From Statistics to Data Science: Implications for Democracy

Ladies and gentlemen,

It is my great pleasure to be back in Switzerland and to address you this afternoon.

Today, I will discuss the transition from statistics to data, and what this might mean for our economies, for our societies and for democracy.

Let me begin with data.

I contend that data is the word that defines our age.

With that in mind, let me start with some introductory remarks about data.

Today, data have assumed a new importance for economies and societies. They are at the heart of almost everything we do, a ubiquitous globalized commodity, easily shared, duplicated and traded.

Data are the glue that binds and drives the digital economy, communications, government, social media, the cloud, blockchain, the internet of things, crypto-currencies and even politics.

For so long, we have thought of data as an input for statistics – a byte we feed into a computer, a tool to inform decision-making. But data are now a policy issue in and of themselves. Given the importance of data to the globalized digital economy, surveillance, politics and AI, there will be few more important geopolitical issues in the coming years.

Ladies and Gentlemen, we live in an information age. Data are central to this age. Data are one of the most critical pieces of infrastructure in modern economies and societies. In common with many other key infrastructures, data require production, transportation, security, storage, refinement and dissemination. They require investment and continual maintenance.

Given the importance of data for today's economy and society – the first key message I would like to leave is that the architectural design of such an important piece of infrastructure should not be left to chance but should be carefully designed and constructed.

I would like to start by discussing a key role of official statistics – that is the provision of public goods.

Until recent decades, official statistics were considered the preserve of government. But this view has changed, as official statistics have increasingly come to be recognised as public goods and prerequisites for democratic dialogue – playing a critical role in safeguarding the deliberative public space.

The acceptance of official statistics as a public good has gone hand-in-hand with the notion of democratic and participatory government, and of political, economic, personal liberty and freedom of agency.

The idea of official statistics as a public good was formalised 30 years ago when the United Nations Statistical Commission adopted the *Fundamental Principles of Official Statistics* in 1994. Ten years ago, in 2014, these principles were endorsed by the United Nations General Assembly.

The notion of official statistics as a public good is set out principle 1 which states:

‘Official statistics provide an indispensable element in the information system of a democratic society, serving the Government, the economy and the public with data about the economic, demographic, social and environ-mental situation. . . are to be compiled and made available on an impartial basis by official statistical agencies to honour citizens’ entitlement to public information’.

Thus, when UNGA endorsed these principles, heads of state from around the world were explicitly saying that official statistics were a public good.

The economic definition of a public good defines public goods as goods as both non-rivalrous and non-excludable; in other words, everyone has equal access, and the use or possession of a statistic by one person does not exclude simultaneous and full possession by another. Thus, an official statistic can be copied, shared and used by many people at the same time.

However, the concept adopted by UNGA was broader; it implicitly incorporated the notion of quality, as they were concerned with the benefits to, and the wellbeing of, the public. In other words, dissemination of poor quality or misleading information would be a public bad, rather than a public good.

This is my second key message – all data and statistics are not made equal. Good quality statistics are essential to qualify as a public good.

The shift from statistics to data, raises interesting questions regarding whether data are public goods? It seems unlikely that all data can be public data – many data are now proprietary. But some data arguably need to be protected as public goods. But which data exactly and who makes that decision?

A third key message follows from this. In an information age – access to data is absolutely critical. This is not a discussion that Governments or the public can hide from. It is a discussion that cuts to the heart of our identities, our security, our wellbeing and our democratic systems.

I would now like to turn my attention to responsibility and accountability. The evolution from statistics to data science has been accompanied by a noteworthy transition, that is the change in language from ‘evidence informed’ to ‘data driven’.

For statistics to be used to inform decisions rests on the ideal that figures are the neutral arbiters in political debate – back to the idea of a deliberative public space I noted above.

Interpretation of a statistic may differ but at least the fact itself is agreed. This idea is being undermined by the emergence of ‘post truth’, ‘alternative facts’ or ‘bespoke facts.’

The emergence of evidence informed decision making, has its origins in the recognition of risk management that emerged with the secularization of society, as people and governments began to act as free agents with responsibility for their own lives and decisions. Changes were also driven by the great historical shifts of industrialisation, globalisation and empire which required information on markets, banking, financial systems, credit and debt, prices and other aspects of commerce that involved risk.

Two events in the early 20th century had a profound impact on official statistics – the great depression and World War 2. Prior to the great depression, the laissez-faire economics that persisted didn’t require any statistics, as no government decisions were required. But in response to the Depression and the emergence of interventionist, countercyclical, social and economic policy that changed – Keynesianism introduced the need for evidence.

The other great driver of change has been war; WW2 in particular. WW2 and subsequent reconstruction efforts led to demands for quantitative evidence. The global scale and massive mobilization of WW2 required intense planning, involving securing and distributing materials and commodities, organizing labour, transport and logistics, imposing price controls and cyphers and cryptanalysis. It also involved balancing the requirements of domestic economies with war economies, including limiting non-essential imports. All of which required data and statistics.

Post war Taylorism or scientific management drove demand for statistics in industry. By the 1980's this morphed into New Public Management and spread this culture to the public sector, as accountability and performance measurement became seen as a guarantee of objectivity.

The digital and ICT revolutions of the past 30 years have brought an avalanche of by-product big data, which have facilitated algorithmic based decision-making and modern AI. This change has been accompanied by a change in rhetoric; from evidence-informed decision making to data-driven decision making. The former acknowledged and made transparent the judgements and trade-offs involved in democratic decision making, the latter implicitly adopts a datacratic approach, where data alone are sufficient to make decisions – i.e. an algorithmic approach.

This important transition is at the heart of concerns around the use of artificial intelligence in decision making. The explosion in data saw the abandonment of axiomatic or symbolic artificial intelligence in favour of data based artificial intelligence. Interestingly, this was to some extent been paralleled by an opening up of rationalist economic theory to include behavioural economics, which too requires evidence.

The data driven approach also coincides and aligns with the 'end of theory' or 'correlation supersedes causation' approach, where the traditional hypothesis-based science is being replaced. This Copernican shift in discovery and decision-making poses profound epistemological questions regarding the meaning of ‘knowledge’ – what does it mean to ‘know’ something if we don’t understand the cause? This distinction lies at the heart of the dichotomy between statistics and data science.

Why is this change important? It is important for transparency and for accountability. For responsibility. Algorithmic decision making is data driven – but those decisions may be difficult to query, understand or challenge. Furthermore, it may remove the politician from decision making. Counter intuitively then, the data deluge may reduce political accountability, not improve it. This has profound implications for Democracy.

This is my next key message. For public debate, accountability, democracy – we must understand the origins of the statistics. They must be reproduceable. Hence the importance of public metadata. They should inform decisions – but they should not drive decisions. That is the role of our elected officials - they must retain responsibility for decisions.

Furthermore, I anticipate that national data sovereignty will be increasingly challenged in the future, making accountability and responsibility more demanding.

Today, digital data can be easily stored, shared, exchanged, and copied. They are a globalised commodity that defy national boundaries and national legislation, they challenge the notion of national sovereignty itself. As a resource, data cannot be managed from a national perspective alone. Some sort of international data governance framework will be required to safeguard the privacy basic human rights of citizens.

This cuts to the very heart of democracy. With the massive computing power available, everyone's data can be stored, matched and linked. Used selectively or unwisely, unrepresentative AI, unsupported by ethical guidelines, may hardcode and amplify biases, or impose unjust decisions on an unsuspecting and defenceless public. The future of privacy itself, as a concept, as a reality, is also at stake. No one's past can be deleted or digitally forgotten.

In an era of governance by numbers, of quantification, it is important that peoples and communities retain control of their data, benefit from their data, and are not dictated to by a small, data elite.

Thus another key message is that approaching data governance from a purely national perspective is a mistake. It is to misunderstand fundamentally what is happening in the data world.

This brings me to my last point.

Paradoxically, the fact that large volumes of data exist does not mean data are available or easy to access. In recent decades we have witnessed a massive, asymmetric concentration of data. This concentration of data holdings introduces obvious risks of abuse and manipulation. Many data now are proprietary and inaccessible to only a few. Simultaneously we have seen a retreat in progress towards open data.

This raises profound questions for any country, for any society. Who has access, and who disseminates data and statistics? Put another way - who controls?

I would like to conclude my remarks today by returning to a few key messages.

In the digital era, data are a key piece of infrastructure – they underpin the entire digital economy, from banking to research, to AI. To manage this infrastructure properly, careful architectural design is required. It is too important to be left to chance.

Not all data and statistics are equal. Quality standards and metadata really matter. Good quality statistics are essential to being a public good. Data and statistics on their own are insufficient. To properly inform public debate, to support accountability and democracy, we must ensure the veracity, progeny and reproducibility of statistics.

Statistics should inform decisions – not drive decisions. Making decisions is the role of our elected officials. In an era of governance by numbers, of quantification, it is seductive to fall for the illusion that data can tell us what to do. But I urge you to remember that data and statistics are not substitutes for judgement: data and statistics demand judgement.

Developments in the data world are not trivial. They will affect every single one of us. We cannot hide. It is not too late to shape the world we want - we should not be dictated to by a data elite. We all share the responsibility to protect our democracies. This I would suggest cannot be addressed by anyone country alone – data governance requires an international solution.

Ladies and gentlemen

Thank you for your attention