Beziehung zwischen X und Y müssen ausgeschlossen werden.

Bedingung 3. bedeutet für unser Beispiel: Finden wir eine alternative Erklärung für die Korrelation Brille/graue Haare, schließt dies die Brille als Ursache für graue Haare aus. Wir müssen also eine bessere Erklärung oder die "wahre Ursache" finden.

Wenig überraschend: Die wahre Ursache für graue Haare (und Brille) ist natürlich das Alter.

Wenn wir älter werden, werden unsere Haare irgendwann grau und viele von uns müssen eine Brille tragen. Das Alter ist eine gemeinsame Ursache für beides, für die Notwendigkeit einer Brille und das Ergrauen der Haare (ein sogenannter Confounder).

[1]Einige Forscher mögen einwenden, dass es in der Tat möglich ist, mit einigen sehr theoretischen Prämissen ein "strukturelles Kausalmodell" zu ermitteln. In der Praxis müssen wir jedoch feststellen, dass wir kausale Wirkungen nicht auf der Grundlage von reinen Beobachtungsdaten nachweisen können - randomisierte kontrollierte Studien bleiben der Goldstandard.

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

Theoretisch könnten wir das beweisen, indem wir alle Passanten nach ihrem Alter befragen und die Korrelation zwischen Brillen und grauen Haaren innerhalb jeder Alterskategorie auswerten. Wenn es in keiner der Altersgruppen eine Korrelation gibt, dann hat das Alter die Brille als direkten Kausalfaktor für graue Haare vollständig "anderweitig erklärt".

Umgekehrt: Wenn wir wider Erwarten feststellen, dass bei gleichaltrigen Passanten immer noch eine Korrelation zwischen Brille und grauen Haaren besteht, kann das Alter nicht der einzige Grund für graue Haare sein. Es muss noch andere Faktoren geben, und der Faktor Brille würde immer noch zu den möglichen Kandidaten für graue Haare gehören.

In diesem Fall müssten wir weiter nach anderen Faktoren (Confoundern) suchen. Befragen wir alle Passanten nach zusätzlichen Daten, wie Geschlecht, Ernährungs- und Rauchgewohnheiten in den letzten Jahren, sportlichen Aktivitäten ... wenn nichts von all den verfügbaren Informationen die Brille als Faktor für graue Haare ausschließt - dann müssen wir die Möglichkeit in Betracht ziehen, dass die Brille graue Haare verursacht.

Wenn ein Faktor nicht durch andere verfügbare Informationen erklärt werden kann (die dem Ziel zeitlich vorausgehen), dann wird dieser Faktor zu einem wahrscheinlichen Kandidaten als kausaler Faktor. Und zwar umso wahrscheinlicher, je mehr andere Informationen verfügbar sind. In diesem Sinne könnten wir "beobachtete Kausalität" definieren als "keine andere verfügbare Erklärung innerhalb des gegebenen Informationsumfangs"[3].

Glück gehabt!? Das klingt wie "viele Informationen helfen viel". Die schlechte Nachricht ist zwar immer noch, dass Ursache und Wirkung nicht auf der Grundlage von Beobachtungsdaten nachgewiesen werden können. Aber die gute Nachricht ist: "Big Data" (im Sinne von komplexen Daten) kann zumindest dazu beitragen, viele Hypothesen auszuschließen und die interessanten schnell einzugrenzen.

EIN PRAKTISCHER ANSATZ FÜR KAUSALE SCHLÜSSE

GANZHEITLICHE SICHTWEISE – EIN MUSS

Also: Ein Faktor ist ein guter Kandidat für einen potenziell kausalen Faktor, wenn er durch keine andere Variable "anderweitig erklärt" werden kann. Je mehr unterschiedliche Daten wir haben, desto intensiver können wir nach "alternativen Erklärungen" suchen. Eine Annäherung an die Kausalität erfordert daher ganzheitliche Daten. Diese wiederum bedeuten ein komplexes Datenmodell. Denken Sie an Patienten-Daten: Sie erfordern zwangsläufig ein Modell mit mehreren Tabellen und zugehörigen Entitäten - keineswegs nur eine flache Tabelle.

Die meisten statistischen Ansätze gehen jedoch von einem Satz von N Variablen aus, d. h. von einer flachen Tabelle mit N Spalten, wobei jede Zeile eine Beobachtung darstellt (ein Passant in unserem Café-Beispiel mit Variablen wie "Brillenträger" oder "Alter").

In Real-World-Datenmodellen gibt es jedoch a priori nichts dergleichen. Komplexe Daten müssen erst in eine flache Tabelle abgebildet werden, was heute noch ein manueller, von Annahmen geprägter Prozess ist, der außerhalb der Möglichkeiten statistischer Lernalgorithmen liegt. ObjectAnalytics schließt diese Lücke: Daten werden als ganzheitliche Objekte anstelle von flachen Tabellen dargestellt. Das erleichtert die Analyse von Objekten

[3]Dies ist noch keine mathematische Definition und unterscheidet sich wahrscheinlich auch von möglichen anderen "Definitionen" der Kausalität. Die hier verwendete Definition von Kausalität oder die von Kenny (1979) bezieht sich nur auf "direkte" Ursachen (d. h. ohne Variablen, die eine indirekte Wirkung "über andere" haben).

(z. B. “Patienten”) als Ganzes. Das bedeutet konkret: Wir können ein ganzes Objekt mit all seinen abhängigen Daten nach “alternativen Erklärungen” durchsuchen. ObjectAnalytics bietet somit neuartige Möglichkeiten, potenzielle Ursache-Wirkungs-Beziehungen zu verstehen.

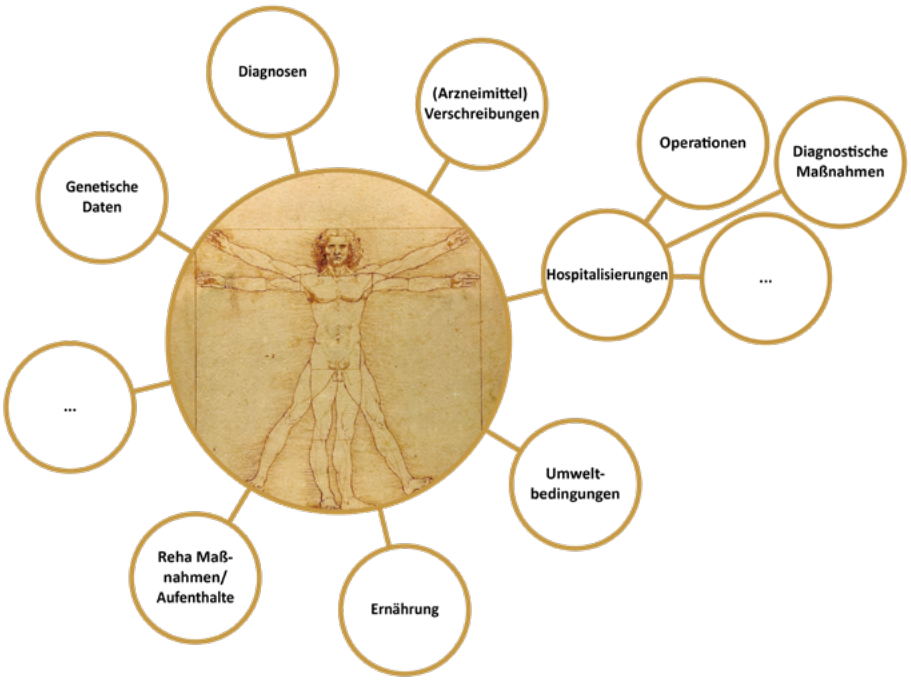
EINBINDUNG VON EXPERTEN

Auf der Grundlage umfassender Daten können wir viele korrelierende Faktoren als nicht kausal ausschließen. Wir erhalten so Hinweise, dass andere wahrscheinlich für eine sehr direkte und potenziell kausale Abhängigkeit in Frage kommen. Aber selbst, wenn wir über eine Fülle von Beobachtungsdaten verfügen, kann die Kausalität nicht nachgewiesen werden.

Ein praktischer Ansatz zur Bewertung der kausalen Auswirkungen muss daher das Feedback von Experten einbeziehen. Diese müssen in der Lage sein, die Ergebnisse intuitiv zu verstehen. Ein vorrangiges Ziel unseres Ansatzes ist es daher, die Abhängigkeit eines Zielereignisses von früheren Ereignissen in einem möglichst einfachen Modell mit intuitiv verständlichen Parametern zusammenzufassen. Diesem Zweck dient ein Netzwerk von so genannten “probabilistischen ODER- und UND-Verknüpfungen”.

Der Gesamtansatz ist auf enge Experten-Interaktion ausgelegt. Anfänglich ist der Blick auf das Wesentliche durch Unmengen von bedeutungslosen Korrelationen verstellt. Der “Nebel” lichtet sich schnell mit dem ersten Modell, das die vielen Korrelationen auf eine kleine Gruppe direkter (potenziell kausaler) Faktoren reduziert. Diese werden von Experten bewertet, indem einige verworfen und andere markiert werden, die potenziell Ursache und Wirkung sind - und die auch mögliche Interventionen darstellen. (Das Alter könnte graue Haare verursachen, aber wir können nicht in das Alter eingreifen. Wenn jedoch Rauchen graue Haare verursacht, können wir mit dem Rauchen aufhören, um graue Haare zu vermeiden).

Unter Einbeziehung von Expertenfeedback werden detaillierte Modelle erstellt, die dann z. B. als Elemente in einem intelligenten Geschäftsprozess eingesetzt werden können.



Beispiel für ein komplexes, “real-world” Objekt: “Der Patient” mit verschiedenen Datenströmen, die an dem Patienten hängen. ObjectAnalytics bedeutet, die verschiedenen Datenbereiche in Relation zueinander analysieren zu können.

EIN BEISPIEL

Die Vorhersage von Gesundheitsrisiken ist wichtig - noch wichtiger ist das Verständnis der Ursachen als Grundlage für gezielte Interventionen. “Präzisionsmedizin” wird erst realisierbar, wenn wir Ursache und Wirkung von Medikamenten und Behandlungen in einzelnen Patientengruppen (inkl. Nebenwirkungen) verstehen.

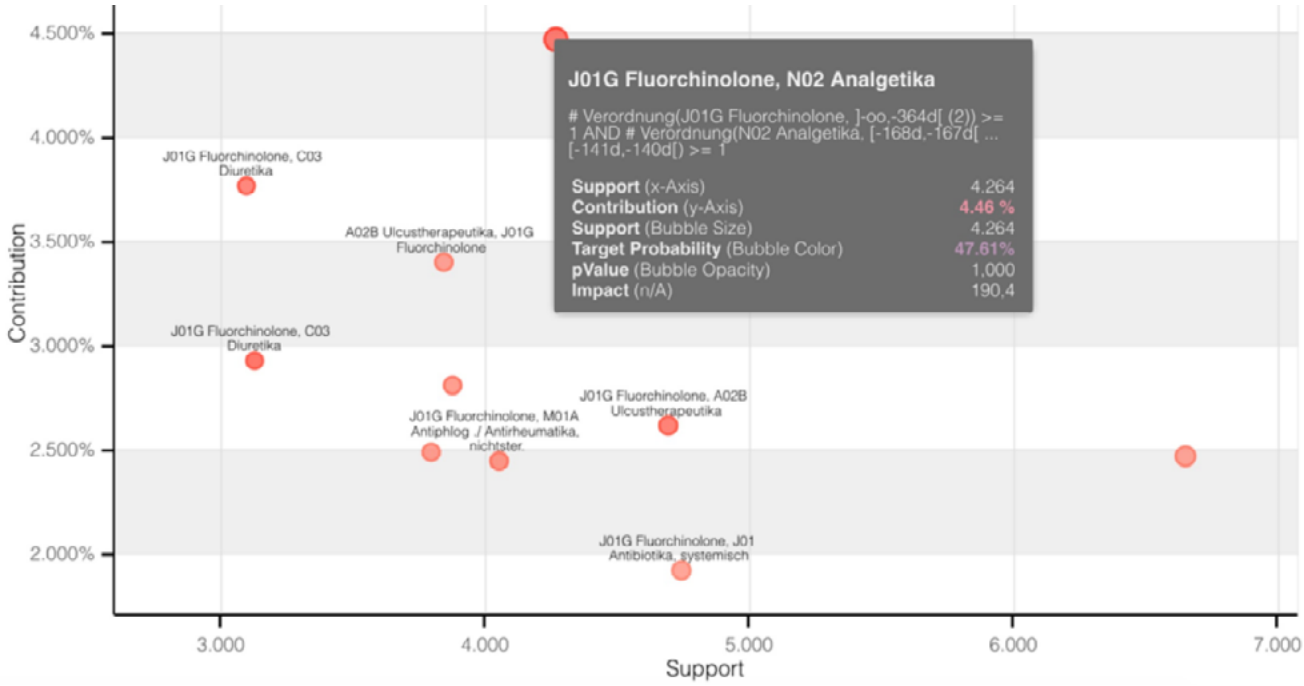
In der Abbildung unten haben wir das Risiko für eine depressive Episode auf der Grundlage der gesundheitlichen Vorgeschichte eines Patienten (Medikamente und Behandlungen vor der ersten Verabreichung eines antidepressiven Medikaments) vorhergesagt. Es wurden Daten von 3 Mio. Patienten mit ~200 Mio. Verschreibungen für Tausende von verschiedenen Medikamentenkategorien verwendet.

Der Algorithmus durchsucht das Objektmodell (alle Datenströme, die mit dem Objekt “Patient” verbunden sind) nach Faktoren, die mit der ersten antidepressiven Behandlung zusammenhängen, aber davor liegen. Bei der Suche im Objektbaum werden hunderttausende von Hypothesen gebildet.

Jede Hypothese für einen potenziellen Kausalfaktor wird gegen alle anderen getestet, um zu bewerten, ob einer der anderen Faktoren als “alternative Erklärung” (Confounder) dienen kann.

Nur die Faktoren, die die umfangreiche Suche nach alternativen Erklärungen überstehen, werden Experten vorgelegt. Je nach Tiefe der Suche verbleiben ~50 bis 100 Faktoren als “nicht anderweitig erklärbar” und potenziell kausale Beiträge zum Depressionsrisiko.

Verschiedene Arten von Antibiotika erscheinen als Risikotreiber, insbesondere die “Fluorchinolone”. Sie haben einen sehr unterschiedlichen Risikobeitrag in verschiedenen Kombinationen mit anderen Medikamenten bzw. für verschiedene Patientengruppen. Stimmungsschwankungen und Depressionen sind eine bekannte Nebenwirkung von Fluorchinolonen[4], werden jedoch evtl. erheblich unterschätzt - und scheinen sich bei verschiedenen Patientengruppen sehr unterschiedlich zu manifestieren. Eine individuelle Entscheidung über die Behandlung mit Fluorchinolonen scheint daher obligatorisch.



Identifizierte Faktoren, die in direktem Zusammenhang mit einer späteren antidepressiven Behandlung stehen, dargestellt in einem Blasendiagramm: x-Achse: Anzahl der Patienten mit diesem Faktor, y-Achse: Beitrag zum Risiko. Es werden nur Faktoren gezeigt, die mit Fluorchinolonen zusammenhängen.

[4]Siehe z. B.: www.akdae.de/Arzneimittelsicherheit/Bekanntgaben/Archiv/2004/20040528.html

ANWENDUNG DER KAUSALITÄT: INTELLIGENTE MESSUNGEN, INTELLIGENTE INTERVENTIONEN

Alle Welt spricht über “Künstliche Intelligenz”. Keiner spricht über “Kausalität”. Würden intelligente Interventionen nicht voraussetzen, dass die durchgeführten Aktionen in einem kausalen Zusammenhang mit dem angestrebten Ziel stehen?

Es liegt auf der Hand, dass die Fähigkeit, wahrscheinliche Kausalbeziehungen zu finden, zahlreiche Möglichkeiten bietet. Das obige Beispiel der “Präzisionsmedizin” ist nur eines davon. Aber wir müssen nicht so weit gehen und uns ein KI-System ausdenken, das autonom die besten Maßnahmen auf der Grundlage von Ursache und Wirkung findet. Es gibt viel nahe-liegendere Anwendungen der Kausalität, zum Beispiel eine angepasste Messung verschiedener Arten von durchgeführten Eingriffen.

Nehmen wir an, es gibt viele Kunden und Sie sprechen diese mit verschiedenen Arten von Aktionen an. Bei jeder Messung der Wirkung verschiedener Arten von Maßnahmen sollte die Kausalität berücksichtigt werden. Sie wollen einer Aktion keine Ergebnisse zuschreiben, die nicht durch diese Aktion verursacht wurden. Eine angepasste Messung ist eine offensichtliche Anwendung von Konzepten zur Kausalität. Auf dieser Grundlage kann man dann damit beginnen, die Ressourcen dort einzusetzen, wo sie die größte Wirkung haben.

Wir stehen erst am Anfang des Verständnisses von Kausalität auf der Basis von realen Daten. Die Produkte von Xplain Data bieten praktikable Konzepte für Kausalität im Bereich der Künstlichen Intelligenz. Wir stellen Algorithmen zur Verfügung, die komplexe Objekte “so wie sie sind” verarbeiten können und in der realen Welt angesiedelt sind, anstatt in einer künstlich aufbereiteten analytischen Umgebung. Bald schon wird Kausalität eine wichtige Säule für Künstliche Intelligenz sein.

Xplain Data
Discover Causality

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