Res Publica (2023) 29:303–319
https://doi.org/10.1007/s11158-022-09544-5

AI and the Social Sciences: Why All Variables are Not Created Equal

Catherine Greene[1]

Accepted: 7 January 2022 / Published online: 17 February 2022

Abstract

This article argues that it is far from trivial to convert social science concepts into accurate categories on which algorithms work best. The literature raises this concern in a general way; for example, Deeks notes that legal concepts, such as proportionality, cannot be easily converted into code noting that 'The meaning and application of these concepts is hotly debated, even among lawyers who share common vocabularies and experiences' (Deeks in Va Law Rev 104, pp. 1529–1593, 2018). The example discussed here is recidivism prediction, where the factors that are of interest are difficult to capture adequately through questionnaires because survey responses do not necessarily indicate whether the behaviour that is of interest is present. There is room for improvement in how questions are phrased, in the selection of variables, and by encouraging practitioners to consider whether a particular variable is the sort of thing that can be measured by questionnaires at all.

Keywords Recidivism · Social data · Nomadic concepts · Survey data · Social science variables

Introduction

The recent literature argues for collaboration between social scientists and data scientists (for example see Miller 2019). This is not just because social scientists can provide general insights on methodology, but because machine learning is increasingly used in social applications, such as analysing voting patterns of US senators or hiring patterns in universities (Wallach 2018). However, as Wallach writes, 'we must treat machine learning for social science very differently from the way we treat machine learning for, say, handwriting recognition or playing chess. We cannot just apply machine learning methods in a black-box fashion, as if computational social science were simply computer science plus social data' (2018, p. 44). This article

✉ Catherine Greene
c.m.greene@lse.ac.uk; catherinegreene76@gmail.com

[1] London School of Economics (CPNSS), Houghton Street, London WC2A 2AE, UK

addresses the problem of social data in recidivism prediction. Philosophers of social science often highlight the oddness of variables and concepts used in the social sciences when contrasted with variables used in the natural sciences. The social sciences are subject to reflexivity when people change their behaviour in response to the ways in which they are described, and to *ceteris paribus* clauses, which reflect the complexity of many social situations. Concepts used in the social sciences, such as happiness or development, are also difficult to define in terms of necessary and sufficient conditions. This is a particular problem for algorithms that seek to predict recidivism, where variables such as criminal associates are difficult to gauge with questionnaires. This article argues that such concepts should be carefully assessed before including them in prediction algorithms.

Questionnaires provide data that appears precise because respondents pick from a number of pre-set responses. This article shows that this precision is often illusory, and that respondents picking the same response may have very little in common with each other or with the characteristics that the criminological literature suggests are predictive of future offending. It applies my Nomadic framework (Greene 2020) to concepts used in recidivism questionnaires and argues that it helpfully distinguishes between variables that may encompass heterogeneous groups of people or behaviour and those that do not. This paper demonstrates why assessing the extent of a person's criminal associates is particularly problematic, and it presents a framework that helps to assess when concepts are problematic and how this can be mitigated.

The focus is on algorithms that calculate the risk of recidivism. Numerous ethical and legal concerns have been raised about the use of recidivism prediction algorithms (see Kehl et al. 2017, Oleson 2011, Re et al. 2019, Starr 2014). This paper also raises a number of normative issues, including how accurate recidivism prediction should be, whether accuracy differs depending on population characteristics, and how the performance of human beings should be compared to that of algorithms. These ethical and normative issues are beyond the scope of this paper, the aim of which is to highlight a potential problem and demonstrate its significance. Nevertheless, these issues are significant and can be fruitfully addressed in subsequent work.

Outline of Article

This article begins with a brief review of the reasons why concepts used in the social sciences are often problematic. The following section introduces the recidivism prediction algorithms and the variables that are considered important for recidivism prediction. The article then explains why the extent to which an offender has criminal associates is of particular importance in predicting future offending before demonstrating how questionnaires are used to gauge whether an offender has criminal associates. It shows how very different people can give the same response to a particular question. The next section discusses qualitative coding, which is often used to code social data, before arguing that this is not a solution to the problem of turning social science variables into data sets. The final section introduces the Nomadic framework and shows how it can help to distinguish between variables that are more

or less heterogeneous. The predictive success of recidivism prediction algorithms is then discussed, before concluding.

What's the Problem with Social Data?

Philosophers of social science note that many concepts social scientists use are problematic when compared to those used in the natural sciences. Natural scientists know what they mean when they talk about temperature, electric charge, and gold. Concepts used in the social sciences, including things such as wellbeing and democracy, are described as heterogeneous and ambiguous (Woodward 2003, 2016). Gasper describes them as 'umbrella terms, which cover many different possible concepts' (2010, p. 359). Wellbeing and democracy include many other concepts, such as happiness and freedom of the press. Someone using the concept democracy can have any number of these concepts in mind, and these often differ from the concepts another social scientist thinks important. Little describes social science concepts as 'cluster concepts', which encompass 'a variety of phenomena that share some among a cluster of properties' (1993, p. 190). In common with Gasper, this description illustrates that when we use these concepts we refer to a loose agglomeration of phenomena. This contrasts with paradigmatic scientific terms, such as gold, whose meanings are clear and which successfully pinpoint specific phenomena. This difference can also be described in terms of necessary and sufficient criteria which can be used to define many scientific concepts, but which are more difficult to specify for many social science concepts.

These differences make it difficult to work with social science concepts because there is no agreed-upon definition of happiness or freedom of the press, and when we use these concepts we often mean slightly different things. This has led some philosophers of social science to argue that the goal of the social sciences is understanding or explanation of single cases rather than prediction. Data science often aims at prediction, even with social data. How can this be reconciled? The remainder of this article is an exploration of the problems posed by using social scientific concepts in prediction algorithms.

Predicting Recidivism

In the US, increasing use is made of algorithms that calculate recidivism risk among convicted offenders. These recidivism scores are used, to varying degrees, by judges to determine sentence length. Similar algorithms are used to recommend whether offenders should be given bail and to recommend interventions that may help to reduce criminals' propensity for criminality.

The major recidivism prediction models are the Level of Service/Case Management Inventory (LS/CMI), Violence Risk Appraisal Gruide (VRAG), Lifestyle Criminality Screening Form (LCSF), General Statistical Information on Recidivism Scale (GSIR), Correctional Officer Management Profiling for Alternative Sanctions (COMPAS), and the Risk Prediction Index (RPI). Oleson (2011) notes that many

of these systems use the same variables. The Handbook of Recidivism Risk/Needs Assessment Tools (Singh et al. 2018) provides a thorough description of each of the main algorithms. The COMPAS system uses machine learning, while the others are based on standard regression models.

Variables Used in Recidivism Prediction

The recidivism models are proprietary; however, Oleson notes that meta-analysis has demonstrated the relevance of 17 variables. These are (listed in order of significance):

1. Criminal companions
2. Criminogenic needs
3. Antisocial personality
4. Adult criminal history
5. Race
6. Pre-adult antisocial behaviour
7. Family-rearing practices
8. Social achievement
9. Interpersonal conflict
10. Current age
11. Substance abuse
12. Family structure
13. Intellectual functioning
14. Family criminality
15. Gender
16. Socioeconomic status of origin
17. Personal distress (Olsen 2011, pp. 1353–1367)

The variables used in recidivism prediction models are usually collected through questionnaires that aim to gauge the degree to which the factors above are relevant. Most usually, an offender will complete a questionnaire, either by providing answers to a criminal justice professional or by filling in the form directly. A comprehensive review of US research examining the validity of assessments made using systems designed to predict recidivism in US correctional facilities shows that 82% of risk assessments were completed by 'professionals in correctional settings', the remainder were conducted by researchers or self-administered (Desmarais et al. 2018, p. 12). The surveys took between 5–10 min and 60 min to complete (2018, p. 10). The proprietary nature of these models makes it difficult to assess the importance of the factors used, and these undoubtedly vary in each of the different systems. This paper argues that there is an in-principle reason why it is difficult to accurately gauge the criminality of someone's companions or associates, which extends beyond data-gathering concerns. As Olsen writes, 'employing those variables in evidence-based sentencing decisions may prove difficult. Some variables will be difficult for

courts to know (e.g. ascertaining intellectual functioning may require clinical assessment)' (2011, p. 1368).

Questionnaires appear to be an excellent way to gather information from individuals because they are usually simple, quick to fill in, and group responses into convenient, mutually exclusive, categories. This yields data that is easy for algorithms to work with. However, while concepts such as age or postcode have clear meanings, others do not. Unfortunately, we cannot exclude problematic variables because they are relevant to predicting recidivism. The following section shows why judging the extent to which an offender has criminal associates is important for recidivism prediction.

Why Do We Care About Criminal Associates?

The *Practitioner's Guide to COMPAS Core* describes 'an involvement with antisocial friends and associates as one of the "big five" risk factors for criminality' (2015, p. 32). The *Guide* cites Gendreau in support of this; he concludes that there is agreement about some predictors of adult offender recidivism, 'age, gender, past criminal history, early family factors, and criminal associates' (1996, p. 576). According to Oleson, criminologists argue that criminal behaviour is learned and therefore adopted principally through contacts within small groups. He notes, however, that whether having criminal associates causes one to be criminal, or whether people with criminal inclinations choose to associate with criminals is an open question (2011, p. 1353).

The *Practitioner's Guide to COMPAS* includes a summary of important criminological theories, among which is Subculture Theory, originally developed from the Chicago School on Gangs (2015, p. 5). The *Guide* writes that behavioural norms are transmitted through social interactions. Particular behaviours, such as shoplifting or drug-dealing, can become the norm in certain subcultures. Membership of a subculture can include adherence to particular values and a common way of life. This is the theory that appears most relevant to understanding the importance of gang membership or association. Let us accept that belonging to a subculture in which criminality is encouraged or obligatory is, to some degree, predictive of future criminality. The extent to which someone associates with criminals is predictive of future offending and should be included in a recidivism algorithm. The following section reviews the questionnaires that are used to gauge the extent to which an offender associates with criminals.

Quantifying Criminal Associates

An offender arrested and charged with a crime is asked to fill in a form with a professional in the penal system. One of the questions asked is, 'Have your friends been in trouble?' The four available answers are 'Mixed, Gang member/associate, Essentially not in legal trouble, and Mostly in legal trouble' (Singh et al. 2018, p. 292). This question is taken from the JAIS Assessment for boys, not from an

adult questionnaire. The COMPAS Probation Assessment Instrument as it was used in New York asks:

Q17. The offender has peers and associates who (check all that apply):

Use illegal drugs	Lead law-abiding lifestyles
Have been arrested	Are gainfully employed
Have been incarcerated	Are involved in prosocial activities
None.	

Q18. What is the gang affiliation status of the offender:
Current gang membership
Previous gang membership
Not a member but associates with gang members
None
(Lansing 2012, p. 23)

The algorithms used in the prediction of recidivism are proprietary, and not all the questionnaires are available without subscribing. Other extracts from questionnaires that are available online have questions similar to these, asking whether the offender's friends are gang members or associates.

In the juvenile questionnaire, the offender chooses between four responses and, therefore, gives a seemingly precise answer to the question. However, this answer is not precise. There is a significant difference between associating with gang members and being a gang member. Associating suggests a much looser relationship, perhaps equivalent to hanging out with, whereas being a gang member suggests a greater level of participation in gang activities. It is possible to associate with gang members but not participate in their activities. The COMPAS questions distinguish behaviours in a different way—pointing at more specific characteristics of an offender's peers, but still requiring an offender to judge whether they 'associate' with gangs. To see how the notion of associating is problematic, consider the example of an inner-city priest, who might fulfil many of the criteria for committing crime, but is presumably at very little risk of doing so. The priest is male, associates with gang members, fits into a high-risk age category, and may have had difficult early life experiences. We could speculate further that the priest's difficult early life could have included criminal activity. The purpose of this example is not to equivocate unnecessarily about semantics, but to suggest that an enormous variety of behaviour is encompassed by the selection of an 'associates with gang members' response, ranging from a person who participates in criminal activities with gang members, to a person who is more loosely acquainted with gang members. While answers seem precise, it is unclear how representative they are of the information we want to know, which is whether the people the offender spends their time with are encouraging or enabling them to commit crime. This is because answers to these questions incorporate a wide range of behaviour, only some of which is predictive of future criminality. Other questions are subject to the same concern; these examples are taken from the

juvenile questionnaire (the questions come first, with the available responses in brackets):

> Q 29: Have you ever tattooed or cut on yourself? (yes/no)
> Q 32B: What do you like and dislike about yourself? (emphasises inadequacy/emphasises strengths/can't describe himself)
> Q 33: In general, do you tend to trust or mistrust people? (basically trusting/mixed or complex view/basically mistrusting)
> Q 40: Can you describe your father's personality? (If answer is unclear, ask youth to describe another person he knows well). (Multifaceted/superficial)
> Q 67: Appearance and hygiene. (Below average/average/above average)
> Q 68: Comprehension. (Below average/above average/average)
> Q 69: Affect. (Average/depressed (sluggish)/animated (hyper))
> Q 70: Self-disclosure. (Evasive/very open/average)
> (Singh et al. 2018, pp. 294–300)

Question 29 suggests an equivalence between self-harm and tattooing. While this may sometimes be the case, it is not clear that these two activities are indicative of the same mental state. In some cultures tattooing may be fairly common, while cutting yourself is more often a sign of distress. The interviewer has the same scope to interpret questions as an offender does when answering Question 32 onwards (quoted above). For example, a person reporting that they generally trust people may fail to do so articulately, resulting in being categorised as having a complex or mixed view. It is also unclear how one is to judge whether a description of a person is superficial without knowing the person being described. A judgement about hygiene depends on the average the interviewer has in mind. Finally, these questions are asked in a stressful environment, which might affect the lucidity of answers and the level of detail given. The argument that interviewers can agree on classifications is addressed below.

The same worries apply to the other COMPAS responses. Using illegal drugs ranges from fairly benign drug-taking to serious addiction. It is also unclear what the significance of arrest is, because arrest is not synonymous with criminality. Conviction is a better gauge of criminality. The positive side of the scale is no better specified-'gainfully employed' and 'pro-social activities' can both mean a variety of things.

Prince and Butters highlight the same worry with several items on the LSI-R tool. They argue that it is easy to see how different questionnaire administrators could struggle to provide the same responses for the same individual. They discuss the judgement required by the assessment tool administrator about whether a person participates in an organised activity. They note that the definition of 'organised activity' is unclear. The handbook states that church counts as an organised activity, but only if participation extends beyond mere attendance. They note that regularly playing football with friends could be considered an organised activity—it is a group activity, with rules and conventions, and encourages social interaction. There are a number of ways in which playing football is like being an active member of a church. How exactly this question is answered is likely to depend on the views of the assessor (2013, p. 24).

The questionnaires given to offenders are supposed to measure characteristics that are predictive of recidivism. These questionnaires contain a variety of questions, some of which appear to encompass a significant range of behaviour. While the answers given to these sorts of questions are precise, the behaviour underlying these answers can be heterogeneous. In the juvenile questionnaire, a person who loosely associates with gang members and a committed gang member both give the same response to the criminal companions question. The behaviour of these two individuals may be very different, as may be the importance of their gang affiliation for their propensity to commit further crimes. Other questions in these questionnaires also leave significant room for interpretation on the part of the interviewer and offender. The next section of this paper reviews some responses to this criticism before arguing that these are unsatisfactory.

Qualitative Coding

Systematising qualitative data in the social sciences is not a new problem. Social scientists often use a process known as qualitative coding to generate codes for behaviour or traits that they analyse. Childs and Demers describe qualitative coding as a 'tool for analysing data involving strings of meaningful words' (2018, p. 1). One oft-used method is to annotate transcripts of interviews and code aspects of people's speech to generate a set of codes that reflect the main themes or issues. Usually, codes are associated with individual words or short phrases. Multiple coders usually code data independently and then discuss, and try to resolve, any differences in codes. (Ganji et al. provide a good summary of the qualitative coding process. Marathe and Toyama (2018) also provide a good review, with a discussion of increased automation of coding.)

Zade et al. (2018) discuss a simple example in which social scientists attempt to research political views by analysing tweets. Coders assign one of five mutually exclusive codes to tweets: support, rejection, neutral, unrelated, and uncodable. After codes have been assigned independently, the social scientists try to resolve any disagreements. Machine learning algorithms can learn from examples to achieve the same end. It seems, therefore, that the social sciences have a way to overcome problems with systematising data. The responses to questions can be coded by social scientists, who reach agreement about what different responses mean. In other words, they agree that a positive response to a question about criminal associates really does mean that someone is subject to this risk factor.

The *Practitioner's Guide to COMPAS* addresses this concern, noting that 'People are complex and multi-faceted. Interpretation is hard, yet it is necessary for understanding behaviour and for determining strategies for intervention' (2015, p. 4). They try to ensure Construct Validity, which they define as 'the extent to which a scale measures what it is supposed to measure' (2015, p. 20). Northpointe (COMPAS's developer) measures construct validity by looking at correlations between measures of the same or divergent constructs. They also assess whether their scales correlate in expected ways with variables in the COMPAS system, as well as those

used in the LSI-R system (which Northpointe describes as the industry leader). The reliability of the COMPAS system has also been checked by a study where criminals were retested to assess the comparability of scores. Correlations ranged from 70% to 100% (2015, p. 25).

To summarise, the people administering these questionnaires are professionals who undergo training before using these systems, so, once the questions have been refined by social scientists, it should be relatively unproblematic to administer them consistently. Furthermore, developers of these systems also ensure that variables that should correlate with each other do in fact do so. While this should reduce the concerns with making qualitative judgements, the following section argues that it does not do so satisfactorily.

Problems with Qualitative Coding

Aroyo and Welty (2015) argue that coding methods do not discover truth. Inter-coder reliability is an attempt to measure the extent of agreement between coders. The reasoning behind this is that if a number of coders agree about the relevant code, this code is likely to be correct. However, Aroyo and Welty take issue with the idea that disagreement is bad, arguing instead that disagreement is a source of information; it can indicate that the text or source being analysed is ambiguous. They asked people to analyse the following statement:

> [GADOLINIUM AGENTS] used for patients with severe renal failure show signs of [NEPHROGENIC SYSTEMIC FIBROSIS] (2015, p. 17)

When asked to decide what relationship exists between the terms in brackets, some said it was causal, and others said that it was a side effect. Both readings are compatible with the text. The distinction matters because a 'side effect represents the possibility of a condition arising from a drug' whereas a causal relationship is suggestive of sufficient causality (2015, p. 17). Disagreement between coders is usually tackled by guidelines illustrating how different texts, or statements, should be understood. Aroyo and Welty argue that while guidelines do generate greater agreement, they do not increase quality. They write that they work by, 'forcing human annotators to make choices they may not actually think are valid, and removing the potential signal on individual examples that are vague and ambiguous' (2015, p. 18). Relating this to the recidivism case, it suggests that while different professionals may be able to fill in the questionnaire in the same way for the same offender, this just masks the underlying variability in behaviour of offenders. It does not alter the fact that behaviours with very different characteristics are lumped together in ways that may not be relevant.

The literature also suggests a degree of arbitrariness in the data collection process. Loza (2018) assesses the Self-Appraisal Questionnaire (SAQ), which predicts recidivism, and suggests rehabilitation programmes for violent and non-violent offenders. This questionnaire is administered by a forensic professional, and the offender fills in the answers. Statements are categorised as 'true' or 'false'. The professional may clarify statements for the offender. For example, Loza writes that if

the offender gives a response implying that both 'true' and 'false' might be correct, they are asked to choose which response applies 'even slightly more to his/her case. If the offender still cannot choose one over the other, then it is considered as a "true" response' (2018, p. 167). For this test, it is important that all questions are answered. Three unanswered items may affect the result of a particular subscale, but not necessarily the whole test. However, if answers that are not clearly 'true' or 'false' are pushed into the 'true' category, a number of answers may be inaccurate. The issue with this is not so much whether a recidivism prediction for a particular offender is inaccurate (although this is clearly a concern) but that the aggregate 'true' responses, although seemingly precise, include a great deal of heterogeneity. This heterogeneity spans from answers that are clearly 'true' answers to those that are almost indistinguishable from 'false'. Two offenders with the same responses to questions may not have all that much in common.

Questionnaires can be used to generate data, but when we ask people to answer questions they may not interpret the questions, or their behaviour, uniformly. Consequently, those giving the same responses may not represent a well-defined defined category of people over and above their pattern of responses. Specifically, they may not accurately indicate the variable that data scientists wish to analyse; in this case the extent to which an offender has criminal associates that enable or encourage them to commit crime. However, it is easy to take this worry too far; not all questions in the recidivism questionnaires work like the examples above. It is relatively easy to give precise and meaningful answers to questions such as: Date of birth, age, date of first conviction, number of previous convictions, and years spent in prison. These can all be answered, and coded, precisely. The following section reviews my framework outlined in Greene (2020), which is helpful for distinguishing between variables and assessing the extent to which answers to questionnaires yield heterogeneous data sets.

Nomadic Concepts

The Nomadic framework helps to categorise variables. Nomadic concepts include a great deal of heterogeneous phenomena within their scope. For example, two people can easily disagree about whether a particular person, John, is happy. Each person can have a different notion of happiness in mind, including whether John seemed happy the last time they saw him, whether he has satisfied his life goals, whether he seems content, or whether he satisfies a list of objective criteria, such as having friends, a job, and hobbies. The concept of happiness includes a great deal of heterogeneous phenomena, and two people discussing happiness may have very different ideas about what it means, while still using the concept in an appropriate way. More specifically, concepts are Nomadic when they meet the following two necessary and jointly sufficient criteria:

1. A wide variety of social phenomena can be included within the scope of the concept. This results from these concepts having many possible meanings, unclear boundaries, and changing over time. These characteristics are not an all-or-noth-

ing matter because these concepts can vary in the number of meanings they have, how unclear their boundaries are, and the extent to which they change over time.
2. The characteristics outlined in criterion 1 mean that disagreements about Nomadic concepts, and arguments making use of them, are difficult to resolve with academic analysis. Over time, the analysis of a Nomadic concept leads to the incorporation of different social phenomena.

I argue that social exclusion is an example of a Nomadic concept, but it also applies to associating with criminals. Associating with criminals has many meanings. It suggests making common cause or joining together. But it can also mean a much looser relationship, such as identifying with or hanging together with. It can cover a whole range of behaviour from participating in illegal activities with gang members to very casually hanging out with criminals. Some people may associate with criminals just because other members of their family are criminals or because their friends are. In other cases it may be a matter of expediency to maintain a good relationship with a particular gang and, therefore, to associate with them. This is not just obfuscation. Given the theories of recidivism above, what we want to know is whether a person associates with criminals in a way that makes them likely to commit crimes.[1] However, this is not what the answers to the questionnaires tell us. They tell us whether a person's behaviour can be described as associating with criminals, not whether this association is significant in any particular way or what the motivation for this association is.

The notion of associating with criminals also has unclear boundaries. Regardless of which meaning we focus on, how much interaction do we need to see to agree that someone associates with criminals? Do we judge this based on the amount of time spent? The activities engaged in while in the company of criminals? If so, is chatting about a football match, or guns, different from chatting about rival gangs or gossiping about recent local crimes? Is associating with them when they visit your house different from visiting criminals in their houses? The boundary between associating and not associating is difficult to draw and is likely to depend on local circumstances and the judgement of the person filling out a questionnaire.

Whether a person is judged to associate with criminals is also liable to change over time, depending on the degree of criminality in the local area. If someone lives in an area that, over time, becomes more dominated by gangs, then a judgement about the degree to which they associate with them will change. For example, if someone who lives in an area in which no known criminals live says that they sometimes meet up with gang members, this is likely to be more significant than if a person who lives in an area where many criminals live gives the same response. The changing make-up of neighbourhoods, and local conditions, will have an effect on how this is interpreted over time.

[1] This would be less of a concern if recidivism prediction based on current questionnaires were successful; the shortcoming with validations studies to date are reviewed below. Furthermore, it is more difficult to be confident that algorithms are discovering causation relationships rather than correlations when underlying behaviour is heterogeneous.

Associating with criminals is a Nomadic concept, but not all concepts on the recidivism questionnaires are. The Nomadic framework allows us to see why this is so. Take age at first conviction. This is the age at which a person was first convicted of an offence. A person is convicted at a trial or after they have pleaded guilty. Their age and date of conviction have clear meanings that are easily defined. This concept has clear boundaries and is stable—its meaning is expected to remain the same over time. There is a clear difference between the concepts 'age at first conviction' and 'associating with criminals'. The Nomadic framework allows us to understand what it is about the concepts that make them different and to compare these concepts with others to understand when heterogeneity of underlying behaviour may be a problem.

Making 'Associating' More Accurate

We can try to make concepts more precise. We could ask: 'Do any of your associates or companions encourage you to commit crime?', or 'Do you belong to any groups in which criminal activity is valued or encouraged?', or 'Do you feel you need to commit crimes to "fit in" with friends, family, or associates?' Greene (2020) addresses the same proposal for clarifying 'wellbeing'. I argue that such attempts will fail when they make use of other Nomadic concepts. This is the case with the present example. For example, encourage, value, and fit in are all Nomadic because there are many things that we can legitimately mean when we use these concepts, and their boundaries are blurred. Even the concept of crime is Nomadic in these questions. We might mean serious crimes or more marginal criminal activity, such as graffiti. Technically, a budding street artist could say (if answering honestly) that they belong to a group in which criminal activity is valued and encouraged. But this is a different matter from a group who encourage robbery. Trying to make this more specific still and asking about 'serious' crimes is beside the point because for some people graffiti is a very serious matter indeed—just not as serious as robbery (one would hope).

This is not to say that all attempts to clarify questions are doomed to failure. One question that seems particularly amenable to clarification is Q 29: Have you ever tattooed or cut on yourself? (yes/no). This question is presumably trying to discover a history of self-harm. If so, the ambiguity over motivations for tattooing or cutting as a form of body art could be removed by rephrasing the question in the following way: Have you ever self-harmed, for example by cutting or tattooing? While the concept of self-harm is imprecise, in this case it works to encompass a wide range of behaviour, including, but not limited to, cutting and tattooing.

To illustrate, we begin with the concept to tattoo on, or cut yourself. This can mean many things, one of which is self-harm. But it can also mean decorating your body for aesthetic or cultural reasons. We can focus on one of these meanings by rephrasing the question in terms of self-harm. Self-harm can mean many things—cutting yourself, burning yourself, or causing harm in a variety of other ways, potentially including restricting food or eating too much. In the context of this question, though, all of these meanings are relevant. We want to encompass this heterogeneous activity. We have not made the concept self-harm entirely clear, however,

because boundary issues still exist. How much self-harm matters? Does a one-off incident count? Is there a limit on how long ago it happened? The answers to these questions are, again, likely to depend on context. However, in the self-harm example, we began with a concept that has many meanings, we rephrased the question to focus on the meaning we intend. In this case, this concept can also mean many things, but they are all things of interest. Despite making the meaning more precise, we have not succeeded in removing all imprecise boundaries. Making concepts used in the social sciences more accurate is therefore not simply a matter of thinking about what we mean but thinking about the nuances and meanings that can be encompassed by a concept. When we succeed in doing this in some ways, we fail in others.

Implications for Data Analysis with Nomadic Concepts

It is natural that data scientists will want to work with variables and concepts used in the social sciences. It is also likely that this analysis will yield some interesting and useful results. This article proposes a way to think about variables and concepts that are used in the social sciences. It argues that applying the Nomadic framework helps researchers think systematically about variables and gauge the extent to which they are Nomadic, depending on the number of meanings they have, the extent of their boundary issues, and how they might change over time.

Data scientists designing algorithms to predict recidivism have good reasons to include the extent to which an offender associates with criminals, or is a gang member. The usual way to gauge whether these risk factors apply is to ask offenders to fill in questionnaires. Unfortunately, these variables are Nomadic. The problem is that a range of heterogeneous behaviour is describable by each of these terms. Offenders giving the same responses are therefore not necessarily the same as regards their risk of offending. What we really want to know is the extent to which their companions enable or encourage criminal acts. Mere association or membership does not tell us this. As argued above, attempts to make questions more precise are unhelpful when alternative variables are also Nomadic.

The remedy is therefore not to avoid using variables that are important for social science, but to think about them systematically and consider when data sets might be heterogeneous. When variables are Nomadic, data scientists can assess the extent to which this is problematic and whether anything can be done to aid clarification. Sometimes, using different questions may help—as in the self-harm case. Conviction could be used more often to assess reoffending, rather than arrest. When clarification is not possible, the extent to which heterogeneous behaviour and characteristics can be encompassed by the same variable should be reflected in the confidence attributed to any regularities or correlations discovered through data analysis. Alternatively, preference can be given to variables that are less Nomadic. Despite the apparent relevance of criminal associates, predictive success might be achievable without using this variable. Additionally, reasons for different interpretation of questions should be considered. For example, varying cultural norms may be relevant for understanding differences between people giving the same responses to

questionnaires. This will determine which populations a model is expected to yield accurate predictions for; some questionnaires may yield homogeneous data sets among some populations but not others. Combining data sets gathered using different questionnaires may be particularly problematic as differences in the way questions are phrased may yield different responses.

Are All Answers Equally Predictive?

A natural reply to the worry that answers to questionnaires do not identify a clearly defined behaviour is that the predictions these algorithms make are accurate. Lin et al. (2020) find that algorithms were more successful than humans in predicting recidivism when presented with a rich data set. Desmarais et al. (2018) conducted a meta review of studies assessing the accuracy of predictions made by recidivism risk assessment instruments for reoffending by general offenders in the US. Their review identified 19 risk assessment tools that had been evaluated in 53 studies between 1970 and 2012. This represented 72 samples of adult offenders in US correctional facilities. Overall, they found these tools to be successful, but conclude that, 'predictive validity may vary as a function of offender characteristics, settings, and recidivism outcomes' (2018, p. 21). They also highlight some gaps in the current validation research.

Firstly, they note that the definition of recidivism differed in these studies. Arrest was used most frequently (31%), followed by conviction (13%), and incarceration (10%) (2018, p. 12). This is concerning because an arrest does not necessarily mean that someone has committed a crime. The likelihood of arrest may rise after a prior conviction, regardless of the likelihood of committing further crimes because the police round up 'the usual suspects'. Eaglin (2017) notes that arrest is often used because arrest data is a 'cheap, easy, and accessible data set for researchers to pull information' (Eaglin 2017, p. 103).

Secondly, Desmarais et al. note that for five instruments, all studies were conducted by an author of the tool under investigation. Nearly a third of all studies they analysed were from research conducted by an author of the tool under investigation, and for the RMS, COMPAS, and SFS family of systems 'at least half of the studies were completed by an author of the instrument under investigation' (Desmarais et al. 2018, p. 12). It would be helpful to have more independent research as this indicates a potential for conflicts of interest. Casey et al. (2014) comment that independent evaluation is not just required to counteract bias (be it intended or unintended), but because the developers of systems may have a deeper understanding of their models than other researchers. This greater understanding may influence how the model is implemented in their testing site compared to other settings (Casey et al. 2014, p. 19).

Finally, inter-rater reliability was only measured in two studies, but in these it was very high—90% agreement (Desmarais et al. 2018, p. 14). Inter-rater reliability is important because it gives an indication of difficulties with interpreting questions or offenders' answers. Consequently, they write that there is a critical need for data on the inter-rater reliability of recidivism risk assessments completed on adult

offenders in US correctional settings (2018, p. 20). Bearing these concerns in mind, Desmerais et al. conclude that the decision to use a predictive tool should be influenced by the evidence, or lack of evidence, of the tool's usefulness within a particular offender population or sub population.

In order to assess the predictive success of recidivism models, the definition of predictive success should be clear. Singh et al. (2013) conducted a second-order review of how the predictive validity of prediction algorithms is measured and concluded that there is little consensus on how predictive validity should be measured. They write that 'the lack of reporting consistency in the description and interpretation of performance indicators across studies suggests the need for standardized guidelines for risk assessment predictive validity studies' (Singh et al. 2013, p. 66). They suggest that these could take a similar form to the reporting checklists for the prognostic risk assessment literature in medical diagnostics. There is a clear need for further validation of recidivism prediction algorithims, and greater appreciation of the difficulties inherent in using social data should form part of this validation.

Conclusion

The desire to systematise and quantify social science data is growing. This article argues that social science data is often difficult to convert into accurate categories on which algorithms work best. It provides a framework that explains why some concepts are difficult to codify and allows practitioners to assess the concepts they wish to use in quantitative analysis.

There are ways in which data scientists can address the concerns raised in this paper. Firstly, it would help public confidence and academic engagement if the recidivism questionnaires were made public. In particular, data scientists might consider making public what they think the data is supposed to measure and the questions which attempt to elicit answers to these questions. As Eaglin writes: 'To ensure fair construction of risk tools, government agencies and tool developers should create democratic accountability measures that invite the public to engage in the tool-construction and selection process' (2017, p. 106). A confidence level, or something similar, might also be included to highlight concerns about heterogeneity in answers.

Funding This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Aroyo, L., and C. Welty. 2015. The truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine* 36 (1): 15–24.

Casey, P. M., J. K. Elek, R. K. Warren, F. Cheesman, M. Kleiman, and B. Ostrom. 2014. Offender Risk & needs assessment instruments: A primer for courts national centre for state courts. Downloaded from https://nicic.gov/offender-risk-needs-assessment-instruments-primercourts

Childs, E., and L. B. Demers. 2018. Qualitative coding bootcamp: An intensive training and overview for clinicians educators, and administrators. *MedEdPortal* 14: 10769.

Deeks, A. S. 2018. Predicting enemies. *Virginia Law Review* 104: 1529–1593.

Desmarais, S. L., K. L. Johnson, and J. P. Singh. 2018. Performance of recidivism risk assessment instruments in US correctional settings. In *Handbook of recidivism risk/needs assessment tools*, ed. S. Desmarais, Z. Hamilton, J. P. Singh, D. G. Kroner, and J. Stephen Wormith, 13–30. Hoboken: Wiley Blackwell.

Eaglin, J. 2017. Constructing recidivism risk. *Emory Law Journal* 67: 59–122.

Ganji, A., M. Orand, and D. W. McDonald. 2018. Ease on down the code: Complex collaborative qualitative coding simplifies with 'Code Wizard'. In: *Proceedings of the ACM on human-computer interaction* 2: 132.

Gasper, D. 2010. Understanding the diversity of conceptions of well-being and quality of life. *The Journal of Socio-Economics* 39: 351–360.

Gendreau, P., T. Little, and C. Goggin. 1996. A meta-analysis of the predictors of adult offender recidivism: What works! *Criminology* 34 (3): 575–607.

Greene, C. 2020. Nomadic concepts, variable choice, and the social sciences. *Philosophy of the Social Sciences* 50 (1): 3.

Kehl, D., P. Guo, and S. Kessler. 2017. Algorithms in the criminal justice system: Assessing the use of risk assessment in sentencing. *Responsive Communities Initiative. Berkman Klein Centre of Internet and Society. Harvard Law School*. Downloaded from: https://dash.harvard.edu/handle/1/33746041

Lansing, S. 2012. New York state COMPAS-probation risk and need assessment study: Examining the recidivism scale's effectiveness and predictive accuracy Downloaded from https://epic.org/algorithmic-transparency/crim-justice/EPIC-16-06-23-WI-FOIA-201600805-LansingNYcompas_probation_report_2012.pdf

Lin, Z., J. Jung, S. Goel, and J. Skeem. 2020. The limits of human predictions of recidivism. *Science Advances* 6 (7): eaaz0652.

Little, D. 1993. On the scope and limits of generalisations in the social sciences. *Synthese* 97 (2): 183–207.

Loza, W. 2018. Self-appraisal questionnaire (SAQ): A tool for assessing violent and non-violent recidivism. In *Handbook of recidivism risk/needs assessment tools*, ed. J. P. Singh, D. G. Kroner, J. Stephen Wormith, S. Desmarais, and Z. Hamilton, 165–180. Hoboken, NJ: Wiley Blackwell.

Marathe, M., K. Toyama. 2018. Semi-automated coding for qualitative research: A user-centered inquiry and initial prototypes. In: *Proceedings of the 2018 CHI conference on human factors in computing systems* 348

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267: 1–38.

Oleson, J. C. 2011. Risk in sentencing: Constitutionally suspect variables and evidence-based sentencing. *SMU Law Review* 64 (2): 1329–1402.

Practitioner's Guide to COMPAS Core. 2015. Downloaded from http://www.northpointeinc.com/downloads/compas/Practitioners-Guide-COMPAS-Core-_031915.pdf

Prince, K., and R. P. Butters. 2013. Recidivism risk prediction and prevention assessment in Utah: An implementation evaluation of the LSI-R as a recidivism risk assessment tool in Utah, Utah Criminal Justice Centre. University of Utah. Downloaded from https://socialwork.utah.edu/_resources/documents/LSI-Implementation-Report-final.pdf

Re, R. M., and A. Solow-Niederman. 2019. Developing artificially intelligent justice. *Stanford Technology Law Review* 22 (2): 242–289.

Singh, J. P., S. L. Desmarais, and R. A. Van Dorn. 2013. Measurement of predictive validity in violence risk assessment studies: A second-order systematic review. *Behavioural Sciences and the Law* 31: 55–73.

Singh, J. P., et al. 2018. *Handbook of recidivism risk/needs assessment tools*. Hoboken, NJ: Wiley Blackwell.

Starr, S. B. 2014. Evidence-based sentencing and the scientific rationalisation of decriminalisation. *Stanford Law Review* 66: 803–871.

Wallach, H. 2018. Computational social science ≠ computer science + social data. *Communications of the ACM* 61 (3): 42–44.
Woodward, J. 2003. *Making things happen*. Oxford: Oxford University Press.
Woodward, J. 2016. The problem of variable choice. *Synthese* 193: 1047–1072.
Zade, H., M. Drouhard, B. Chinh, L. Gan, and C. Aragon. 2018. Conceptualising disagreement in qualitative coding. In: *Proceedings of the 2018 CHI conference on human factors in computing systems* 159.