Big data for mechanistic explanations

VIRGINIA GHIARA

CENTRE FOR REASONING, UNIVERSITY OF KENT

Outline

- Introduction: big data, new data sources and the end of theory
- Causal mechanisms in the social sciences: what is the nature of data? Multilevel data and process tracing
- Big data for mechanistic explanations: big data can be used to cast light on the nature of causal mechanisms and to study the interactions between mechanisms at different levels. Furthermore, new methodological issues and solutions have emerged in big data studies.

Outline

- **Introduction**
- Causal mechanisms in the social sciences
- Big data for mechanistic explanations
- Conclusion

Introduction

Over the last decade, many disciplines have experienced a "data deluge"

- Massive amounts of data generated with high frequency and high speed
- Data automatically generated by machines (e.g. sensor data automatically collected from a device)
- New sources of data (e.g. big data from the Internet)
- Variety of data (e.g. geospatial data, audios, videos, unstructured texts, 3D data…)
- Large amounts of digital information about people's behavior

Recently, it has been argued that this data deluge will transform the way in which we do science: big data will make data-driven research possible and causal studies redundant

> "The new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all." (Anderson 2008)

Anderson C (2008) The end of theory: The data deluge makes the scientific method obsolete. Wired, 23 June. Available at http://www.wired.com/2008/06/pb-theory/

This and other similar claims have generated an intense debate both inside and outside the academia

Many authors have defended the role that causality play in scientific research

In particular, some biomedical projects such as EnviroGenomarkers and EXPOsOMICS, based on various sources of data and large datasets, have been used by many scholars as case studies to examine researchers' engagement with causality and the types of causal evidence that can be gathered from big data studies (Russo and Williamson, 2012; Illari and Russo, 2013; Canali, 2016)

In this paper, I focus on the possibility of using big data for studying causal mechanisms in the social sciences

Canali S. (2016). Big Data, epistemology and causality: Knowledge in and knowledge out in EXPOsOMICS. *Big Data & Society* 3(2): 1–11.
Illari P. and Russo F. (2013) Information channels and biomarkers of disease. *Topoi* 35: 175–190.
Russo F, Williamson J (2012) Envirogenomarkers. The interplay between difference-making and mechanisms. *Med Stud* 3:249–262.

Outline

- Introduction
- **Causal mechanisms in the social sciences**
- Big data for mechanistic explanations
- Conclusion

Causal mechanisms in the social sciences

Many proponents of the mechanistic approach appear to claim that mechanistic discovery is the most important aim of the social science

> "The main message of this book is that the advancement of social theory calls for an analytical approach that systematically seeks to explicate the social mechanisms that generate and explain observed associations between events." (Hedström and Swedberg 1998, p. 1)

> "Sociology seeks to identify social structures harbouring causal mechanisms that generate empirically observable effects" (Brante 2001, p.178)

> "Identifying causal mechanisms is a fundamental goal of social science. Researchers seek to study not only whether one variable affects another but also how such a causal relationship arises." (Imai et al., 2011, p. 765)

Hedström, P. and Swedberg, R. (ed.) (1998). *Social Mechanisms: An Analytical Approach to Social Theory*. Cambridge University Press, Cambridge.
Brante, T. (2001). Consequences of realism for sociological theory-building. *Journal for the Theory of Social Behaviour* 31 (2): 167-194.
Imai, K., et al. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review,* 765-789.

Causal mechanisms in the social sciences

Despite this common idea, however, agreement is lacking concerning the exact definition of a 'causal mechanism'

This agreement is made more difficult by the fact that, although the nature of causal mechanisms in the social sciences has been studied both by social scientists and by philosophers of science, the discussions within these disciplines have often proceeded mostly independently of each other

Some basic ideas, however, are quite similar:

- Process or complex system?
- Multilevel mechanisms

Processes or systems?

- Both in the social sciences and in philosophy, one of the main question concerns the way in which causal mechanisms should be defined. Two different accounts have been developed. According to someone, mechanisms are causal processes producing causal effects; according to other mechanisms are complex systems made of entities organized in such a way to produce an effect

Processes

- Both Salmon and Dowe saw processes as world lines of objects, and causal processes as those processes that transmit conserved quantities (e.g. mass-energy, linear momentum, or charge) after an interaction between two (causal) processes (Salmon 1984, 1997; Dowe 1992)

- In the social sciences, some authors like Bennet and George (1997), claimed that mechanisms are causal processes through which causal or explanatory variables produce causal effects. Rather than considering just physical processes, however, they included in their notion what they call **social processes**, composed of intentions, expectations, information, strategic interaction and so on

Many authors did not find the process approach exhaustive

Bennett, A., & George, A. L. (1997). *Process tracing in case study research* (pp. 17-19). MacArthur Program on Case Studies.
Dowe, P. (1992). 'Wesley Salmon's process theory of causality and the conserved quantity theory'. *Philosophy of Science*, 59(2), 195–216.
Salmon,W. C. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton University Press, Princeton.
Salmon, W. C. (1997). Causality and explanation: a reply to two critiques. *Philosophy of Science*, 64(3), 461–77.

Complex systems

"These two concepts [causal interaction and causal process] […] provide the foundation of our understanding of causality. But what about poverty as a cause of delinquency? Or maternal education as a causal factor for child survival? I am not questioning the plausibility of those causal statements. Indeed they are plausible, but—I ask—is aleatory causality a viable approach to causality in the social sciences? Salmon thinks it is. However, it seems to me that it is not self-evident to 'see' causal processes and interactions in social science scenarios." (Russo 2009, p. 18)

Russo, Federica (2009). *Causality and Causal Modelling in the Social Sciences. Measuring Variations*. Methodos Series. Springer, New York.

Complex systems

For this reason, some authors have developed a mechanistic account more suitable to other sciences, according to which mechanisms are complex systems

> "Mechanisms are entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions." (Machamer et al. 2000, p. 3)

> "A mechanism for a behavior is a complex system that produces that behaviour by the interaction of a number of parts, where the interactions between parts can be characterized by direct, invariant, change-relating generalizations." (Glennan, 2002, p. 344)

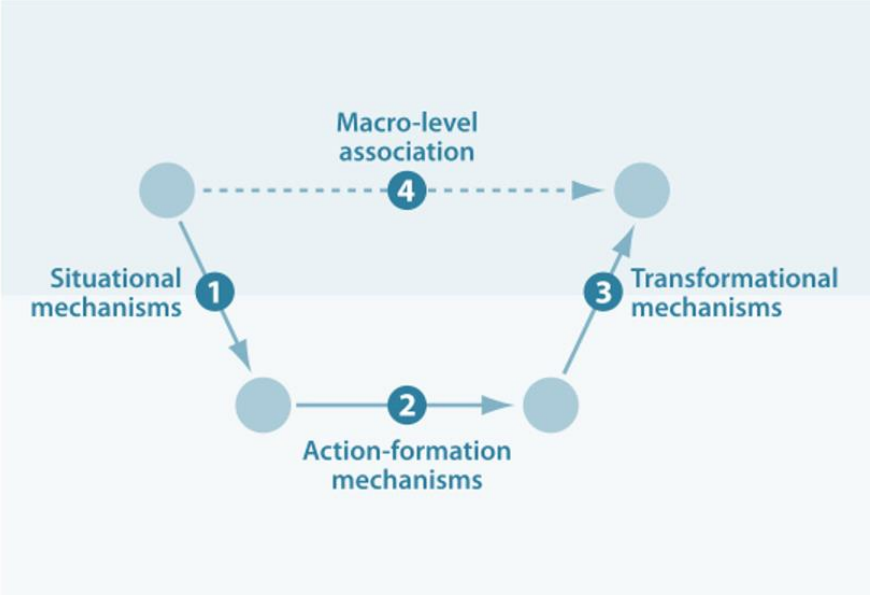
> "A mechanism for a phenomenon consists of entities and activities organized in such a way that they are responsible for the phenomenon." (Illari and Williamson, 2012, p. 120)

Glennan, S. (2002). Rethinking mechanistic explanation. *Philosophy of Science*, 69, S342–S353.
Illari, P. M., & Williamson, J. (2012). What is a mechanism? Thinking about mechanisms across the sciences. *European Journal for Philosophy of Science*, 2(1), 119-135.
Machamer, P., et al. (2000). Thinking about mechanisms. *Philosophy of Science*, 67, 1–25.

Multi-level mechanisms

Many scholars share the idea that causal mechanisms in the social sciences are structured at different levels:

- Macro level: macro and aggregate factors
- Micro level: individual factors



This assumption is shown by the so-called

Coleman Boat (1990)

Coleman JS. (1990). *Foundations of Social Theory*. Cambridge, MA: The Belknap Press

Given the existence of such levels, what is the relationship between them?

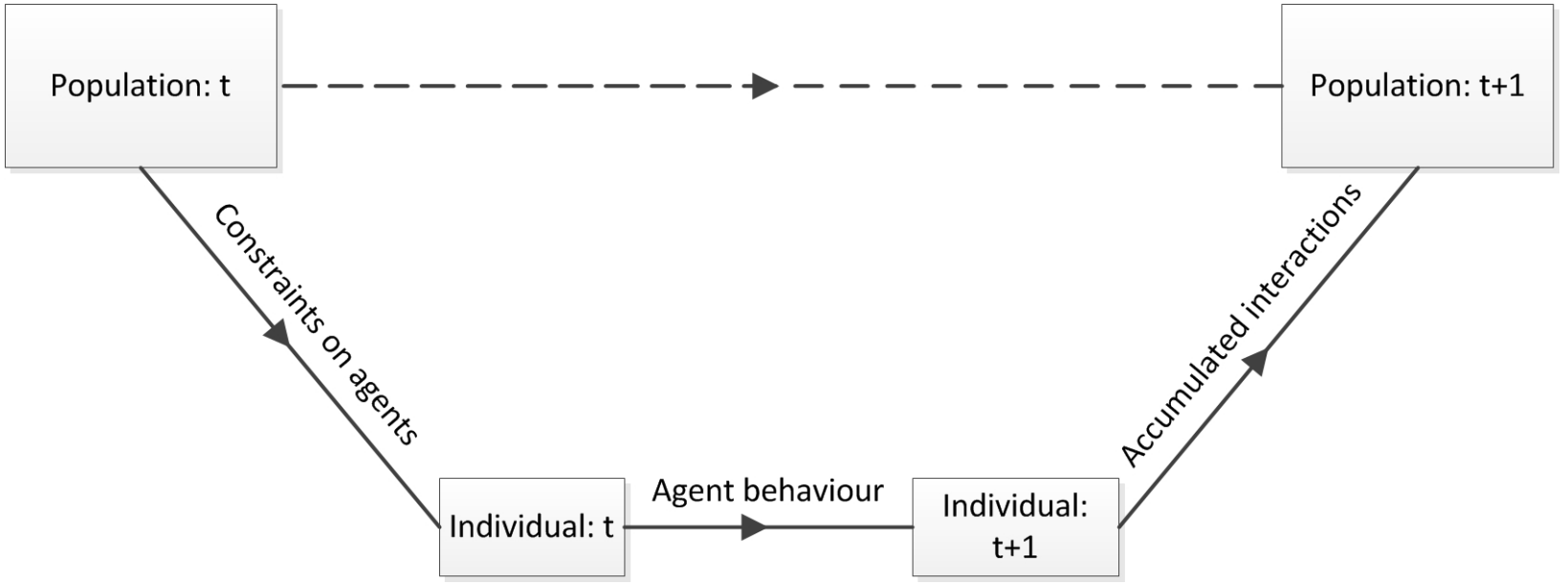
- Macro-level should be explained by considering mechanisms at the micro-level (Coleman 1990, Hedström and Swedberg 1996)

> "There exist no macro-level mechanisms; macro-level entities are linked to one another via combinations of situational, individual action, and tranfsormation mechanisms, i.e., al macro-level change should be conceptualized in terms of three separate transitions (macro-micro, micro-micro, micro-macro)." (Hedström and Swedberg 1996, p. 299)

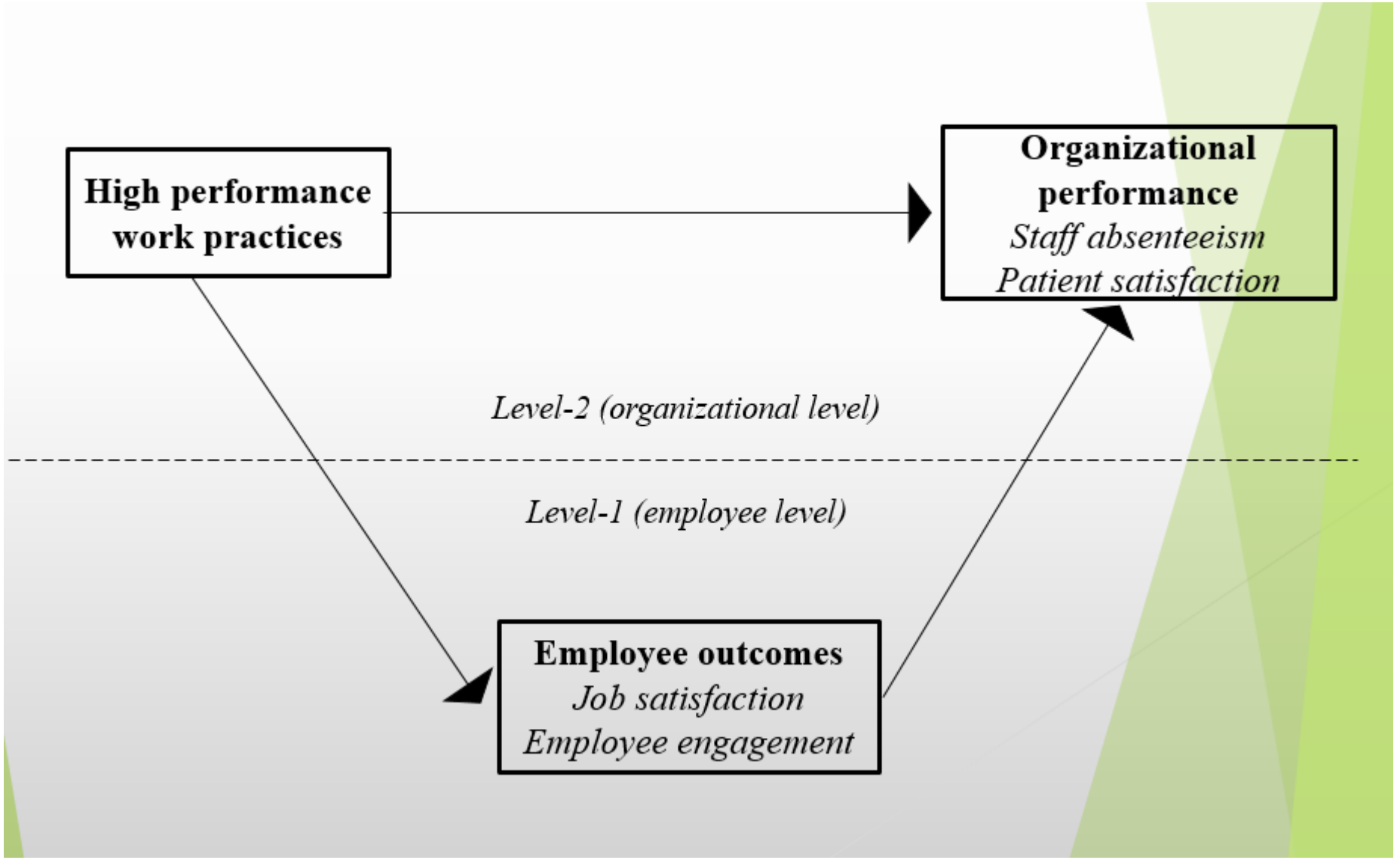
Coleman JS. (1990). *Foundations of Social Theory*. Cambridge, MA: The Belknap Press
Hedström, P., & Swedberg, R. (1996). Social mechanisms. *Acta Sociologica*, *39*(3), 281-308.

The Coleman's boat

- To explain a social phenomenon it would not be enough to follow arrow 4, linking together two phenomena at the macro level: it is necessary to specify the causal mechanisms by which macro properties are related to each other.
- One should identify the situational mechanisms by which social structures constrain individuals' action (arrow 1)
- The action-formation mechanisms, by which individuals choose how to act, should be described (arrow 2)
- Scientists should understand the transformational mechanisms according to which individuals, through their actions and interactions, generate social outcomes (arrow 3)



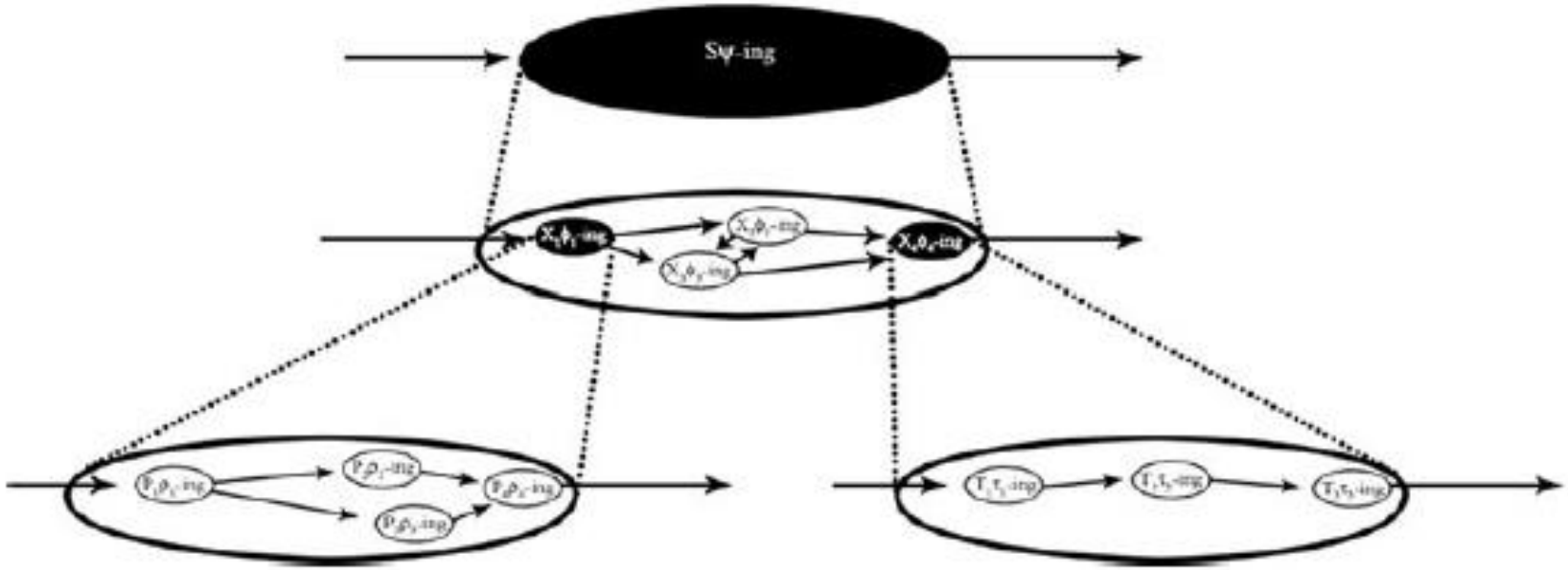
Hassan, S., et al. (2013). Asking the oracle: Introducing forecasting principles into agent-based modelling. *Journal of Artificial Societies and Social Simulation*, *16*(3), 13.


High performance work practices
Organizational performance
Staff absenteeism
Patient satisfaction
Level-2 (organizational level)
Level-1 (employee level)
Employee outcomes
Job satisfaction
Employee engagement

- Macro-level can be explained by considering mechanisms at the macro-level: micro-level factors just constitute the macro-level factors (Vromen 2010, Ylikoski 2014)

> "Individual persons, their causal powers and their interactions are always constitutively involved in the operation of a macro-level mechanism, regardless of the level at which we choose to look at the operation of the mechanism […] but if we look at the causal chain connecting two macro-phenomena at the macro-level, we can in principle see the complete causal chain" (Vromen 2010, pp. 373 and 375)

Vromen, J. (2010). Micro-foundations in strategic management: squaring Coleman's diagram. *Erkenntnis*, 73(3), 365-383.
Ylikoski, P. (2014). Rethinking Micro-Macro Relations. In *Rethinking the Individualism-Holism Debate* (pp. 117-135). Springer International Publishing.

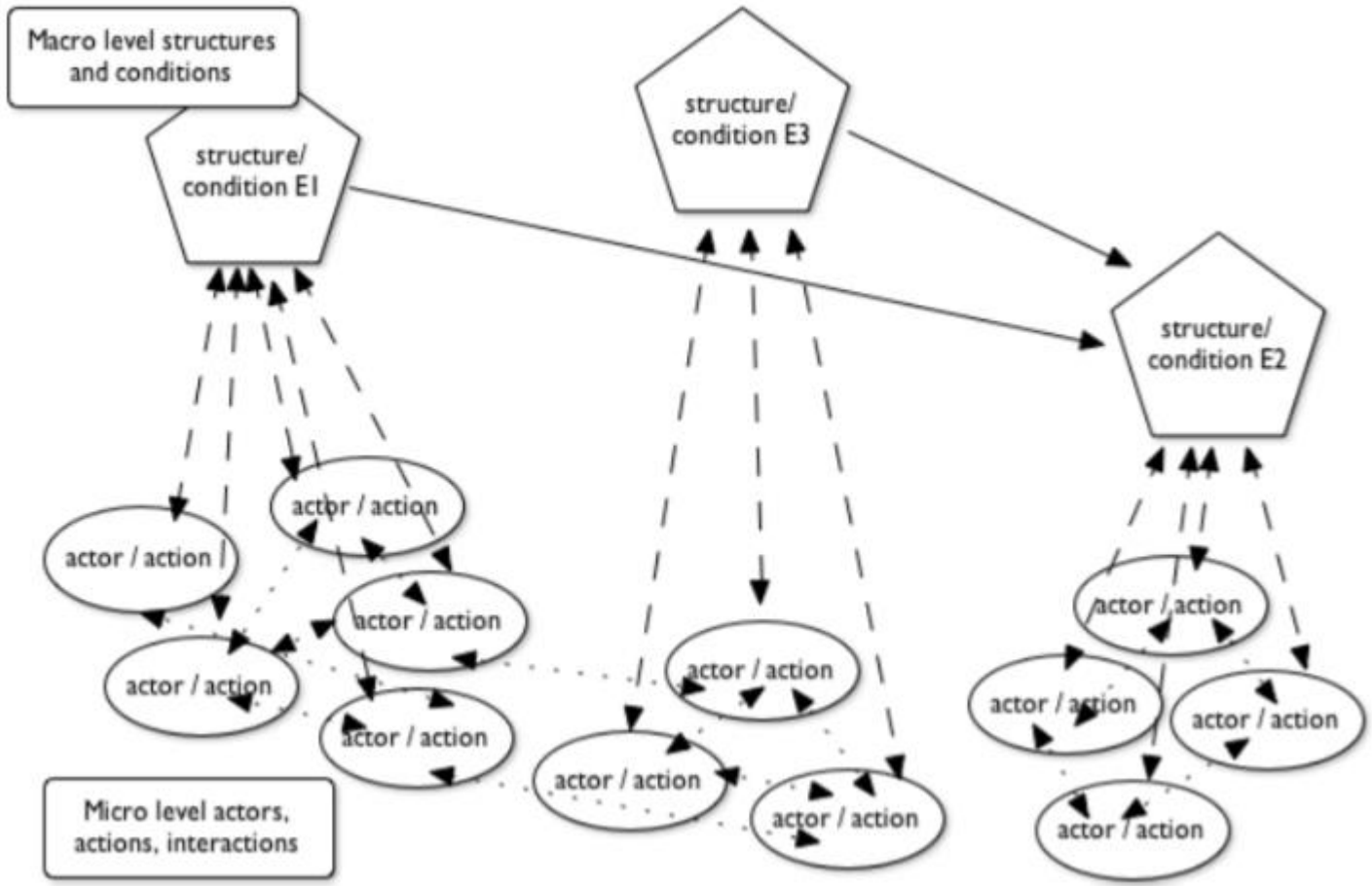


Vromen, J. (2010). Micro-foundations in strategic management: squaring Coleman's diagram. *Erkenntnis*, 73(3), p. 377

- Meso-level properties (such as relational and structural ones) may explain social phenomena (Little 2012)

> "We can settle on a level of meso or macro explanation without dropping down to the level of the actor. We need to be confident there are microfoundations, and the meso properties need to be causally robust. But if this is satisfied, we do not need to extend the explanation down to the actors." (Little 2012, p. 146)

Little, D. (2012). Explanatory autonomy and Coleman's boat. *THEORIA. Revista de Teoría, Historia y Fundamentos de la Ciencia*, 27(2), 137-151.



Little, D. (2012). Explanatory autonomy and Coleman's boat. *THEORIA. Revista de Teoría, Historia y Fundamentos de la Ciencia*, 27(2), 147.

- What social scientists are trying to perform is often called process tracing
- In the last decades, PT has become an "umbrella term" widely discussed both in the social sciences and philosophy of the social sciences
- At least 20 main definitions of PT can be identified according to the notion of causal mechanism, the discipline, the purpose and the methods involved (Trampusch and Palier 2016)
- Although these differences, scholars tend to agree that PT is performed to detect causal mechanisms

Trampusch, C., and Palier, B. (2016). "Between X and Y: how process tracing contributes to opening the black box of causality". *New Political Economy*, 21(5), 437-454.

Beach and Pedersen's classification

Recently, Beach and Pedersen (2013) advanced the debate on PT by identifying three variants of this methodology

1. Theory-testing PT: evaluating whether evidence shows that a hypothesized causal mechanism is present in a particular case and that it operates as theorized.

To test the hypothesis, scientists have to translate a theoretical expectation into a specific prediction of what observable manifestations of the mechanism should be present if the hypothesized mechanism worked. Then, empirical evidence can be collected and, according to what is observed, social scientists should be able to infer whether their confidence in the hypothesis should be updated and whether the mechanism functioned as predicted

Beach, D., and Pedersen, R. B. (2013). *Process-tracing methods: Foundations and guidelines*. University of Michigan Press.

2. Theory-building PT: building a theory about a causal mechanism between X and Y that can be generalized to a population of a given phenomenon.

To build a theory, scientists investigate the empirical material available to infer whether it reflects an underlying causal mechanism at work. Once the theory is built, it has to be tested.

3. Explaining-outcome PT: crafting a sufficient mechanistic explanation of a particular outcome.

Sufficiency can be confirmed when it can be validated that there are no important aspects of the outcome for which the explanation does not account

Beach, D., and Pedersen, R. B. (2013). *Process-tracing methods: Foundations and guidelines*. University of Michigan Press.

Outline

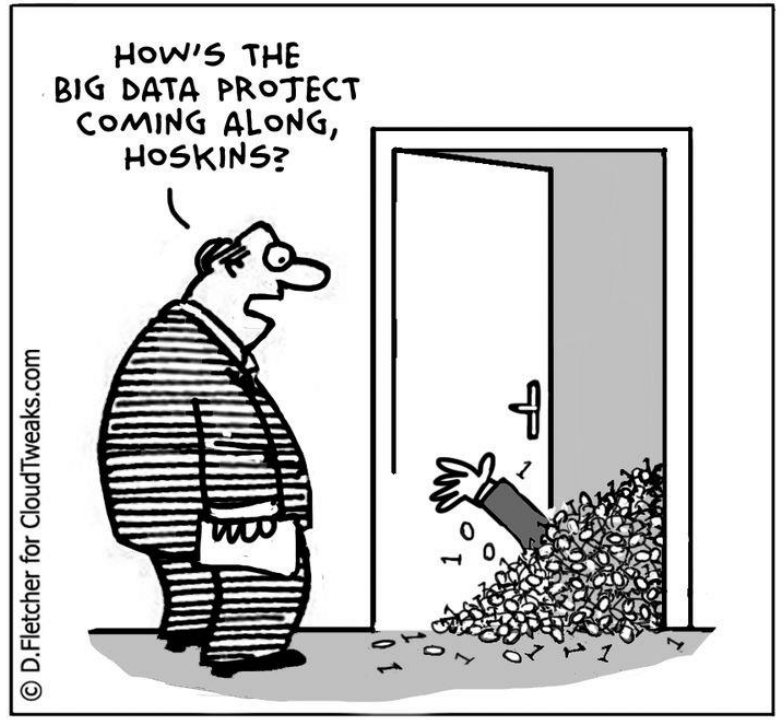
- Introduction
- Causal mechanisms in the social sciences
- **Big data for mechanistic explanations**
- Conclusion

Big data for mechanistic explanations

Can big data enhance social scientists' ability to discover causal mechanisms and to explain given phenomena by appealing to mechanisms ?

The new data sources, together with the large amount of data now available, can offer new opportunities but also new challenges in seeking causal mechanisms:

- New insights on the nature of social mechanisms
- New forms of data collection
- Internal and external validity problems and solutions



New insights on the nature of social mechanisms

Process or complex system?

- Big data studies can cast light on the presence of organized mechanistic entities, but can also help to recognize both the presence of processes and complex systems

- The studies on the causal relationships between exposure and disease show that there are processes leading from repeated exposures to disease formation, but that such processes interact with some complex-systems mechanisms whose task is to maintain the integrity of the body. Moreover, when such complex systems do not manage to maintain this integrity, further processes as well as further complex systems can be involved, and this combination can finally lead to disease (Russo and Williamson 2012).

Russo F, Williamson J (2012) Envirogenomarkers. The interplay between difference-making and mechanisms. *Med Stud* 3:249–262.

New insights on the nature of social mechanisms

Process or complex system?

- Instances of exposure can be seen as world lines that carry and exchange conserved quantities
- Complex systems mechanisms for maintaining the integrity of the body (mechanisms for cell metabolism, cell repair, cell death…)
- Unstable and irregular processes can be activated when complex systems fail
- Further complex systems may interact with such processes

Russo F, Williamson J (2012) Envirogenomarkers. The interplay between difference-making and mechanisms. *Med Stud* 3:249–262.

New insights on the nature of social mechanisms

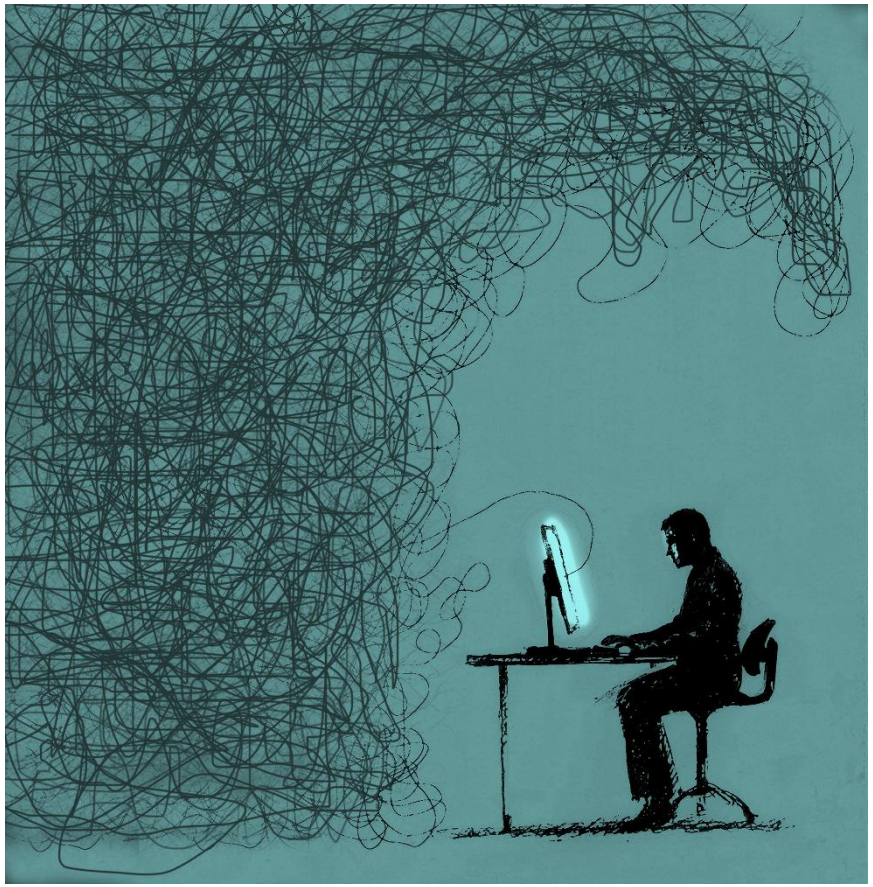
Process or complex system?

The big data project EnviroGenomarkers, helped to recognize that, in order to capture the mechanisms operating in cases of disease formation, one needs to appeal to both Salmon-Dowe processes and complex-system mechanisms

New insights on the nature of social mechanisms

Multilevel mechanisms

- Web-based experiments as well as large cohort studies appear particularly useful to collect data at different levels and to understand the relationship between macro phenomena and individual actions, and interactions that brought them about.

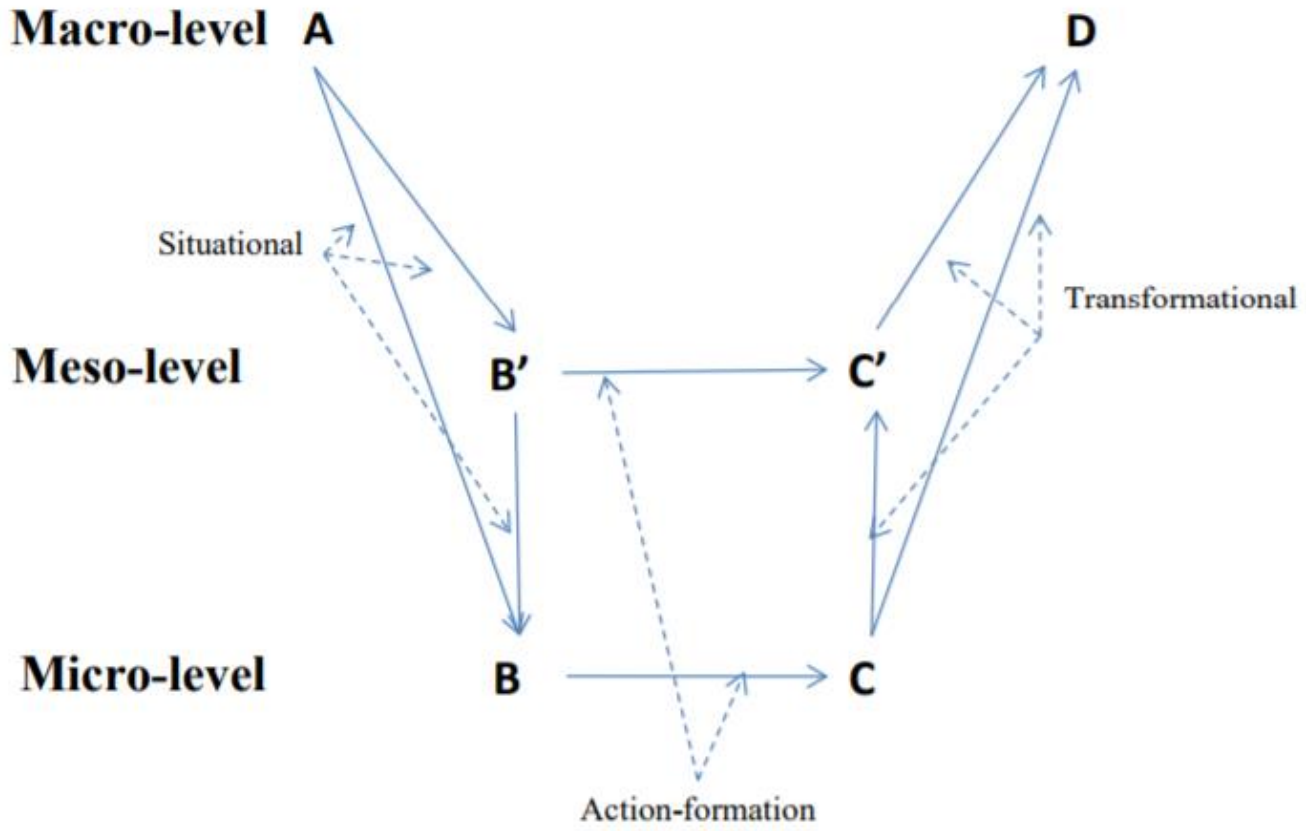


Salganik et al. (2006) studied the role of social influence in individual behaviour and macro phenomena through a web-base experiment. More than 14,000 participants listened to previously unknown songs, rated them, and freely downloaded them if they desired to do that. Subjects were randomly assigned to different groups. Individuals in only some groups were informed about how many times others in their group had downloaded each song.

They discovered that individuals' music preferences were altered when they were exposed to information about the preferences of others. Furthermore, they found that the extent of social influence had important consequences for the collective outcomes that emerged. The greater the social influence, the more unequal and unpredictable the collective outcomes became

Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762), 854-856

New insights on the nature of social mechanisms



Kim, P. H., Wennberg, K., & Croidieu, G. (2016). Hidden in Plain Sight: Untapped Riches of Meso-Level Entrepreneurship Mechanisms.

Data collection

Observational studies: field study, case studies

"Case studies are often useful for the purpose of elucidating causal mechanism."(Gerring 2013, p. 65)

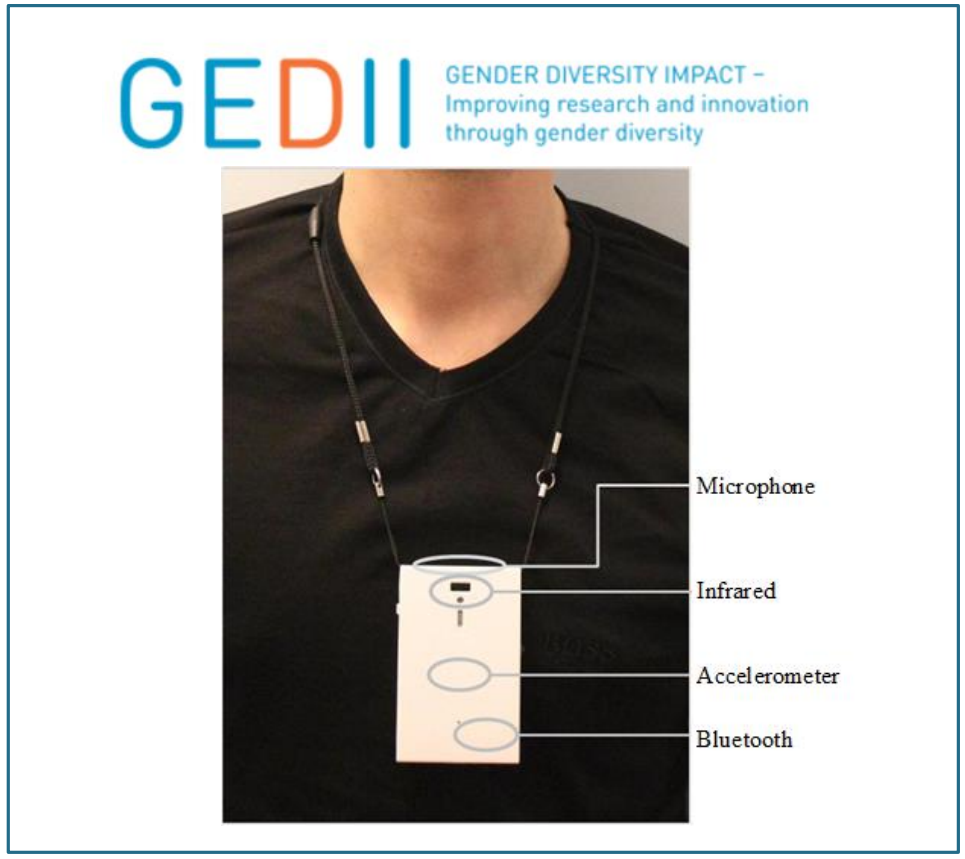
- New data sources such as satellites, smartphone based systems and wearable sensor devices help to develop a comprehensive image of the phenomenon under study

- In some cases, these forms of data can help to increase the internal validity of the study (measurement of physical activity/inhalation in EXPOsOMICS are more accurate than self-reported data)

Gerring, J. (2007). The case study: what it is and what it does. In *The Oxford handbook of comparative politics.*

Data collection

- In other cases, the truthfulness of data can be questioned. For instance, can people wearing sensor devices experience the Hawthorne effect (i.e. people modify an aspect of their behaviour in response to their awareness of being observed)?



Data collection

Experimental studies

"Since the quality of field data is usually too poor to answer specific questions concerning causal mechanisms and processes, physiologists and social scientists find it useful to try and answer them first in the laboratory" (Guala 2002, p. 11)

- A web-based experiment is an experiment that is conducted over the Internet. The Internet can be a medium through which scientists target large and diverse sample with reduced administrative and financial costs; or a field of social science research in its own right.

Guala, F. (2002). Models, simulations, and experiments. In *Model-based reasoning* (pp. 59-74). Springer US.
Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311(5762), 854-856

Data collection

Experimental studies

Pros

- Ease of access to a large number of demographically and culturally diverse participants
- Access to rare and specific participant populations (ecstasy consumer, drug dealers…)
- Reduction of experimenter effects

Cons

- Subjects taking the experiment less seriously and behaving with less risk-aversion
- Researcher cannot control for significant distractions occurring during the course of the experiment
- Dropout

Data collection

Large-scale studies

- The growing amount of data enables scientists to work without model or representative populations: scientists have data from so many individuals that they can be sure to have the average features of the population under study

- This is the case of the EnviroGenomarkers and the EXPOsOMICS projects, based on large datasets involving several countries

External and Internal Validity

External validity

- No need for model populations, therefore no need for extrapolation (EXPOsOMICS)

- When a mechanistic hypothesis is moved from one population to another, in-depth data can help to compare the two social environments

External and Internal Validity

Internal validity

Is big data objective?

- Data automatically produced are not raw!
- Nevertheless, sometimes this is better than having just self-reported data
- Data on human behaviours could still be biased because of some experiment effects
- Data cleaning

Outline

- Introduction
- Causal mechanisms in the social sciences
- Big data for mechanistic explanations
- **Conclusion**

Conclusion

- Big data can be used both to obtain correlational evidence and to investigate causal mechanisms
- Big data can show the presence of causal processes, complex systems or both
- Fine grained data casts light on the presence of micro-meso-macro levels interacting with each other
- New data sources offer the opportunity to gather mechanistic evidence in different ways by using experiments, case studies and large-scale studies
- External validity can be increased thanks to the large amount of data about the population under study
- In some cases, internal validity can be increased (measurement data), in other cases, it can be decreased (experiment effects)

Thank you for your attention!