Mainstreaming Data Science across Domain Applications

Lessons from interdisciplinary research collaborations

Gammeltoft-Hansen, Thomas; Moeslund, Thomas B.

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

Mainstreaming Data Science across Domain Applications: Lessons from interdisciplinary research collaborations

Thomas Gammeltoft-Hansen[1] & Thomas B. Moeslund[2]
June 2023

I. Introduction

Computer science and data-driven research methods today have developed a hitherto unheard-of capacity to select, analyze and utilize the large amount of data that is available to us. Furthermore, a wide variety of collaborations harnessing modern digital technologies have been established in all segments of Danish society with the intent and hope to benefit both public and private entities in a broad sense. Thus, it is becoming increasingly clear that some of the most complex societal challenges can only be solved through interdisciplinary efforts, and that data science, in combination with other disciplines, has a huge potential to create new insights and scientific breakthroughs.

Reflecting this, targeted funding for this type of interdisciplinary research is increasingly made available from both Danish and international funding bodies. In Denmark, e.g., the Novo Nordisk Foundation and the Villum Foundation had a strategic focus on data science and established specific funding pools for interdisciplinary research (e.g., Villum Synergy) for several years, and in 2023, DKK 456,1 million was allocated to research related to digitalization and new technologies in the agreement on the Danish Research Reserve (Forskningsreserven 2023). A similar trend exists for international research bodies, where a growing proportion of e.g. ERC and NordForsk grants (e.g. NordForsk Interdisciplinary) support research projects combining data science and domain expertise. Over the last five years, large-scale grants have further enabled the establishment of dedicated hubs for interdisciplinary engagements between data science and other disciplines, e.g., the Pioneer Centre for AI (co-funded by Ministry of Higher Education and Science, Danish National Research Foundation, the Carlsberg Foundation, the Lundbeck Foundation, The Villum Foundation, and the Novo Nordisk Foundation), the Algorithms, Data & Democracy Project (Velux & Villum Foundations), and Digital Research Centre Denmark (Innovation Fund Denmark).

In practice, however, the application of data science methods is still new to many disciplines. For grant recipients, the expansion of funding opportunities creates new opportunities, but also hard-earned lessons in engaging with and managing interdisciplinary projects. For research foundations, cross-cutting feedback on pitfalls and best practices is similarly important to attune funding schemes and work with universities, industry and public institutions to create better infrastructures. The present report seeks to offer some initial reflections in this regard, drawing on the authors' own experience leading a Villum Synergy project together, as well as a number of other interdisciplinary projects with other partners[3]. In addition, we convened a one-day workshop to facilitate a broader exchange on experiences, obstacles and best practices for interdisciplinary collaborations involving data science. Held in early 2023, the workshop involved 44 participants, including peers

[1] Professor and Director, Center of Excellence for Global Mobility Law (MOBILE), Faculty of Law, University of Copenhagen.

[2] Professor and Director, Center for AI for the People & Visual Analysis and Perception Lab, Faculty of IT and Design, Aalborg University. The workshop on which this report is based was sponsored by a Villum Foundation Synergy Initiation grant "Explainable AI and Fairness in Asylum Law (XAIfair)", and the report informed by two additional projects funded by a NordForsk Interdisciplinary grant and a UCPH DATA+ grant.

[3] For overview, see: AI for the People (AAU); Nordic Asylum Law & Data Lab (UCPH).

involved in similar projects from five Danish universities with varying roles, expertise and seniority levels, as well as a number of stakeholders from Danish research foundations and industry[4]. The report is an attempt to capture some of the issues and insights raised, as well as set out some recommendations for funding bodies and host institutions.

II. Initiating cooperation

Several participants raised the issue that initiating cooperation for this type of research project is often difficult, namely due to disciplinary divides and limited networks between scholars across different research traditions. A number of project ideas had emerged as a result of "lucky accidents," due to chance encounters with colleagues from other disciplines at informal or private events. In a number of cases, participants had been dependent on "bridge-builders" – colleagues within their own discipline already engaged in interdisciplinary collaboration and offering to open up their network.

Scoping relevant cooperation partners can further be difficult due to a lack of sufficiently detailed knowledge of each other's expertise and field. Some participants reported initially having a somewhat "blackbox" understanding of the corresponding discipline, or finding out too late that the collaborator's methodological or thematic expertise was a less than ideal fit for the specific project idea.

A key point raised at the workshop was that good ideas for this type of research should ideally emerge as a result of genuine co-creation, opposed to simply being led by interests in one or the other discipline. Yet, this also requires that participants have a baseline understanding of each other's fields of research and establish a sufficiently "common language" (see further below) already at the application phase. Consequently, several participants reported that this type of project application was also more time-consuming to write and develop. Some had invested significant time sitting together in the development phase, with some even partaking in collective writing retreats. Expanding opportunities for pre-award exploratory cooperation, as well as meeting spaces and regular events for researchers with an interest in interdisciplinary research involving data science, may thus help raise both the quality and quantity of applications.

III. Data access, scoping and sharing

Another core issue raised by many participants relates to accessing, scoping and sharing involved datasets. In many cases, domain experts provide or facilitate access to the relevant data through their professional networks. Yet, in many cases, project partners underestimate the time needed to ensure access and transferring of data, as well as the work needed to pre-process or clean data before analysis can commence. Where data scientists have not had opportunity to scope datasets in advance, domain experts can tend to overestimate what can be achieved based on limited quantity or quality of data. Another issue that arose in the discussions is the increasing need for powerful compute resources in this type of research. This is often underestimated when drafting research proposals and hence why ad hoc solutions often need to be organized after the project commences. Nearly all participants reported delays in data access, even when pre-approval from data partners had been secured prior to submitting the project application – something creating a particular impediment for short-term or exploratory projects. In some cases, delays or obstacles emerged internally at the research institutions, especially where domain

[4] "New Frontiers for Data Science in Domain Applications: Insights from interdisciplinary collaborations", Aalborg University, 6 February 2023. A full participant list is included in Appendix I.

field support staff were not sufficiently well-resourced or experienced in handling large and/or sensitive datasets. Several participants further pointed to a lack of common approaches among both public and internal officials responsible for GDPR and data transfer agreements, leading to arbitrary outcomes or a "no culture" based on risk aversion.

Interdisciplinary projects of this kind tend to cross institutional boundaries. The need for data to be accessible to a broader group of people can be even more challenging when exceeding departments, institutions or national borders. While most research institutions have developed secure internal infrastructures for data storage and analysis, secure data sharing across institutions is often much more problematic. This can cause particular problems where the domain partner serves as the interlocutor for public or private data partners, but the technical expertise and processing infrastructure is provided by another institution. Discussions further revealed that data access and sharing are impacted by different rulesets and legal interpretation. Even among the Nordic countries, the legal and ethical frameworks differ substantially in terms of their approach and requirements for this type of research, requiring careful planning when undertaking comparative research. Some participants likewise reported different experiences when engaging with Danish municipalities.

Last, but not least, obtaining and cleaning data is often both time-consuming and expensive. For junior scholars involved in interdisciplinary projects, this can entail significant risks if data access is not secured before the beginning of a project; if pre-processing ends having consumed a large part of the project period; or if employments are not coordinated between project partners. Several participants noted that their projects had experienced significant periods where, e.g., PhDs or postdocs were unable to move forward with planned work due to unforeseen circumstances in this regard. Similarly, several PIs had felt the need to find additional resources towards the end of the project period to make up for lost time.

Thus, significant gains could further be made through better infrastructures for sharing and re-using preprocessed data across different research projects. In the current phase of interdisciplinary data science research, different projects often apply for access to and spend significant time cleaning the same data for different purposes. In many instances, this work is simply lost at the end of the project life cycle, requiring new projects to "reinvent the wheel." Funding bodies currently supporting this type of research would do well exploring the opportunities for addressing this issue together with public and private data partners, for example, by requiring grant recipients to give pre-processed/cleaned data back to data partners, or by establishing a national model for access to pre-cleaned public data akin to that operated by Statistics Denmark.

IV. Day-to-day interdisciplinarity

Data scientists and domain experts often have very different scientific languages and disciplinary epistemologies and communication is far from trivial when two disciplines are attempting to merge with the aim of addressing a common problem. Many have assimilated their own disciplinary norms so thoroughly it may become hard to articulate why things are done the way they are or to replace one's own professional language with a common language or a common frame of reference. Research findings and explanations from one domain might be experienced by the researchers outside that particular science domain as complex and impenetrable, while this might be very basic knowledge and seem trivial to the domain experts and vice versa. This can make it hard to meet in equal, have constructive research conversations or to disseminate the results jointly to find a common ground for what constitutes interesting research questions and findings.

A key insight from starting up interdisciplinary projects is devoting sufficient time for onboarding each other across disciplines. Creating a space for true collaboration and essential frameworks of understanding is often more time consuming than initially imagined. Many participants had to accept that they needed to devote a significant amount of time preparing the project group properly for collaboration. Several participants stressed the need to gradually develop a "common language" or 'merged scientific language' within the group, noting that such endeavors will inevitably have to move beyond some of the traditional technical terms and specific research concepts that are expected within each discipline, and may thus entail a tradeoff in terms of academic prestige within the home discipline.

Unless work plans and work packages are carefully designed to maximize interaction, both disciplinary and/or institutional silos can easily lead to fragmentation of projects and lack of synergy between sub-projects. Perhaps due to data science being an almost intrinsically interdisciplinary research area; other science domains might have a tendency to view data science as an add-on component rather than an equal collaborating partner. Another key point stressed by several participants is the importance of maintaining "epistemic parity" in the sense of each discipline having equal worth and value in both the formulation of research questions and defining the parameters for outputs and contributions. To benefit the most from interdisciplinary research, collaboration should avoid being founded on adding methods from one research field to another, and instead strive to synergistically combine theories, methods and concepts from different fields to arrive at new results of broad interest.

Physical proximity probably provides the most ideal framework for collaboration across disciplines, as this allows for daily exchange of ideas, mutual sparring and close coordination. Good examples have been seen where project groups shared both offices and PhD students across disciplines. Where this is not a realistic option, frequent meetings and, e.g., daily online check-ins between junior scholars was highlighted as possible ways to pull the team closer together and pave the way for more in-depth dialogues and co-authoring.

Physical proximity or regular check-ins between team members, however, is only part of the equation. Creating a collaborative research environment within an interdisciplinary group often demands leadership that goes beyond normal mono-disciplinary leadership skills. Arrangements must be made to ensure mutual openness and curiosity between the project group members. Members should feel comfortable asking any question without fear of being subsumed or sidetracked by the others and must be willing to share their knowledge with patience and generosity; many participants stressed the high importance of good social relations and a respectful working environment.

Even when a collaborative undertaking is thriving and produces research with a clear impact, structural barriers to the interdisciplinary production of research and individual career paths must be kept in mind. Few journals are genuinely geared towards interdisciplinary outputs, and differences in methodology and the presentation of evidence can make co-published works difficult to get accepted vis-à-vis formal requirements, word limits and peer reviewers. Likewise, few fields recognize publications in other disciplines as significant when it comes to tenure and promotion decisions. Even funding schemes with an articulated intention to support interdisciplinary research can have an assessment practice where bias in favor of mono-disciplinary research is present, e.g., through the selection of reviewers. While the importance of keeping a strong link to each discipline while embarking on interdisciplinary collaboration was mentioned by several participants,

striking this balance, however, is challenging. It requires careful planning of projected outputs, balancing the overarching project with individual goals. In practice, many projects assigned lead authors for different articles targeting their respective disciplines. Several participants further recommended that e.g., PhD students engaged in interdisciplinary projects should be allocated additional time, e.g., six months, to complete their degree.

V. Recommendations

The present report is mainly intended as a set of reflections and insights for current and future peers engaged in interdisciplinary data science research. Nonetheless, based on our own experience and the broader discussions at Aalborg University in February 2023, we offer the following six recommendations directed to funding bodies and research institutes sponsoring and hosting this type of research:

- Expand opportunities for pre-award exploratory cooperation, e.g., by creating regular events and meeting places for researchers with an interest in interdisciplinary research; or small grants to enable potential PIs more time together to develop larger research applications.

- Create awareness around challenges and best practices related to ensuring data access, data pre-processing and data sharing, e.g., by requiring applicants to submit clear protocols and timelines for data access and sharing; or encourage PIs to take more time between the awarding of a grant to commencement of projects to finalize data transfer agreements and set up internal infrastructures.

- Expand research time for junior scholars engaged in interdisciplinary projects, e.g., by affording additional time for PhD students, or encourage data cleaning and pre-processing to be conducted by non-research staff, e.g., graduate students or data specialists.

- Ensure that host institutions are aware of and sufficiently support the computing and data management requirements for this type of research, e.g., through upstart dialogues between the funding body, the PIs and support staff at the host institution before projects are commenced.

- Facilitate the sharing of best practices conducting interdisciplinary research, e.g., by organizing recurrent seminars between new PIs and more experienced grant holders.

- Improve long-term structures for data-driven interdisciplinary research, e.g., through regular dialogues between funding bodies and public and private sector data partners to ensure better possibilities and more unified approaches for sharing datasets across research projects.

- Expand possibilities for data reuse and centralised handling of pre-processing and cleaning of data, e.g., through establishing national or cross-university structures for secure storage and mutual recognition, and/or requiring grant holders to transfer back processed data to public/private data partners at the end of the project period.