**McNally, Notes for Socialising Big Data article, January 2016**

# Overview

How can we ‘do’ Social Intelligence and an Ethic of Care? Or to put it another way – Where did the idea of Social Intelligence and Ethic of Care come from? Answer to both questions: Collaboratories

# Collaboratories

**Collaboratories as a method for the Big Data Problem Space – Isomorphism between Big Data and Collaboratories**

‘*Big Data involves a redistribution of data collection and research methods expertise and the restructuring of infrastructures, which necessitates engagement with a wider range of collaborators’* (SBD WP pp. 28-29 Summary of Key Conclusions).

*‘When problematized, Big Data requires particular forms of collaboration between different stakeholders and practitioners’* (SBD WP pp. 28-29 Summary of Key Conclusions).

*‘In this regard, collaboration provides a useful way to understand the problems of working with Big Data. Through this approach, for example, we can identify who the important players are and ask questions about this burgeoning topic. The Final Collaboratory has revealed some of the key players in the field, but certain stakeholders were absent, such as, the users and producers of Big Data*.’ (SBD WP pp. 28-29 Summary of Key Conclusions).

**Collaboratories and the role of Provocations**

*‘Many genomic researchers are well-versed in the main ethical and social issues associated with genomics. ….. This familiarity with ethical and legal discussions can make it difficult to initiate other topics of dialogue. …. We found it necessary to put discussions on a different footing by working in visual terms (graphics, tables), making use of genomics researchers own databases and software tools, and generally trying to re-purpose genomic researchers own data literacy in the conversation by showing them data gathered from databases about their own data. This approach leads to mixed results. On the one hand, it certainly overcomes some problems of distance and unfamiliarity. …. On the other, this data is now presented with a view to challenging them to think about their own metrics and their own ways of talking about the value of sequence data. Some robust discussion usually arises’* (SBD WP Genomics and Collaboratories pp. 25-26)

‘*That said, more provocation and controversy could have been introduced’* (SBD WP Final Collab p. 29)

**Collaboratory as an agonistic space for the articulation of differences and productive misunderstandings**

Evelyn makes the point that the (NSI) collaboratory was not organised to share skills, develop methods or analyse Big Data (i.e. not for instrumental purposes), but to pose critical questions. She asks:

* *how can collaboratories be left open to ‘productive misunderstandings’? That is, rather than seeking consensus, how might we state the values or benefits of bringing together and allowing for the tensions and different perspectives and interests to be expressed and engaged?* (SBD WP pp. 24-25 National Statistics and Collaboratories).

This reminds me of Barry et al’s logics of interdisciplinarity. [[1]](#footnote-2) The authors identify 3 logics that motivate interdisciplinary collaboration: the logic of innovation (linked to the economy); the logic of accountability (linked to democratisation); and the logic of ontology. It is the latter that Evelyn seems to be appealing to in asking for how collaboratories can be left open to ‘productive misunderstandings’. The mode of interdisciplinarity that is associated with the logic of ontology is that of agonism/antagonism which also seems to align with the notion of ‘productive misunderstandings’. Thus the goal of a collaboratory could be to create an agonistic space where different kinds of stakeholders can enter into debate with each other rather than speaking to themselves.

*[Will Spooner] proposed that collaboratories are distinguished from other kinds of collaborative research because they examine how things in the world are constituted as objects (e.g. Big Data as a Digital Data Object (DDO). In this regard, collaboratories could be said to disrupt existing hierarchies and to challenge the status quo (Supp Appendices p. 11).*

**But is there a will to participate in such collaboratories – and who would fund them?**

* *in the face of economic and practical constraints such as time, the relevance and value of collaboratories that are more exploratory and conceptual rather than directly instrumental (e.g. developing specific applications/methods) need to be outlined rather than assumed* (SBD WP pp. 24-25 National Statistics and Collaboratories).

# Social Intelligence

Social Intelligence is awareness of the ways in which BD is constituted through and deeply implicated in social relations in both its composition and in its social effects.

**Against the fallacy of Data-Driven Anything**

Narratives about Big Data as the driver of socio-economic change – whether for the better or the worse – tend to attribute Big Data with intrinsic properties – the ability to act in a deterministic way – to have impacts on individuals and collectivities with positive or negative outcomes.

Social Intelligence problematizes the concept of data-driven anything – data-driven economies, data-driven science, data-driven societies.

**Collaboratories as windows onto the Social Intelligence of Big Data**

By bringing together a diversity of stakeholders, collaboratories are spaces in which it is possible to foreground the articulation of counter-narratives instead of over-optimistic and over-pessimistic narratives of Big Data that tend to be ballistic, deterministic and unrealistic.

In terms of social intelligence, our genomics collaboratory tended to focus on the social composition dynamics of Big Data rather than on their societal effects. [Was this also true of our other collaboratories, which also lacked or under-represented participants representing individuals or groups who are affected by, rather than composers of, Big Data?]

The genomics collaboratory brought into focus the various collective efforts that are necessary to bring Big Data into being and into a state in which it could potentially have impacts on the social.

As Hannah stated, collaboratories provide windows on the Big Data problem space – the barriers and hurdles to data composition. These insights are antidotes to the ballistic narratives of Big Data autonomously spouting into being and self-propelling itself throughout society.

Even though those ‘impacted’ upon were not direct participants in our collaboratory, the collaboratory on waste in particular problematized the ability of Big Data to have a predetermined effect or impact on society. Whilst chips in bins had a positive effect on householders in Sweden, there was no guarantee that this technology and the associated Big Data would have the same impact in the UK. In other words, even social impact is not an inherent property of Big Data that can be baked into Big Data by careful composition; it too is relational.

# Ethic of Care

**Genomics already operates an ethic of care**

‘*A minor academic and professional industry has developed around the ethical, social and legal implications of genomics as data-intensive science. In Europe, UK, North America and several other countries, government-funded research has extensively researched ethical issues associated with genomics, mainly in the interests of protecting patients, citizens and the public in general from either losing control of their own data, or in the interests of helping various social groups manage potential disadvantages or discrimination associated with genetic data. … We would suggest that many framings of ethical issues associated with genomic data have been narrowly individualistic, and they have paid little attention to ethics already implicit to data practices’* (SBD WP pp. 22-23 Genomics and Ethics).

The above translates into a framework for interrogating our existing data on genomics from SBD and also into questions we could ask of future studies / collaboratories on Big Data and genomics such as:

* *how deeply are ethical concerns carried into data practice?*
* *in what ways does an ethic of care already operate in data curation?*
* *do not assume that ethics only relates to human research subjects or patients but also plays out in many different forms of relationships* (SBD WP pp. 22-23 Genomics and Ethics).

# PROTEE, Collaboratories, Social Intelligence and an Ethic of Care

The PROTEE indicators provide a framework for relating what can be learnt from collaboratories to enrich our social intelligence and develop understanding of what constitutes an ethic of care.

Using this approach, Social Intelligence focuses on taking into account the myriad entities that are implicated in the social life of Big Data, whilst an Ethic of Care focuses on how relations between entities are, and are not, being taken care of.

Realism, Strategy and Falsifiability are the names of the classes of PROTEE indicators. Whilst related, each class provides a different perspective and each draws out a different dimension of Social Intelligence and an Ethic of Care.

**Realism –** as an antidote to the ballistic fallacy of data-driven anything. Realism asks for attention to be paid throughout the entirety of the social life of BD. It looks at the social life of BD throughout its life course. In place of fragmented narratives of Big Data emerging like water from a fire hose and then proceeding to have utopian or dystopian socio-economic impacts, the objective here is to enrich collective Social Intelligence and understanding regarding the multiplicity of entities and their actions and interactions that are involved in the composition of Big Data and its ability to have social effects. It seeks a thick description – or rather – recognising the impossibility of consensus - a multitude of descriptions.

Here, an ethic of care would be to highlight the myriad actions that are, and can be, undertaken to care for the social relations of Big Data – to make it fit and available for others to use. It highlights the actions of curation that attend to the qualities of Big Data. Who undertakes such actions of care? And who benefits from them? What is made possible through them? What is impossible without them? And what about the social effects of Big Data. They do not just happen by themselves. How are they being taken care of, when and by whom?

**Strategy –** collaboratories also add to our Social Intelligence by highlighting what entities are *outside of* the dominant narratives and collectives – the stakeholders who are absent from the narratives (or the collaboratories) because they choose to be (e.g. commercial companies), or because they have been excluded. Yet although absent, they may still be implicated. Some of them deeply. Collaboratories enrich our Social Intelligence by attending to what kinds of entities and relations Big Data is excluding – highlighting the antagonistic relations and entities that cannot coexist with Big Data, or are being driven into extinction by it – or which exist in a different data economy.

Regarding an ethic of care, collaboratories are occasions to imagine the collaboratory from the outside – to imagine what strategies might involve the entities that are not participating? What could be done to relate to them – to include them – to entice them to participate?

Such imaginings would highlight which kinds of Big Data actions tend to be inclusive and which exclusive, and to which kinds of stakeholders.

It would cultivate recognition of what kinds of social relations Big Data is replacing and consideration of their worth.

It would also invite consideration of how their participation could alter the socio-economic dynamics of Big Data – for better or for worse. What would be displaced by their inclusion? Should they – can they – be involved at any cost? What are the opportunity costs of bringing them in and leaving them out?

‘*As applications of sequencing have broadened, uses and techniques for of analysing sequence data have expanded, but often in tension with existing scientific expertise (for instance, plant breeding vs. genetic modification; ecological field study vs sequence-based studies) (SBD WP Genomics Economies p. 18).*

**Falsifiability –** Falsifiability is about not closing down opportunities. It is about resisting homogenisation and taking responsibility for the future. It is about daring to keep innovating and resisting blindly falling into hegenomic ways of knowing and being. It is about constantly seeking alternative ways to truth through Big Data rather than relying on a few algorithms to dominate how we constitute our future worlds.

*‘Needless to say, algorithms and techniques developed initially for bioinformatics and genomic research have filtered out into other data-intensive sciences’ (SBD WP Genomics Economies p. 18).*

*‘Need to ensure that a single form of data practice does not homogenise or dominate domains to the exclusion of other ways of acting and knowing’ (SBD WP Genomics Economies p. 18).*

From a Social Intelligence perspective, the collaboratories would be occasions to interrogate how social meaning is being attributed to Big Data? Are new ways being sought? Are old ways being challenged? Are the social meanings of Big Data being challenged as time progresses, or is a settlement being quietly reached without contestation?

From an Ethics of Care perspective, this class of indicator would highlight the risks of not being vigilant, and the need for responsibility – for the need for critical voices that are a constant ‘stone in the shoe’.

# Social Intelligence and the Ethic of Care in the Collaboratory

## Social Intelligence

The very composition of the participants in our collaboratory highlighted how the composition of Big Data is generated by the interactions of diverse actors (and technologies). We even applied an ideal typology to categorise them into three groups with respect to their roles in relation to NGS data:

1. Gleaners and cleaners who create sequence data and make it fit for others to use (Sarah Ayling TGAC; Neil Hall, Advanced Genomics Univ of LiverpoolRe)
2. Action heroes who use and apply NGS BD in novel ways (Lucy Raymond, 100,000 Genome Project Rare Diseases; Gurdeep Sagee, Public Health Genomics; Will Spooner, Eagle Genomics)
3. Packers, stackers and cataloguers who are concerned with the bulk handling and warehousing of NGS data and with making it available and discoverable for community use (Laura Clarke, Re-sequencing Informatics EBI; Rasko Leinonen, European Nucleotide Archive EBI; Chris Hayman, Amazon Web Services).

Large scale data collection projects, like the 1000 Genomes Project, are inherently social, involving many laboratories, instrument vendors and vendors of software and reagents, the quality control activities of people like Laura at the EBI, and archiving activities of people like Rasko at the EBI, and the data services of companies like Eagle Genomics and Amazon Web Services. (Plus the people who provide the genomes and those that interact with them in order to do so.)

Thus there was a general awareness at our collaboratory that NGS BD is inherently social. There was also awareness – Social Intelligence - of how the social composition of NGS BD has changed over time, with NGS being in its middle phase.

### Social Intelligence in the archive

* ‘ENA measures the growth of NGS data using multiple metrics’ (Rasko)

The metrics used in the ENA capture some of the social composition of NGD BD. They are a form of Social Intelligence. We used this social intelligence in our graphical provocations at the collaboratory. Repository as informant.

### Social effects

Our collaboratory did not have any representatives from the Social Effects side.

## Ethic of Care

### An Ethic of Care towards others involved in the Social Composition of BD

The Social Framework for BD states that ‘Acting with an ethic of care entails taking responsibility for potential social effects’. At the genomics collaboratory, Social Effects as envisaged in the top half of the Social Framework diagram, only came into play in relation to the activities of Action Heroes (Lucy on Rare Diseases in particular, and Gurdeep on Public Health Genomics).

However, the myriad actions undertaken by the gleaners and cleaners and the packers and stackers to take care of BD and ensure its quality of BD can be seen as of taking responsibility for relations with other kinds of actors involved in the Social Composition of BD, e.g:

* TGAC developed KAT, the quality assessment tool to improve analysis of the quality of assemblies, and undertakes pilot sequencing to assess the nature of genomes before working with them;
* EBI, Laura Clarke & Re-Sequencing Informatics – doing various checks (quantity, quality, identity and consistency) to ensure BD is of sufficient quality for the community to use.

Speaking about Identity and Consistency, Laura states ‘These require more specialised methods that are created for, and then applied to, the project in question’. This (and piloting by TGAC) would seem to exemplify an Ethic of Care – ‘An ethic of care for BD is not applying universal principles, but acting responsibly in light of particular experiences and actual situations ….’ This was also illustrated by Rasko speaking about the ENA:

* The rate of growth must be monitored to assess future sustainability of current technologies, to assist in choice and adoption of new technologies, and to factor in the upper limit to identify and store data’ (Rasko)

This quote also illustrates the use of metrics when making ethical decisions about BD (see below).(But was this just because we focused on metrics in our collaboratory?)

### Amazon Web Services, Social Intelligence and an Ethic of Care? Or Business Intelligence and a Value Proposition?

Amazon has analysed the Social Composition of BD and identified a business opportunity.

* ‘The complexity of genomic research has introduced further challenges, such as data management and access to valuable data sets. Previously, large data sets (e.g. HGP) proved difficult to locate, download, customise and analyse. AWS has devised a solution to this problem, introducing new technologies … that provide scalable, cost-effective, flexible and secure storage services … to manage and analyse genomic data … in addition to hosting data …’

This seems to demonstrate Social Intelligence. (Eagle Genomics also uses Social Intelligence to identify its niche in the Social Composition of NGS BD). What is the difference between Business Intelligence and Social Intelligence. AWS is taking care of certain aspects of BD – is this an example of an Ethic of Care? The other participants at the collaboratory don’t think so! This is in contrast to how they relate to the EBI’s archiving and analysis services. How does an Ethic of Care differ from a Value Proposition?

### A Variable Standard of Care – Demonstrates Social Intelligence; Illustrates Social Composition

NGS data are produced for a wide variety of uses, communities and applications. The standard of care varies with imagined social effects. This applies to the rigour of the results and analysis. It is highest when the social effects are the clinic. This demonstrates Social Intelligence – awareness of feedback between Social Composition and Social Effects. Also different kinds of project exercise different levels of care with regard to how NGS datasets are neatly packaged and stored, and when and whether they are made publicly available with many commercial NGS datasets being held in private. This illustrates the effect of Social Composition on the Social Effects of NGS; public repositories seek to maximise the potential for positive social effects.

### Problems with existing ethical approaches: Informed consent

This was expressed when discussing the ethical issues arising out of the clinical implementation of NGS, including 100,000 Genomics Project, specifically in relation to incidental findings. Gurdeep asked whether informed consent is ever possible? There are significant barriers to acquiring informed consent, particularly as it applies to logistics around time and cost.

### Problems with existing ethical approaches: Loss of anonymity - and an Ethic of Care proposal for how to deal with it?

Gurdeep – anonymity can be lost through data linkage.

Group 2 in the Discussion called for a political discussion about the risks and vulnerabilities around data integration, especially as this applies to health. Is this an Ethic of Care?

### Problems with existing ethical approaches: Metrics and cost-benefit analyses

Gurdeep spoke about how cost-benefit analysis is used in health care. Cost-benefits uses metrics to compare the costs and consequences of two alternative programmes in order to make a decision over which to use. Cost-benefit analysis is an approach from utilitarian ethics.

But in the case of NGS, ‘many different metrics are needed’. Whilst these can be developed, there is still a lack of consensus over which method or analytical approach to use. ‘At present, NGS technology raises more questions than answers’.

It suggests that it is not just more metrics that are needed, but that NGS is challenging the limits of the applicability of existing universal ethical approaches (utilitarianism) for public health care.

### Metrics and the Ethic of Care

Because we framed our collaboratory around the topic of ‘metrics’, metrics featured prominently in the presentations and discussions.

Nonetheless, there did seem to be an affinity between metrics and ethics – or the way that NGS BD is looked after, curated and cared for so that it is fit for use and available for use by others. If an Ethic of Care is maximising the utility of BD, then metrics function as an ethical tool for caring for BD.

However, there seems to be a distinction between how metrics are used in utilitarianism (cost-benefit analyses) and how they feature in what might be called an ethic of care. Whereas in cost-benefit analysis metrics are used to weigh up the pros and cons of two alternative courses of action, in an Ethic of Care they seem to be a proxy for measuring quality – like an inherent property of BD.

### Veracity

A key challenge of working with genomics BD is establishing its veracity, or truthfulness.

* ‘One of the problems of working with BD is that there is no truth set and this gives rise to the issue of veracity’ (Will).
* ‘Assessing the veracity of genomic data raises …. ethical concerns regarding how diagnostics developed from “omics” technologies should be regulated in the clinic (Will).

Ethic of Care in genomics. Is trying to deal with uncertainty at the same time as accepting that there is no skyhook – no privileged position from which to decide on the ‘true sequence of the DNA’ (Raymond, p. 18).

These are concerns that troubled Lucy. How to distinguish correlation from causation? The decision calls for a ‘balancing judgement’ (Raymond p. 19). But also for greater social participation in data composition (Raymond p. 19).

### An Ethic of Care for the Clinic?

Lucy was the most focused on the Social Effects of NGS BD. Indeed, the Social Effects are the objective of the 100,000 Genomes Project. As far as possible, clinicians use metrics to help them to make decisions:

* ‘At present, the key metrics that clinicians rely on daily vary from the quality assurance of gene analysis, gene variation, quantity, the cost of testing, and whether the variant has been reported before’ (Lucy).

But there are missing metrics.

* ‘At present we cannot measure the cause of the disease. Just because a variant is detected in an individual with a disease, the correlation does not imply causation. When looking for the important variants associated with intellectual disability, for example, many factors can be a play (e.g. genetically highly heterogeneous, multiple possible causes. Processing this data into meaningful information consequently introduces a series of risks, which include balancing judgement of variant causation with the risk of harm if the data is incorrect or misinterpreted’ (Lucy)

In the absence of metrics, professional judgement is used to assess the evidence, including assessing the reputation of other scientists. But Lucy anticipates the production of new metrics to reduce uncertainty by altering the Social Composition of the data – see below.

### Social Intelligence and an Ethic of Care

* ‘To improve the quality of analysis we need a greater volume of genomic data pertaining to patients with rare diseases’ (Lucy)

This demonstrates Social Intelligence – the feedback between social effect and social composition. For rare diseases there is a risk of confusing correlation with causation with potentially harmful effects. However, according to Lucy, this risk can be reduced by including more patients with rare diseases in the social composition of BD. But is this really the case? Or is it just self-serving? If it is the case, does this apply to other domains of BD – i.e. including more suspect bodies in the social composition reduces the risk of mistaken social effects? Is this just the same as saying that if we profile everyone there won’t be any risks of misclassification?

### Finally - Will Spooner on social composition at the final collaboratory

* ‘Within the field of genomics, the focus is on establishing new definitions about BD and novel ways to express and classify BD that recognise its social context rather than merely focusing on its technical qualities’ (Spooner).
* ‘For Will, this Final Collaboratory led to the growing realisation that BD is inherently social compared to other sorts of data. The value of BD is maximised by integrating multiple data sources to arrive at a multidimensional viewpoint of the problem at hand. BD is characterised as a data object that has relevance to multiple stakeholders beyond the use for which it was originally collected.’

Will took up the concept of a data object, and found ‘boundary object’ useful, and said he would have like more social science concepts from us to help him.

1. Barry, A., Born, G., Weszkalnys, G. (2008) Logic of interdisciplinarity *Economy and Society*, 37:1, 20-49 [↑](#footnote-ref-2)