**OFFICE OF THE PRIME MINISTER'S CHIEF SCIENCE ADVISOR**

Professor Sir Peter Gluckman, ONZ KNZM FRSNZ FMedSci FRS
Chief Science Advisor

**Keynote address**
**E-Research 2020 Workshop**

**26th June, 2015**
**Wellington**

***The intended and unintended consequences of e-research: why scientists must engage openly with the community***

The topic of today's meeting is a good reminder of just how rapidly science is changing. But while the technologies of science change rapidly, so does the relationship between science and the society in which it is embedded. Indeed, this is a key point that gets much less attention from the science community than it deserves. Yet I would argue that understanding and responding to it is fundamental to ensuring the value of e-research as it evolves. Your success in exploiting this technology will depend as much on considering this second dimension as it does on the science that will emerge and the infrastructure technologies that will be employed.

At its heart, e-research is about having access to, and combining, large amounts of data and then applying statistical and analytical techniques requiring computational power of considerable sophistication. In fact, it is entirely likely that the 'e' prefix may soon be seen as superfluous, as more and more types of research make use of increasingly available datasets and as scientists apply more probabilistic questions requiring sophisticated computational models to interrogate these datasets. These are methods and questions that were unimaginable not so long ago.

When such data relate to areas like crystallography, astronomy or particle physics, the issues that arise lie almost entirely within the scope of the scientific community. But much of the data for which e-research shows great potential will come from the social and environmental sectors where many more contentious issues come into play. It is these that I will want to focus on in the latter part of my address.

It is important to recognise that the advances made possible by e-research arise because of the advances made in technologies, computation and in systems linkages in what has become known as the 'internet of things'. We need to recognise that while these technologies are, on one hand, capable of doing much good for society, there are fair and reasonable concerns for society that those engaged with the technology and the research need to consider.

Increasingly 'big data' is being linked to potential use of Artificial Intelligence – for example it has been suggested to have use in interrogating very large multiple layered "omic" data sets in medicine. We have seen recently scientists of the standing of Lord Martin Rees and Stephen Hawkings express their reservations about the potential risks associated with AI being applied to drive machines and robots. Similarly, when our movements, habits, service usage, engagements, physical health and more can tracked, analysed, combined with other data etc., it is easy to see the many areas of such technologies where the issues of social

license will come to the fore and where the separation between what is research and what is application will not be clear or convincing.

Thus, just as biomedical and biological scientists have learnt that the question of social license for new technologies cannot be ignored, so too will the computer and big data scientist have to learn. Snowden may have focused our attention on the world's intelligence services but there is a lesson in the response to his leaks for all of us: publics do want to know and understand how data about them and their world is collected and used. And this goes beyond privacy concerns: Indeed issues of data ownership and sovereignty; protecting any cultural value of data assets and finding suitable methods of benefit sharing, and even about the inferential leap about what constitutes 'evidence' in an e-research context are all now part of the social license conversation – a conversation that is evolving quickly.

As we think about the massive changes that computation and data storage have allowed for science, it is worth remembering how quickly these changes have emerged. All said and done, the basics of statistics are not that old. While Pascal did his work on probability in the 17$^{th}$ century, it was not really until Fisher invented analysis of variance in 1918 that the modern basis of data-based approaches to science took off. Until technology allowed analysis at scale, the types of statistical approaches that could be undertaken were very limited. But now we face the challenges that arise because the range of statistical techniques that could be applied and the data sets to which it can be applied are enormous.

And this brings fundamental challenges that ultimately will have implications for both scientific and public trust in data-based technologies and analysis. Are the scientists doing the research well placed to appreciate the nuances of the different methodologies? Are the approaches and decision tools they have for dealing with the multiplicity of analyses fit for purpose? What little research I still do is largely in the field of epigenetics and here I see multiple errors potentially being made because many scientists do not understand the limitations of the mathematics underlying the statistics of their large data sets or the multiple issues in data-driven experimental design, data-set integration and so forth.

These are not trivial issues because if big data, data mining and e-research are to become common tools across multiple scientific disciplines, as is inevitable, then fundamental changes in the way we train all scientists will be needed. This is especially important as computational approaches generate and test probabilistic models. Indeed, as the products of science are less empirical and increasingly model-driven and dataset-set tested, we need to know that models are robust and their outputs can stand up as evidence for the complex questions science is asking today.

And deeper issues are arising in the scientific culture as a consequence of e-research. There are fundamental infrastructural, training and capacity needs to be addressed; access to expertise will become a core capacity in its own right; and standards will be needed in many disciplines for how data is captured and filed, with clear data governance protocols to be established and likely tailored to particular disciplines.

Indeed, the core practices and conventions of science are also evolving in response to data-driven research. As data-sets become larger, so too do authorship issues. Already we are seeing papers with over 1000 authors – how can we reconcile this with the traditional individualistic approach to academic assessment and recognition? This is also linked to issues of accessibility and data mining, some of which are quite sensitive to our culture of science. For example if clinicians have put considerable effort and resources into phenotyping their subjects in a clinical trial from which genomic data then gets amalgamated into some much larger study and data-base and they become seen simply as technicians providing the samples for (say) genetic analysis, they can become fairly

aggrieved if their intellectual effort in phenotyping is not recognised. There is the potential for culture wars within the walls of our enterprise.

E-research is also changing another fundamental of science with consequences for peer review and ultimately for funders: What now is the primacy of the hypothesis? Which comes first, experiment or hypothesis? Arguably ever since those who first followed Bacon and established hypothesis-testing at the centre of the scientific method (but certainly over the past 50 years) the scientific enterprise has been built on hypothesis development. In turn, funding bodies have built their assessments of what science to fund based on how the aims, hypothesis and experiment fit together. Hypothesis-free research or phenomenology was considered very suspect and unfundable when I was an emerging scientist. But now it is not uncommon to see big data and data mining proposals put forward as "hypothesis generating" rather than "hypothesis testing". Again this changes the nature of scientific evaluation. Should we now rely more on the relevance and potential impact of the proposed study? These issues are changing the way we think about science, but done right, the gains for society may be very significant.

Though I have raised these issues as a source of concern, I am nonetheless excited by the possibilities of e-research and particularly excited because I see NZ as having some comparative advantages in this space - but the issues that I have raised are active issues in the science policy community globally. For example the EU has had a major project in this area known as Science 2.0. Working in association with them and the OECD, my Office, on behalf of the Small Advanced Economies Initiative, is working on a paper considering how these and other changes will impact on public science funding systems.

Let me turn now to an area that I think is both an opportunity and a challenge – namely that of governmental and related data. I note that one of today's themes is to find ways to build bridges between the academy, government and industry in the sharing and use of data, and indeed, this is the question that is of real interest to government as you refine the value proposition of the e-research agenda in terms of "impact".

Governments collect an enormous amount of data – from businesses, from individuals, on their health, their education, their social circumstance and so forth. This is all correctly and properly protected and, until now, each data-set has been largely isolated from other sets. We have privacy laws and standards that protect individuals from its misuse. These data are essentially collected without consent (or consent is presumed) in that it is automatically collected in relation to various government systems and services. The opportunity to optimise the use of these data both for administrative and research purposes is pressing, but the challenges involved are not inconsequential.

Hence last year the Data Futures Forum issued a discussion paper that disappointingly did not get the attention it deserved to consider how we can reap the benefits of public sector data while protecting the individual's rights and addressing concerns about data sovereignty and responsible use and research. This has led to ongoing work and some of you who are here have been involved in those ongoing discussions. And those who have been involved will recognise that getting the balance right to allow for good use of the aggregated data while protecting individuals' and societal interests is complex, especially given the significant multi-sectoral and multi-cultural interests.

Gradually the NZ government's data sets are being aggregated. This is easier in some areas of public data collection (e.g. weather, transportation, seismic monitoring etc.) than in others. In the social sector – where much of the government and academic interest lies - it is indeed no small task given we are a country that does not give every person a unique

identifier. Do we need to set agreed standards for how future social sector databases are formed?

As the concept of absolute and enduring anonymity in a database is probably no longer realistic, no matter how much we try and anonymise the data, what other tools and provisions do we need? For instance, I have argued that we should make it a serious individual and corporate offence to secondarily identify an individual from a government or perhaps any database without the individual's consent. In any event, we must be transparent about what we allow and what we do, so that there is a clear social license for using – for whatever reason – social sector data held by government.

It is undeniable that there are opportunities in the government datasets for research of both national and global importance. When we combine various forms of detailed data about ourselves with geographical and other environmental data, we can ask questions that could address issues such as why rheumatic fever is so difficult to deal with, we could address issues around the social sector that at the moment are largely the basis of bias and speculation, and so forth. And in cities like Auckland and Christchurch we could use sensing and tracking data to make many and better decisions about planning of infrastructure and resources while protecting our resident's health and meeting their social needs. Harnessing the potential of the Data Futures Forum in *collaboration* with the academic community is something that I can say with assurance is a priority for government.

Similarly, we need to consider how to build bridges with the private sector. New Zealand firms both hold and seek access to rich databases that, shared and combined could enhance productivity on the one hand, but could also lead to beneficial social innovations. But what rules should apply across public and private spaces? Do businesses, governments and academics operate under the same data governance and responsible usage standards, and if not why not? If the private sector is engaged, the issues of secondary identification and opt-out mechanisms are even more important to be clear and controlled. On the other hand, the private sector – from Google and Amazon to our supermarkets – is already in possession of and using much of our data with a passive form of consent. Social license, if it exists at all, has been simply achieved through market forces, but this model cannot and should not apply in all contexts.

These issues are emerging as private sector–sourced information is combined in various ways and linked to sensing technologies and the internet of things. As information from our mobile devices becomes aggregated, fundamental changes in our societal expectations and societal responses will be exposed.

While all this may seem a long way from your immediate interests in big data and e-research, it is not. The public will not finely parse these different forms of e-research. It is critical that we are open about what science and technology is doing and has the potential to do, and to communicate this meaningfully and responsibly to diverse publics. Here the Singapore Statement on Research Integrity and similar statements provide the essential and meaningful guidance that has proven supple enough to apply across disciplines and with regard to new and emerging technologies. As the Royal Society of NZ reviews its code of ethics, it will need to consider some of these issues.

NZ can be at the cutting edge of applying an integrated approach to data and computational infrastructure from which all sectors of society can benefit. Indeed this may be where we could have a competitive edge in areas like social science and environmental research, in being a test bed for new analytical approaches and technologies, but most of all in marrying science and technology with community-relevant questions.

There are challenges though: our size; our geographical position; our relatively low science funding levels (and we do need to reflect on the cultural reasons why NZ society continues to have a low investment in science – perhaps we have spent too much time asking for money and not enough time explaining to society, in a credible manner, the benefits and impacts of the research that we have done).

But we also need to reflect on our many advantages: Our small size allows us to get all the key players in one room; the short distance between actors means we should be able to develop and commit to strategies; and our geographical boundaries make for a valuable analytical unit. Like other small advanced economies, we have also thought a lot about how to make the most of what we have got and hence the trend to research aggregations like the CoREs and National Science Challenges.

In NESI, RIANNZ and NZGL we have seen institutional interests put aside, in favour of more national scientific interests. This has not been easy to achieve and I congratulate all those involved in these less-institutionally focused approaches to science. We live in an age of trans-disciplinarity, and perhaps the most important boundaries to cross are those between our own institutional arrangements and between our technologies and the larger community.