COEURE workshop on

"Developments of data and methods for economic research"

Université libre de Bruxelles, July 2-3, 2015

Final report

Rapporteurs : Stefan Bergheimer and Estelle Cantillon

Contents

Objectives of the workshop

The "data and methods" workshop is one of 12 thematic workshops that were organized to assess the state of the art in economic research. Its objective is to explore developments and prospects (1) in the production, funding, access and use of data for economic research and policy, and (2) in the methods of inquiry used in economic research. The workshop brought together high level scientists, representatives of statistical offices and other producers of data, research funders as well as policy-makers to discuss these issues, identify good practices and derive recommendations for European research and research infrastructure funding.

Scientific Committee

Estelle Cantillon (ULB and COEURE), **Christine De Mol** (ULB), **Bram De Rock** (ULB), **Domenico Giannone** (Federal Reserve Bank of New York), **Georg Kirchsteiger** (ULB) and **Laszlo Matyas** (Central European University and COEURE).

Programme

Thursday July 2

8:30 - 9:00 **WELCOME COFFEE**

9:00 - 9:15 **Welcome, description of objectives of COEURE and of the day, Estelle Cantillon (ULB) and Marianne Paasi (European Commission, DG RTD)**

9:15 - 11:20 **Session 1: Organizing data access for research (Chair: Estelle Cantillon)**

Issues to be addressed: Two key sources of new data on which recent economic research draws are linked individual or firm-level administrative data, and private datasets. Protocols for secure access, de-identifying and confidentiality are essential. What are the keys to the Scandinavian success in making quality administrative data available to researchers? How transposable are these models to other countries, and to other types of confidential data (financial data, private data, …)? What are the remaining limitations? Is the European legal framework adequate? What are the implications for funding mechanisms?

Speakers: **Roberto Barcellan** (Eurostat, decentralized remote access to confidential micro-data), **Caterina Calsamiglia** (CEMFI) and **Frederic Udina** (Statistical Institute of Catalonia) for the Catalan case / education data, **Vigdis Kvalheim** (Norway Social Science Data Service), **Luke Sibieta** (IFS and UK Administrative Data Research Network)

11:20 - 11:40 **COFFEE BREAK**

11:40 - 13:00 **Session 2: Researcher-generated databases (Chair: Bram De Rock)**

Issues to be addressed: Partly in response to frustration with cross-country comparison of data or simply lack of centrally collected data, a number of databases have been created and made public by teams of researchers. What are the benefits and costs of such researcher-generated databases? To what extent

have they been adopted by other researchers? How are they funded? Should funders impose conditions such as methodologies, quality and open access? How can their survival be organized?

Speakers: **Julia Lane** (New York University), **Arie Kapteyn** (University of Southern California), **Guglielmo Weber** (University of Padua)

13:00 - 14:10 LUNCH

14:10 - 15:30 Session 3: Data generation in (quasi)controlled environments (chair: Georg Kirchsteiger)

Issues to be addressed: These past 20-30 years have seen the development of new datasets generated in controlled or quasi-controlled environments: lab experiments, field experiments, randomized controlled trials, and more recently biomedical (neuro-imaging) data. What have these new types of data allowed us to discover that other more traditional datasets had missed? How robust are the results generated from these data? To what extent can such methods be used for policy design? Can such datasets be useful beyond the specific research project they were designed for? How?

Speakers: **Bruno Crépon** (CREST and JPAL) and **Colin Camerer** (Cal Tech)

15:30 - 15:50 COFFEE BREAK

15:50 - 18:00 Session 4: Data standards and cross-country datasets (chair: Bram De Rock)

Issues to be addressed : Cross-country variations in data standards and data definition are big obstacles to multi-country research and comparative analyses that are nevertheless crucial for better understanding the scope for replicability of policies across borders (is the experience of country X relevant for country Y?). Are researchers' needs different from other users of such data? What are the main obstacles to greater data comparability? Where does Europe stand relative to the US? Is there need for greater coordination? What are interesting developments in multi-country quality datasets and what are their success drivers? How are they funded? What new insights have such quality multi-country datasets generated (or could generate)?

Speakers: **Roberto Barcellan** (Eurostat) on the G20 data gap initiative, **Peter Bøegh-Nielsen** (Statistics Denmark) on two int'l projects for harmonized datasets, **Laszlo Halpern** (Hungarian Academy of Science and coordinator of MAPCOMPETE), **Joseph Tracy** (NY Fed) on international banking data, **Lisa Wright** (Bureau Van Dijk, Zephus) on creating reliable international firm-level merger data.

Friday July 3

8:45 - 9:00 COFFEE

9:00 - 10:30 Session 5: The changing face of research-policy / research private sector collaborations (chair: Estelle Cantillon)

Issues to discuss: Researchers have long been involved in policy as ex-post evaluators (policy assessment) or as advisors but new, more collaborative models of interactions are emerging where researchers and policy-makers are partnering, with the benefits of access to data and possibly funding for the former and quality advice for the latter (such partnerships are also developing between researchers and data-intensive firms such as Yahoo, Microsoft, Google, and financial exchanges to name just a few). What are the implications of these developments for the type of research being carried out, or for the organisation of this research? What are the benefits and risks? (e.g. in terms of scientific integrity and independence, data confidentiality and thus non replicability of the results, …). What impact is this likely to have for the current "publication as quality certification" model?

Speakers: **Liran Einav** (Stanford), **Asim Khwaja** (Harvard) and **Markus Moebius** (Microsoft)

10:30 - 10:45 COFFEE BREAK

10:45 - 12:00 **Session 6: Panel on the implications of the developments in data and methods in economics for research funding** **(chair: Véronique Halloin,** Head of the Belgian French-speaking Research Fund (FNRS))

Participants: **Dominik Sobczak** (European Commission, DG RTD), **Angelika Kalt** (Swiss National Science Foundation) and **Paul Sanderson** (Economic and Social Research Council, UK)

12:00 - 13:00 **LUNCH**

13:00 - 14:00 **Session 7: Big data: Definition, challenges and opportunities** **(chair: Christine De Mol)**

Issues to be addressed: What is big data, where has it or is it expected to emerge in economics, and how is it likely to change economics? What are examples of big data applications delivering fundamentally new research insights? What are the challenges? How is Europe positioned to take advantage of big data?

Speakers: **Sendhil Mullainathan** (Harvard) and **Lucrezia Reichlin** (London Business School)

14:00 - 14:15 **COFFEE BREAK**

14:15 - 16:15 **Session 8: How will big data change econometrics?** **(chair: Domenico Giannone)**

Issues to be addressed: What challenges do high dimensionality data create for econometrics? What developments in methods are needed to address them? What are the implications of big data for model fit and model selection? What progress can be expected in dimensionality reduction? Is there a common ground to be found between the standard econometric approach and the need to let the data speak? What synergies may arise with statistics, computer sciences and other disciplines?

Confirmed speakers: **Eric Gautier** (Toulouse), **Jeff Wooldridge** (Michigan State University), **Herman van Dijk** (Rotterdam)

Minutes of the workshop

The conference started with an opening statement by **Marianne Paasi** (DG RTD). She explained that research in Europe remains fragmented and that the predominantly national sources of research funding meant that topics of European relevance did not receive as much attention. COEURE has been implemented to discuss economic research at a European level and ultimately provide research insights for policy-making at the EU level.

Estelle Cantillon (Université Libre de Bruxelles) then briefly outlined the objectives of the workshop. The workshop aims at exploring the developments and prospects in the production, funding, access and use of data for economic research, and in the methods of inquiry used in economics. All types of data are concerned: administrative, private, researcher-collected, researcher-generated data. The subject is broad and the audience is diverse. Discussion and reactions from the audience are therefore all the more appreciated to confront the different perspectives.

Session 1: Organizing data access for research

The first session was dedicated to data access, with a focus on administrative data. These data are a rich source for research but their access should be properly organized to maintain their confidentiality. Different models are (or are being put in place) in different countries. The goal of the session was to better understand their strengths and weaknesses and especially their transposability in other countries.

The first speaker was **Vigdis Kvalheim** who is the Deputy Director of the Norwegian Social Science Data Services (NSD). Scandivanian countries are leaders in providing access to de-identified microdata. The models of access differ across countries but do share a number of facilitating factors:

1. Collecting data for statistical purposes has a long-standing tradition in the Nordic countries. As a result, those countries now possess an infrastructure for data management and a first rate legal framework to organize access (every relevant law has a provision to ensure access).
2. Many data registers (e.g. health, tax) cover the entire population and can be linked thanks to a unique personal identification number for each citizen.
3. Society sees a value in analysing these data: analyses based on these data have had a big impact on policy-making.
4. A trusted and sustainable research infrastructure, like the NSD in Norway, is of utmost importance. Norway combines some of the strictest protection of personal data with the highest level of access.

She described in detail the Norwegian model of access which consists of both mediated access (through NSD) and direct access to Statistics Norway's data. In both cases, the researcher can use the data on his/her own site. Currently NSD serves around 1,200-1,400 projects per year whereas Statistics Norway serves about 250 projects in direct access. She described the pros and cons of each mode of access. Mediated access (through NSD) is a specificity of Norway. Its value added lies in the establishment of procedures for access that safeguard personal information while allowing for a relatively extensive use of individual data.

She discussed two ongoing developments. The first one is a pan-Scandinavian initiative, NordForsk, to facilitate the use of all Nordic registers for research purposes. Indeed, cross Nordic register research is still rare due to the lack of data harmonization, the need to go through the authorization process in each country, and the differing legal, technical and organisational practices. The idea is to converge to

a common application and agreement for data transfer, with remote access as the main channel. The second development is the introduction of remote access in Norway (Remote Access Infrastructure for Register Data, RAIRD). The objective is to increase research on data from administrative registers, by facilitating access (including by foreign-based researchers) and reducing the costs of serving researchers' requests.

After the presentation, a participant raised concerns about the potential misuse of confidential data and asked whether there were any incidences in the past. Vigdis Kvalheim told the audience that she could not report any significant misuse regarding privacy rights. The greater problem is researchers' use of the data for further studies that had not been authorized.

Roberto Barcellan (Eurostat) talked about decentralized remote access to confidential micro-data which was recently put in place at Eurostat. Confidential micro-data available at Eurostat are data coming from surveys such as the Labour Force Survey. The main motivation for introducing remote access is that on site data access (in Luxemburg) is proving to be too big a barrier for researchers. He described the existing European legal framework for access to confidential data for research (Regulation 557/2013). It is a two-step procedure where the research entity has first to be recognized and then has to submit a research proposal. 448 research projects were granted access in 2014, mostly in labour market and social studies. Current developments at Eurostat include the creation of new datasets, public use files, automatic processing of requests, and decentralized and remote access to confidential data (DARA). All these initiatives should speed up and facilitate access to data for researchers.

Luke Sibieta (Institute of Fiscal Studies) talked about the efforts in the UK to increase access and use of administrative data. Greater use of administrative and linked data could both benefit society and researchers. He cited the tax system and education as examples where research using administrative data has had a significant impact on policies in the UK. However, negotiating access had often been a lengthy procedure due to legal, cultural and institutional barriers. Different forms of data access exist (secure transfer to researchers, virtual access or dedicated secure environments). Data linkage remains almost impossible in the UK. However things are changing. The two main drivers of this change of attitude are the perceived benefits of research on linked data (as exemplified by the tax and education studies mentioned above) and the concern that the UK may be losing out in research if it does not grant access to data to its researchers. As a result, the UK Administrative Data Research Network was established in 2014 with funding from the Economic and Social Research Council to foster the access and use of de-identified linked administrative data. An administrative data service (ADS) and four administrative data centres (ADRCs) were set up at the end of 2014. ADS negotiates for access to datasets and trains researchers. The ADRCs provide the secure access to linked data from the government departments and agencies owning the data. Things therefore look promising even though challenges remain, namely legal barriers, social acceptance, willingness of data owners to share data and the feasibility of linkage (the UK does not have national ID numbers).

A participant raised concern about the current requirement in the UK to delete datasets after five years. This limits replicability and is thus a problem for research. Luke Sibieta confirmed that this is indeed a legal requirement and it would require a change in the legal framework to allow longer storage.

Caterina Calsamiglia (CEMFI) talked about her experience securing access to linked administrative and school choice data in Catalunya, where there was no tradition of researcher access to such data and no systematic organization and storage of some of these data. Local authorities initially showed no interest for her research project and were reluctant to provide data. Her experience serves as a case study for other researchers who want to access administrative data. First, one should take every possi-

ble path to the data. Ultimately, she got the data when "threatening" to talk to the media. She recommended involving the local authorities as soon as possible by presenting findings, providing technical assistance, and getting them interested in what could be learned from the data. Her project eventually required linking her data with census data for which she benefitted from a supportive new head of the Catalan Statistical Office (Idescat). Her research required putting in place protocols for data access at Idescat. The contracts were carefully crafted to guarantee that all participants felt secure. When getting other administrations involved, it turned out to be crucial that the research had clear policy implications and that the advice was provided for free.

After recalling the changing environment and demands on public statistics, **Frederic Udina** (Institute d'Estadística de Catalunya - Idescat) described recent developments in providing access to confidential data to researchers in Catalunya. Even though the big surveys like LFS or EU-SILC are conducted at the national level, Idescat is responsible for administering key administrative data from the region. It benefits from its young age and small size which makes it easier for them to adapt and innovate. Frederic Udina described the new integrated data platform (Plataforma Cerdà) Idescat is putting in place, where new data are automatically integrated with pre-existing data, avoiding the traditional stove-pipe model. At the same time, the Statistical Law of Catalunya is being adapted to integrate EU regulation 557/2013 on access to confidential data for scientific purposes and Idescat is developing partnerships with research centres such as the Barcelona Graduate School of Economics to facilitate knowledge about existing data and its access on the basis of a win-win model.

In the subsequent general discussion, a participant asked why data access was overly constrained in economics and the social sciences relative to other fields in science. This is especially the case for data collected by statistical offices. One potential explanation is that the compulsory nature of the statistical offices' data collection requires trust (e.g. firms need to be willing to report the requested data) and that this made statistical offices conservative about data access. On that topic, one participant regretted the opt-out clause present in Germany that allowed some data to be withheld from potential access. This creates potentially biased samples, eventually hurting the quality of research that can be performed. Another point of discussion concerned the role of private data providers as alternatives to statistical offices. They may charge for access but access is usually easier and more user-friendly.

Session 2: Researcher-generated databases

The second session was devoted to researcher-generated databases. Such databases, which are made public, are often developed by researchers when existing data are inexistent or of poor quality. The goal of the session was to understand the strengths and weaknesses of such databases and discuss their funding and perennity.

Julia Lane (New York University) talked about the lessons she has learned from her experiences building data infrastructures (LEHD[1], STAR METRICS, and the data enclave at the University of Chicago to name a few): (1) identify a gap in the available data; (2) create a team, as building a database requires different skills; (3) build coalitions of people in different institutions (universities, states, government etc.) who share the same goals, and (4) develop products like indicators, reports etc. that show the value of the data. At the end, she also emphasized the key role of private foundations. Private foundations are more mission-oriented, less risk averse and tend to have a longer time horizon than public funding bodies. They also are comfortable investing in people instead of only investing in pro-

[1] The Longitudinal Employer Household Dynamics program.

jects. Interestingly LEHD which has been 20 years in the making is now a national statistical program, publically funded since 2008.

Arie Kapteyn (University of Southern California) discussed his experience with the Measurement and Experimentation in the Social Sciences project (MESS). MESS was launched in 2006 at CenterData (Tillburg University). MESS is built around an internet panel (LISS), connected to several other smaller panels (experiments, immigration etc.). The panel uses innovative methods (such as time use apps or internet bathroom scales) and is more cost-effective and flexible than traditional surveys. The resulting data are open to scientists worldwide, free of charge. Scientist can propose new questions to be included in the survey. During its first phase (2006-13), 131 projects have been conducted in MESS involving 85 universities and institutes around the world and Arie Kapteyn described the extensions and developments foreseen for phase 2 (2013-2023).

Funding is a sore point however. MESS was built with infrastructure funding and its financing was not renewed because "infrastructure is not expected to incur recurring costs." In Arie Kapteyn's view this reflects a fundamental misunderstanding of the nature of data infrastructure in social sciences: the value of a panel is in observing the same people over time and so it necessarily entails recurring costs. The lessons from MESS for research funding can be summarized as follows:

- Funding should not be tied to today's policy problems because this leads to short term funding which is disruptive for researchers building infrastructure. In fact, funding should be granted for a longer horizon (with periodic reviews) and focused on building infrastructure that is linked to broad topics (e.g. ageing). If policy-makers want researchers to work on relevant problems, they should make sure relevant data exist and is easily accessible.
- Researchers should be in charge of building and maintaining data infrastructure for surveys. Statistical agencies' role is to collect, pool and harmonize data (the Scandinavian example) and link survey data collected by scientists to administrative data. This will safeguard innovation and create the most value for research. As an example, Arie Kapteyn contrasted the highly successful German social science survey (run by scientists) with a similar project run by Statistics Netherlands that is barely used by the research community.

Guglielmo Weber (University of Padua) presented the Survey of Health, Ageing and Retirement in Europe (SHARE), which tracks a representative sample of individuals age 50+ in different European countries. Population aging is one of the grand demographic changes facing Europe and it is therefore important to understand its likely impacts. Guglielmo Weber described the data collection process, which relies on face-to-face interviews as well as the methodological challenges they faced (different languages, different interpretations). The data is freely available to researchers; they do not allow commercial use. SHARE can be considered a success with users in both the bio-medical and social sciences and more than two publications based on SHARE data per year. The database is also used for evidence-based policy making by Member States and on European and international level.

Despite this success, the main problem the project faces is its complex and unstable funding. In the beginning, the European Commission funded (almost) the entire project whereas currently, the majority of funding must be provided by the Member States (following its recognition as a European Research Infrastructure Consortium - ERIC). This decentralized funding turned out to be very unstable and several partners in a number of countries had to leave the SHARE project. This is highly damaging for the value of the SHARE database, as the value of a panel lies in its ability to track individuals over time.

The three presentations were followed by a very lively discussion. One question concerned the opportunity of charging researchers for data access since funding was so crucial (in all the researcher-generated databases presented, the data was available for free). The answer was that these data are public goods and so their optimal price is zero. Moreover, charging for it would likely reduce usage by a significant amount. Another discussion point was the requirement that administrative datasets provided for research purposes be deleted after a short period of time. This makes it more attractive for researchers to collect their own data rather than using administrative data. A speaker replied that the retention period is usually long enough to publish at least a working paper version. Collecting data that is comparable to administrative data is not feasible and thus not an option for researchers.

Session 3: Data generation in (quasi)controlled environments

The third session was devoted to data generation in controlled and quasi-controlled environments. **Colin Camerer** (California Institute of Technology) spoke about the use of laboratory experiments in economic research. Laboratory experiments are experiments where human subjects are asked to carry out tasks or make decisions in stylized situations designed to capture relevant features of the economic problem of interest. By exogenously varying some of the parameters of the economic environment they are facing, we can better understand the effects of these changes and the mechanisms through which they work. Lab experiments have been around for many years but have been slow to gain acceptance among the economic academic community. This is no longer the case and Colin Camerer documented the increasing importance of experimental economics in terms of publications. Experiments have also found their way to the classroom to bring theories to life. There are still room for improvement however and, in particular, he pointed to the lack of training on how to produce (experimental) data during graduate (Ph.D.) studies.

The most important contributions of experimental economics have been in the areas of decision theory, supply and demand of goods, game theory, auctions and public goods. Most lab findings have been found to be robust to plausible variations and in most cases have been replicated many times. In his view, there are still opportunities for development and he viewed large-scale political economy experiments (in the flavour of experimental research about the 'poverty trap') as one of those. There are also interesting developments in methods for experimental economics, including eye-tracking and fMRI. Their impact in economics will mostly depend on attention and take-up by economists.

When comparing laboratory experiments to field experiments, he first stressed that laboratory experiments are not designed to inform policy but rather to provide replicable evidence about general phenomena. Nevertheless, when the experimental design closely matches the field situation, the lab results have usually closely replicated the field results. For him, this suggests that lab experiments and field experiments can be complementary. Some findings are first discovered in the lab, then replicated in the field. Likewise some findings are first discovered in field data and then dissected in lab experiments.

Bruno Crépon (ENSAE Paris Tech and J-PAL) spoke about the lessons learnt from randomized controlled trials (RCTs). RCTs are field experiments where the people being studied as randomly allocated to face one economic environment or the other (called "treatment"). The outcome enables researchers to infer *causal* effects of policies or more generally changes in the environment. Bruno Crépon briefly described several studies in the US and France (mainly social programmes) that illustrated how RCTs can be put to work to evaluate the efficiency and effectiveness of different programmes. Another field in which many RCTs are conducted is development economics. Beyond evaluating policies,

evidence generated by RCTs has also changed the way economists think about different policies (for example on the effectiveness of microcredit). A third use of RCTs is to test and quantify theoretical predictions (in that sense RCTs are close to lab experiments). As an example, the speaker described that in development aid, there was often a price for goods instead of giving them away for free. The theory behind is that goods are only valued if they have a price. This turned out to be not the case. Asking for money had no effect on the usage and only reduced the demand.

The subsequent discussion focused on the external validity of RCTs, ethical aspects of RCTs, and data access. A participant raised doubts about the external validity of the results found in RCTs because the RCTs are conducted in another environment than the actual policy (e.g. corruption may play a role in the actual implementation but less in a RCT). Bruno Crépon acknowledged that this could be an issue. Some participants raised concerns about the ethical aspects of RCTs. This may be important in case of health issues. RCTs can potentially lead to conflicts within societies if they touch cultural or politically controversies. Bruno Crépon answered that there existed strong ethical guidelines about experiments with human beings. These are usually set at the university level. A participant asked about the availability of the data generated by lab experiments or RCTs. Colin Camerer replied that a number of journals now require the publication of the dataset. The American Economic Association has also launched an initiative to archive RCTs. A participant also noted that laboratory experiments or RCTs tended to be shorter (or at least the main results) and were thus relatively more published in non-economics journals like Science or Nature.

Session 4: Data standards and cross-country datasets

Cross-country variations in data standards and data definition are big obstacles to multi-country research and comparative analyses. The objective of session 4 was to investigate recent developments in data harmonization and their drivers.

Roberto Barcellan (Eurostat) discussed the G20 data gaps initiative. The statistical system learned several lessons during the financial crisis: (1) there exist several data gaps. For instance, there were no or insufficient data on real estate prices or on wealth and income distribution; (2) data were available with delay (e.g. output data is provided with a delay of three years); (3) there was a lack of global indicators; (4) communication and systematic cooperation between statistical agencies is inadequate. This hinders the ability of policy-makers to anticipate crises, analyse their diffusion and consequences and devise suitable solutions. As a reaction, statistical agencies have started to collaborate more, combine data from public and private sources (data warehouse approach); in short, to transition from the 'data collectors' paradigm to the 're-users of data' paradigm. One output is the G20 data gaps initiative, a set of 20 recommendations on the enhancement of economic and financial statistics. There is also the Inter-Agency Group (IAG) consisting of central banks and statistical offices, which seeks to identify areas that would benefit from interagency cooperation and coordinate work among international agencies. The main outcomes are the Principal Global Indicators that consists of different macroeconomic indicators on a global level. The speaker concluded by giving an outlook to the future work of the G20 data gaps initiative and by briefly describing three case studies on real estate prices, quarterly sectoral accounts and distributional information.

Peter Bøegh Nielsen (Statistics Denmark) talked about how to establish internationally harmonized statistical databases. The main challenge when creating firm-level databases is the traditional stove pipe production system in statistical offices, where different units work in isolation. The results are separate databases. In order to create integrated data, statistical registers must be linked at the firm

level. He described the organization and work of several European micro data linking projects involving different countries and data sources. He then illustrated the questions that could be addressed by such harmonized international micro datasets by presenting case studies on SME's and trade. At the end, he listed the challenges related to the linking of databases. Among these, the requirement that the identical enterprise is included in the different databases is the most important one.

László Halpern (CERS-HAS) shared the result of the FP7-funded MAPCOMPETE project ("Mapping European Competiveness") that he coordinated. The aim of the project was to provide a thorough assessment of data opportunities and requirements for the comparative analysis of competitiveness in European countries. Competitiveness is frequently measured on the basis of macro variables but it is important to recall that firms compete against each other and not countries. It is therefore crucial to focus on firms. MAPCOMPETE defined a number of macro indicators of competitiveness as well as a range of micro-indicators of competitiveness covering productivity, firm dynamics, international and R&D activities and ownership. The project assessed the computability and accessibility of each indicator for each country. The speaker went on to discuss preconditions (advances in technology and methodology, opening up of official data repositories) and potential benefits (address more complex research questions, higher data quality). The main barriers to matching micro data can be grouped in three categories: (1) factual restrictions, e.g. lack of harmonization in data definition or lack of data, (2) technical restrictions, e.g. no common identifiers in the data, and (3) legal restrictions, e.g. data access is restricted or matching not allowed due to privacy protection. He concluded by giving a brief overview of other initiatives aiming at improving access to cross-country micro-level databases and pathways to improve data.

Joseph Tracy from the Federal Reserve Bank of New York talked about the Financial Stability Board (FSB) DataGaps Project which aims at generating a common data template for global systemically important banks (the project builds on several recommendations of the G20 data gap initiative described by Roberto Barcellan). He first reminded the audience about the pre-crisis data environment, which was largely focused on the safety of domestic banks (micro-prudential approach). Even there, several data gaps already existed such as the inability of banks to compute their exposure to Lehman Brothers during its default. As the financial system became increasingly global, the existing data structures to supervise financial markets turned out to be inadequate. To produce more relevant data, the FSB DataGaps initiative created a data hub at the Bank of International Settlements where all home supervisors submit their data. The data now cover the structure and interconnections in the global financial network in order to identify risk concentration and potential spill-overs (measures of micro-*and* macro-prudential risk). Joseph Tracy described the kind of questions the project needed to address. For example in terms of frequency of data reporting, there were tensions between monitoring supervisors who considered that a less frequent data collection was adequate and agencies mandated with crisis management who pushed for higher frequencies to ensure that banks had the capability to report at a high frequency during a crisis. A big challenge was to overcome the reluctance of national supervisors to share their data. It turned out to be important to follow the principle that the national supervisors still 'own' their data and there is no direct access to the pooled data, only reports are shared between supervisors. Beyond that, the database is reciprocal meaning that they only receive reports if they share their data. The speaker concluded by briefly describing the organization and responsibilities of the data hub. The project will reach its last phase in 2016.

The session ended with the presentation of **Lisa Wright** (Bureau van Dijk). Bureau van Dijk (BvD) is a private firm that provides information and business intelligence. One of its products is Zephyr, a database that provides harmonized data on all M&A transactions in the world. She described the chal-

lenges for data collection that they are confronted with. A transaction involving two public companies does not pose any difficulties because it follows formal information rules set by the stock exchanges. If a private company buys a public company (or vice versa), the data availability depends on the rules of the stock exchange and on the transaction value and stake. For the case in which a private company buys another private company, data availability depends on the local legislation. For its data, BvD relies on official sources as well as news services. She described how primary information about an M&A deal was treated so as to create reliable and complete records, that are consistent across countries. To ensure reliability and accuracy, BvD employs external information providers that have local expertise. Interestingly BvD is both a user of data from statistical agencies and a supplier to statistical agencies.

The plenary debate that followed the presentations focused on the quality of data and statistical services. One participant noted the low quality of Eurostat data and reported technical problems with their databases, especially in comparison with the Federal Reserve who maintains a user-friendly statistical database (FRED). He concluded that macroeconomic research was therefore difficult with European data. Roberto Barcellan (Eurostat) acknowledged the current limitations even though they are working on data visualization and some of these other technical issues. Their current focus is data harmonization. Another participant noted that the push for data harmonization was all very good but worried that the fact that it was solely policy driven could lead to yet again a failure to predict the next crisis because the relevant data were not collected. A panellist answered that, while data gaps could of course prevent us from predicting the next crisis, the main reason for our inability to predict crises is that reports will overlook the relevant data. The data for policy use is in its nature very sensitive and access must to be restricted, limiting its usefulness for research. Another participant wanted to know to what extent statistical offices were prepared for the future crisis which the participant saw in the insurance and pension market. A speaker expressed his hope that the data hub at the BIS would be extended in this direction and said that preliminary work was undertaken.

Session 5: The changing face of research-policy and research-private sector collaborations

Researchers have long been involved in policy as ex-post evaluators (policy assessment) or as advisors but new, more collaborative models of interactions are emerging where researchers and policy-makers are partnering, with the benefits of access to data and possibly funding for the former and quality advice for the latter. Likewise such partnerships are also developing between researchers and data-intensive firms such as Yahoo, Microsoft, Google, and financial exchanges to name just a few. The goal of session 5 was to identify the implications of these developments for the type of research being carried out or for the organisation of this research, as well as discuss the benefits and risks involved in these developments (e.g. in terms of scientific integrity and independence, data confidentiality and thus non replicability of the results, …).

Asim Khwaja (Harvard University) offered his perspective on the changing relationship between researchers and policy makers in the context of research in development economics. There has been a significant increase in the number of empirically rigorous studies (such as RCTs) in the recent past but he noted their failure to actually influence policies or to deliver the expected results once implemented "for real." He saw five reasons for this. First, many research questions are not directly relevant to policy needs. Second, research findings may not generalize to other contexts than the one in which they

were generated. Third, proposed solutions are not always comprehensive and in particular fail to account for incentives of stakeholders. Fourth, most research studies are designed as one-time evaluations without built-in mechanisms to readjust given the observed results. This makes them less useful for actual implementation. Fifth, much of the existing research is solution-driven rather than problem-driven: it seeks to identify the problem that the solution the researchers have in mind addresses. Khwaja presented the 'Smart policy design' approach that he and colleagues at the Harvard Kennedy School of Government have developed in response to these concerns. Their proposed approach takes the "design dimension" of policy seriously by applying both theory and data and anticipating, from the very beginning, the need for continuous redesign in the face of new information. He illustrated this approach with a recent project on education in Pakistan. He had four recommendations for successful research–policy collaborations: (1) Be truly problem-driven, (2) Build trust: this requires genuine engagement on all sides (instead of simply viewing the partner as a data stream); (3) Allow for longer term relationships with multiple research outputs rather than one-off projects; (4) Avoid "Consulting relationships", which means that it is best if the researcher has its own funds and the partners contribute resources as well.

Liran Einav (Stanford University) talked about private sector collaborations in economic research. He first documented the rise in proprietary data used in published research. He then described some of his own research projects using proprietary data from different industries. These projects illustrate common themes and issues he saw when working with private proprietary data. First, relationships are essential. The project needs an inside cheer-leader to be successful. One also needs to understand that there will be some constraint on research topics when working with private data. There will also always be some risks that the partner pulls out or veto publication but good relationships can help. Second, all of these research projects with private data were financed by public grants and Einav viewed the independence that such public funding brought as an advantage. Third, access to private data tends to be more effort intensive. These data are initially collected for the needs of the business, not for research. This means that it may not cover all the variables one would want and that their formatting may not be optimized for research use. Einav predicted that greater use of private data in economics will lead to larger research teams. Einav concluded by stressing that, despite all the challenges with private data, there were no good public substitutes for such data and that the rewards from such data were worth the risks for the researchers and the costs in terms of lack of reproducibility. He predicted that standard practices will emerge as more researchers figure out their own ways to this type of data.

Markus Moebius (Microsoft Research Lab) spoke about the empirical economics program at Microsoft Research. Microsoft Research Labs fund basic research in computer science and social sciences through their in-house team of researchers or their visiting researcher program. Their in-house researchers are free to choose their research agenda but one of the benefits of working at MSR is access to unique proprietary data. MSR is also facilitating access to their data by external researchers through specific programs. As with all proprietary data, there are issues: datasets are very large, some data are very sensitive and access requires specific permissions, datasets are largely unstructured and poorly documented, the format frequently changes, and data get deleted after a while. These require suitable software and skills to work with them. Moebius argued that these challenges have limited economists' use of the data. As an illustration of the potential of these data, he presented an article using browsing data to analyse the question whether news aggregators (e.g. Google News) are complements or substitutes for news consumption. He concluded by describing how the empirical economics program at MSR works and stressing Microsoft's interest in having innovative research carried out on the basis of their data.

In the subsequent discussion, the discussion mainly revolved around the question of how to assess the quality of research when data are not available for replication. Khwaja argued that the project partners (firms, public institutions) were interested in accurate results alleviating the problem of purposeful cheating. He also suggested that an independent researcher could be granted access to the data before publication. Einav did not think that the quality of research might be harmed by not having the possibility of replicating results. Replication might be helpful but it is not necessary. He said that other types of empirical fraud ('cooking' results etc.) were a greater concern. Moebius agreed and added that similar databases could be used as a substitute for replication.

Another participant noted that some universities do not allow their researchers to sign non-disclosure agreements and wanted to know how the speakers dealt with issues like this. Einav agreed that this was a problem and that he found himself in continuous struggle with the legal office of his university. Khwaja said that, in his experience, legal constraints coming from his partners, regarding the form and time of publication and regarding the results that are presented, were a much bigger constraint.

Session 6: Panel on the implications of the developments in data and methods in economics for research funding

Véronique Halloin (FNRS) introduced the panel. Funders are key actors of research and they need to be able to accompany developments, including changes in research support needs, in the different disciplines. She raised a number of questions that she hoped the panel could address, namely whether data standards should be imposed on researchers collecting their own data, whether funders should fund individual researchers collecting their data or data infrastructure, the consequences of new sources of data for research and research funding, whether researchers should provide input to data definitions of statistical agencies, the tension between open access and confidential data, the future of neuroeconomics and other data-related initiatives at the interface with other disciplines, which are often a challenge for funding agencies.

Dominik Sobczak (European Commission) talked about research funding at the EU level. Within the "Horizon 2020 – Framework program for Research and Innovation", data projects fall in the category of "Research Infrastructures." The goals are to develop new world-class research infrastructures and to integrate existing national and regional research infrastructures. For the latter, the greatest challenge to data integration and data exchange is the comparability of data. Sobczak highlighted several specificities of social science and humanities data: the prevalence of subjective or even objective data subject to different interpretations (including concepts like unemployment which varies across European countries), the diversity of tools and methods used for analysis, and the relatively low cost of collection. He also discussed a few implications of these specificities: the commonality of conflicting results – even based on the same data, the need for flexibility to research needs, the importance of standards for comparability. According to him, the potential for "big data" is still largely unexplored in social sciences and humanities. He stressed that research infrastructures need a constant financial support to be sustainable. For this reason, research infrastructures need to demonstrate their impact on policy formulation, society and economy. As a positive example in that respect, the speaker mentioned the 'Macroeconomic model database' that turns out to be useful for policy design.

Paul Sanderson (Economic and Social Research Council (ESRC), UK) discussed the impact of 'Big Data' on the way research is funded in the UK. ESRC wants to ensure the independence of research and to promote collaboration between researchers and data providers. Furthermore, it wants to develop the skills necessary for "Big data" projects. These goals are pursued by several initiatives. A first initi-

ative is the creation of the Administrative Data Research Network (which Luke Sibieta discussed in session 1) whose goal is to facilitate access to administrative data, and the associated setting up of an Administrative Data Service and four Administrative Data Research Centres (ADRCs) as points of entry and service for researchers working with these data. A second initiative, which started in February 2014, is the creation of Business and Local Government Data Research Centres which enable and facilitate academic research access to specific business and local government data. Their next big initiative will likely focus on social media and real time analytics. Sanderson also briefly outlined efforts of the ESRC to develop understanding and capabilities to make the most of Big Data (National Centre for Research Methods, Centre for Doctoral Training in New forms of Data). On the question of governance of the different centres mentioned in the presentation, Sanderson said that such centres were funded but not run by ESRC and that their funding was the result of a competition between different universities.

Angelika Kalt (Swiss National Science Foundation) presented the SNSF's perspective on data. From the SNSF's perspective, data and results must be reproducible to ensure the effectiveness of funding and scientific integrity, and there are ongoing discussions about how to store the data that the agency has funded. While noting the tension between SNSF's strong bottom up funding tradition (about 80% of the funded research is bottom up), Kalt acknowledged the growing importance of research infrastructures. Her view was that research infrastructures should not be the sole responsibility of researchers but also of political authorities who have a role in setting priorities and common standards, and ensuring sustainable funding. For these reasons, she was also cautious about the competitive funding model of research infrastructures. She presented several funded research infrastructures in Switzerland such as cohort studies in medical science and FLARE (international infrastructure in physics and astrophysics). In the social sciences, the SNSF partially funds FORS, the Swiss Centre of Expertise in the Social Sciences which has the mandate to produce and safeguard data from surveys, to develop survey methods and to advice and collaborate with researchers. She also described ongoing efforts to combine separate sources of information and create a central place with science-related digital content. Kalt concluded with an outlook on the impact of Big Data. Big Data will change the nature of research projects and might lead to stagnation or decrease of certain research fields like hands-on science or fundamental research. It has the potential to replace lab costs and expenses for field studies. However, it also creates new legal and ethical problems.

The debate revolved around the question of governance and funding of research infrastructures. There was a consensus that the fragmented and conditional funding of the SHARE project discussed in session 2 could be seen as an example how funding should not be organized. A participant was of the opinion that research infrastructures should never be run by the government and that interdisciplinary research was common business and did not need special support. Kalt agreed with the concerns raised by the participant but reported that there had been cases where research infrastructures were also mismanaged by researchers who may not have the management skills to run them. Sobczak replied that research infrastructures were run with a kind of shareholder model. Regarding interdisciplinary research, he was convinced that there was still great potential as only close fields had cooperated so far. With respect to funding European-wide projects, he emphasized that the Commission could be only serve as facilitator and that Member States had to stand up to their funding obligations.

Session 7: Big Data: Definition, challenges and opportunities

The purpose of session 7 was to explore the potential of big data for economics. Is big data going to change economics? What are the examples of fundamentally new insights generated by big data?

According to **Sendhil Mullainathan** (Harvard), big data is not only characterized by the size of datasets but also by new kinds of data (e.g. satellite data, social media) and new ways to analyse them (e.g. machine learning). Machine learning constitutes an alternative to the general principle of *deduction*. It instead takes data and sees what works best, i.e. *inducing* features from data. According to Mullainathan, machine learning has several benefits with the new types of data being generated. It enables researchers to handle wide data (i.e. more variables than data points) or very rich functional forms, and do so without over-fitting. The speaker then contrasted traditional econometrics to machine learning. The key difference is that traditional econometrics focuses on the unbiasedness of estimators and on the inference on estimated coefficients, whereas machine learning is all about predictive accuracy of the dependent variable and about being able to handle high dimensional functional forms and variables. He argued that prediction is interesting for three reasons. First, prediction can be used beneficially for policy. To illustrate, he presented some of his recent research with Kleinberg, Ludwig and Obermeyer using machine learning to forecast the rate of recidivism of released prisoners. Second, it offers new ways to test theories. Here, he briefly outlined his work on 'theory completeness' and 'inductive theory testing'. Lastly, the speaker noted that machine learning offered new ways to handle data (for example in the process of data cleaning).

Lucrezia Reichlin (London Business School) talked about the potential of big data in macroeconomics. At its origin are early independent works by Mark Watson and her-self on dynamic factor models. Three main ideas underlie the big data research program in macroeconomics: (1) the existence of unexploited data of potential interest for macro, including micro-data, (2) the fact that economic data tend to be correlated (business cycles), and the need to rethink the concepts of convergence and parsimony when we increase both the number of variables and the length of the sample. She discussed the curse of dimensionality that such large models raise: the proliferation of parameters is likely to lead to high estimation uncertainty making predictions based on traditional methods poor or unfeasible. She gave an overview of how researchers have addressed this curse of dimensionality ranging from factor models, which limit complexity by focusing on few sources of variation, to penalized regressions. When data are correlated, these alternative methods have been found to have similar performance in macro. One of the most successful applications of big data in macroeconomics is "now-casting", the exploitation of the continuous real-time data flow to make real-time predictions. She concluded by saying that there were still unexploited available, traditional data sources. Regarding new sources of data (e.g. Google), she thought that they were potentially useful but not yet convincingly used. She stressed the need to develop specific tools for economic data, as has been developed in macroeconomics and cautioned against blindly importing methods from other fields.

Session 8: How will big data change econometrics?

The goal of the session was to discuss the challenges that high dimensional data create for econometrics.

Eric Gautier (Toulouse School of Economics) recalled that in classical statistics the dimension of the unknown parameters is small compared to the sample size and that inference in that case often relies on the assumption that the sample size approaches infinity. In contrast, the number of parameters is

large and possibly exceeds the number of observations in the case of big data. When the number of parameters is larger than the number of observations, some of them are likely to be equal to zero and the challenge then is to figure out which ones are zero. This adds to the curse of dimensionality. He illustrated the problem using the genetic determinants of a medical condition where the number of possible of DNA sequences was much larger than the data. Even if we have reliable coefficient estimates, using p-values to select significant coefficients would lead to a result in which many true zeros are declared significant. Gautier described some of his recent work that is targeted at solving this problem.

Herman van Dijk (Erasmus University Rotterdam) talked about computational challenges and Bayesian inference with big data. The availability of large datasets allows predicting variables of interest more accurately than before, but forecasting with many predictors requires new modelling strategies, sequential updating and extra computing power. He illustrated the computational challenges and the potential of parallel computing with a practical example of forecasting S&P500 using 3,700+ financial time series. Simulation-based Bayesian econometrics is particularly potent for this type of problems because (1) it can deal with complex economic issues, e.g. nonlinearities in data, (2) information reduction is natural in the Bayesian approach, (3) it has high practical relevance as it allows for estimation of impulse responses after shocks and thus estimation of policy effects. He briefly summarized key points about Bayesian econometrics before outlining the main challenges he saw ahead. First, more micro-data need to be used to improve simulation-based Bayesian econometrics. These new data will require to model large choices. Second, we need to further develop methods that take advantage of developments in hardware.

Jeffrey Wooldridge (Michigan State University) discussed how big data could improve causal inference. The key idea is that while the number of predictors can be very large, the causal variable (e.g. the program intervention) has typically low dimension. LASSO or other approaches can be used for dimension reduction and generate a prediction of the outcome, absent the policy. Such methods could in theory be used to improve precisions in RCTs. He discussed how the approach could be generalized to include instrumental variables, heterogeneous treatment effects, panel data and the case in which many observations are available. In his concluding remarks, Wooldridge warned against ignoring basic lessons, such as forgetting about practical significance, over-controlling or assuming serial independence.

A participant noted that a random sample was needed to learn something about the underlying population and wanted to know in which big data problems this assumption was fulfilled. Gautier agreed that there was often self-selection in big data (e.g. users of specific websites).

List of participants

Roberto	Barcellan	European Commission, EUROSTAT
Stefan	Bergheimer	Université libre de Bruxelles
Michiel	Bijlsma	Netherlands Bureau for Economic Policy Analysis
Peter	Bøegh-Nielsen	Statistics Denmark
Lidia	Brun	Université libre de Bruxelles
Caterina	Calsamiglia	CEMFI
Colin	Camerer	Cal Tech
Estelle	Cantillon	Université libre de Bruxelles
Angela	Capolongo	Université libre de Bruxelles
Laurens	Cherchye	KU Leuven
Michele	Cincera	Université libre de Bruxelles
Bart	Cockx	Ugent
Bruno	Crépon	CREST and JPAL
Rembert	De Blander	KU Leuven
Olivier	De Groote	KU Leuven
Christine	De Mol	Université libre de Bruxelles
Bram	De Rock	Université libre de Bruxelles
Koen	Declercq	KU Leuven
Thomas	Demuynck	Maastricht University
Catherine	Duverger	Université libre de Bruxelles
Tilemahos	Efthimiadis	European Commission, Joint Research Centre
Liran	Einav	Stanford University
Mariachiara	Esposito	Science Europe
Anders	Fredriksson	Université de Namur
Esmeralda	Gassie-Falzone	Independent
Eric	Gautier	Toulouse School of Economics
Domenico	Giannone	Federal Reserve Bank of New York
Michel	Goldman	Université libre de Bruxelles
Anna	Grochowska	European Commission, DG FISMA
Pegah	Haji Mirza Khoshnevis	KU Leuven
Véronique	Halloin	FNRS (Belgian French Speaking Research Fund)
Laszlo	Halpern	Hungarian Academy of Science
Adel	Hatamimarbini	Université catholique de Louvain
Joerg	Heining	Institute for Employment Research
Bart	Hertveldt	Federal Planning Bureau
Marc	Ivaldi	Toulouse School of Economics
Angelika	Kalt	Swiss National Science Foundation
Arie	Kapteyn	University of Southern California
Athina	Karvounarakis	European Commission, Joint Research Center
Asim	Khwaja	Harvard University
Georg	Kirchsteiger	Université libre de Bruxelles
Vigdis	Kvalheim	Norway Social Science Data Service
Julia	Lane	New York University
Thomas	Lejeune	National Bank of Belgium
Maria-Cruz	Manzano	European Commission, DG FISMA
Laszlo	Matyas	Central European University
Eunate	Mayor	Toulouse School of Economics
Markus	Moebius	Microsoft Research
Sendhil	Mullainathan	Harvard University
Rama Lionel	Ngenzebuke	Université libre de Bruxelles
Gilles	Nisol	Université libre de Bruxelles
Laura	Nurski	KU Leuven
Marianne	Paasi	European Commission, DG RTD
Claudia	Pacella	Université libre de Bruxelles
Davy	Paindaveine	Université libre de Bruxelles
Thi Thu Hien	Pham	KU Leuven
Francisco	Pino	Université libre de Bruxelles
Lucrezia	Reichlin	London Business School

Lorenzo	Ricci	Université libre de Bruxelles
Bettina	Ryll	Université libre de Bruxelles
Paul	Sanderson	UK Economic and Social Research Council (ESRC)
Owusu	Sarpong	KU Leuven
David	Schiller	Institute for Employment Research
Claudio	Schioppa	Université libre de Bruxelles
Hansjakob	Schlaich	European Central Bank
Luke	Sibieta	Institute for Fiscal Studies and UK Administrative Data Research Network
Lode	Smets	KU Leuven
Dominik	Sobczak	European Commission, DG RTD
Stefano	Soccorsi	Université libre de Bruxelles
Jann	Spiess	Harvard University
Harald	Stieber	European Central Bank
Peter	Struijs	Statistics Netherlands
Joseph	Tracy	Federal Reserve Bank of New York
Alessandra	Tucci	European Commission, DG TRADE
Frederic	Udina i Abelló	Statistical Institute of Catalonia
Herman	van Dijk	VU University Amsterdam and Erasmus University Rotterdam
Karina	Véliz	Université libre de Bruxelles
Frank	Verschaeren	Statistics Belgium
Guglielmo	Weber	University of Padua
Philippe	Weil	Université libre de Bruxelles
Jeffrey	Wooldridge	Michigan State University
Lisa	Wright	Bureau Van Dijk
Florian	Ziel	Université libre de Bruxelles

Speakers' bios

Roberto Barcellan is Head of the Unit in charge of Methodology and Corporate Architecture at Eurostat, the statistical office of the European Union. He covered different managerial positions in Eurostat: Head of the Units in charge of Price Statistics, Purchasing Power Parities, Housing Statistics and National Accounts – Production and secretary of the Committee on Monetary, Financial and Balance of Payments Statistics (CMFB). His current portfolio includes methodology, research and innovation, confidentiality and enterprise architecture and the programme for the modernisation of the European Statistical System production processes. He holds a PhD in Statistics from the University of Padua, Italy.

Peter Bøegh-Nielsen, PhD, is head of division for statistics on structural business statistics, including statistics on globalization, at Statistics Denmark. He has been chairing several European development projects on measuring economic globalisation, global value chains and international sourcing. He currently leads a European project on linking micro data of business statistics, including the design of a database containing harmonized information at firm level from 10 European countries. He is a member of the Bureau of the OECD Working Party on Globalisation of Industry and has been chairman of the bureau of the United Nations city group (Voorburg Group on services statistics) and chairman of OECD's Working Party on Indicators on the Information Society.

Caterina Calsamiglia is a research professor of economics at CEMFI and affiliated professor from the Barcelona GSE. Her research focuses on public economics, with an emphasis on school choice, affirmative action and welfare economics. She is the first researcher in Catalunya to get access to linked administrative data on education outcomes and related socio-economic data. She recently received an ERC grant for her work on school choice.

Colin Camerer is the Robert Kirby Professor of Behavioral Finance and Economics at the California Institute of Technology where he teaches cognitive psychology and economics. A pioneer in behavioral economics and neuroeconomics, he is interested in how psychological forces and their deeper neuroscientific foundations influence economic decisions involving individuals and markets. In his research, he uses experiments to better understand how individuals and markets function, neuroscience to gain insight into the neuroscientific drivers for decision making and behavior, and game theory. He is chair of the Russell Sage Foundation Behavioral Economics Roundtable and was named a MacArthur Fellow in 2013.

Bruno Crépon is Professor of economics at ENSAE. He is a board member of the Abdul Latif Jameel Poverty Action Lab (J-PAL), a research network devoted to randomized evaluations of social programs, where he co-chairs the Employment

Program. Crépon is also an IZA and CEPR Research Associate. Crépon completed his undergraduate studies at Ecole Polytechnique in Paris in 1986 and ENSAE in 1988. He completed his PhD in Economics at Université de Paris I in 1994. His research focuses on program evaluation, especially on employment and youth employment programs in both developed and developing countries. He has conducted many randomized evaluations in France, Morocco, Egypt, Cote d'Ivoire and South Africa.

Liran Einav is a Professor of Economics at Stanford University and a Research Associate in the National Bureau of Economic Research. Einav's areas of specialization are industrial organization and applied microeconomics. An important strand of his work is focused on health and other insurance markets, including the development of empirical models of insurance demand and pricing, and empirical analyses of the implications of adverse selection and moral hazard. Einav has also studied credit markets and more recently online markets, often making use of large proprietary datasets. He is co-editor of Econometrica.

Eric Gautier is professor of econometrics at the University of Toulouse I. His research interests include statistics and econometrics with high dimensional parameters.

Domenico Giannone is Senior Economist at the Federal Reserve Bank of New York and Research Fellow of the Centre for European Policy Research (CEPR). His general fields of research are forecasting, monetary policy, business cycles and growth. He has designed econometric models that are routinely used to inform policy decisions in many institutions, was co-founder and director of Now-Casting.com (a web-based forecasting company), and was a member of the CEPR Business Cycle Dating Committee. He is associate editor for the Journal of Applied Econometrics, the International Journal of Forecasting and Empirical Economics.

Véronique Halloin is Secretary General of the Belgian French-speaking Research Fund (FNRS) whose goal is to promote fundamental research in the French-speaking Community of Belgium. She holds a Ph.D. in Civil Engineering and was a full professor of Chemical Engineering at the Université Libre de Bruxelles and Vice-Rector for Research (2006-08) prior to her current appointment. As Secretary General of the FNRS, she also represents Belgium in various international research and research-related bodies such as the OECD's Global Science Forum, CERN, and the European Foundation for Science.

László Halpern is director at the Centre for Economic and Regional Studies of the Hungarian Academy of Sciences in Budapest and a research fellow in the International Macroeconomics and International Trade research programmes of the Centre for Economic Policy Research (London). He has written widely on exchange rate and exchange rate policy in Central and Eastern Europe; on enterprise behaviour, microeconomic environment and economic policy; on multinational enterprises, foreign direct investment and economic development. He coordinates the FP7 project MAPCOMPETE whose objectives are to analyze data opportunities and requirements to analyze and compare competitiveness in EU countries.

Angelika Kalt is Deputy Director of the Swiss National Science Foundation, where she is responsible for quality and development issues concerning research funding and assessment. Dr. Kalt is also the head of the SNSF division that promotes interdisciplinary and cooperative research. She was previously professor in Earth Sciences (petrology and geodynamics) at the University of Neuchâtel.

Arie Kapteyn is professor of economics and the founding Executive Director of the Dornsife Center for Economic and Social Research at the University of Southern California. Much of Professor Kapteyn's recent research is in the field of aging and economic decision making, with papers on topics related to retirement, consumption and savings, pensions and Social Security, disability, economic well-being of the elderly, and portfolio choice. He is the founder of several internet panels including the CentERpanel at Tilburg University, the American Life Panel, a nationally representative sample of 6,000 households maintained at RAND, and the Understanding America Study (2,000 households) at USC.

Asim Khwaja is the Sumitomo-Foundation for Advanced Studies on International Development Professor of International Finance and Development at the Harvard Kennedy School, and Co-Director of Evidence for Policy Design (EPoD), a research group that uses a "smart policy design" approach to foster collaborations between researchers and policy makers that can help systematically identify, diagnose and then design and test solutions to pressing policy problems. Khwaja's own areas of interest include economic development, finance, education, political economy, institutions, and contract theory/mechanism design. His research combines extensive fieldwork, rigorous empirical analysis, and microeconomic theory to answer questions that are motivated by and engage with policy.

Vigdis Kvalheim is Deputy Director of the Norwegian Social Science Data Services (NSD) and manages, among others, the Data Protection Services and the Individual Level Data Unit at NSD. She is or has been involved in several Norwegian, Nordic and international data management activities, EU projects, working groups and advisory boards. She is a member of the Board of Directors of the International Federation of Data Organizations for the Social Sciences (IFDO), the NordForsk's high-level group on research infrastructures and chairs the Project Board for the Norwegian Remote Access Infrastructure for

Register Data (RAIRD). She was, until recently acting director of CESSDA, the umbrella organization for the European national data archives.

Julia Lane is Professor of Public Policy at the Wagner School and Professor of Practice at the Center for Urban Science and Progress at New York University. She has carried out a number of data-driven projects over her career. She is, among others, the founder of the Longitudinal Employer-Household Dynamics Program at the US Census Bureau (which evolved into a permanent Census Bureau Program); she has created and developed the NORC/University of Chicago data enclave to provide remote and protected access to sensitive microdata and has overseen a number of social surveys in the US.

Markus Moebius is principal researcher at Microsoft Research New England. His primary research interests include the economics of social networks where he seeks to develop models of learning, coordination and cooperation in social networks and to test those using field and lab experiments. His recent work explores how browsing data can be leveraged to analyze news consumption.

Sendhil Mullainathan is Professor of Economics at Harvard University. His research interests span a broad spectrum including behavioral economics, labor economics, the market for media and corporate finance. His latest research focuses on using machine learning and data mining techniques to better understand human behavior. He helped co-found a non-profit to apply behavioral science (ideas42), co-founded a center to promote the use of randomized control trials in development (the Abdul Latif Jameel Poverty Action Lab), and has worked in government in various roles, including most recently as Assistant Director of Research at the Consumer Financial Protection Bureau.

Lucrezia Reichlin is Professor of Economics at the London Business School, co-founder and director of Now-Casting Economics ltd, non-executive director of UniCredit Banking Group and AGEAS Insurance Group. She is Chair of the Scientific Council at the Brussels based think-tank Bruegel as well as a member of the Commission Economique de la Nation (advisory board to the French finance and economics ministers). Between March 2005 and September 2008 she served as Director General of Research at the European Central Bank. She is an expert on forecasting, business cycle analysis and monetary policy. The econometric methods she has developed for short term forecasting (now-casting) are widely used in central banks around the world.

Paul Sanderson leads on economics and finance at the UK Economic and Social Research Council (ESRC), where he is responsible for the development of the ESRC's research portfolio in these areas and for developing academic engagement and collaboration with the financial services sector. Paul has been with the ESRC since 2010. He began his career as an academic economist but has subsequently held positions in management and research at the Bank of England, HM Treasury and with a number of firms in the UK private sector.

Luke Sibieta is programme director of the education and skills sector at the Institute for Fiscal Studies. He has wide experience in accessing and conducting research using administrative data. He is also a non-executive board member of the UK's new Administrative Data Research Network (http://www.adrn.ac.uk/), an initiative that is seeking to make more administrative and linked data available to researchers (funded by the Economic and Social Research Council).

Dominik Sobczak holds a degree in Finance and Banking as well as a degree in European Studies, both from Warsaw School of Economics where he also worked as an academic publishing on monetary and fiscal integration in the EU as well as on financial markets. Since 2005 he works at the European Commission in the DG Research and Innovation. In 2005-2014 he was responsible for funding research in economics, finance and demography. Since September 2014 he is Executive Secretary of the European Strategy Forum on Research Infrastructures.

Joseph Tracy is Executive Vice President and Senior Advisor to the President at the Federal Reserve Bank of New York. He is, among others, a member of the Global Legal Identifier Project to create a global system of unique identifiers for legal entities and chair of Data Workstream for the Data Gaps Implementation Project sponsored by the Financial Stability Board. Tracy holds a Ph.D. in Economics from the University of Chicago. His research interests include unions and collective bargaining as well as housing and urban economics.

Frederic Udina is the general manager at the Statistics Bureau of Catalonia (Idescat) since 2011 and president of the Catalan Institute of Public Policy Evaluation (Ivàlua) since 2013. Frederic is also professor in math, probability and statistics at Universitat Pompeu Fabra and at the Barcelona Graduate School of Economics. He holds a PhD in Mathematics from the Universitat Politècnica de Catalunya and a degree in Mathematics from the Universitat Autònoma de Barcelona. He has published numerous articles, has participated in multiple international research groups and performed consulting in statistics for Morgan Stanley, among others.

Herman van Dijk is professor of econometrics at VU University Amsterdam and professor emeritus at Erasmus University Rotterdam where he was director of the Tinbergen Institute and director of the Econometric Institute. He is the co-founder of

the EC2 meetings of European econometricians; cofounder of the European Seminar on Bayesian Econometrics (ESOBE) and cofounder of the Econometric and Tinbergen Institute lectures that are published by Princeton University Press. His research interests cover a range of topics in econometrics with, as common themes: Simulation-based Bayesian Econometric Techniques for Inference, Forecasting and Decision analysis. His recent research involves econometric forecasting with large data sets and advanced algorithms using parallel computing.

Guglielmo Weber has investigated various aspects of consumer behaviour (consumption over the life cycle, consumer demand, portfolio choice), starting with his PhD (LSE, 1988), and continuing with publications in international academic journals including the American Economic Review, Econometrica, the Review of Economic Studies and the Journal of Political Economy. His recent research interests include the economics of ageing and retirement – and have led him to be country team leader for the Survey on Health Ageing and Retirement in Europe (SHARE) in Italy and responsible for the economics contents of the questionnaire. He has held tenured positions at University College London, Università Ca' Foscari di Venezia and Università di Padova. He is an International Research Associate of the Institute for Fiscal Studies and a Research Fellow of the CEPR (London). He is deputy-coordinator of the SHARE-ERIC, the European consortium that manages SHARE.

Jeffrey Wooldridge is University Distinguished Professor of Economics at Michigan State University where he has taught since 1991. His research interests include econometrics of cross section and panel datasets. He is a Fellow of the Econometrics Society, of the Journal of Econometrics and a recipient of the Alfred P. Sloan Fellowship.

Lisa Wright is managing director at Bureau Van Dijk, a publisher of global business information (ORBIS, Bankscope, Zephyr). She is responsible for BvD's M&A products globally and manages Zephus Ltd, the BvD subsidiary that specializes in researching and creating their Global M&A, PE, IPO & Venture Capital deals database, Zephyr.