Towards a Data Science for General Collective Intelligence

Andy E. Williams, Nobeah Foundation, Nairobi, Kenya
awilliams@nobeahfoundation.org

Abstract

The recently emerging science of Human-Centric Functional Modeling hypothesizes that for systems whose behavior can be seen as occurring within different domains, all possible states of those systems and all possible behaviors through which those systems might transition from one state to another in a given domain of behavior can be represented in terms of functional state spaces. In this sense functional state spaces are then potentially a complete representation of data regarding systems. If complete representations of the human meaning of data are complete semantic representations, then these representations of information in functional state space are also complete semantic models of information. Implementations of such functional state spaces are hypothesized to be graph databases. A General Collective Intelligence is a hypothetical platform that leverages such semantic models to gain a predicted exponential increase in the general problem-solving ability of groups, which entails a significantly increased capacity to navigate data towards a solution to any problem that can be defined in such data. This paper explores the implications of these emerging fields in data science.

Keywords:

General Collective Intelligence, functional state space, dynamic decentralization, data science, graph database

Introduction

A General Collective Intelligence or GCI [1] is a hypothetical platform that self-assembles individuals, and/or intelligent agents in a self-sustaining way into potentially massive networks of cooperation able to execute collective reasoning with exponentially greater general problem-solving ability, and therefore exponentially greater problem-solving ability that might be targeted at solving any problem in general. A GCI leverages one functional state space (the collective conceptual space) to represent all possible collective reasoning processes and all possible computing processes which might automate that reasoning, and a GCI leverages another functional state space (cooperation state space) to represent all possible patterns of cooperation through which outcomes of that collective reasoning or computing might be scaled in order to exponentially increase the capacity for groups to execute any possible reasoning or processing [2].

Functional State Space as the Graph Database for Individual User Data

As each individual thinks, Human-Centric Functional Modeling or HCFM represents that individual's cognition as using reasoning to move through a space of concepts or a "conceptual space that represents the functional state space of the cognitive system. This conceptual space is then a representation of all information within the knowledge of that individual. A complete representation of conceptual space as a graph has not yet been implemented, since properties such as distance between concepts have yet to be precisely defined. However, it is hypothesized that any subset of conceptual space defines an ontology (figure 1), and subsets of conceptual space relevant to specific problem domains have been successfully been approximated by such ontologies used by GCI platforms. One example of such a platform design is the Social Impact Marketplace, which is intended to automate cooperation capable of radically increasing impact on the livelihoods of agricultural communities as a collective goal [3].

As in the example provided by the design of this platform, only three patterns are required to model a very large number of ways that cooperation can be used to exponentially increase impact on a wide variety of problems. These patterns are represented by an ontology within this platform [4]. The question is how these patterns can be applied in the design of data science tools?

Collective Functional State Space as the Graph Database of a General Collective Intelligence

Consider a simple query to the GCI from some individual who asks “are you interested in this movie review I’ve made?”, and who directs that question to all individuals in their immediate network. In conceptual space (the functional state space of the cognitive system), a question, or any other problem, is a gap between one concept and another. A solution is the reasoning providing the path between those two concepts. In this case, this question is represented as the gap between the concept of the movie review, and the concept of the personal interest in that review of the individual being asked the question. In the GCI algorithm the value created by answering this question is predicted, and then this question is sent by the GCI to the network of the individual along with an offer of some portion of that projected value as an incentive to participating in answering the question, and along with an offer of some other portion of that projected value as an incentive to recruit other individuals to answer the question. Each individual in the first person’s network might then respond by answering whether they are interested in that movie review, and/or respond by asking others in their own network the same question. After interacting with the second order network (the network of people within the network of the people first asked the question), this interaction proceeds to the third order network, and so forth on towards the Nth order network (the network of people N times removed from the individual who originally asked the question). Since the information available to each of these individuals in this Nth order network is represented by a graph of conceptual space, then the composite of all these graphs represents the information available to the entire group. That is, the information available to the entire group represents the collective conceptual space that acts as the functional state space of their collective intelligence.

This question “are you interested in this movie review I’ve made?” might be answered approximately with information that might be contained in a simple ontology. One simple ontology as an example, might confine answers to “yes” or “no”, so that other answers like “unsure” are not possible. It is hypothesized that any subset of conceptual space forms an ontology when for each concept in that subset of conceptual space there is a reasoning process also in that conceptual space that connects that concept to other concepts. contains one reasoning process For questions for which an approximating ontology exists, a complete implementation of the individual and collective conceptual space might not be required.

Figure 1: Any subset of functional states can form an ontology.

However, consider a more complex query to the GCI from some individual who asks “how does the Acacia plant affect depression”? In conceptual space this question is represented as the gap between the concept of the Acacia plant, and the concept of the depression of the subject of the question. Now the Acacia plant might be combined with other ingredients in some yet to be discovered ways that radically increase its impact in conquering depression. In order for a GCI to be able to discover these potential medications, the GCI must have the capacity to discover any combination of processes that might create medications containing the Acacia plant. Assuming each of these processes has a representation in conceptual space, then this search for any combination of processes that might create medications containing the Acacia plant can be accomplished through a search of conceptual space.

Each process in any given combination of processes that is being searched will be linked to the previous process by an input ingredient that is the output of the previous process. Each set of inputs or outputs will be represented by a set of concepts. Since each set of concepts is itself a concept, any set of concepts will be represented by a bounded region in conceptual space corresponding to that single concept representing the group of concepts. Each reasoning process in the chain of reasoning defining the processing steps in the Acacia plant based cure for depression will then be represented by a segment in a path through conceptual space from one concept potentially representing a set of ingredients, to another concept representing a set of products. The search of all the collective conceptual space for all possible cures must have the capacity to discover any combination of such processes (figure 2).

Figure 2: Searching conceptual space for a cure

Since it is hypothesized that there are a number of categories of regions in the collective conceptual space that humans as groups can't systematically explore without GCI, there are a number of categories of regions in the collective conceptual space of a group that are potentially hidden without GCI. Assuming it is true that the self-assembling processes possible through GCI can use three basic patterns of solutions [?] to explore an exponentially greater region of the collective conceptual space, this would

be expected to result in a significantly greater number of potential solutions and significantly increased impact on collectively targeted outcomes (for example in one conceptual case study a projected 750 times increase in agricultural livelihoods per donor program dollar spent [3]). If so, GCI is then deeply tied to data discovery. The process of mining for all possible chains of collective reasoning is then an important pattern with which GCI might be used to increase the general problem-solving ability of the group.

Functional State Space as the Query Language for User Data

As of the writing of this paper, no universal graph query language has been adopted in the same way as SQL was for relational databases. However, a proposal to create a new standard graph query language (ISO/IEC 39075 Information Technology — Database Languages — GQL) was approved by members of ISO/IEC Joint Technical Committee 1 (ISO/IEC JTC 1) in 2019 [5].

With the conceptual space however, natural language is the universal graph query language, since any well-formed natural language question is a valid problem in terms of representing the lack of a path between one or more concepts in that space. In the role of the conceptual space as the functional state space for the cognitive system, information is retrieved from conceptual space through reasoning processes that actively direct the cognitive system from one concept to another. Or information is retrieved from conceptual space through understanding processes that passively direct the cognitive system from one concept to another. The answer to the question “am I interested in this movie review that a particular person [Name of Person] has made?” might be given by any one of a great many reasoning or understanding processes. If these processes are semantically modeled, then the query language to find them is naturally spoken language. Any query with a functional model that is represented by the same input concept and the same output concept, with the same concepts defining the context of execution, is a valid query expression in natural language. In figure 3, a query “what is the individual’s favorite movie” is used to discover a reasoning process P_1 in the conceptual space of user A that is able to answer that query for user A. That same query is used to discover a reasoning process P_2 in the conceptual space of user B that is able to provide an answer for user B.

Individual Conceptual Spaces

Figure 3: (Left) Conceptual space of person A uses reasoning process P_1 to find favorite movie. (Right) Conceptual space of person B uses reasoning process P_2 to find favorite movie.

Paths in Collective Functional State Space as the Query Language of a General Collective Intelligence

In order to answer the query for multiple individuals we construct a “collective conceptual space” that defines a semantic model for all the information in the hypothetical “collective cognition”. In order to be able to answer the question “what is your favorite movie” for all individuals, the collective cognition (the GCI) must determine which reasoning process (P_1 in the case of person A, and P_2 in the case of person B) is most fit in answering the query for each individual as in figure 4.

Figure 4: A GCI operating in the collective conceptual space of persons A and B combined must determine which reasoning process is most “fit” for each individual in order to correctly use reasoning process P_1 to find the favorite movie of person A and in order to correctly use reasoning process P_2 to find the favorite movie of person B.

GCI defines a framework for modeling the fitness of any reasoning process in achieving any targeted outcome, so once all processes are represented in the collective conceptual space, a GCI might use this metric for fitness to discover responses to such queries.

Architecture Level Services for General Collective Intelligence as a Platform

In addition to executing business logic that users might select or otherwise determine through collective reasoning processes, there are a number of architecture level services that a GCI platform might execute. These services, like networking, security, as well as data access and storage, result in changes to architecture service level characteristics like scalability, maintainability, security, and resilience. These services are provided through modeling the underlying functionality using a separate functional state space for each domain of behavior, then taking any existing software which provides that functionality and decoupling it into a library of functional components according to the functional model of that state space, and then integrating that library with the GCI so the GCI can choose to implement the required functionality with whichever library function is most fit in each given situation [12]. These architecture level services also need to interoperate with any data science methods. This might potentially be achieved by abstracting such services out of any tools with which any data science methods are implemented. How this might be achieved remains to be elaborated.

Case Study: Healthcare

As a hypothetical example, consider the concept of the GCI based “National Electronic Health Record/ Single Virtual Patient Record” (EHR/SPR) platform that is proposed to be deployed during the second phase of a planned Collective Intelligence based Program to Accelerate Achievement of the Sustainable Development Goals (CIPAA-SDGs). The theory of GCI, and conceptual case studies of this GCI based healthcare application concept as well as other GCI based application concepts, suggest that sustainability (e.g. climate change), poverty, and other systemic problems are properties of functional state space in that they appear to be represented by the same patterns in every functional state space representing some domain of behavior for every relevant system [14]. Case studies of various GCI based solutions proposed for this CIPAA-SDGs program, including of this concept of a GCI based

EHR/SPR platform have demonstrated the potential to exponentially increase our collective capacity to address the problems targeted by those solutions. By understanding these patterns, and by defining how an increase in impact can be quantified as "exponential" these general patterns of solutions might increase the collective capacity solve such wicked, existential challenges until they are predicted to be reliably solvable where the model suggests this has never before been the case at any time in human history.

In the case of the EHR/SPR, though this platform might initially define a limited ontology for patient data that is able to represent only an incomplete model of handwritten records, over time this limited ontology might be evolved into a "human state space" that provides a complete semantic model of the state of the human system and is therefore able to capture the complete meaning of handwritten records. Using Human-Centric Functional Modeling as required by GCI, each functional component of the human organism moves through it's own functional state space, the body moves through a sensory-motor state space, the emotions move through an emotional space, the mind moves through a conceptual space, and the consciousness moves through a space of awarenesses, or an awareness or consciousness space. These functional state spaces form a composite "human state space". Furthermore, the motion of groups of individuals through this collective human state space might be used to describe the impact on collective well-being as a result of any health care policy.

One of the fundamental principles by which collective outcomes might be significantly increased over those achieved by any individual is collectively intelligent cooperation that aligns outcomes with collective well-being as opposed to aligning outcomes with the interests of shareholders in some pharmaceutical company or with the interests of some other subset of individuals in the group. A fundamental part of collectively intelligent cooperation in the General Collective Intelligence platform is putting individuals in charge of their own data so that decision-making can be decentralized. By creating a personal repository for personal healthcare data which resides in each patient's sole control, and by creating trusted intelligent agents within that repository that can automatically answer queries on behalf of that individual according to policies the individual sets, a collectively intelligent healthcare platform might directly pay individuals to have their personal agent submit anonymized healthcare data in response to queries. By semantically modeling this data in terms of a "human state space" so it has intrinsic human meaning, those agents could potentially answer queries even if they have incomplete data, or if the data is in a different format from that expected by the query. With millions of intelligent agents working on behalf of each user in this way, it might be possible to begin to conduct massive decentralized health care or medical research studies that can be completed in microseconds rather than in months or years. These studies might facilitate observing the effects of all of a much wider range of potential medicines.

By semantically modeling the processes by which interventions are executed, by predicting the outcomes of each intervention, and by creating a means to objectively compare and select outcomes, a GCI might enable simulations to optimize the outcomes of those interventions. Predicting outcomes of each intervention is possible wherever outcomes are a known function of the intervention (wherever outcomes are computable). Furthermore, predicting outcomes of each interventions is possible using pattern detection where there is an observable relationship between intervention and outcome but the mechanism isn't known (wherever outcomes are not computable). Finally, by measuring the difference between predictions and actual results, those predictions could be continuously improved.

Adding decentralization to the semantic data storage, and to the semantic process storage, as well as to the capacity for simulations created by defining this human state space, also introduces the possibility of maximizing outcomes through massive decentralized simulation of policy outcomes on a population

wide scale. It is the position of this author that no existing simulation or other decision-support system is able to model every single individual and to decide on interventions that maximize outcomes in terms of well-being for each individual, as opposed to modeling some subset of individuals or even modeling a single average individual and assuming that when it comes to interventions “one size fits all”. The result applied to health care for example, is interventions that very predictably are completely unable to even become aware of the state of health in some individuals or unable to become aware of the failure to address ill health in some individuals. A simulation that leverage GCI to become massively decentralized on the other hand could potentially select a representative subset of the population and simulate outcomes of the healthcare policy on those individuals. Rather than having to invade privacy by collecting each individual's data in order to do such a simulation, a decentralized simulation approach in effect might just asks the agent governing each individual’s personal data repository to anonymously "simulate the given circumstance", and then might compile all the answers into a demographic profile for the geography, which might be used to calculate a metric of collective well-being for each policy intervention.

Discussion

One researcher cited a number of research challenges in data science they believed to be relevant to science, to technology, and to society [6]. How General Collective Intelligence might help address all of these challenges when applied to data science is discussed in table 1.

Research Area	Predicted Impact of General Collective Intelligence
Scientific Understanding of Learning, Especially Deep Learning Algorithms.	Whether trying to understand type 1 (intuition or pattern detection based) reasoning, or whether trying to understand why a deep learning model as an artificial implementation of pattern detection based or type 1 reasoning produces one result and not another, leveraging a GCI to combine all available data into a single semantic data model, and leveraging GCI to model all possible theories is predicted to exponentially increase the number of theories it is possible to validate while also exponentially increasing the size of the data set that might be used validate those theories against. Such theories might concern how robust or fragile models are to perturbations to input data distributions, how to verify that deep learning will perform the intended task well on new input data, how to characterize or measure the uncertainty of a model’s results, or about other questions such as deep learning’s fundamental computational limits, or at what point does more data and more compute not help [6], [7]. If the collective capacity to test such theories can be represented in terms of the volume of the collective conceptual space that can be navigated per unit time, GCI is predicted to exponentially increase that volume, and hence is predicted to exponentially increase that capacity.
Causal Reasoning	By leveraging GCI to semantically model all possible causal reasoning techniques, data scientists might leverage new methods to make causal inference estimation more efficient and flexible [8], and might explore multiple causal inference as is relevant to most real-world observations which are due to multiple factors that interact with each other [9].
Precious Data	It has been hypothesized that data can be categorized as “precious” for one of three reasons: the data set is expensive to collect; the data set contains a rare event (low signal-to-noise ratio); or the data set is artisanal—small, task-specific, and/or targets a limited audience [6]. General Collective Intelligence again has the potential to combine all data, and all data science methods into a single semantic model, so that for each of the different kinds of precious data, it might be possible to discover optimal new data science methods and algorithms that take into

Multiple, Heterogeneous Data Sources	<p>consideration the domain and the intended uses and users of the data.</p> <p>While models might be improved and knowledge increased for some problems by collecting lots of data from different data sources, it has been claimed that state-of-the-art data science methods cannot as yet handle combining multiple, heterogeneous sources of data to build a single, accurate model [6]. Furthermore, the difficulty of understanding the uncertainty in a data model is exacerbated when built from multiple, possibly unrelated data sources. Using GCI to define a single semantic data model might provide a mechanism for the standardization of data types and data formats, and in doing to might reduce undesired or unnecessary heterogeneity. If as has been said, “focused research in combining multiple sources of data will provide extraordinary impact” in this area, then if GCI is a means for achieving such focus GCI this suggests the impact of GCI is potentially important.</p>
Inferring From Noisy and/or Incomplete Data	<p>Data scientists must build models from data that is at times noisy and from uncertain sources in order to do prediction and inference [6]. The problem is then how to better separate noise from signal and how to improve the efficiency and accuracy of the inferences we make. Again, GCI creates the potential to semantically model all data as well as all possible inferences and theories about the efficiency, accuracy, or any other properties of those inferences, so that we can exponentially increase our ability to solve that problem.</p>
Trustworthy AI	<p>The theory of GCI suggests that AI is reliably untrustworthy in a number of well-defined areas, that include wherever the AI is aligned with the interests of its designers [13]. This same theory also suggests that GCI is the only known way to reliably overcome that fundamental untrustworthiness [13].</p>
Computing Systems for Data-Intensive Applications	<p>Traditional designs of computing systems are more focused on increasing computational speed and power in processing a given set of data, than in processing more data [6]. In some cases it can still take weeks to build accurate predictive models where real-time predictions are ideal. GCI has the potential to maximize capacity for heterogeneous processing of massive amounts of data for fast access. Since GCI can be targeted at any problem in general, it also has the potential to maximize communication and network capability, energy efficiency, and other attributes required for improved data processing.</p>
Automating Front-End Stages of the Data Life Cycle	<p>Data needs to be prepared for analysis through techniques that are still often labor intensive and tedious before deep learning or other data science techniques can be applied [10]. GCI might be used to define semantic models of any automated methods that address data collection, data cleaning, and data wrangling, without losing other desired properties, for example, accuracy, precision, and robustness, of the end model, and GCI might exponentially increase the capacity to search for the optimal model that might be applied.</p>
Privacy	<p>Because GCI is a pattern of cooperation organized between individuals or individual agents working under each individual’s sole control, which each access the user’s private repository of data, identities, and applications under that user’s sole control, privacy can be ensured by the user themselves before the data is submitted. In this way GCI maximizes the collective capacity to preserve the confidentiality of each party’s data set in ways that are not reliably achievable otherwise.</p>
Ethics	<p>The theory of GCI suggests that the current understanding of the ethical issues surrounding data science is misleading and incorrect. These points can be framed along three axes: (1) concerning the ethics of data: how data are generated, recorded, and shared. The theory of GCI suggests that without an additional layer of</p>

decentralization in terms of all user data, identities, and processes coming from private repositories within the user's sole control, the use of data cannot be free to align with collective user well-being as opposed to aligning with the interests of centralized decision-makers. 2) the ethics of algorithms: how artificial intelligence, machine learning, and robots interpret data. The theory of GCI suggests again that without this additional layer of decentralization any algorithms cannot be free to align with collective well-being. (3) the ethics of practices: devising responsible innovation and professional codes to guide any data science, GCI or any other emerging science [11]. The theory of GCI suggests that no professional codes or other policies can evolve to ensure collective well-being at the rate at which centralization can evolve to exploit collective well-being in order to serve the interests of some centralized decision-maker. Furthermore the theory of GCI predicts that this capacity for exploitation must increase with time in the absence of GCI, and that any such policies provide a dangerously false sense of protection. GCI however, is predicted to protect collective well-being at even an exponentially faster pace of innovation in data science.

Table 1: *Predicted impacts of General Collective Intelligence in various research areas considered to be important to data science.*

Conclusions

If (as hypothesized) an exponential increase in ability to solve any problem in general is achievable through GCI, and if that exponential increase in ability is not reliably achievable otherwise, then GCI is potentially the most important area of future innovation in all disciplines, including data science. This article has discussed the potential application of GCI to address research challenges in data science believed to be relevant to science, to technology, and to society. However, GCI is still very new and almost completely unknown. As many universities and colleges are creating new data science schools, institutes, centers, and so on [10], General Collective Intelligence is not yet on the radar. At this early stage the applications of GCI to data science are purely theoretical and speculative. However, it is hoped that this paper, as an exploration of this potential for impact, might inspire actual experimentation.

References

- [1] Williams, A. E., Defining a Continuum from Individual, to Swarm, to Collective Intelligence, to General Collective Intelligence, International Journal of Collaborative Intelligence, in press (2022)
- [2] Williams, Andy E., A Generalized Model for Distributed, Autonomic and Self-Organizing Computing, Accepted to the Proceedings of the 2022 International Symposium on Computing and Artificial Intelligence (ISCAI 2022).
- [3] Williams, A. E., Approximating an Artificial General Intelligence or a General Collective Intelligence, International Journal of Collaborative Intelligence, in print (2021)
- [4] Williams, A. E. (2022, February 27). Using Models to Capture Patterns for Radically Increasing Social Impact. <https://doi.org/10.31730/osf.io/rqs72>
- [5] Sharma, Chandan. "FLUX: From SQL to GQL query translation tool." 2020 35th IEEE/ACM International Conference on Automated Software Engineering (ASE). IEEE, 2020.

- [6] Wing, J. M. (2020). Ten Research Challenge Areas in Data Science . Harvard Data Science Review. <https://doi.org/10.1162/99608f92.c6577b1f>
- [7] Thompson, N. C., Greenewald, K., Lee, K., & Manso, G. F. (2020). The computational limits of deep learning. <https://arxiv.org/abs/2007.05558>
- [8] Athey, S. (2016). Susan Athey on how economists can use machine learning to improve policy. Stanford Institute for Economic Policy Research. <https://siepr.stanford.edu/news/susan-athey-how-economists-can-use-machine-learning-improve-policy>
- [9] Wang, Y., & Blei, D. M. (2019). The blessings of multiple causes. Journal of the American Statistical Association, 114(528), 1574-1596, <https://doi.org/10.1080/01621459.2019.1686987>
- [10] Wing, J. M. (2019). The data life cycle. Harvard Data Science Review, 1(1).
- [11] Floridi, L., & Taddeo, M. (2016). What is data ethics? Philosophical Transactions of the Royal Society A, 374(2083), Article 20160360. <https://doi.org/10.1098/rsta.2016.0360>
- [12] Andy E. Williams, "Cognitive Computing and its Relationship to Computing Methods and Advanced Computing from a Human-Centric Functional Modeling Perspective", In: Pal R. and Shukla K. P. (eds), SCRS Conference Proceedings on Intelligent Systems, SCRS, India, 2021, pp. 107-122. <https://doi.org/10.52458/978-93-91842-08-6-9>
- [13] Williams, A. E. (2021, January 11). AI Safety and General Collective Intelligence. <https://doi.org/10.31730/osf.io/gw3ks>
- [14] Williams, A. E. (2022, February 1). Is Sustainability a Pattern of Solution in a Functional State Space?. <https://doi.org/10.31730/osf.io/cb2xh>