Understanding the link between computational behavioral economics and the Big Data revolution. Towards a new paradigm in Macroeconomics?

Paola D'Orazio

Chair of Macroeconomics - Ruhr University Bochum (Germany)

Paper submitted to the Duke Forest Conference 2016 (Durham, NC - USA)
Session on Big Data and Economic Methodology

November 12, 2016

Introduction

The term Big data is ubiquitous

Despite its widespread usage, it seems that there is no mature discussion about BD in academia as well as no univocal definition for it $\rightarrow$ *confusion* regarding concepts as well as *skepticism* with respect its use in academic research.

Research questions:

- What are Big Data?
- Why are they important for economic research? (. . . why should economists care about them?)
- In presence of the *Big Data revolution*, is Economics (in particular, Macroeconomics) heading to a paradigm shift?

Aim of the paper

- Report how the debate on BD in social sciences is developing and potentially affecting research in economics, and, in particular, in Computational Behavioural Economics (CBE)
  $\rightarrow$ By CBE we mean Agent-based computational models "informed" by findings reported by Behavioural Economics (see Akerlof, 2002; Colander et al., 2008; Velupillai and Kao, 2014, among others).
- Discuss the extent to which the increasing production and availability of BD for social researchers can make it possible a shift from *data-scarce* to *data-rich* investigations

Motivation of the paper

We highlight two main reasons for drawing the attention on this interlinkage:

1. Because from the interaction of BD and CBE, a new research paradigm is emerging in economics, and in particular in macroeconomics $\rightarrow$ **Agent-based computational economics offers the possibility to store and analyze large datasets, but also to use them and find hidden patterns**

Motivation of the paper

2 Because computational behavioral economic models need Big Data

- Since researchers **lack micro level data** to empirically *microfound* their models, BD can contribute to the improvement on the modeling of the behaviors of the agents that populate the micro level of Agent-based models (ABM)
- The massive sample size and the *granularity* that characterize BD increase the power of data in revealing individual features and/or actions
- They can help researchers observing aspects of the human behaviour (e.g., social links, preferences and so on) beyond the traditional micro data availability.
- This in turn could help in solving the issue of *limited observation* we have in traditional statistics.

The spectrum of data

- **Observational data** (survey data) → Data collected from surveys can be very large and complex, but they are collected in a systematic way over a small sample of the entire population.
- **Experimental data** → similarly to observational data, are collected over a small known sample and in a systematic way because they respond to a specific investigation that aims at testing a specific (theoretical) hypothesis. Differently, from observational data, they are relatively small in size and this reduces the complexity of the data structure. → **partially true in case of proper Macroeconomic experiments (high number of participants)**
- **Big data** → different possible definitions

Definitions of Big Data (I)

a. Ward and Barker (2013) try to collate the various definitions that have been developed in the past 4 years: "*BD is a term describing the storage and analysis of large and or complex data sets using a series of techniques*"

b. Laney (2011)
   1. *volume*: magnitude of data
   2. *variety*: structural heterogeneity
   3. *velocity*: rate at which data are generated

c. Gandomi and Haider (2015):
   1. *veracity*: unreliabity of data
   2. *variability (and complexity)*: variation in the data flow and different sources
   3. *value*: low value with respect to the volume of data

Definitions of Big Data (II)

d Kitchin (2013):

1. *exhaustiveness*: the extent to which BD strive to capture the entire population or systems
2. *relationality*: because BD contain common fields that allow for the connection of different datasets
3. *scaleability*: because BD can expand rapidly in size

e Our view: The term Big Data entails a phenomenon which is based on the *interaction between technologies and data analysis development* and both *co-evolve* with the present technological frontier, i.e., the magnitude of the attribute "big" is related to the computational power currently available for research, data storage and processing

Definitions of Big Data (III)

The definition of BD

should thus involve at least three concepts:

- the *size* (magnitude) of data
- the *complexity* of the data structure
- the increasing rate at which they are produced and stored thanks to the available *technologies*

Are Big Data a *panacea* for social scientists?

1. Having a **big** dataset composed of millions of observational points does not necessarily mean having a "good" dataset $\Rightarrow$ *Quality* and *quantity* should go indeed hand in hand.
2. The **representativeness** of the sample is an issue that is often **not** canceled out by the big dimension of the sample (e.g. analyses carried out by considering the network "traffic" on Twitter and Facebook).
3. BD are usually rather **context-dependent** than generic.
4. Because BD are often **proprietary**, they are usually very expensive. This in turn implies high costs for getting the access to the datasets, which in turn may result in a new form of digital divide in the research community.
5. BD are **complex and unstructured**, implying challenge for statistics and potential new developments for econometrics (Varian, 2014)

Other features of Big Data

- **heterogeneity** $\rightarrow$ heterogeneity that stems from sub-population data and different sources. Sometimes the sub-population data in small samples are considered as outliers because of their insufficient frequency but they can also be considered as an opportunity to model differences in the sample.
- **noise accumulation** $\rightarrow$ accumulation of estimation errors due to the simultaneous estimation of several parameters at a time, so that some variables with significant explanatory power can be overlooked.
- **spurious correlation and incidental endogeneity** $\rightarrow$ in presence of big and complex datasets it can be that uncorrelated variables (because of their independence) are found to be correlated due to the high dimensionality of the dataset.

Methods to dig into "the wilderness" of BD (I)

- In presence of large datasets which could be composed of a very large number of observations (billions) and a large number of covariates (millions), resorting to **machine learning** could be thus the only option to dig into "the wilderness" of BD and carry out the empirical research.
- Machine learning is particularly gaining a lot of attention (Varian, 2014) because, *caeteris paribus*, of its a**bility to deal with large datasets and the possibility of continuos "retraining" of the algorithm over time as the environment changes**
- It does not exist a unified framework yet $\Rightarrow$ large set of algorithms among which researchers could choose.
- Fan et al. (2014) suggest the **adoption of new methods** such as: penalized quasi likelihood, sparset solution in high confidence set, independence screening.

Methods to dig into "the wilderness" of BD (II)

MACHINE LEARNING ALGORITHMS		
Supervised learning	**Regression**	- Linear regression - OLS - LOESS - Logistic Regression - Stepwise regression
	Decision tree	- Classification and regression tree - Conditional decision tree
	Random Forest	
Unsupervised learning	**Clustering**	- k-means - k-medians - hierarchical clustering
Supervised learning	**Markov decision processes (deep learning)**	- Boltzmann Machine - Belief network
	Neural networks	
	Bayesian learning	

Figure: *Machine learning algorithms: an overview.* Source: Author's elaboration.

Which Big Data for Economics? (I)

- **Government administrative data**
  - Data on education
  - Social insurance
  - Local government spending
  - Medicare
  - IRS (Internal Revenue Service): instituted in 1913 that offer a large micro-level dataset of tax revenues. It has been used by Piketty and Saiz (2003) to derive historical series of income shares for the top percentiles earners among US households
- Data that result from **online activity** (queries and/or social media)
- Data on **consumer spending and sentiment**: e.g., MasterCard distributes "SpendingPLus" that provides a real-time consumer spending data in different retail categories

Which Big Data for Economics? (II)

Researchers engaged in **financial agent-based modeling** already make use of very complex and large datasets (LeBaron, 2000; Hommes, 2006).

- Over the past decades, extremely **high frequency data** has become available
- Researchers have a detailed picture of exactly **how the market is unfolding**, as well as the exact **dynamics of trade clearing**.
- There are also series that show detailed **holdings of institutions** and that record the **flows** coming in and out of these funds
- The availability of high frequency data, pricing and volume data from financial markets **enhances the transparency of the trading behaviours of financial agents and allows for a detailed modeling (and/or calibration) of the features of artificial financial markets**

Towards a new empirical revolution

Even if Big Data present some drawbacks and are not a panacea ...

The distinction among the different types of data available for social research lead us to claim that **the process of diffusion and use of Big Data**, thanks to the availability and development of new computational and storing technologies, could eventually **lead to an empirical revolution** akin to the Keynesian revolution that led to the availability of macro data (Patinkin, 1976).

The relationship between Data, Information and Knowledge

Data are unorganized and unprocessed facts about events, which constitute the prerequisites to information. *Information* can be considered as an aggregation of data (*processed data*) which implies a meaning and a purpose. From information we derive *knowledge*; i.e., human understanding of a subject matter that has been acquired through study and experience.

Science paradigms: evolution and features

SCIENCE PARADIGMS		
PARADIGMS	**FEATURES**	**TIMING**
I **Experimental Science**	***Empiricism*** *"data speak for themselves, free of theory"* (*induction*)	before XVIII century
II **Theoretical Science**	***Knowledge-Driven*** Modeling, abstraction (*deduction*)	before advent of computational power
III **Computational Science**	***Simulation of complex systems*** (*induction+deduction*)	before Big Data
IV **eScience**	***Data-intensive*** (*abduction+deduction+induction*)	present

Figure: Source: author's elaboration based on Hey et al. (2009) and their report of Jim Gray's speech on *eScience*.

Paradigms evolution in Macroeconomics (I)

- Even if paradigmatic accounting is usually problematic - and this especially true in the case of Economics - a similar evolution in research and methodology as the one sketched in Figure 2 can be found also in the Economics research field.

Taking an evolutionist perspective . . .

macro methodology has been climbing a *fitness landscape*, where recombination, which is a key feature in evolution, has been playing an essential role in the development of new methodologies in macroeconomics.

Paradigms evolution in Macroeconomics (II)

We follow **Colander, Howitt, Kirman, Leijonhufvud, Mehrling (2008) Beyond DSGE Models: Towards an Empirically Based Macroeconomics, AER** that offer an overview of the evolution of the research methodology in Macroeconomics

1. *Up until the 40s and 50s*, macroeconomics proceeded without a formal theory: macroeconomics policy was based on a loose and largely empirical understanding of the macro economy. The **models** used by economists were **simple and not formalized in rigours mathematical terms**: this phase relied mainly on **induction** $\Rightarrow$ **Paradigm I**.

Paradigms evolution in Macroeconomics (III)

2 With the development of macro-econometric models *in the 50s*, many of the Keynesian models were presented as having formal underpinnings of microeconomic theory; then *in the 70s*, the **formal modeling** began in the spirit of General Equilibrium theory with the aim to build a general equilibrium model of the macro economy based on **explicit and fully formulated micro foundations**:
research methodology based mainly on the use of **deduction**
$\Rightarrow$ **Paradigm II**.

Paradigms evolution in Macroeconomics (IV)

3 However, since analytical models are technically difficult and constrained by the mathematical strait-jacket, it is not clear which, if any, will provide a meaningful advance. Colander et al. maintain that, in order to progress, macro theory should move on to models that take agents' heterogeneity and interaction into account: thanks to the *increase in computing power* over the past decades, there is an alternative approach that can **cut the Gordian analytical knot**; it uses *agent-based computational economics (ACE) and* **simulations** to analyze the macro economy
$\Rightarrow$ **Paradigm III**.

Simulations as a third way of doing science

Simulation **differs from standard deduction and induction** in both its implementation and its goals. (Axelrod and Tesfatsion, 2006)

- Scientists use deduction to derive theorems from assumptions, and induction to find patterns in empirical data.
- Simulation, **like deduction**, starts with a set of explicit assumptions; but **unlike deduction**, simulation does not prove theorems with generality.
- Instead, simulation generates data suitable for analysis by induction; but **unlike typical induction**, the simulated data come from a rigorously specified set of assumptions regarding an actual or proposed system of interest rather than direct measurements of the real world.
- Simulation thus permits **increased understanding of systems through controlled computational experiments**.

Paradigms evolution in Macroeconomics (V)

4 The *data revolution* (Kitchin, 2014) is recently leading to the so-called "computational turn" (Berry, 2011)
$\Rightarrow$ implies **alternative epistemologies in humanities and social sciences and changing of research practices**, also in Economics.

Paradigms evolution in Macroeconomics (VI)

Paradigm shift in Macroeconomics

In the Economics field, we emphasize that **a paradigm shift (IV) implies a further shift from the mainstream deductive approach as well as the inductive approach** (Lawson, 1989; Arthur, 1994), **to an even more articulated shift where induction and deduction, together with abduction will contribute to the understanding of an economic complex phenomenon**

$\Rightarrow$ this is at the **core of agent-based computational economics** where theory, experiments and simulations are the essential elements of the research methodology.

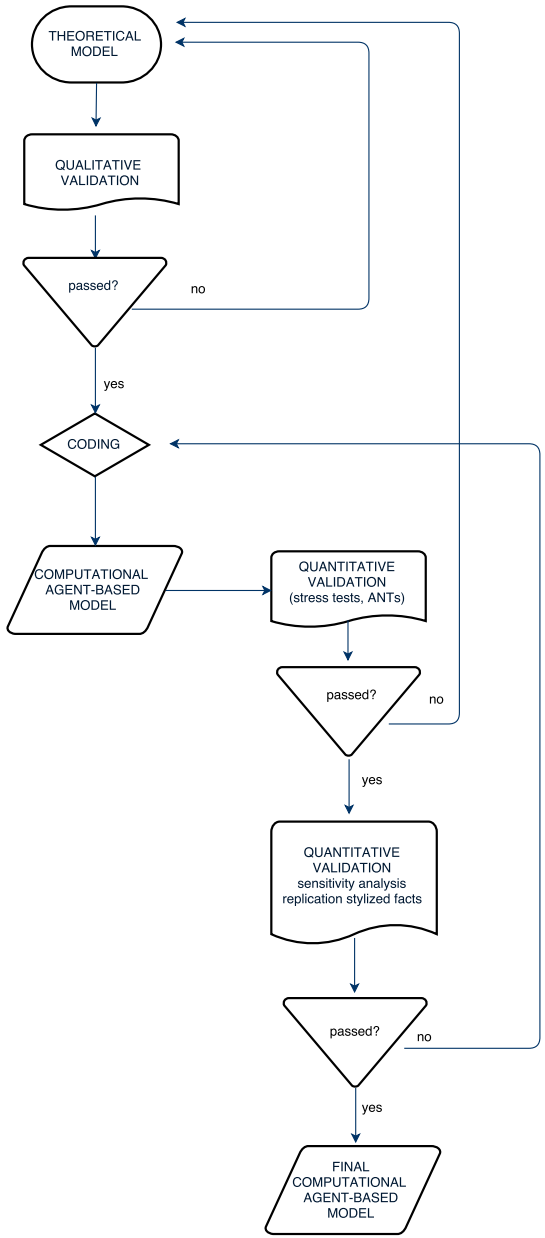

Figure: *Macroeconomic agent-based model: building methodology.* Author's elaboration

Behavioral heterogeneity

- Because of their features, ABMs are thus suitable tools to study complex systems: *systems with many interacting entities and non-linear interactions among them*.
- The value added, in term of modeling approaches and economic insights, of ABMs is that the dynamics of aggregate variables are the result of complex, continuously and endogenously changing *micro-structure* (Epstein, 2007) $\Rightarrow$ **they allow for modeling a large range of economic behaviors**

Behavioral heterogeneity

- Standard Macroeconomics relies on *microfoudations*: the gap between the micro behavior and the macro "flavor" is filled by resorting to the **representative agent** assumption and to the process of **aggregation**.
- However, the reduction of the behavior of a group of heterogeneous agents to the behavior of a single representative agent can lead to **misleading** - or even wrong - conclusions (Kirman, 1992).

Behavioral heterogeneity

The theoretical debate and investigation on the issues arising from the use and implementation of rational expectations (RE) representative agent models has gradually led to the "**behavioral revolution**" (starting from (Simon, 1957)): has provided models with **more realistic psychological foundations** (Thaler, 1994; Akerlof, 2002; Kahneman, 2003; Camerer and Loewenstein, 2004; Camerer et al., 2004) and "practical" ways in which behaviors can be incorporated in them

Open issues in the "alternative" microfoundation of Agent-based models

- economists are still tangled up with **several competing options in the choice of the microeconomic model** (Chen and Wang, 2011).
- when the researcher wants to bring his model to the data, the building of the empirical microstructure still suffers from the **paucity of individual micro data** compared to aggregate (field) data.

The contribution of experimental economics

- By observing human subjects behaviors in the **experimental laboratory**, experimental economics has been helping in bridging the gap between the micro and the macro level in economic models.
- The u**se of experimental micro data to calibrate artificial agents** is among the most important innovations that helped improving the self-referential nature of macroeconomic models (Duffy, 2006).

Improvement of the experimental method through the interplay with Big Data availability and production

Critiques to the experimental method are usually related to the so-called *internal* and *external* validity of experiments.
Advances:

1. **thanks to the possibility of collecting and storing big datasets from several realistic contexts, allows researchers to consider the field as their own experimental laboratory and themselves as "consumers of data"** (Fan et al., 2014, p.293)
   $\Rightarrow$ there is a wide *digital landscape*: collect data on users' opinions from online questionnaires, on consumers preferences via online buying, online auctions, social networks, etc.

Improvement of the experimental method through the interplay with Big Data availability and production

Example: the Billion Prices Project (BPP)

Developed by Alberto Cavallo and Roberto Rigobon: provides an alternative measure of **retail price inflation** relying on data from hundreds of online retail websites in more than fifty countries and data are used to construct price indices that can be updated in real time.
The authors then use the data gathered to *construct the BPP index to document patterns of price changes* in the same way that researchers have used the data underlying the CPI (Cavallo, 2015).

b Compared to the "standard" experimental lab, in the case of BD the researcher can work in a **more natural experimental environment** where
   a **his direct action is hidden to the agents**; this is potentially very important because it helps in avoiding that agents' - directly or indirectly - try to meet the expectations (about the results' of the experiment) of the researcher,
   b **framework effects** no longer exist because there is no need to ask questions (regarding e.g. preferences) directly to the subjects; preferences are just tracked in data,
   c **sample size**, **sample bias** and **representativeness** are no more an issue the researcher should be concerned about
   d for some cases, there is no need to have testable hypothesis of the **theory** (as required in standard experiments); the researcher observes data and grasps possible theoretical insights after the discovery of recurring *patterns*,

Concluding remarks

The different types of data used in social research represent

- noticeable opportunities in specific areas of investigation
- they co-evolve with the different scientific paradigms that have been developing and unfolding over the last decades.

Implications of the *Big Data revolution* for Computational Behavioral Economics

- It implies a **deep epistemological change** which will lead (is leading) economists to consider (or re-consider)
  a. the process of research
  b. the way in which they engage with data analysis, data processing and data gathering
  c. the constitution of knowledge.
- "Inanimate data can never speak for themselves, and we always bring to bear some **conceptual framework**, either intuitive and ill-formed, or tightly and formally structured, to the task of investigation, analysis, and interpretation" (Gould, 1981, p.166).
- Computational (behavioral) economics provides researchers with a set of tools that enables the transition towards a **new scientific paradigm**, which is based on the availability of new kind of data, new analytic tools and new research methodology.

Data as the new starting point for additional analysis and research

- CBE allows researcher to take into account the specific *context* and *contingency* and BD can be used to *refine* the understanding of some features (e.g., agents' preferences) or network structures.
- The massive sample size and high dimensionality of BD collected from various sources could be exploited to provide a very **detailed and empirically-grounded specification of the agents' behaviors** $\Rightarrow$ allows for a more complete and empirical *heterogeneity* of agents already at the initialization of the model, while heterogeneity is usually treated as an *emergent property* of computational agent-based models.

Improvements of the experimental (economic) method

- The field becomes the experimental lab and researchers become "consumers of data"
- Researchers can work in a more natural experimental environment and potentially mitigate the internal and external validity critiques

Thank you for your attention!

Comments are particularly welcome!

E-mail: paola.dorazio@rub.de

Personal website: http://econhoratio.org

Akerlof, G. A. (2002). Behavioral macroeconomics and macroeconomic behavior. *American Economic Review*, 92(3):411–433.

Arthur, W. B. (1994). Inductive reasoning and bounded rationality. *The American economic review*, 84(2):406–411.

Axelrod, R. and Tesfatsion, L. (2006). Appendix a. a guide for newcomers to agent-based modeling in the social sciences. *Handbook of computational economics*, 2:1647–1659.

Berry, D. M. (2011). The computational turn: Thinking about the digital humanities. *Culture Machine*, 12(0):2.

Camerer, C. and Loewenstein, G. (2004). *Behavioral economics: Past, present, future*. Princeton: Princeton University Press.

Camerer, C. F., Loewenstein, G., and Matthew, R. (2004). *Advances in Behavioral Economics*. Princeton University Press, Princeton.

Cavallo, A. (2015). Scraped data and sticky prices. Technical report, National Bureau of Economic Research.

Chen, S.-H. and Wang, S. G. (2011). Emergent complexity in agent-based computational economics. *Journal of Economic Surveys*, 25(3):527–546.

Colander, D., Howitt, P., Kirman, A., Leijonhufvud, A., and Mehrling, P. (2008). Beyond DSGE Models: Toward an Empirically Based Macroeconomics. *American Economic Review*, 98(2):236–40.

Duffy, J. (2006). Agent-based models and human subject experiments. In Tesfatsion, L. and Judd, K. L., editors, *Handbook of Computational Economics*, volume 2 of *Handbook of Computational Economics*, chapter 19, pages 949–1011. Elsevier.

Epstein, J. M. (2007). Agent-based computational models and generative social science. In *Generative Social Science Studies in Agent-Based Computational Modeling*, Introductory Chapters. Princeton University Press.

Fan, J., Han, F., and Liu, H. (2014). Challenges of big data analysis. *National Science Review*, 1(2):293–314.

Gandomi, A. and Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2):137 – 144.

Gould, P. (1981). Letting the data speaking for themselves. *Annals of the Association of American Geographers*, 71(2):166–176.

Hey, T., Tansley, S., and Tolle, K. (2009). *The fourth paradigm: data-intensive scientific discovery*, volume 1. Redmond: Microsoft Research.

Hommes, C. (2006). Heterogeneous agent models in economics and finance. In Tesfatsion, L. and Judd, K. L., editors, *Handbook of Computational Economics*, volume 2 of *Handbook of Computational Economics*, chapter 23, pages 1109–1186. Elsevier.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review*, 93(5):1449–1475.

Kirman, A. P. (1992). Whom or What Does The Representative Individual Represent. *Journal of Economic Perspective*, 6:117–36.

Kitchin, R. (2013). Big data and human geography opportunities, challenges and risks. *Dialogues in human geography*, 3(3):262–267.

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage.

Laney, D. (2011). S3-d data management: Controlling data volume, velocity and variety. Technical report, Application Delivery Strategies by META Group Inc. Retrieved from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

Lawson, T. (1989). Abstraction, tendencies and stylised facts: a realist approach to economic analysis. *Cambridge journal of Economics*, 13(1):59–78.

LeBaron, B. (2000). Agent-based computational finance: Suggested readings and early research. *Journal of Economic Dynamics and Control*, 24(5-7):679–702.

Patinkin, D. (1976). Keynes and econometrics: on the interaction between the macroeconomic revolutions of the interwar period. *Econometrica: Journal of the Econometric Society*, pages 1091–1123.

Simon, H. A. (1957). *Models of man: social and rational; mathematical essays on rational human behavior in society setting*. John Wiley Sons.

Thaler, R. H. (1994). *Quasi rational economics*. Russell Sage Foundation.

Varian, H. R. (2014). Big data: New tricks for econometrics. *The Journal of Economic Perspectives*, 28(2):3–27.

Velupillai, K. and Kao, Y.-F. (2014). Computable and computational complexity theoretic bases for herbert simons cognitive behavioral economics. *Cognitive Systems Research*, 2930:40 – 52.

Ward, J. S. and Barker, A. (2013). Undefined by data: a survey of big data definitions. *arXiv preprint arXiv:1309.5821*.