

IDW 23- Abschlussbericht

Jurij Schönfeld

Die International Data Week (IDW) fand 2023 unter dem Motto *A Festival of Data* in Salzburg statt. Auf dem Programm standen verschiedene Themen aus den Bereichen Data Science und Data Management. Dank des Stipendiums der RDA-DE war es mir möglich daran teilzunehmen. Ich möchte daher an dieser Stelle die Gelegenheit nutzen um einige interessante Erkenntnisse der IDW zu teilen. In erster Linie geht es mir darum junge Wissenschaftler*innen mit einer ähnlichen Perspektive auf Daten, wie ich sie habe, auf Themen aufmerksam zu machen, die abseits der täglichen Arbeit liegen.

Es ist wohl beinahe jeder Mensch täglich von einer Unmenge an Daten umgeben, als Student der Naturwissenschaften bin ich mir dessen halbwegs bewusst, denn sie prägen meinen Arbeitsalltag ungemein. Messdaten auswerten, ein Modell kalibrieren, Simulationen laufen lassen und Chat-GPT konsultieren sind nur eine kleine Auswahl von Situationen in denen sich viele Wissenschaftlerinnen täglich wiederfinden. Nur die wenigsten von uns denken dabei aktiv darüber nach, woher die benutzten Daten eigentlich kommen und welche Infrastruktur hinter der Verwaltung dieser steckt. Wir ärgern uns nur darüber, wenn etwas nicht funktioniert, wir etwas nicht finden können, wir uns durch Datenschutzbestimmungen kämpfen oder kreative Workarounds finden müssen um einen bestimmten Datensatz für unser Projekt nutzbar machen zu können. Parallel dazu beschäftigt sich die IT-Abteilung unseres Institutes oder Arbeitgebers, gemeinsam mit Data Managern und Data Stewards sowie Expert*innen auf der ganzen Welt, genau mit dieser Thematik. Ein wichtiges Kriterium ist dabei, dass die Daten FAIR sein müssen, das bedeutet, dass die Daten auffindbar, zugänglich, interoperabel und langfristig wiederverwendbar sein müssen. Unter Umständen ist es aber schon lange her, dass sie an der Auswertung der Daten selber beteiligt waren oder sie haben keinen Überblick mehr über neue Methoden zur Datenauswertung und deren datentechnische Anforderungen. Aus genau diesem Grund sind Events wie die International Data Week eine große Chance um Menschen die alle mit Daten, aber aus verschiedenen Perspektiven, arbeiten, an einen Tisch zu holen. Um gemeinsam über die eigenen und die fremden Herausforderungen zu sprechen und die Probleme des jeweils anderen zu verstehen und Lösungen zu entwickeln. Gelegenheit dazu bot bereits die Eröffnungssitzung *Spatial Data Science: Geographic Context Matters*. Hier wurden Herausforderungen aber auch Möglichkeiten bei der Analyse von Daten mit räumlicher Auflösung diskutiert. Dazu ergänzend konnten sich bei der Session *Data Cubes as a tool to organise data in data* Wissenschaftler*innen und Data Managers über das Speichern und Auswerten von Daten, die in sogenannten Hypercubes organisiert sind austauschen. Besonders interessant sind räumlich (und zeitlich) aufgelöste Daten natürlich im Kontext der Erdwissenschaften. In der Session *Climate Risk and Vulnerability Mapping* wurden dazu verschiedene Ansätze diskutiert, wie sich Auswirkungen des Klimawandels auf die Gesellschaft quantifizieren lassen und wie sich diese in den Prozess einbeziehen lässt. Gerade das evaluieren von Klimabezogenen Daten ist in der Regel mit massivem Rechenaufwand verbunden. Die Session *Bridging the Data/HPC divide* thematisiert, wie in Zukunft eine bessere Zusammenarbeit zwischen Data Management- und High Performance Computing-Zuständigen gewährleistet werden kann und wie sich Aufmerksamkeit für die Herausforderungen der anderen Seite schaffen lässt. Auswertung von Daten und damit verbundene Probleme wurden während der *Data Science Lightning Talks* mit Blickwinkel auf verschiedene Themen, wie Statistik, Nichtlineare Dynamik und Machine Learning, diskutiert. Neben der technischen Umsetzung und aktuellen Methoden des Machine Learning/AI, war ein elementarer Bestandteil der IDW auch deren ethische Bedeutung z.B. in den Sessions *AI Ethics Presentations* und der abschließenden Sitzung *Closing Plenary Session and, Ethics, data science and AI in dialouge with one another and society*. Als

Empfänger des RDA Reisestipendiums war es meine Aufgabe bei der Abschlussitzung die zentralen Diskussionspunkte festzuhalten und diese im Folgenden aufzuarbeiten.

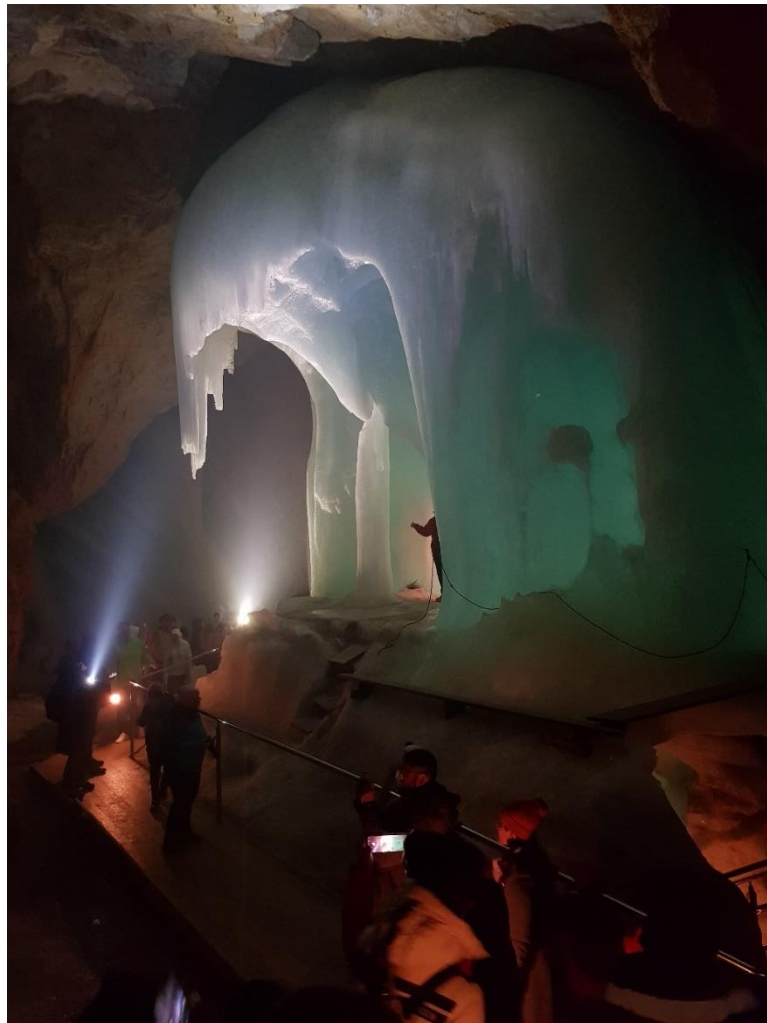
Die Nutzung von AI ist, in der Data Science, in den letzten Jahren vielleicht das relevanteste Thema gewesen. Es gibt viel Diskussion darüber, ob das Intelligent in AI wirklich gerechtfertigt ist oder nicht, aber sicher ist, dass der wissenschaftliche und gesellschaftliche Hype um dieses Thema noch einige Jahre anhalten wird und auch wenn in kurzen Abständen immer wieder neue Verfahren und Architekturen vorgestellt werden, stimmt es bis jetzt, dass AI nur so gut ist, wie die Daten mit denen sie trainiert wird. In der wissenschaftlichen Community geht es in der Regel allerdings nicht darum einen Datensatz nach den FAIR/CARE Prinzipien bereitzustellen, stattdessen werden in erster Linie Paper über AI-Architekturen und Anwendungen geschrieben und zwar mit dem Fokus darauf in einem möglichst renommierten Journal veröffentlicht zu werden. Speziell dieser Bereich der Wissenschaften ist allerdings sehr kurzlebig und meist eher eine Meinung oder Momentaufnahme des aktuellen wissenschaftliche Standes. Die entstanden oder benutzten Daten, die seltener der Fokus des Papers sind, werden oft nicht oder nur unzureichend zugänglich gemacht, obwohl diese über Jahrzehnte hinweg wissenschaftliche Relevanz haben können. Das Wiederbenutzen anderer Daten und deren qualitative Aufbereitung spielen im wissenschaftlichen Belohnungssystem nicht die Rolle, die sie bei der Entwicklung von AIs einnehmen.

Dies bringt Probleme bei der ethischen Nutzung von AI mit sich. Die Daten mit denen wir unsere AIs trainieren sind ein Spiegel der Gesellschaft, genau darin wird oft ihre Qualität gemessen. Allerdings spiegeln sie eben auch zunächst alle Gesichtspunkte der Gesellschaft wieder, insbesondere strukturellen Rassismus und Diskriminierung, vor allem wenn Daten aus dem Internet stammen, in



dem diese Strukturen im Allgemeinen stärker ausgeprägt sind. Die Gesellschaft wird in absehbarer Zeit diese Strukturen nicht aus eigener Kraft überwinden. Es ist auch keine Lösung die Ethik für die AI-Herausforderung zur Rechenschaft zu ziehen. Sie ist kein Akteur dem man die Schuld geben kann und der etwas nachzubessern hat. Sie ist eine Sammlung moralischer Werte, die wir, als Wissenschaftler*innen und Gesellschaft, verpasst haben umzusetzen. *Garbage In, Garbage Out!* Ist ein Satz der im Zusammenhang mit AI, auch auf der IDW, regelmäßig wiederholt wird und er

beschreibt genau dieses Problem. Es gibt keinen Ethik-Patch der das AI-Problem lösen wird, stattdessen müssen diskriminierende und rassistische Muster in Daten erkannt und aufgelöst werden. Dies kann allerdings nur gelingen, wenn dafür auch die entsprechenden Anreize geschaffen werden und die allgemeine Aufmerksamkeit vorhanden ist.



Ein weiteres Thema, das Gesellschaft und die IDW wesentlich bewegt ist der Klimawandel. Globale Initiativen, die Daten sammeln und bereitstellen, sind essenziell um Wetter- und Klimaphänomene zu verstehen und zahlreich auf der IDW vertreten. Abseits der Konferenz bot die IDW 2023 die Gelegenheit ein einmaliges Zeugnis der Veränderung unseres Planeten zu besuchen. Etwa eine Stunde mit dem Zug sowie ein ca. zweistündige Wanderung von Salzburg entfernt befindet sich die Eisriesenwelt, ein ca. 42 km langes Höhlensystem und die größte Eishöhle der Welt, die dem Klimawandel noch einige Jahre trotzen wird. Sie ist einen Besuch wert, weil sie die Auswirkungen des Klimawandels anschaulich macht und uns vor Augen führt wie kostbar unsere Umwelt ist. In den Eiskristallen und den gewaltigen Formationen der Höhle spiegeln sich die Schönheit, Verletzlichkeit, und Komplexität der Natur wider, ganz so wie komplexe Datensätze, die Schönheit und die Verletzlichkeit unseres Wissens und unserer Erkenntnisse reflektieren. Es ist unsere Pflicht, verantwortungsbewusst mit der Natur und diesen Daten umzugehen, denn beide bergen den Schlüssel zur Erhaltung unserer Welt und unseres Wissens.