Sylvia KRITZINGER[1], Katharina PFAFF, Julia BARTA, Jana BERNHARD, Hajo BOOMGAARDEN, Anja EDER, Nikolaus FORGÓ, Filip PASPALJ, Claudia PLANT, Barbara PRAINSACK, Dimitri PRANDNER, Simon RITTEL, Martin TEUFFENBACH & Sebastian TSCHIATSCHEK (Wien, Graz und Linz)

Digitize! – Computational Social Science in der digitalen und sozialen Transformation

Zusammenfassung

Die Digitalisierung bringt neben vielen Vorteilen für die sozialwissenschaftliche Forschung und die Lehre auch neue Herausforderungen mit sich. Die Weiterentwicklung digitaler Datenerhebungs- und Analyseverfahren zur Analyse und Gestaltung gesellschaftlicher und politischer Transformationsprozesse muss im Fokus der Zukunftsausrichtung sozialwissenschaftlicher Hochschullehre und Forschung stehen. Neue Datenformate und Praktiken erfordern neue forschungsethische und datenschutzrechtliche Praktiken und Standards. Im Zentrum von Computational Social Science steht die juristisch und ethisch fundierte, reflektierte Nutzbarmachung von digitalen Forschungsdaten und Analyseverfahren und ihre Vermittlung.

Schlüsselwörter

Computational Social Science, Datafizierung, Forschungsethik, Datenschutz

1 E-Mail: sylvia.kritzinger@univie.ac.at

DOI: 10.21240/zfhe/SH-F/11

Digitize! – Computational social science in the digital and social transformation

Abstract

Next to the many benefits that digitalisation creates for research and teaching, it also poses new challenges in the social sciences. The development of digital data collection and analysis methods for examining and shaping social and political transformation processes must be a priority for teaching and research in the social science. New data formats and practices require new standards and practices within research ethics and data protection. Computational social science focuses on the legally and ethically grounded and reflective use of digital research data and analysis methods and their communication.

Keywords

datafication, computational social science, research ethics, data protection

1 Kurzdarstellung *Digitize!*

Digitize! ist ein interdisziplinäres Projekt, bei dem die Herausforderungen der Digitalisierung in Zusammenarbeit zwischen den Sozialwissenschaften, Data Science, den Rechtswissenschaften und der Forschungsethik bearbeitet werden. Im Fokus stehen innovative Analyseverfahren sowie die juristisch und ethisch fundierte Nutzbarmachung von digitalen Forschungsdaten im Rahmen einer digitalen Infrastruktur für sozialwissenschaftliche Daten in Österreich. Des Weiteren werden Möglichkeitsräume für innovative Lehre im Rahmen der *Computational Social Science* (CSS) analysiert und entwickelt.

Konkret beschäftigen sich Forscher:innen des *Digitize!*-Projekts mit den Möglichkeiten zur Anwendung von Data-Science-Methoden auf sozialwissenschaftliche Daten. In der interdisziplinären Zusammenarbeit können die unterschiedlichen, für die jeweilige Forschungsdisziplin eigenen, Sichtweisen auf die Daten gegenübergestellt werden. Daraus entstehen Synergien einerseits für die methodische Konzeption der

Datenerhebung, andererseits für die unterschiedliche Analyse der Daten selbst. Die daraus gewonnenen Erkenntnisse tragen direkt dazu bei, innovative Forschungsarbeit in den involvierten Forschungsdisziplinen voranzutreiben. Rechtliche und ethische Compliance-Fragen werden in *Digitize*! kontinuierlich reflektiert und in Anwendung gebracht. Die sich aus der CSS-Forschung ergebenden Themen wurden kritisch und fundiert evaluiert und reflektiert.

Im Folgenden werden die unterschiedlichen disziplinären Zugänge und Forschungsergebnisse von *Digitize!* vorgestellt sowie mögliche daraus resultierende zukünftige Entwicklungen in Forschung und Lehre aufgezeigt.

2 Ergebnisse und Lessons Learned

Die Frage nach den bisherigen Ergebnissen und Erfolgen von *Digitize!* wird einerseits aus Sicht der beteiligten Forschungsdisziplinen, andererseits aus interdisziplinärer Sicht im Hinblick auf zukünftige Entwicklungen von CSS beantwortet. Ausgehend von den Zielen von *Digitize!* können die Ergebnisse der einzelnen Arbeitspakete (AP) reflektiert werden.

2.1 Perspektive der Sozialwissenschaften

2.1.1 Der *Digitize!*-Online-Panel-Pilot

Ein primäres Ziel von *Digitize!* ist es, ein repräsentatives, offline-rekrutiertes Online-Panel für Bevölkerungsbefragungen aufzubauen. Hierfür werden mehrere Piloterhebungen durchgeführt, die innovative und experimentelle Forschungsfragen zur Umfragemethodik aufgreifen (PFAFF et al., 2022). Von Februar 2022 bis Juni 2023 wurden bisher vier Befragungswellen durchgeführt. Die Umfragen wurden über die cloudbasierte Onlineplattform *Qualtrics* realisiert und sind von verschiedenen digitalen Geräten wie Computern, Laptops, Tablets und Mobiltelefonen aus zugänglich. Am Ende jeder Befragung wurde die Bereitschaft zur Teilnahme an Folgebefragungen im Rahmen des *Digitize!*-Panels abgefragt. Insgesamt haben sich bisher 2.773 Personen für das Online-Panel angemeldet. *Abbildung 1* zeigt die Anzahl der Teilnehmer:innen pro Befragungswelle sowie die Anzahl der Personen,

die neu ins Panel eingestiegen sind (= Panelist:innen). Die Rücklaufquote und die Bereitschaft zu Folgebefragungen variiert über die Wellen.

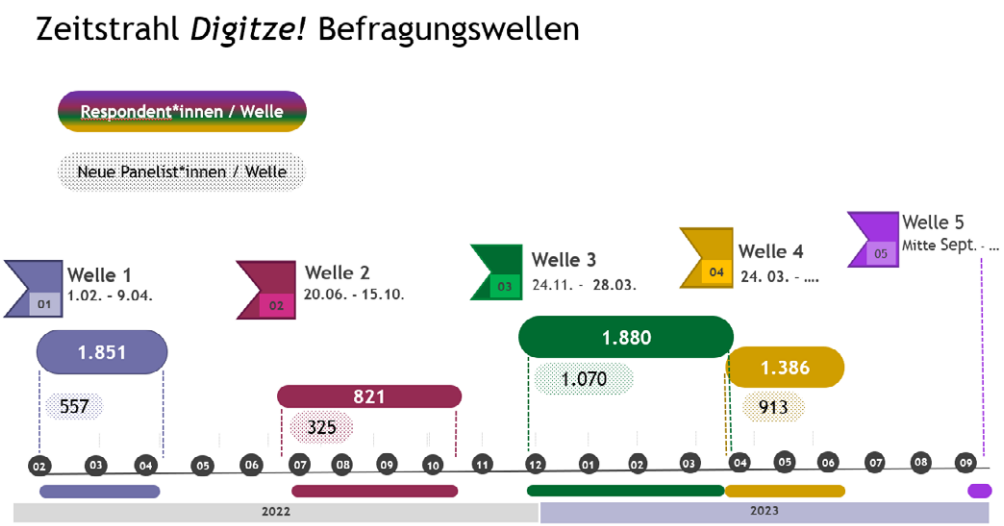

Abb. 1

In verschiedenen experimentellen Settings wurde geprüft, welche Faktoren die Rücklaufquote und die Panelrekrutierung beeinflussen. Es wurden folgende Aspekte getestet:

1. Die Art der Incentivierung
2. Unterschiedliche Einstiegsvarianten zum Fragebogen
3. Länge und inhaltliche Ausführlichkeit des Einladungsschreibens

Die bisherigen Ergebnisse zeigen, dass unterschiedliche Faktoren verschiedene Personengruppen motivieren, bei der Umfrage bzw. beim Panel mitzumachen.

Ad 1) Die 8.000 Personen, die über eine repräsentative ZMR-Stichprobe eingeladen wurden, wurden in vier unterschiedliche Anreizsysteme zufällig eingeteilt: (1) 2.000 Personen erhielten mit der Einladung einen Gutschein im Wert von 5 Euro für die Supermarktkette Spar (‚unconditional voucher'), (2) 2.000 Personen erhielten erst nach Abschluss der Befragung einen 5 Euro-Gutschein (‚conditional voucher'), (3) 2.000 Personen erhielten die Möglichkeit, nach Abschluss der Befragung 5 Euro an eine von fünf Spendenorganisationen zu spenden (‚conditional donation'), (4) 2.000 Personen konnten nach Abschluss des Fragebogens zwischen einer 5-Euro-Spende oder einem 5-Euro-Spargutschein wählen (‚conditional choice'). Die Ergebnisse zeigen einen Trade-off zwischen der Maximierung der Rücklaufquote einerseits und der Maximierung der Panelrekrutierung andererseits. Die *Rücklaufquote* ist bei Personen, die einen ‚unconditional voucher' erhalten haben, am höchsten. Das Interesse an Folgebefragungen – und somit wichtig für die Panelrekrutierung – ist jedoch bei Personen, denen der Gutschein erst bei Beendigung der Umfrage in Aussicht gestellt wurde, am höchsten. Während sich beim ‚unconditional voucher' 38% für das Panel angemeldet haben, sind dies beim ‚conditional voucher' mehr als die Hälfte. Für das primäre Ziel eines Online-Panel-Aufbaus wurde basierend auf diesem Ergebnis ab der folgenden Welle, ein ‚conditional voucher' in Form eines 5-Euro-Gutscheins als regulärer Anreiz für alle Teilnehmer:innen eingesetzt.

Ad 2) In Welle 3 wurde getestet, ob die Umfrageumgebung sowie eventuelle Einstiegshürden die Teilnahme beeinflussen. 3.333 Personen erhielten einen direkten Zugang zur Umfrage über einen gekürzten Link; 3.334 Personen mussten neben dem gekürzten Link einen individualisierten Zugangscode zusätzlich eingeben; 3.333 Personen stiegen über die Projektwebsite in die dort eingebettete Umfrage mittels eines Zugangscodes ein. Die Rücklaufquote für alle Treatments war ähnlich: Es gab keinen statistisch signifikanten Zusammenhang zwischen der Art der Einstiegsvariante und der Panelrekrutierung.

Ad 3) In Welle 4 wurden unterschiedliche Arten von Einladungsschreiben getestet. 4.973 zufällig ausgewählte Personen erhielten ein einfach formuliertes, kurzes Einladungsschreiben, 4.973 Personen ein längeres und ausführlicheres. Die Rücklaufquote und die Panelrekrutierung unterscheiden sich zwischen diesen beiden Gruppen nicht. Des Weiteren wurden den Respondent:innen unterschiedliche Links übermittelt. Auch hier zeigen vorläufige Ergebnisse keinen Unterschied.

Zusätzlich zum Online-Panel-Pilot wurden im Rahmen von *Digitize!* der *Soziale Survey Österreich* (SSÖ) und zwei Module des *International Survey Programme* (ISSP) erstmals seit 1986 in einem Push-to-web-Design erhoben.[2] Die Befragten wurden offline rekrutiert; die Bruttostichprobe lag bei 4.150 gezogenen Adressen, wovon letztlich 1.387 Personen an der Umfrage teilnahmen. 1.056 konnten online befragt werden, 331 entschieden sich für die alternative Teilnahme via Papierfragebogen. Der Papierfragebogen wurde von Befragten ab 70 Jahren, Befragten ohne Matura und von Frauen vergleichsweise häufig genutzt. Die Umstellung der Erhebung des SSÖ/ISSP auf das neue Push-to-web-Design ist ein wichtiger Meilenstein in der Modernisierung von Bevölkerungsumfragen in Österreich. Der erfolgreiche Umstieg bedingte die Einführung erweiterter Qualitätskriterien für die Stichprobenziehung, eine sorgfältige Anpassung des Fragedesigns sowie entsprechende Incentivierung der Befragten. Die SSÖ-Erhebung im Rahmen von *Digitize!* hat zu einer nachhaltigen Veränderung zukünftiger SSÖ-Befragungen beigetragen, welche aufgrund der hohen Teilnahmebereitschaft auch zukünftige im Push-to-web-Design erfolgen werden.

2.1.2 Automatisierte Textanalyseverfahren für deutschsprachige Texte

Die Anwendung von computergestützter Textanalyse in den Sozialwissenschaften ist weit verbreitet, jedoch oft unsystematisch und ad hoc in der Anwendung. Ziel von *Digitize!* ist es, die Validität von Textanalyseverfahren für die deutsche Sprache zu verbessern und Routinen und Protokolle für eine kritische und standardisierte Nutzung zu entwickeln.

Zunächst wurde eine Übersicht von automatisierten Textanalysemethoden (ATAM), die in der Forschung mit Texten in deutscher Sprache wissenschaftlich verwendet wurden, erstellt. Alle Methoden, die in der Literatur mehrfach angewendet wurden, sind in der online Datenbank METEOR eingetragen und so einfach auffindbar und miteinander vergleichbar. Dort sind 44 Tools mit dazugehöriger Dokumentation gelistet. Da einige dieser Methoden nicht für den österreichischen Kontext erstellt wurden, sondern für den deutschen Sprachraum im Allgemeinen, wird daher in einem zweiten Schritt ein Sprachmodell entwickelt, welches speziell für die Analyse

2 Der SSÖ wurde von einem Team von Soziolog:innen der Universitäten Graz, Linz und Salzburg durchgeführt.

von Texten mit österreichischem Kontext geeignet ist. Das Sprachmodell wird derzeit auf Basis von Millionen von österreichischen Nachrichtenartikeln trainiert und Forscher:innen frei zur Verfügung gestellt. Damit ist es zwar deutlich kleiner als die Sprachmodelle, die derzeit von den großen Playern (Google, Facebook und OpenAI) zur Verfügung gestellt werden, weist dafür aber eine transparente, wissenschaftliche Dokumentation auf und birgt kein Risiko der Abhängigkeit von Konzernen. Das dritte Ziel ist die wissenschaftliche Forschung zu ATAM, um Empfehlungen für die systematische und valide Anwendung spezifischer Herangehensweisen zu formulieren.

In Bezug auf den letzten Punkt beschäftigt *Digitize!* sich vor allem mit der Methode der Themenmodellierung. Diese erlaubt es, große Sammlungen von Textdaten zu analysieren und darin vorhandene Themenstrukturen zu entdecken. Die Themenmodellierung ist in der Politik- und Kommunikationswissenschaft weit verbreitet, aber das Fehlen eines standardisierten Validierungsrahmens stellt derzeit eine große Herausforderung dar. Im Rahmen von *Digitize!* wurden daher 656 Forschungsarbeiten, die Themenmodelle angewendet hatten, analysiert, um Bewertungen und Metriken als Schlüsselkomponenten der Validierung zu identifizieren. Zweitens wurden die Auswirkungen der Validierungsmethoden auf die Modellauswahl bewertet. Dabei wurde festgestellt, dass unterschiedliche Methodiken der Themenmodellierung zu unterschiedlichen Ergebnissen in Bezug auf die Qualität der Anwendung führen, was sich auf die Resultate und dadurch auch die Theorieentwicklung auswirken könnte. Dies unterstreicht die Notwendigkeit eines standardisierten Validierungsrahmens, um eine hohe Zuverlässigkeit von Forschungsergebnissen zu gewährleisten. Dieser soll unterschiedliche Aspekte einer Validierungspipeline umfassend und transparent darstellen.

Des Weiteren wird die automatisierte Auswertung von offenen Fragenkategorien in Umfragen getestet, und zwar mithilfe von Themenmodellierung und semantischen Wörterbüchern. Diese Methoden des Natural Language Processing (NLP) können wertvolle Unterstützung in den Sozialwissenschaften bieten. Es werden vorab trainierte Sprachmodelle verwendet, um Muster in den Antworten zu identifizieren. Semantische Wörterbücher werden verwendet, um die Tonalität der Antworten zu bewerten. Automatisierte Textmethoden können ein leistungsfähiges Werkzeug für die Analyse offener Antwortkategorien sein, welches es Forscher:innen ermöglicht,

Muster und Einsichten induktiv aufzudecken, die durch manuelle Analyse allein nur schwer oder gar nicht zu erkennen wären.

Schlussendlich wurden Themenmodelle in Kombination mit Sprachmodellen zur Messung politischer Agenden untersucht. Durch die Einbeziehung verschiedener politischer Textarten (soziale Medien, Parlamentsreden und Presseaussendungen) wird versucht, eine umfassendere und genauere Beschreibung der öffentlichen Agenda von politischen Parteien zu liefern, was Forschenden helfen wird, Agenda-Setting-Prozesse jenseits begrenzter Zeitrahmen, Kommunikationsarten oder vordefinierter Themen zu untersuchen. Dabei wird erörtert, wie computergestützte Methoden zur Untersuchung von Medien-Bias eingesetzt werden können, indem die Ähnlichkeit zwischen der Sprachnutzung politischer Parteien und der Zeitungsberichterstattung gemessen wird.

Die automatisierte Textanalyse bietet der Sozialwissenschaft viele Möglichkeiten, die Forschung voranzutreiben. Während viele Methoden vor allem aus der Informatik und Computerwissenschaft importiert werden, ist eine sozialwissenschaftliche valide Implementierung notwendig, die über eine unkritische Ad-hoc-Anwendung von Verfahren hinausgeht.

2.1.3 Computational Social Science und die sozialwissenschaftliche Methodenlehre

Die Arbeit mit neuen Datenbeständen, die durch Digitalisierung entstanden sind, erfordert die Kenntnis von maßgeschneiderten Daten- und Messtheorien, Qualitätskriterien sowie entsprechenden Qualitätssicherungsverfahren und Strategien für die Datenanalyse. Dies bedingt, dass sich auch die sozialwissenschaftliche Lehre – insbesondere die Methodenausbildung – mit diesen Themenstellungen auseinandersetzt.

Dies wird in unterschiedlicher Form umgesetzt. Einerseits wurden zu Beginn der 2020er-Jahre Studiengänge wie Digital Society (Universität Linz) oder Computational Social Systems (Universitäten Graz und TU Graz) aufgebaut, die sich explizit mit Fragen der CSS auseinandersetzen. Gleichzeitig bedarf es aber auch entsprechender Kurse, die die Computerwissenschaftler:innen in die Forschungslogiken der Sozialwissenschaften einführen (e.g. LEITGÖB et al., 2023).

Entsprechend wurde in *Digitize!* in mehreren Schritten der Frage nachgegangen, welche Positionen sozialwissenschaftliche Lehrende zu digitalen Methoden haben, welche offenen Bildungsressourcen in dem Themenbereich existieren und welche Schlüsse sich daraus ableiten lassen.

In vier Wellen wurden zwischen 2020 und 2022 die Methodenlehrenden an Österreichs Universitäten quantitativ über ihre Zugänge zur Lehre digitaler Forschungsmethoden befragt (n = 124). Im Sommer 2021 wurden außerdem neun qualitative Interviews durchgeführt (e.g. PRANDNER & HASENGRUBER, 2021). Es zeigte sich, dass CSS für die Lehrrepertoires der Befragten nur geringe Relevanz hat. In der 4. Befragungswelle gaben nur 5,2% an, dass die CSS Teil ihrer Lehre seien. Knapp 20% beurteilten jedoch prozessgenerierte Daten als zukünftig (sehr) wichtig für die Sozialwissenschaften. Die qualitative Studie zeigte auch, dass die Digitalisierung der Lerntechnologie für pädagogische Zwecke als unmittelbar relevant gesehen wird (HASENGRUBER & PRANDNER, 2022), während die Inklusion von CSS-Inhalten in der Methodenausbildung eher ein Zukunftsthema darstellt.

Die Erfassung von deutschsprachigen *Open Educational Resources* (OERs) zu diesem Thema bildete diese Skepsis ebenso ab. Einerseits konnten generell nur wenige Inhalte identifiziert werden, die eine OER-Klassifizierung erfüllen – oftmals waren kommerzielle Interessen oder fehlende Lizenzen Grund dafür –, andererseits waren vor allem kritische Themen wie die Basis von sozialwissenschaftlicher Forschungsmethodologie und fortgeschrittene Auswertungstechniken nicht abgedeckt (PRANDNER & FORSTER 2022). Detaillierte OERs existieren aber für Datenerhebungstechniken.

Vor diesem Hintergrund hat der in Kooperation mit iMOOX und dem Center for Teaching and Learning der Universität Wien produzierte *MOOC Computational Social Science* besondere Bedeutung. Dieser umfangreiche MOOC behandelt Innovationen und Potenziale interdisziplinärer Forschung in den CSS aus unterschiedlichen Blickwinkeln. In sechs Lektionen werden u.a. neue Datenerhebungs- und -analyseverfahren in den interdisziplinären CSS, Herausforderungen und Chancen der Zusammenarbeit von Social & Data Scientists, relevante rechtliche, ethische und gesellschaftliche Fragen sowie konkrete Anwendungsmöglichkeiten in der universitären Forschung erläutert. Der MOOC richtet sich an Studierende sowohl der sozialwissenschaftlichen Disziplinen als auch der Data Science; aber auch an alle, die an interdisziplinärer Forschung und Lehre interessiert sind.

2.2 Perspektive der Data Science

In der letzten Dekade konnte die Forschung zu künstlicher Intelligenz wegweisende Durchbrüche erzielen, die unser tägliches Leben nachhaltig prägen werden. Allein im Jahr 2022 erfuhren Anwendungen wie ChatGPT, Stable Diffusion oder Midjourney eine unvergleichbare breite mediale und öffentliche Aufmerksamkeit. Umso wichtiger ist es, den Austausch zwischen den Sozialwissenschaften und der Informatik gezielt zu stärken und eine produktive Zusammenarbeit sowohl in der Forschung als auch in der Lehre nachhaltig zu etablieren.

Das Potenzial von Data Science ist in den Sozialwissenschaften bei Weitem noch nicht ausgeschöpft und verspricht zahlreiche neue Anwendungs- und Forschungsmöglichkeiten. In *Digitize!* wird erforscht, welche Data-Science-Methoden sich in den Sozialwissenschaften anwenden lassen und wie diese hierfür angepasst werden müssen. Der Fokus liegt dabei auf der Entwicklung skalierbarer Algorithmen zur Anwendung auf sozialwissenschaftliche Forschungsdaten. Im Rahmen von interdisziplinären Projekten wurden Problemstellungen und Anwendungsgebiete diskutiert, um das vorhandene Fachwissen der Datenwissenschaftler:innen gezielt für die Forschung der Sozialwissenschaftler:innen einsetzen zu können.

Ein vor allem durch ChatGPT bekannt gewordenes Teilgebiet der künstlichen Intelligenz, das derzeit außergewöhnliche Erfolge verzeichnet, ist die Textverarbeitung, auch bekannt als Natural Language Processing (NLP). Hierbei werden neuronale Netzwerke auf riesigen Datenmengen trainiert, um Zusammenhänge und Muster in Texten zu erkennen und die extrahierten Informationen dann für verschiedene Anwendungen einsetzen zu können. Auf diese Weise lassen sich beispielsweise Themen, sogenannte Topics, automatisch erkennen und deren zeitlicher Verlauf analysieren.

Gleichzeitig wird vertiefend an Modellen zur automatisierten Kategorisierung von Texten geforscht. Durch Anwendung verschiedener Methoden des maschinellen Lernens wurde ein Algorithmus zur hierarchischen Generierung von Topics entwickelt. Mit diesem Forschungsvorhaben ist es gelungen, die Grundlagenforschung im Bereich der Datenwissenschaften voranzutreiben und diese anschließend direkt auf Problemstellungen im Bereich der Sozialwissenschaften anzuwenden.

Neben dem Schwerpunkt im Bereich NLP werden neue Algorithmen zur Analyse von Umfragedaten und einer effizienteren Durchführung von Umfragen entwickelt.

Einerseits wird an der Anwendung erfolgreicher Techniken wie neuronalen Netzwerken auf ordinale Daten gearbeitet. Dabei handelt es sich um Daten mit geordneten kategorischen Werteskalen, beispielsweise Skalen mit Werten von *„Stimme sehr zu“* bis *„Stimme überhaupt nicht zu“*. Da im maschinellen Lernen die meisten Algorithmen entweder für rein numerische oder rein kategorische Daten entwickelt wurden, besteht ein großes Verbesserungspotenzial für Algorithmen, die auf diesen hybriden Datentyp angewendet werden können. Deren Entwicklung wird zu neuen Erkenntnissen bei der Analyse von Umfragedaten führen und dadurch einen wertvollen Beitrag in den Sozialwissenschaften liefern.

Daneben werden Modelle und Algorithmen entwickelt, die es ermöglichen, besser mit unvollständigen Daten bei Umfragen umzugehen. Fehlende Eingaben können auftreten, weil Teilnehmer:innen an Umfragen gewisse Fragen nicht beantworten wollen oder aus Zeitgründen Teile der Umfragen auslassen. Insbesondere wurden generative neuronale Netzwerke verbessert, sodass diese besser an unvollständigen Daten trainiert werden und akkuratere Vorhersagen für fehlende Eingaben treffen können. Dies ermöglicht eine verbesserte Nutzung der erhobenen Daten und eröffnet gleichzeitig das Potenzial, Fragebögen gezielt zu verkürzen, um dadurch sowohl die Qualität der erhaltenen Antworten als auch die Rücklaufquoten zu erhöhen – ein Ansatz, der im *Digitize! Online Panel* evaluiert wird. Primäres Ziel dieses Testlaufes ist es, die Anzahl der Fragen und damit die Dauer der Befragung für Befragte zu reduzieren, ohne dadurch den Informationsgehalt der Antworten zu verringern. Die Aufteilung der Fragen in mehrere Blöcke basiert hierbei auf dem Prinzip der Informationsmaximierung. Die Zuweisung der Teilnehmer:innen zu den Blöcken erfolgte randomisiert; zur Validierung erhielt eine Kontrollgruppe alle Fragen aus allen Blöcken. Eine Erweiterung ist der dynamische Fragebogen, welcher basierend auf den bisherigen Eingaben der Teilnehmer:innen live über die Auswahl nachfolgender Fragen entscheiden.

Darüber hinaus wurden Algorithmen für die Identifikation von kausalen Zusammenhängen aus Beobachtungsdaten weiterentwickelt, die bereits ohne randomisierte Zufallsexperimente einige Ursache-Wirkung-Beziehungen aufdecken können. Insbesondere ist es gelungen, bestehende Algorithmen zu erweitern, um Expert:innenwissen in den Identifikationsprozess einzubringen und dadurch die Anforderungen bezüglich der notwendigen Datenmengen für die akkurate Erkennung von kausalen Zusammenhängen zu reduzieren. In weiterer Folge ist die Evaluierung der

entwickelten Algorithmen auf Umfragedaten geplant, um dadurch beispielsweise Zusammenhänge zwischen Items in Umfragen zu identifizieren.

Auch die Lehre im Bereich der Sozialwissenschaften soll durch eine Einführung in maschinelles Lernen gestärkt werden. Der Grundstein hierfür wurde durch mehrere Video-Beiträge zu dem oben erwähnten neu konzipierten MOOC gelegt. In diesen Lehrvideos vermitteln Forscher:innen der Data Science die Grundkonzepte maschinellen Lernens für einen ersten Überblick. Darauf aufbauend wurde ein Konzept für eine neue Lehrveranstaltung ausgearbeitet, die den Studierenden dann anschließend die Möglichkeit bietet, ihre Fähigkeiten im Bereich des maschinellen Lernens gezielt zu vertiefen. Durch die integrative Herangehensweise wird hier eine Brücke zwischen den Sozialwissenschaften und den neuesten Entwicklungen im Bereich der Data Science gebaut, um Studierenden optimal auf die Anforderungen in der Forschung sowie der Privatwirtschaft hin auszubilden.

2.3 Die ethische Perspektive

Neben der forschungsethischen Begleitung unterschiedlicher Aktivitäten im *Digitize!*-Projekt beschäftigt sich *Digitize!* auch mit ethischen Fragen digitaler Praktiken in unserer Gesellschaft. Ein bis heute ungelöstes Problem ist die Tatsache, dass die meisten rechtlichen Normen und regulatorischen Instrumente den öffentlichen Wert der Datennutzung nicht berücksichtigen. Ein Resultat davon ist, dass die Nutzung von Daten für öffentliche und/oder gemeinnützige Zwecke denselben – und manchmal auch noch strengeren – rechtlichen und ethischen Richtlinien unterworfen sind wie kommerzielle Forschung, die keinen oder nur geringen öffentlichen Nutzen generiert.

Digitize! beteiligt sich daher an federführender Stelle an der Entwicklung des Ansatzes der Datensolidarität, die – ganz allgemein gesprochen – das Ziel hat, eine gerechtere Verteilung der Nutzen und Risiken, die sich aus digitalen Praktiken ergeben, zu erreichen.[3] Sie beruht auf drei Säulen: Erstens sollen jene Formen der Datennutzung, von denen große Vorteile für weite Teile der Öffentlichkeit (und insbesondere marginalisierte Gruppen) zu erwarten sind, ohne irgendjemandem signi-

3 https://www.tandfonline.com/doi/abs/10.1080/15265161.2023.2256267#:~:text=Data%20solidarity%20encompasses%20a%20radical,data%2C%20different%20rules%20should%20apply

fikanten Risiken auszusetzen, erleichtert werden – durch rechtliche Erleichterungen und öffentliche Förderung. Zweitens soll Schaden dadurch vermieden werden, indem besonders gefährliche Formen der Datennutzung, die es nur gibt, weil sie große Profite schaffen, verboten werden. Dort, wo Schaden nicht effektiv vermieden werden kann, soll es bessere und niedrigschwellige Unterstützung für die Betroffenen geben.[4] Drittens muss ein größerer Teil der finanziellen Profite privater Unternehmen, die durch digitale Daten ermöglicht wurden, auch wieder in die öffentliche Sphäre zurückfließen.[5]

Um diese Ziele umzusetzen, ist ein grundlegendes Umdenken in einigen wesentlichen Bereichen notwendig. So stellt der Ansatz der Datensolidarität auf den öffentlichen Wert ab, den unterschiedliche Formen der Datennutzung generieren, anstatt Risiken als bestimmten Daten*typen* inhärent zu betrachten (z. B. personenbezogene Daten oder Gesundheitsdaten). Er sieht vor, Datennutzung, von der plausibel erwartet werden kann, großen öffentlichen Wert zu generieren, ethisch und regulatorisch anders zu behandeln als Datennutzung, die keinen öffentlichen Wert generiert – und sogar hohe Risiken für Individuen oder Gesellschaften beinhaltet. Die Art der Daten, die hierfür verwendet werden, ist einer von mehreren Faktoren, die in der Bewertung des öffentlichen Nutzens berücksichtigt werden.

Der öffentliche Wert einer Datennutzung ergibt sich aus der Abwägung zwischen den Vorteilen, die durch eine bestimmte Datennutzung für bestimmte Gruppen oder die Allgemeinheit zu erwarten sind, mit den Risiken für individuelle Menschen oder ganze Gruppen, die die Datennutzung beinhalten. Dazu wurde von Forscher:innen des Arbeitspakets 8 des *Digitize!* Projekts im Rahmen von Kollaborationen mit Wissenschafter:innen im In- und Ausland ein Online-Tool entwickelt, das eine strukturierte Beurteilung des öffentlichen Nutzens ermöglicht. Das Tool, das seit Oktober 2023 öffentlich zugänglich ist (pluto.univie.ac.at), beinhaltet 21 Fragen, die eine gewichtete Abwägung von Nutzen und Risiken ermöglichen (so werden etwa sowohl Nutzen als auch Risiken für marginalisierte Gruppen schwerer gewichtet, als wenn sie privilegierte Gruppen betreffen). Die auf diese Weise durchgeführte Abwägung von Risiken und Nutzen führt zu einer numerischen Angabe des öffentlichen Wertes der Datennutzung. Die „Datennutzung" umfasst sowohl die Daten, die für die Ent-

4 https://academic.oup.com/medlaw/article/28/1/155/5543530

5 https://www.thelancet.com/journals/landig/article/PIIS2589-7500(22)00189-3/fulltext

wicklung oder das Training eines Modells verwendet werden, als auch das Modell, wenn es in der realen Welt eingesetzt wird.

Je nach auf diese Weise errechneten öffentlichen Wert unterscheidet Datensolidarität zwischen vier Arten der Datennutzung: Typ A (siehe *Abbildung 2*) schafft wahrscheinlich einen bedeutenden öffentlichen Wert und birgt keine hohen Risiken für Individuen oder Gruppen. Ein Beispiel aus dem Bereich der Socieal ist die Verwendung öffentlich verfügbarer Daten zur Analyse der Effekte struktureller Diskriminierung in Online-Publikationen (WAGNER et al., 2015). Der Grundsatz der Datensolidarität gebietet es, solche Formen der Datennutzung entweder durch eine Lockerung rechtlicher Anforderungen oder durch die Bereitstellung öffentlicher Mittel zu unterstützen.

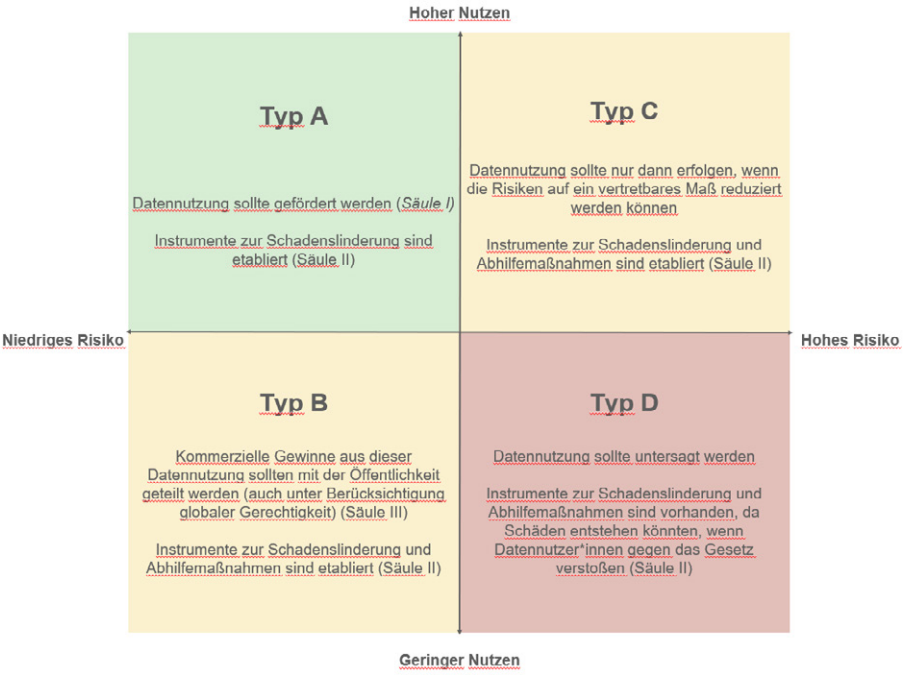

Abb. 2: adaptiert von PRAINSACK et al., 2022

Die zweite Kategorie der Datennutzung (Typ B) umfasst jene Praktiken, die wahrscheinlich keinen bedeutenden öffentlichen Wert schaffen, aber auch keine großen Risiken bergen. Ein Beispiel ist die Verwendung von Kundendaten durch ein Marketingunternehmen, um Werbung auf Zielgruppen zuzuschneiden. Bei dieser Art der Datennutzung sollte sichergestellt werden, dass ein Teil der erzielten Gewinne in den öffentlichen Bereich zurückfließt.

Die dritte Kategorie (Typ C) umfasst Datennutzungen, von denen zu erwarten ist, dass sie erheblichen öffentlichen Wert schaffen, die aber auch unannehmbar hohe Risiken bergen. Wenn beispielsweise in einem CSS-Projekt zu politischem Autoritarismus die Gefahr der Re-Identifizierung einzelner Studienteilnehmer:innen besteht, dann ist dies ein Fall des Typs C. Eine solche Datennutzung sollte nur dann erfolgen, wenn die Risiken auf ein akzeptables Maß reduziert werden können. In diesem Fall würde sich die Datennutzung des Typs C in eine Datennutzung des Typs B verwandeln und in weiterer Folge als solche behandelt werden.

Die vierte und letzte Kategorie umfasst Datennutzung (Typ D), von der nicht zu erwarten ist, dass sie öffentlichen Nutzen schafft, obwohl sie trotzdem hohe Risiken birgt. Ein Beispiel ist die Art der Datenanalyse, die von Cambridge Analytica zum Zweck der politischen Manipulation von Nutzer:innen der digitalen Plattformen durchgeführt wurde (SCHNEBLE et al., 2018). Datennutzung des Typs D sollte durch bindende rechtliche Normen untersagt und von genuin punitiven Rechtsfolgen begleitet sein. Die (grenzüberschreitenden) Durchsetzungsmechanismen müssen zudem effektiv genug sein, um selbst große multinationale Unternehmen von Gesetzesverstößen abzuhalten.

Für alle Arten der Datennutzung sollten in den Fällen, in denen die Schadensvermeidung versagt hat und ein Schaden eintritt, Instrumente zur Schadensbegrenzung zur Verfügung stehen. Diese Instrumente müssen einfach und schnell zugänglich sein und dürfen nicht davon abhängen, dass bestimmte Gesetze gebrochen werden oder dass die geschädigte Partei beweisen kann, wessen Handlung oder Unterlassung den Schaden verursacht hat (MCMAHON et al., 2020).

2.4 Die juristische Perspektive

In *Digitize!* stehen zwei Aufgaben im Fokus – eine nach innen und eine nach außen gerichtete: Nach innen wirkt die Sicherstellung der Compliance des Projekts; nach außen richtet sich die juristische Grundlagenforschung, aufbauend auf den Erfahrungen aus *Digitize!*.

Im Rahmen der Compliance sollte die Einhaltung der gesetzlichen Bestimmungen bei der sozialwissenschaftlichen Forschung garantiert und sollten so Projektrisiken minimiert werden. Dazu wurde im Februar 2021 ein Helpdesk für ethische und rechtliche Fragen eingerichtet. Dieser ist projektöffentlich, sodass sich grundsätzlich alle Projektpartner:innen an den Helpdesk wenden und mitlesen können. Der Helpdesk erlaubt, einerseits für die Einhaltung rechtlicher Vorschriften zu sorgen und andererseits die Aktivitäten des Projekts aus ethischer und rechtswissenschaftlicher Sicht zu begleiten. Mithilfe des Helpdesks konnten zahlreiche rechtliche Probleme behandelt und die Einhaltung der einschlägigen Rechtsbestimmungen in der Forschungspraxis gewährleistet werden. Beispielsweise wurden folgende Themen im Rahmen des Helpdesks behandelt: Aufbau und Begleitung des digitalen Onlinepanels, Beischaffung von Daten aus dem Zentralen Melderegister und deren rechtmäßige Verarbeitung, Vorlage für Einwilligungserklärungen, Verarbeitung besonderer Kategorien personenbezogener Daten sowie urheberrechtliche Fragen zur Analyse von Textdaten.

Ein Beispiel für das Zusammenspiel zwischen nach innen und außen gerichteter Arbeit wurde ab dem Herbst 2021 möglich: Im September 2021 stellte das Justizministerium seinen Entwurf für die Umsetzung der Richtlinie (EU) 2019/790 vor. Die Richtlinie sah Bestimmungen für die Privilegierung von Text- und Data-Mining (TDM) für wissenschaftliche Zwecke vor, welche in § 42h UrhG umgesetzt werden sollten. TDM ist eine Methode zur automatisierten Auswertung von Informationen über Muster, Trends und Korrelationen und wird auch im Rahmen von *Digitize!* angewendet. Verschiedene Projektpartner:innen beteiligten sich mit einer Stellungnahme an dem Gesetzgebungsverfahren. Erfreulicherweise wurden einige von *Digitize!* angeregte Änderungen des Entwurfs in der finalen Fassung des Gesetzes berücksichtigt. So werden nunmehr einzelne Forscher:innen neben Forschungs- und Einrichtungen des kulturellen Lebens im Gesetz erwähnt. Klargestellt wurde ferner, dass die Weitergabe der im Rahmen von TDM erstellten Vervielfältigungen zu Zwecken der Überprüfung der wissenschaftlichen Erkenntnisse in Forschungsgruppen

erlaubt ist. Gerade im Lichte der letzten Wochen, in denen das Potenzial großer AI-Sprachmodelle allgemein bekannt wurden („ChatGPT"), zu deren Entwicklung eine große Menge von Trainingsdaten benötigt wird, stellt sich erneut die Frage nach der Angemessenheit der bestehenden Regeln zum TDM, sodass die Thematik *Digitize!* – und die sozialwissenschaftliche Forschung – weiter beschäftigen wird.

Computational Social Science kann zu neuen Erkenntnissen und Innovationen sowohl im Bereich der Sozialwissenschaften als auch der Data Science beitragen. Aus rechtlicher und ethischer Perspektive werden bei zunehmender Verfügbarkeit großer Datensätze und neuer Analyseverfahren Fragen zur Durchführung und Anwendung von Forschung dringender und komplexer. Diese betreffen insbesondere den Schutz der Privatsphäre, des geistigen Eigentums und der Datensouveränität. Datensouveränität wird als Konzept über die Herrschaft über Daten herangezogen. Dies betrifft einerseits die Frage nach der Hoheit über Daten im territorialen Sinne und die Gewährleistung europäischer Datenschutzstandards, andererseits aber auch die persönliche Kontrolle über Daten durch natürliche Personen, welche insbesondere durch Wissen über Verarbeitungsvorgänge und Datentransparenz gewährleistet werden soll.

Bei zukünftigen Entwicklungen der CSS muss daher sichergestellt werden, dass insbesondere die einschlägigen datenschutzrechtlichen, datenrechtlichen und AI-rechtlichen Bestimmungen und ethischen Richtlinien eingehalten werden. Zu berücksichtigen ist ferner, dass sich aus der Nutzung fremder Datenbanken, Texte und anderer Werke auch potenzielle immaterialgüterrechtliche, insbesondere urheberrechtliche Implikationen ergeben. Bei künftigen Entwicklungen der CSS ist zu beachten, dass der Schutz des geistigen Eigentums Innovationen nicht behindert und der Zugang zu Daten und Ergebnissen für wissenschaftliche Forschung erhalten bleibt.

3 Perspektiven für zukünftige Entwicklung der Computational Social Science

Die erwähnten Forschungsergebnisse bezüglich der Entwicklung und Nutzung neuer Algorithmen an der Schnittstelle der Daten- und Sozialwissenschaften zeigen das hier bestehende immense Forschungs- und Lehrpotenzial auf. Konkrete Ziele umfassen beispielsweise den Aufbau eines offline-rekrutierten Onlinepanels, die bessere Vorhersage von Teilnahme an Panelumfragen basierend auf demografischen Merkmalen sowie dem Antwortverhalten von Teilnehmer:innen an Umfragen, als auch die systematische Verkürzung von Umfragen ohne Informationsverluste hinnehmen zu müssen. Methoden des Clusterings und des Lernens von niedrig-dimensionalen Repräsentationen können genutzt werden, um Zusammenhänge und Abhängigkeiten zwischen Fragen zu erkennen. Darüber hinaus können die Methoden für die Identifikation von kausalen Zusammenhängen genutzt werden, um die Sozialwissenschaften bei der Erstellung von Umfragen zu unterstützen und relevante Aspekte bzw. fehlende Details in den Fokus zu rücken.

Die computergestützten Sozialwissenschaften integrieren sozialwissenschaftliche Forschung und computergestützte Techniken. Die Vermittlung von Computerkenntnissen wie das Lesen und Schreiben von Code, das Verstehen der Auswirkungen von computergestützten Methoden und die Verknüpfung technischer Herangehensweisen mit sozialwissenschaftlichen Theorien sind für die sozialwissenschaftliche Hochschulbildung von entscheidender Bedeutung. Darüber hinaus sind ein ethisches und rechtliches Bewusstsein sowie die Fähigkeit, über weiterreichende Implikationen in Bezug auf Diversität und Ungleichheiten nachzudenken, wichtig. Der Einsatz von künstlicher Intelligenz und sogenannten Large Language Models wie beispielsweise GPT von OpenAI, auf dem ChatGPT basiert, kann die sozialwissenschaftliche Forschung bereichern, muss aber sinnvoll, realistisch und ethisch vertretbar eingesetzt werden. In Zukunft ist es wichtig, dass Forschende und Lehrende der Sozialwissenschaften über technologische Entwicklungen informiert bleiben und sich mit ihnen auseinandersetzen, um sicherzustellen, dass ihr Einsatz nachvollziehbar und vernünftig bleibt. Forschende der Datenwissenschaften können aus den theoretischen Ansätzen und Anwendungen der Sozialwissenschaften wichtige Forschungsfragen für ihr Feld ableiten. Die *Computational Social Science* wird also

in den kommenden Jahrzehnten eine gewichtige Rolle in der disziplinären und interdisziplinären Forschung und Lehre spielen.

4 Literaturverzeichnis

Hasengruber, K. & Prandner, D. (2022). Rebuilding our Toolkits for the future – How social science research educators changed their teaching in 2020 and 2021 to be fit for a digital future. In *8th International Conference on Higher Education Advances* (HEAd'22) (S. 371–378). Editorial Universitat Politècnica de València.

Leitgöb, H., Prandner, D. & Wolbring, T. (2023). Editorial: Big data and machine learning in sociology. *Front. Sociol., 8*, 1173155. https://doi.org/10.3389/fsoc.2023.1173155

McMahon, A., Buyx, A. & Prainsack, B. (2020). Big data governance needs more collective responsibility: The role of harm mitigation in the governance of data use in medicine and beyond. *Medical law review, 28*(1), 155–182.

Pfaff, K., Weitzel, D., Assenbaum, L., Brandl, D., Kvir, N., Meinel, J., Perner, W., Voith, V., Windisch, F. & Kritzinger, S. (2022). *Digitize!* Online Panel Survey (SUF edition). https://doi.org/10.11587/8SFV2L, AUSSDA, V3, UNF:6:033G0FjXOtdYab2QhPoMMw== [fileUNF].

Prainsack, B., El-Sayed, S., Forgó, N., Szoszkiewicz, Ł. & Baumer, P. (2022). Data solidarity: a blueprint for governing health futures. *The Lancet Digital Health, 4*(11), e773–e774.

Prandner, D. & Forster, M. (2022). Are There Enough Open Educational Resources Dealing With Social Science Research Methods? Insights From the D-A-CH Region. *Front. Educ. 7*, 902237. https://doi.org/10.3389/feduc.2022.902237

Prandner, D. & Hasengruber, K. (2021, July). Embracing the digitalization of research education? How social science research education was influenced by the COVID-19 pandemic. In *7th International Conference on Higher Education Advances* (HEAd'21) (S. 413–420). Editorial Universitat Politècnica de València.

Schneble, C.O., Elger, B.S. & Shaw, D. (2018). The Cambridge Analytica affair and Internet-mediated research. *EMBO reports, 19*(8), e 46579.

Theocharis, Y. & Jungherr, A. (2021). Computational social science and the study of political communication. *Political Communication, 38*(1–2), 1–22.

Wagner, C., Garcia, D. Jadidi, M. & Strohmaier, M. (2015). It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. In *The International AAAI Conference on Web and Social Media* (ICWSM2015), May 2015. Oxford, UK.

Autor:innen

Sylvia KRITZINGER || Universität Wien, Institut für Staatswissenschaft || Kolingasse 14–16, A-1090 Wien

https://staatswissenschaft.univie.ac.at/ueber-uns/wissenschaftliches-personal/sylvia-kritzinger/

sylvia.kritzinger@univie.ac.at

Katharina PFAFF || Universität Wien, Institut für Staatswissenschaft || Kolingasse 14–16, A-1090 Wien

https://staatswissenschaft.univie.ac.at/ueber-uns/wissenschaftliches-personal/katharina-pfaff/

katharina.pfaff@univie.ac.at

Julia BARTA || Universität Wien, Institut für Staatswissenschaft || Kolingasse 14–16, A-1090 Wien

https://staatswissenschaft.univie.ac.at/ueber-uns/projektadministration/julia-barta/

julia.barta@univie.ac.at

Jana BERNHARD || Universität Wien, Institut für Publizistik- und Kommunikationswissenschaft || Kolingasse 14–16, A-1090 Wien

https://publizistik.univie.ac.at/institut/mitarbeiterinnen-mitarbeiter/praedocs/bernhard-jana/

jana.bernhard@univie.ac.at

Hajo BOOMGAARDEN || Universität Wien,
Institut für Publizistik- und Kommunikationswissenschaft ||
Kolingasse 14–16, A-1090 Wien

https://publizistik.univie.ac.at/institut/mitarbeiterinnen- mitarbeiter/professuren-senior-staff/boomgaarden-hajo/

hajo.boomgaarden@univie.ac.at

Anja EDER || Universität Graz, Institut für Soziologie ||
Universitätsstraße 15/G4, A-8010 Graz

https://homepage.uni-graz.at/de/anja.eder/

anja.eder@uni-graz.at

Nikolaus FORGÓ || Universität Wien, Institut für Innovation und Digitalisierung im Recht || Schenkenstraße 4, A-1010 Wien

https://id.univie.ac.at/team/univ-prof-dr-nikolaus-forgo/

nikolaus.forgo@univie.ac.at

Filip PASPALJ || Universität Wien, Institut für Innovation und Digitalisierung im Recht || Schenkenstraße 4, A-1010 Wien

https://id.univie.ac.at/team/univ-prof-dr-nikolaus-forgo/team/paspalj-filip/

filip.paspalj@univie.ac.at

Claudia PLANT || Universität Wien, Fakultät für Informatik || Währinger Straße 29, A-1090 Wien

https://informatik.univie.ac.at/fakultaet/leitung/dekaninnen/person/59835/

claudia.plant@univie.ac.at

Barbara PRAINSACK || Universität Wien, Institut für Politikwissenschaft || Universitätsstraße 7, A-1010 Wien

https://politikwissenschaft.univie.ac.at/ueber-uns/mitarbeiterinnen/prainsack/

barbara.prainsack@univie.ac.at

Dimitri PRANDNER || Johannes Kepler Universität Linz, Institut für Soziologie, Abteilung für empirische Sozialforschung || Altenberger Straße 69, A-4040 Linz

https://www.jku.at/institut-fuer-soziologie/abteilungen/empirische-sozialforschung/team/dimitri-prandner/

dimitri.prandner@jku.at

Simon RITTEL || Universität Wien, Fakultät für Informatik || Währinger Straße 29, A-1090 Wien

https://dm.cs.univie.ac.at/team/person/114318/

simon.rittel@univie.ac.at

Martin TEUFFENBACH || Universität Wien, Fakultät für Informatik || Währinger Straße 29, A-1090 Wien

https://dm.cs.univie.ac.at/team/person/114104/

martin.teuffenbach@univie.ac.at

Sebastian TSCHIATSCHEK || Universität Wien, Fakultät für Informatik || Währinger Straße 29, A-1090 Wien

https://dm.cs.univie.ac.at/team/person/109359/#info

sebastian.tschiatschek@univie.ac.at