Scientific understanding through big data:

From ignorance to insights to understanding

María del Rosario Martínez-Ordaz
martinezordazm@gmail.com

Instituto de Investigaciones Filosóficas - UNAM
This work was supported by UNAM's PAPIIT IN406225 and Proyect CBF2023-2024-55.

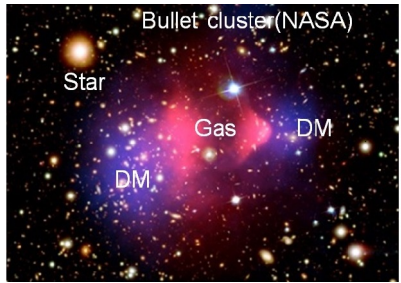
June 17, 2025

- Science has moved from being computationally *aided* to being data-*driven* (Cf. Zhou et al., 2019, p. 1018).

  Implementation of **Big Data** and tools like Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL), neural networks, etc.

- Expectation: **Big data** $\rightarrow$ (increment in the amount of data $\rightarrow$ increment in the scope/depth of sc. knowledge and scientific understanding)

- Expectation: **Big data** $\rightarrow$ (increment in the amount of data $\rightarrow$ increment in the scope/depth of sc. knowledge and scientific understanding)



however...

- An increment in the scope/depth of sc. knowledge and scientific understanding would depend on the agents' capability to assess the reliability of the source of the new information...

- But... with BD and computational tools high degrees of epistemic opacity can prevent this from happening...

Aim(s)

Identify the circumstances under which agents can overcome their ignorance and achieve understanding when using big data in their disciplines.

Main thesis can be summarized by the following two observations:

1. The ignorance associated present in cases of BD-implementation is *ignorance of theoretical structure*.
2. The (partial) overcoming of this ignorance is associated with the reliable use of specific insights that, later on, will lead to further epistemic achievements such as (modal) understanding.

+ **Question the epistemic consequences of such implementation.**

Scientific Understanding

"consist of knowledge about relations of dependence. When one understands something, one can make all kinds of correct inferences about it" (Ylikoski, 2013: 100).

If one legitimately understands an X-phenomenon, one would be able to 'explain' X-phenomenon in many different ways... (under which circumstances X is the case, what does it mean for X to be the case, etc).

If one legitimately understands a Y-method, one would be able to tell how this method behaves in different circumstances, what warrants its reliability in each of the applications, etc...

Objections to understanding through BD: Explanation, factivity, relations of dependence

Epistemic opacity

- *O* is *epistemically opaque* to an agent, in a particular context, if the agent ignores all the features of O that are relevant to a specific task within the context.
  * Products
  * Procedures

Much of the success in big data practices depends on computational processes that cannot be fully scrutinized, examined, and justified by human agents (see Humphreys 2009).

- *Essential Epistemic Opacity*: “A process is essentially epistemically opaque to X if and only if it is impossible, given the nature of X, for X to have access to and be able to survey all of the relevant elements of the justification” (Duran and Formanek, 2018, p. 651).

Heavy technological implementation

Heavy technological implementation $\rightarrow$ epistemic opacity

Much of the success heavy technological implementation-scenarios depends on computational processes that cannot be fully scrutinized, examined and justified by human agents (see Humphreys 2009).

Procedural EO

(i) Algorithmic functionally opacity, (ii) Procedural opacity, (iii) Run opacity (Cf. Creel 2020)

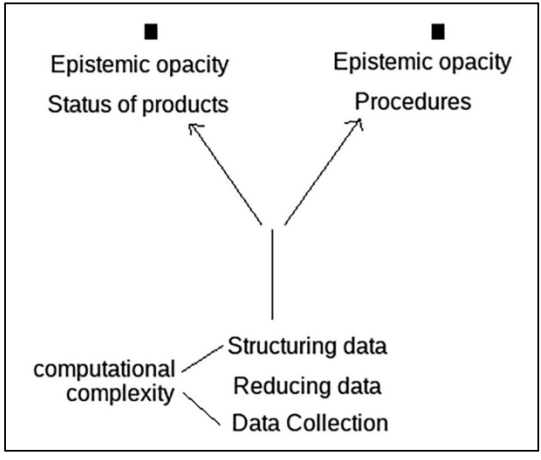
$\therefore$ No procedural K, no explanatory K
$\therefore$ No clear J for trusting the procedures/outcomes

EO regarding the status of products

Simulations, experiments, abstractions, fictions, ...

1. Correlations $\rightarrow$ No-causal explanations
2. Understanding requires causal exp.

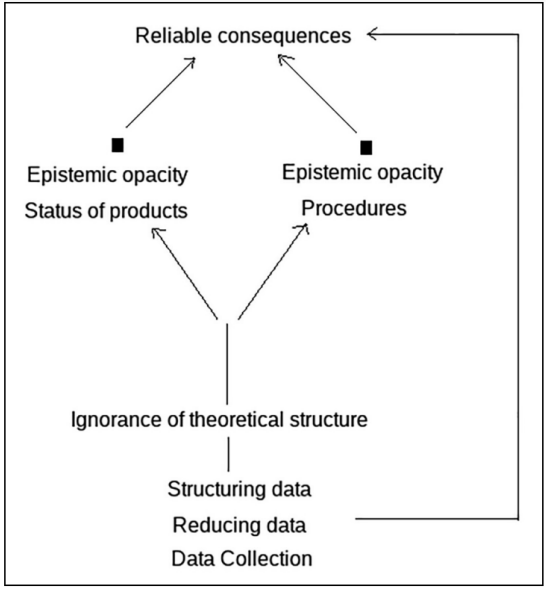
$\therefore$ No understanding


Epistemic opacity
Status of products
Epistemic opacity
Procedures
computational complexity
Structuring data
Reducing data
Data Collection

Ignorance

Epistemic opacity $\neq$ ignorance
EO indicates a deeper problem than just ignoring procedures/products...

When scientists “cannot provide inferential explanations about why an output obtains, they are not ignoring *only* a specific recipe, they are ignorant of how the bits of data relate to one another —at least, inferentially; and this is indicative of *ignorance of theoretical structure*”(Martinez-Ordaz, 2022, p. 127).


Reliable consequences
Epistemic opacity
Status of products
Epistemic opacity
Procedures
Ignorance of theoretical structure
Structuring data
Reducing data
Data Collection

What is the epistemic status of such **reliable consequences of BD.**?

we don’t know/ we are not so sure

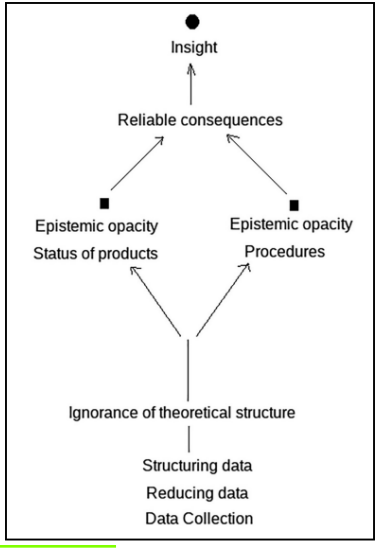
Why? Because we are not sure about how they were obtained... even though we are sure they are reliable...

Insight

The word “insight” is commonly used to indicate either an epistemic product or a belief-formation process.

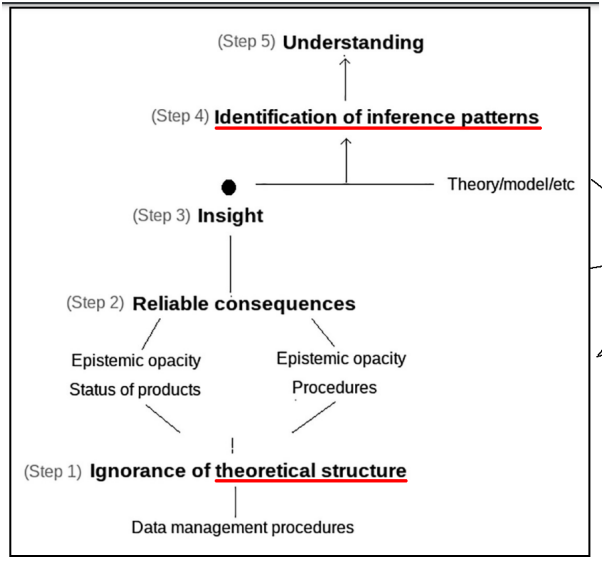
- Process: Insight consists of a sudden realization or discovery of a solution path that allows one to solve a problem.
  1. Ampliative reasoning for problem-solving, **the solution** obtained via **an unclear path**
- Epistemic product: Type of commitment (epistemic) agents have toward the solution that is produced through a given inferential path.
  1. Beliefs that are formed through an unclear or unrigorous process as a response to a given problem,
  2. they appear to be strong and robust enough to guide our acceptance or rejection of other beliefs.
  3. Additionally, insights are an indication of our grasping of a specific problem, object, domain, or phenomenon.

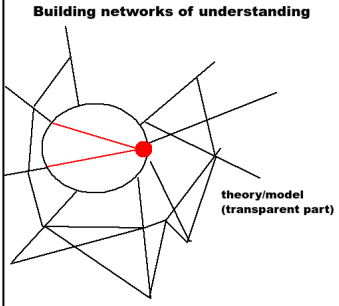
Reliable consequences and insights



what is the role of insights in moving to scientific understanding?

Reliable consequences and insights





While the identification of inference paths might lead scientists to a cohesive picture of a particular domain using the outputs of big data implementation, this does not mean that this particular picture is in any relevant sense connected to the target phenomena.

Modal understanding. It is often said that someone has a modal understanding of X when that person knows how to navigate the possibility space associated with X (Cf. Le Bihan, 2017, p. 112).

An example from science

Predicted 21 phosphine ligands using unsupervised ML with only five experimental data, 2021, Schoenebeck Research Group (Germany)[1]

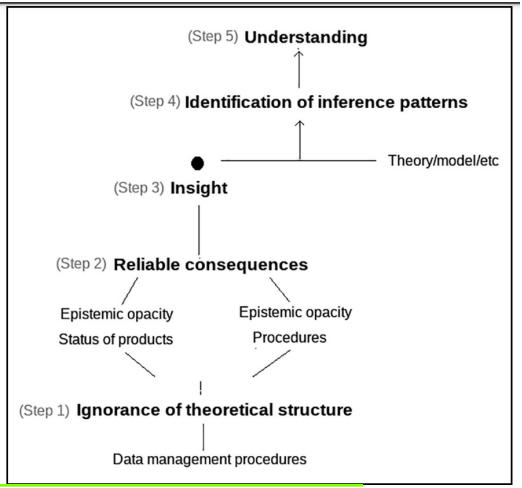
1. Problem: getting a new catalyst with only five experimental data points

2. The solution to a given problem.
   - formed through an unclear/unrigorous process in response to a specific problem; UNSUPERVISED ML
   - strong and robust enough to guide NEW RESEARCH

[1] Hueffel, J.A., T. Sperger, I. Funes-Ardoiz, J.S. Ward, K. Rissanen, and F. Schoenebeck (2021): "Accelerated dinuclear palladium catalyst identification through unsupervised machine learning". Science 374, 1134-1140

(Semi)Final remarks

Sketched an epistemological landscape of the detection and overcoming of highs degrees of ignorance to gain knowledge and understanding in BD-contexts



An example from science

"Catalysts are used in many industrial processes. Traditionally, the optimal design of catalysts has been empirical or has mostly depended on experimentation. (...) **With the rapidly increasing amount of available experimental and computational data, as well as the development of catalysis informatics, catalyst structure and activity relationships can now be well described using ML models, which are very useful for catalyst development** (...) It has been shown that compared with traditional computational and experimental trial-and-error approaches, ML methods possess great potential for accelerating the discovery of high-performance heterogeneous catalysts." (Cf. Zhou *et al*. 2019: 1023).

An example from science

In this context, ML has become key for the development of the so-called *data-intensive materials design* (or *inverse design*) protocol, in which the desired function of a material is specified beforehand, and then candidates are extracted from a database (which can be either computational or experimental).

An example from science

"AI is a rapidly evolving field that involves various domains, such as reasoning, knowledge representation, and machine learning (ML). Machine learning has been widely implemented for numerous drug discovery applications pertaining to large data sets. It uses various algorithms and techniques to recognize templates and patterns within the given data set (...)ML methods have been classified under two broad subcategories, supervised learning and unsupervised learning methods." (Tripathi et al. 2021: 1440).

An example from science

Predicted 21 phosphine ligands[2] **using unsupervised ML with only five experimental data**, 2021, Schoenebeck Research Group (Germany)[3]

1. Problem: getting a new catalyst with only five experimental data points

2. The solution to a given problem.
   - formed through an unclear/unrigorous process in response to a specific problem; UNSUPERVISED ML
   - strong and robust enough to guide NEW RESEARCH

[2] Los *ligandos de fosfina* son moléculas orgánicas que contienen fósforo y son ampliamente utilizados en química organometálica y catálisis debido a su versatilidad y capacidad para estabilizar y activar centros metálicos.

[3] Hueffel, J.A., T. Sperger, I. Funes-Ardoiz, J.S. Ward, K. Rissanen, and F. Schoenebeck (2021): "Accelerated dinuclear palladium catalyst identification through unsupervised machine learning". Science 374, 1134-1140

An example from science

Predicted 21 phosphine ligands[2] **using unsupervised ML with only five experimental data**, 2021, Schoenebeck Research Group (Germany)[3]

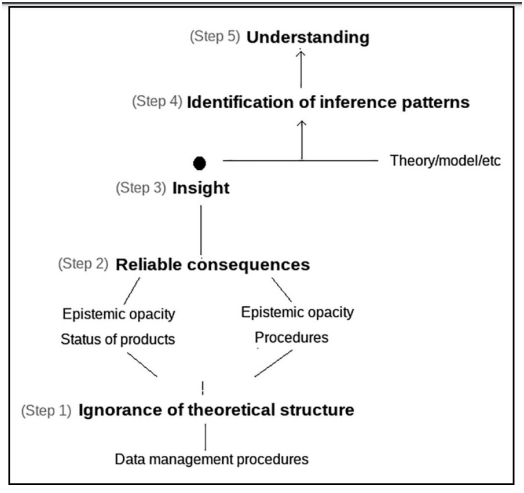
1. Problem: getting a new catalyst with only five experimental data points

2. The solution to a given problem.
   - formed through an unclear/unrigorous process in response to a specific problem; UNSUPERVISED ML
   - strong and robust enough to guide NEW RESEARCH

Insight?
Understanding?

[2] Los *ligandos de fosfina* son moléculas orgánicas que contienen fósforo y son ampliamente utilizados en química organometálica y catálisis debido a su versatilidad y capacidad para estabilizar y activar centros metálicos.

[3] Hueffel, J.A., T. Sperger, I. Funes-Ardoiz, J.S. Ward, K. Rissanen, and F. Schoenebeck (2021): "Accelerated dinuclear palladium catalyst identification through unsupervised machine learning". Science 374, 1134-1140


(Step 5) Understanding
(Step 4) Identification of inference patterns
Theory/model/etc
(Step 3) Insight
(Step 2) Reliable consequences
Epistemic opacity
Status of products
Epistemic opacity
Procedures
(Step 1) Ignorance of theoretical structure
Data management procedures

Agents have slowly adopted new types of insight-formation strategies/mechanisms that privilege novelty, computational efficiency, and solution adequacy –over the intuitiveness of the solution.

Can the same happen in less-scientific contexts?

Stockfish and chess-learning

Before engines like Stockfish, chess players primarily learned strategies through books and games played by masters, often absorbing the "human" intuitions of pattern recognition, sacrifices, and long-term planning from past champions.

Stockfish and chess-learning

Stockfish and other chess engines prioritize computational efficiency and often uncover moves or strategies that defy traditional human logic. Stockfish can evaluate millions of positions in seconds, quickly finding moves that maximize material or positional advantage, often choosing lines that would seem risky or unintuitive to a human player.

This has changed how newer players approach the game, even at the beginner level. Instead of following established patterns or "rules of thumb," they might adopt more aggressive, counterintuitive moves that engines recommend, or develop a deep understanding of complex, non-human positions (like sacrifices that only pay off 20 moves later) that engines evaluate as optimal.

Stockfish and chess-learning

Stockfish and other chess engines prioritize computational efficiency and often uncover moves or strategies that defy traditional human logic. Stockfish can evaluate millions of positions in seconds, quickly finding moves that maximize material or positional advantage, often choosing lines that would seem risky or unintuitive to a human player.

This has changed how newer players approach the game, even at the beginner level. Instead of following established patterns or "rules of thumb," they might adopt more aggressive, counterintuitive moves that engines recommend, or develop a deep understanding of complex, non-human positions (like sacrifices that only pay off 20 moves later) that engines evaluate as optimal.

Insight?
Understanding?

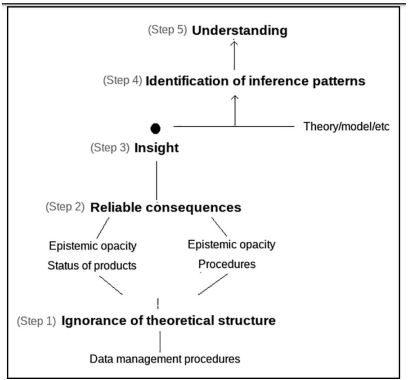
Epistemology, broadly speaking, tends to treat knowledge as a universal phenomenon rather than something tethered to any specific discipline or method; however, the increment in the use and dependency of new technologies (such as Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL), among others) has significantly modified how knowledge is pursued and achieved –not only within the realm of scientific inquiry but also in everyday contexts.

With the emergence of new technologies and our constant interaction with them, questions should arise regarding whether our philosophical characterization of some epistemic and cognitive phenomena should remain applicable.

Ex. The source/role of epistemic feelings in insight and understanding is very different in pre-AI cases than in AI-aided scenarios

Final remarks

Sketched an epistemological landscape of the detection and overcoming of highs degrees of ignorance to gain knowledge and understanding in BD-contexts



Due to the interaction between humans and AI, an important part of our philosophical understanding of some cognitive and epistemic phenomena urges to be modified.

Thanks!

Scientific understanding through big data:

From ignorance to insights to understanding

María del Rosario Martínez-Ordaz
martinezordazm@gmail.com

Instituto de Investigaciones Filosóficas - UNAM
This work was supported by UNAM's PAPIIT IN406225 and Proyect CBF2023-2024-55.

June 17, 2025