# Assignment 4 – Praktikum Smart Data Analytics, Group 4

## Task 2

### 2a)

**Einsatz der Methoden PCMCI und TCDF auf COVID-19 Datensätzen für Deutschland, Italien und die Vereinigten Staaten von Amerika.** (Dokumentation in den Jupyter-Notebooks im Methoden-Unterverzeichnis, vgl. Readme-Datei!)

### 2b)

**Visualisierung der Kausal-Resultate von PCMCI und DCTF.** (Siehe Resultate-Verzeichnisse im Methoden-Unterverzeichnis, vgl. Readme-Datei!)

### 2c)

**Validierungverfahren der Methoden.**

Zur Validierung der Ergebnisse stehen in diesem Falle leider keine Ground-Truth Daten bereit. Daher werden die Ergebnisse durch manuelles Betrachten mit Hilfe von Domänenwissen bewertet.

Eine Möglichkeit, zusätzlich die Stabilität der gefundenen kausalen Beziehungen zwischen den verschiedenen Algorithmen zu Bewerten, wäre die paarweise *Intersection over Union* der gefundenen Beziehungen zu bilden und diese Werte in einer Matrix abzutragen.

**Manuelle Auswertung mit Domänenwissen**

Es werden im Folgenden jeweils die interessantesten Eigenschaften und Besonderheiten genannt, insbesondere im Vergleich zu den zuvor genannten Verfahren und Datensätzen.

1. **PCMCI**

Deutschland, Tau 8:

* Einige offensichtlich falsche Kausalitäten, beispielweise bedingt die Schließung von Schulen nicht die Schließung von Geschäften
* Es wird die interessante Kausalität aufgezeigt, dass die Schließung von Schulen zu einer Verringerung der Mobilität der Bevölkerung führt
* Die Gesamtzahl der Covid-Test bedingt die geschätzte Zahl der Infektionen – diese wird auf Grundlage der Tests erstellt
* Die Gesamtzahl der benötigten Krankenhausbetten bestimmt wie erwartet mit einer Verzögerung von 5 bis 6 Tagen die Zahl der benötigten Beatmungsmaschinen. Patienten werden also nicht sofort beatmet, nachdem sie in das Krankenhaus eingewiesen werden
* Krankenhaus-Aufnahmen bedingen sowohl die Zahl der Gesamtbetten, als auch der Intensivbetten und der Beatmungsmaschinen, die benötigt werden

Italien, Tau 8:

* Die Gesamtzahl der Toten Bedingt mit Verzögerung einer die Zahl der bestätigt infizierten. Dies möglicherweise ist der Fall, weil erst nach dem Tot von Personen die Infektion an Corona als Todesursache festgestellt wird

USA, Tau 8:

* Bestätigte Infektionen werden durch die mittlere Geschätzte Zahl der Infektionen bedingt. Dies kann dadurch erklärt werden, dass die Zahl der gemessenen Infektionen mit der Dauer bis zum Symptom-Beginns (4-7 Tage) nach Schätzung des Infektionsbeginns interherhinken. Personen werden also erst einige Tage nach ihrer Infektion positiv auf Covid-19 getestet

Deutschland, Tau 14:

* Die Mittlere Zahl der Tode wird bedingt von der Zahl der beatmeten mit einer Verzögerung von fünf Tagen

Italien, Tau 14:

* Die Mobilität der Bevölkerung bedingt die mittlere Zahl der Tode mit einer Verzögerung von 11 Tagen

USA, Tau 14:

* Die Zahl der bestätigten Fälle bedingt mit einer Verzögerung von 10 Tagen die notwendige mittlere Zahl von Beatmungsmaschinen

1. **TCDF**

* Es werden weniger kausale Zusammenhänge gefunden, die Kanten im Graphen sind spare
* Einige der kausalen Verzögerungen, insbesondere für größere Kernel-Größen sind unrealistisch lang und daher als falsch anzunehmen

### 2d)

**Schlussfolgerungen.**

Mit den gegebenen Verfahren konnten wir einige interessante Kausale Zusammenhänge identifizieren, die bei einer manuellen Betrachtung realistisch erscheinen. Allerdings enthalten die Ergebnisse häufig auch wenig sinnvolle Ergebnisse.

Eine detaillierte manuelle Betrachtung ist daher in jedem Fall erforderlich, Kausale Inferenz kann hierfür jedoch spannende Impulse generieren.

### Bonusaufgabe

**Datensatz-Augmentation.**

Um bessere und aussagekräftigere Ergebnisse durch weitere Einbringung von Domänenwissen zu erzielen, erweitern und kombinieren wird die vorhandenen Datensätze des *Institute of Health Metrics and Evaluation.*

1. Integration von Bundesländer-Daten für Deutschland, aus dem IHME-Summary-Datensatz

Regionale Maßnahmen spielen für die Analyse der Bekämpfung von COVID-19 eine entscheidende Rolle. Daher augmentieren wir den Datensatz um die Zeitserien der Anzahl der Bundesländer in Deutschland, in denen an einem Datum

* Kontaktbeschränklungen herrschten
* Veranstaltungen untersagt sind
* Geschäfte geschlossen wurden
* Nicht-essentielle Geschäfte geschlossen wurden
* Schulen geschlossen sind

Systematik: Manuell wurden die Bundesländer Deutschlands in den IHME-Daten identifiziert und Zeitserien erzeugt, die als Zähler fungieren. Tritt an einem Tag eine der Beschränkungen in einem Bundeland ein, wird der Zähler inkrementiert, wird jedoch eine Einschränkung Regional aufgehoben, so wird der Zähler dekrementiert.

**Diese Ergebnisse sind für Deutschland bereits in die Notebooks und Visualisierungen für Aufgabe 2a) integriert!**

*Exploration - Weitere mögliche Datensätze zur Augmentation:*

1. Wetterdaten

Aus wissenschaftlichen Studien ist bekannt, dass Kälte und verringerte Luftfeuchtigkeit die Verbreitung von COVID-19 begünstigen. Daher ist es sinnvoll, Wetter-Datensätze zu betrachten.

1. Tageweise mittlere Lufttemperatur für Deutschland, basierend auf DWD Open Data
2. Tageweise mittlere Luftfeuchtigkeit für Deutschland, basierend auf DWD Open Data