Capturing and Preserving the AI process as paradata for accountability and audit-trail purposes

Scott Cameron

University of British Columbia, Canada

InterPARES research assistant

Presented at InterPARES Abu Dhabi symposium, February 19th, 2023

What is paradata?

"information about the procedure(s) and tools used to create and process information resources, along with information about the persons carrying out those procedures."

(InterPARES, qtd in Davet et al., "Archivist in the Machine", 2023)

Paradata precedents

The use of paradata to monitor and manage survey data collection

Frauke Kreuter*, Mick Couper†, Lars Lyberg‡

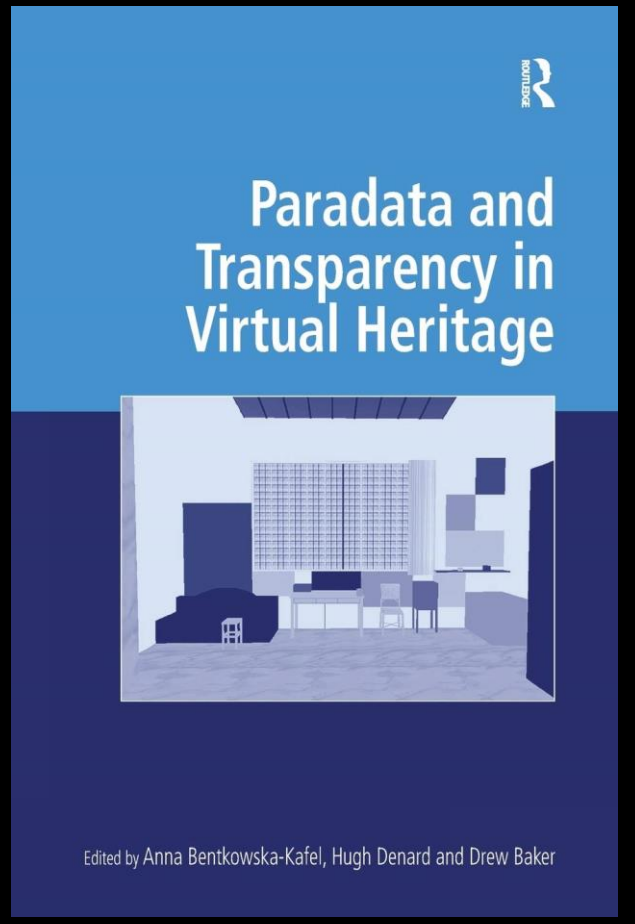
Abstract

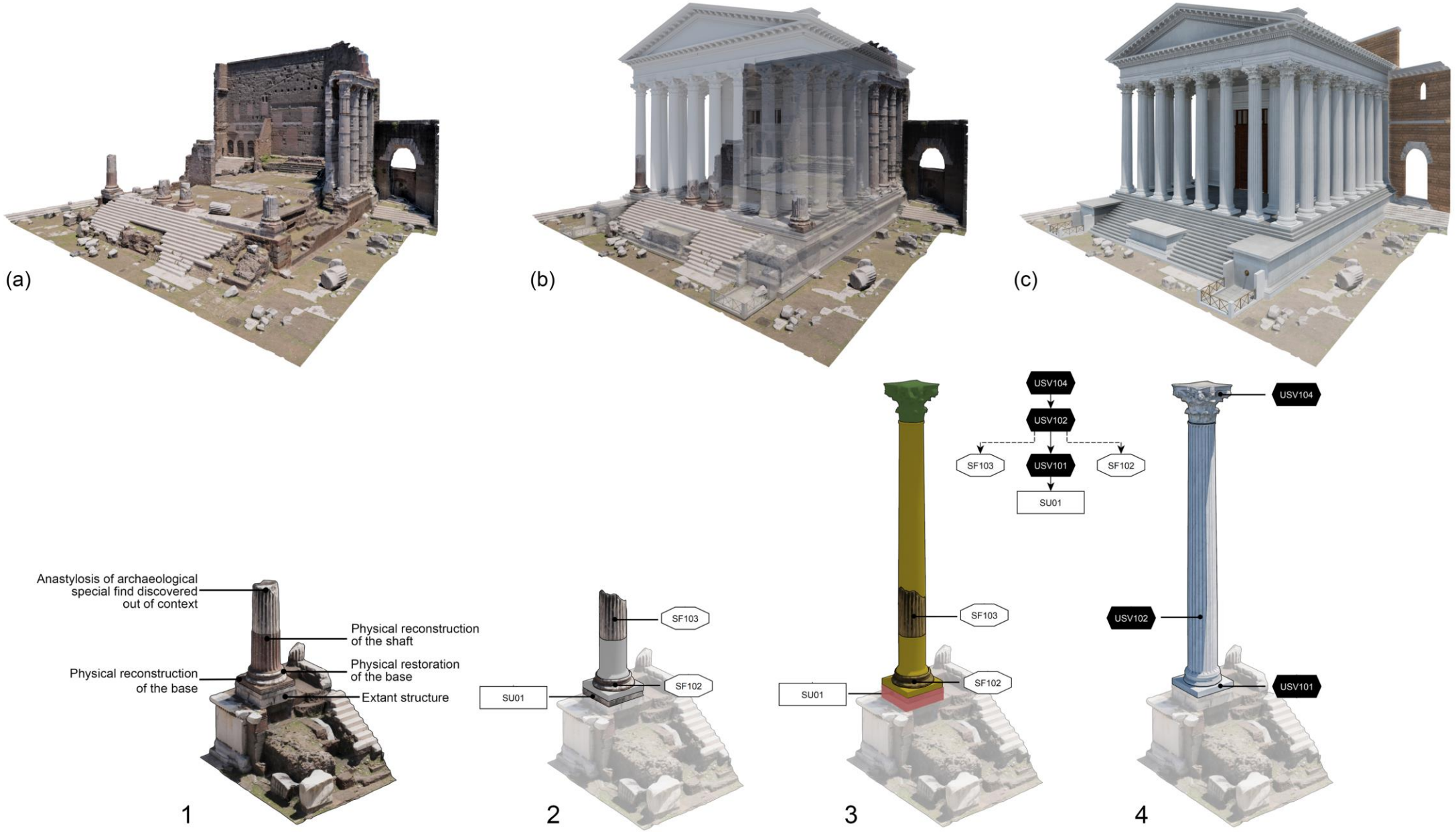
Paradata are automatic data collected about the survey data collection process captured during computer assisted data collection, and include call records, interviewer observations, time stamps, keystroke data, travel and expense information, and other data captured during the process. Increasingly such data are being used in real time to both monitor and manage large scale data collection processes. In this paper we use a statistical process control perspective to describe how such data can be used to monitor the survey process. Process control charts and statistical models can be used to identify areas of concern during data collection, and can lead to further investigation of the problem and (if necessary) intervention. We describe the data and analyses that are available and present several case studies of paradata use in different types of surveys and organizations.

Key Words: Paradata, Process Control, Survey Management

1. Introduction

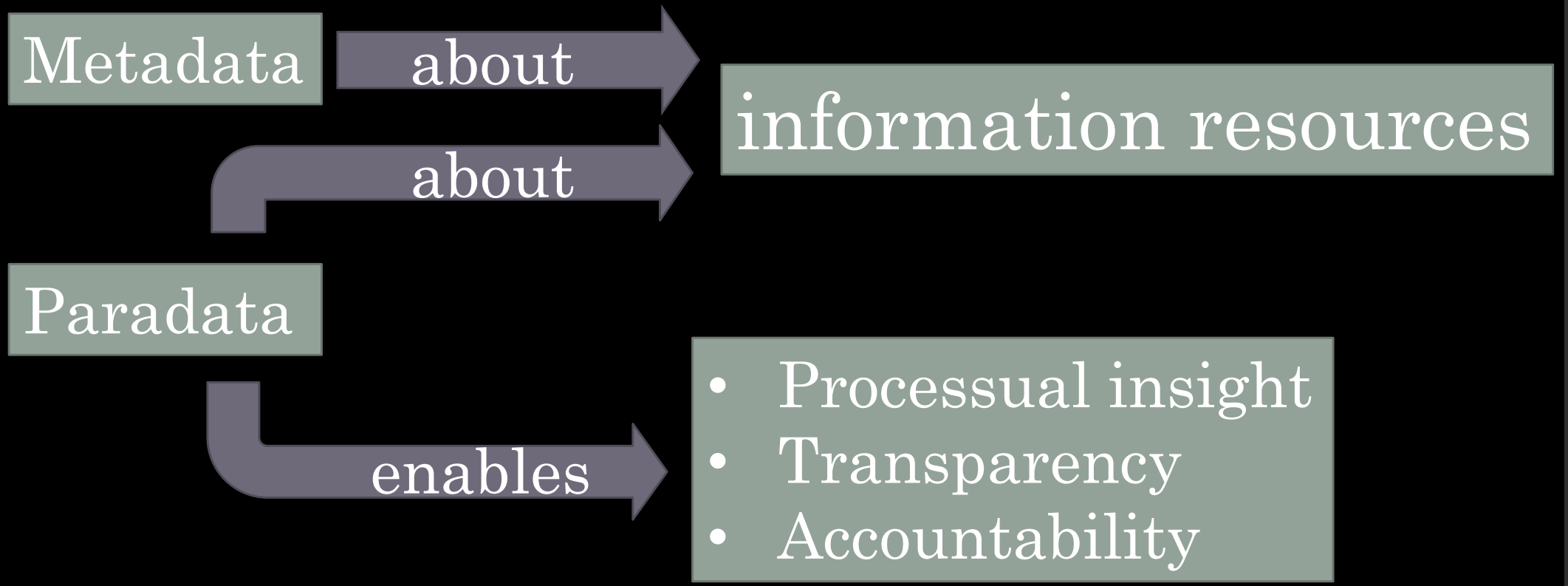
Researchers all around the world nowadays use computer assisted methods in one form or another to collect survey data. Such computer assistance is obvious in web surveys, but is equally present in telephone surveys supported by automated call schedulers or mail surveys that take advantage of logs provided by postal services. All of these systems produce auxiliary data about the survey process as by-products. While discussing the use of keystroke data, Couper (1998) originally coined the term paradata as a general notion of such by-product process data produced by a computer-assisted data collection system. Since then survey methodologists have broadened the paradata concept to other aspects of the survey process and other





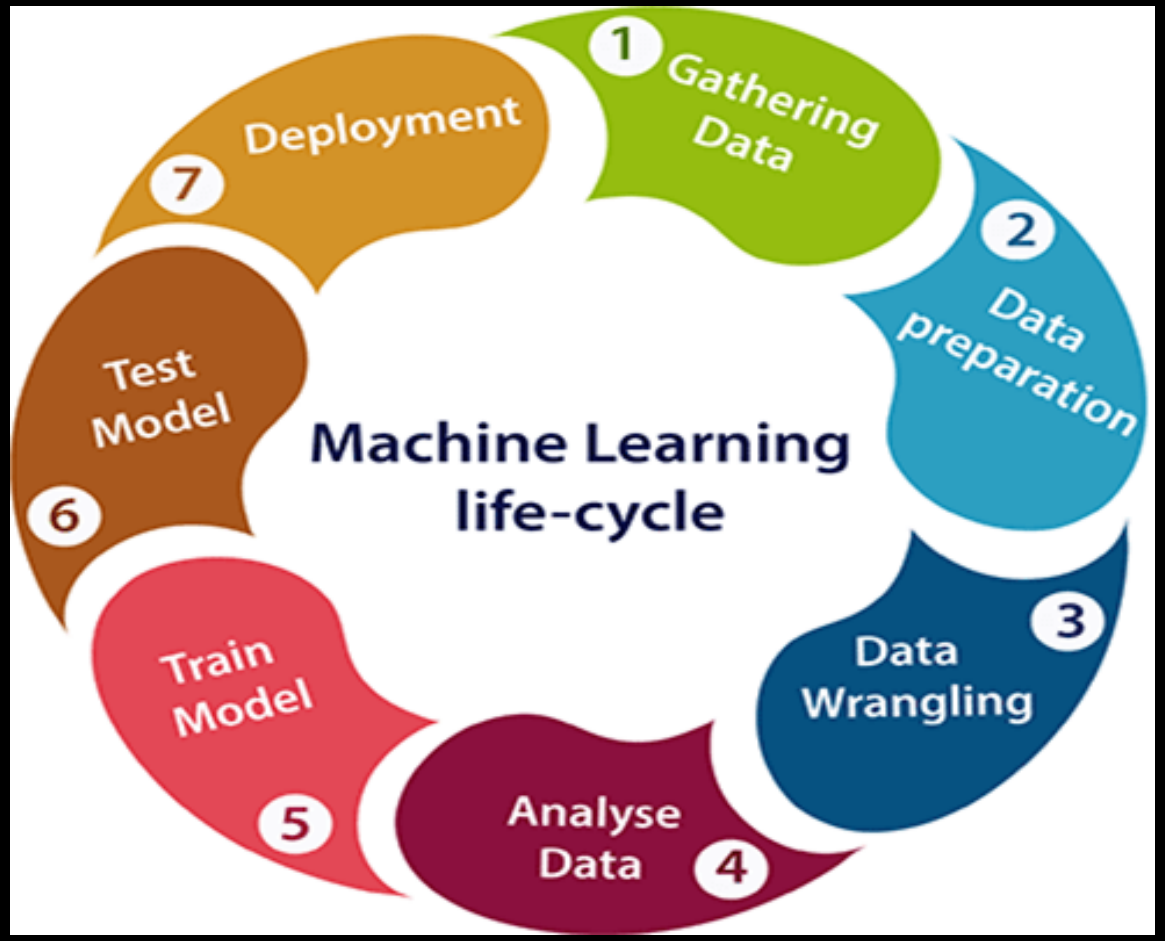
Demetrescu, Emanuel, and Daniele Ferdan.. (2021). From Field Archaeology to Virtual Reconstruction: A Five Steps Method Using the Extended Matrix, *Applied Sciences* 11(11), 5206. https://doi.org/10.3390/app11115206

Paradata distinguished by relationship + purpose



Defining artificial intelligence

- The capacity of computers or other machines to exhibit or simulate intelligent behaviour; the field of study concerned with this. Abbreviated *AI*.
  - Oxford English Dictionary. https://www.oed.com/view/Entry/271625
- Artificial intelligence is the capability of a computer system to mimic human cognitive functions such as learning and problem-solving. Through AI, a computer system uses math and logic to simulate the reasoning that people use to learn from new information and make decisions.
  - Microsoft Azure documentation. https://azure.microsoft.com/en-us/solutions/ai/artificial-intelligence-vs-machine-learning/#introduction



The machine learning (ML) life cycle

- Obtain and format dataset
- Obtain or produce ML model
- Train model with dataset prepared
- Evaluate model performance
- Implement model
- May continuously improve model with new data

Challenges in documenting AI

- Computer exercises decision-making capacity
- Complex AI processing/ML not always comprehensible
- Interaction of training process, model selection, and social context creates unintended consequences

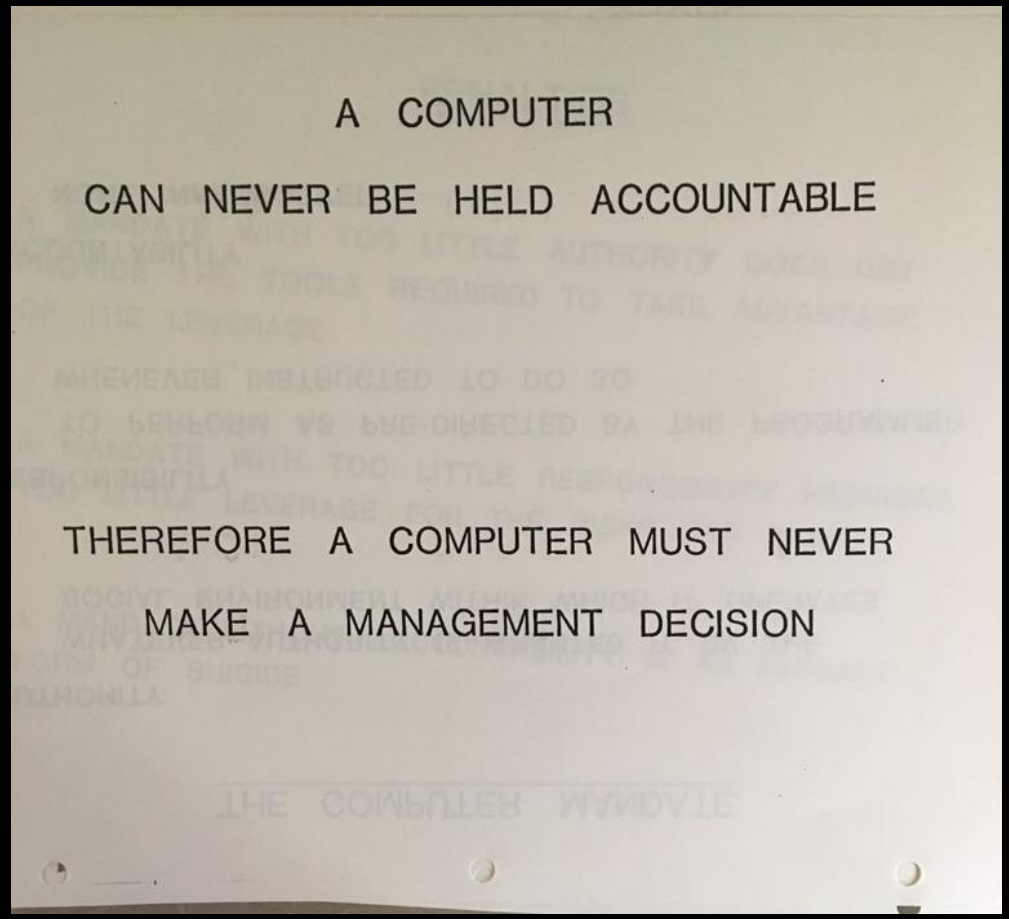


Image: Slide reportedly from an IBM training document, 1979. Source: https://twitter.com/bumblebike/status/832394003492564993/

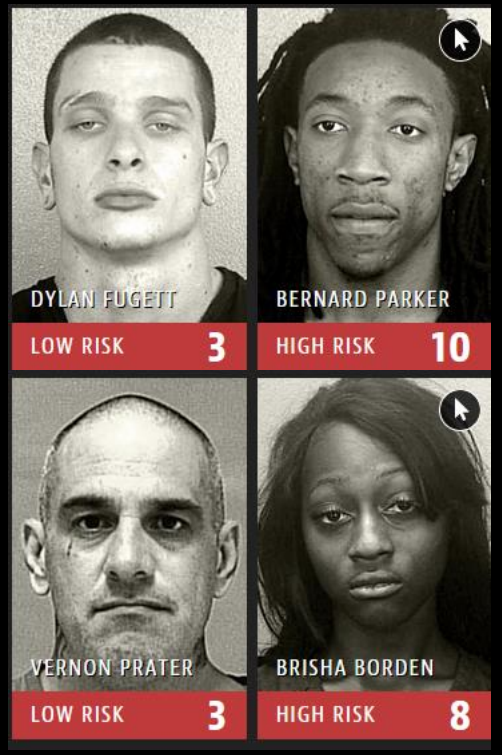
Amazon hiring via machine learning

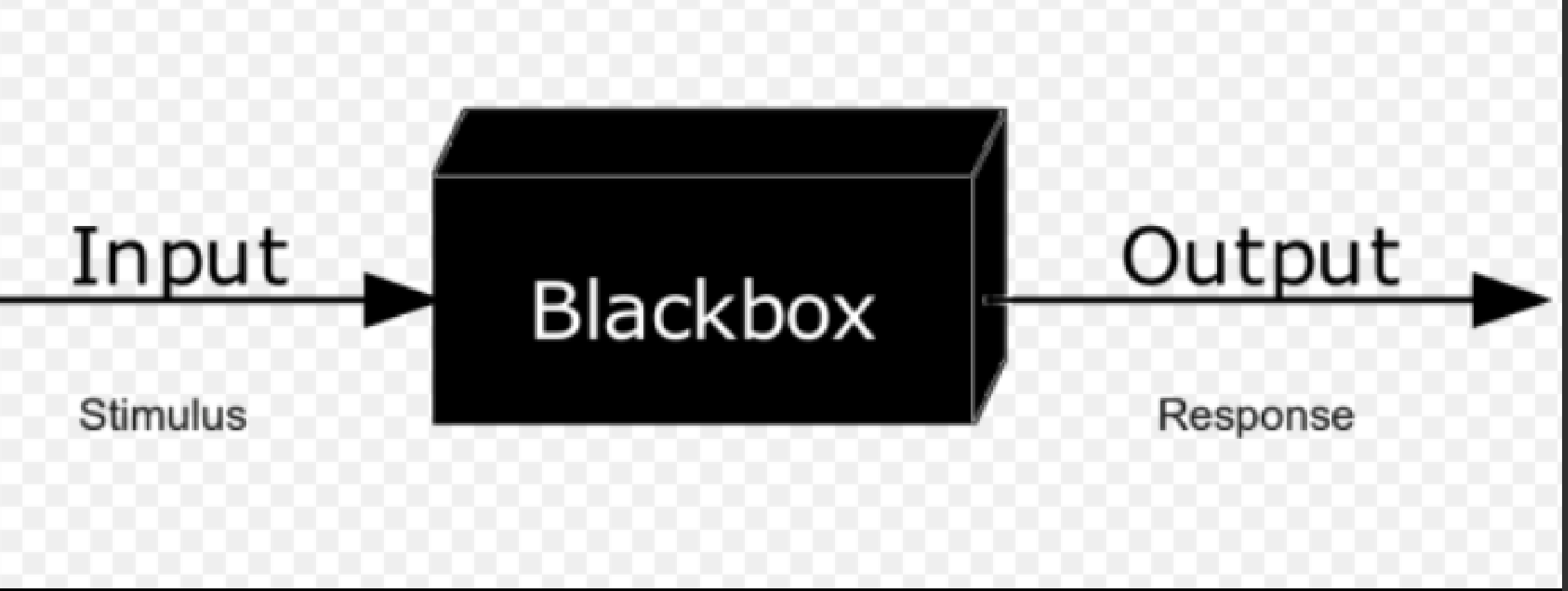
TECH / AMAZON / ARTIFICIAL INTELLIGENCE

Amazon reportedly scraps internal AI recruiting tool that was biased against women / The secret program penalized applications that contained the word “women’s”

COMPAS software example

- **Private sector black box algorithm used in US courts to predict risk of recidivism for parole assessments**
- **ProPublica charged in 2016 that the tool was twice as likely to falsely predict recidivism for black subjects vs. white**
- **Dressel and Farid (2018) were unable to demonstrate racial bias in the tool; they did demonstrate that COMPAS was no more effective in predicting recidivism than participants from Mechanical Turk using a basic subjective analysis of age, sex, prior count, crime, degree of crime, and juvenile offenses.**





The black box problem

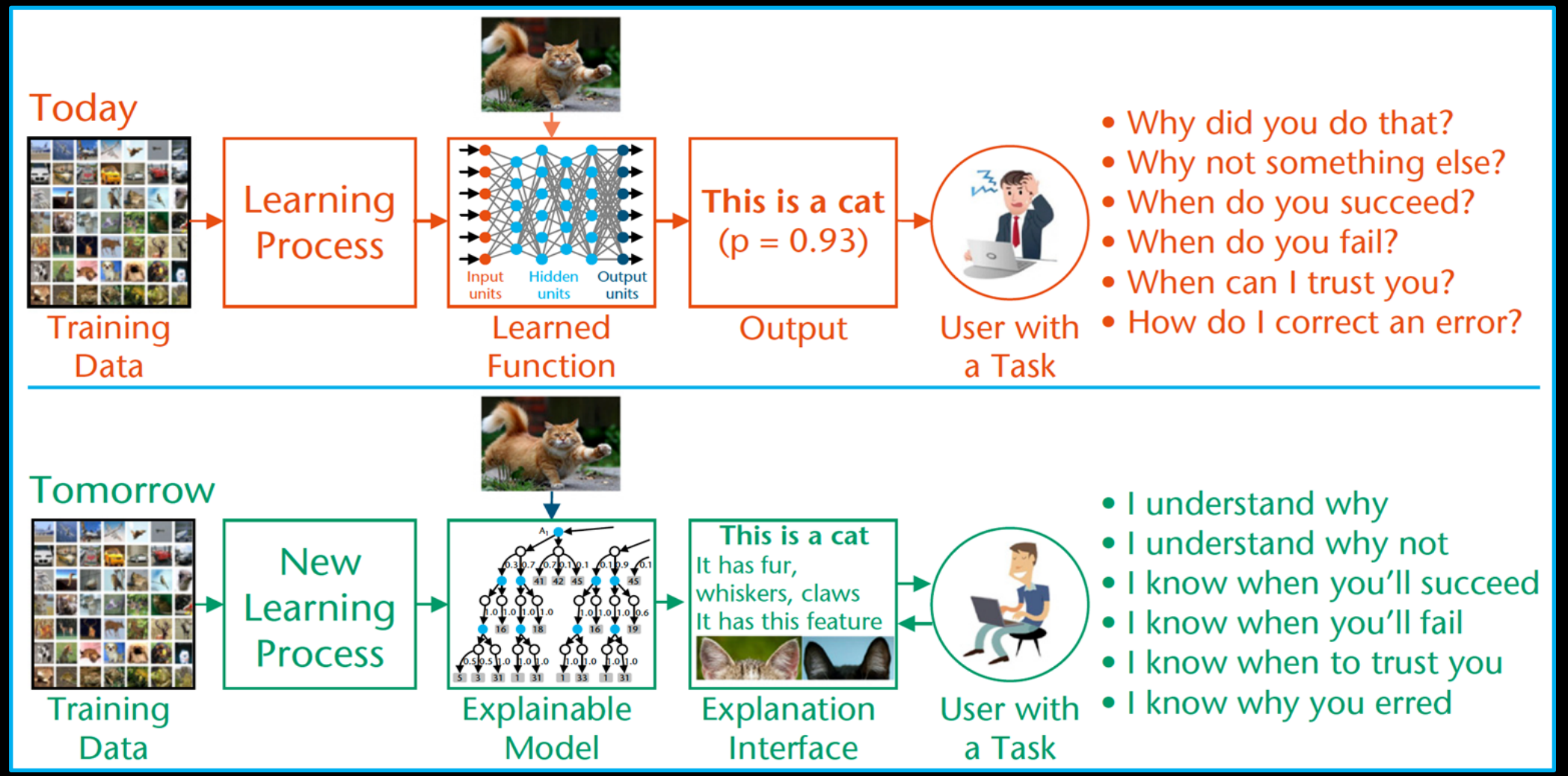
Opaque AI processes

- Basic AI models – e.g. decision trees – may be understood easily and may be considered self-explanatory.
- Complex machine learning (ML) models pose greater challenges to understand or document their processes.
- In many cases, ML tools' "underlying structures are complex, non-linear and extremely difficult to be interpreted and explained to laypeople" (Vilone and Longo, 2020)
- More sophisticated ML tools can analyze sophisticated problems with greater accuracy, at the expense of interpretability (Arrieta et al., 2020)

The black box paradox

- More sophisticated AI tools are more effective BUT
  - less explainable
  - less predictable
  - **less usable in high-risk applications**
- **Opening up the black box is therefore necessary for AI applications requiring accountability.**

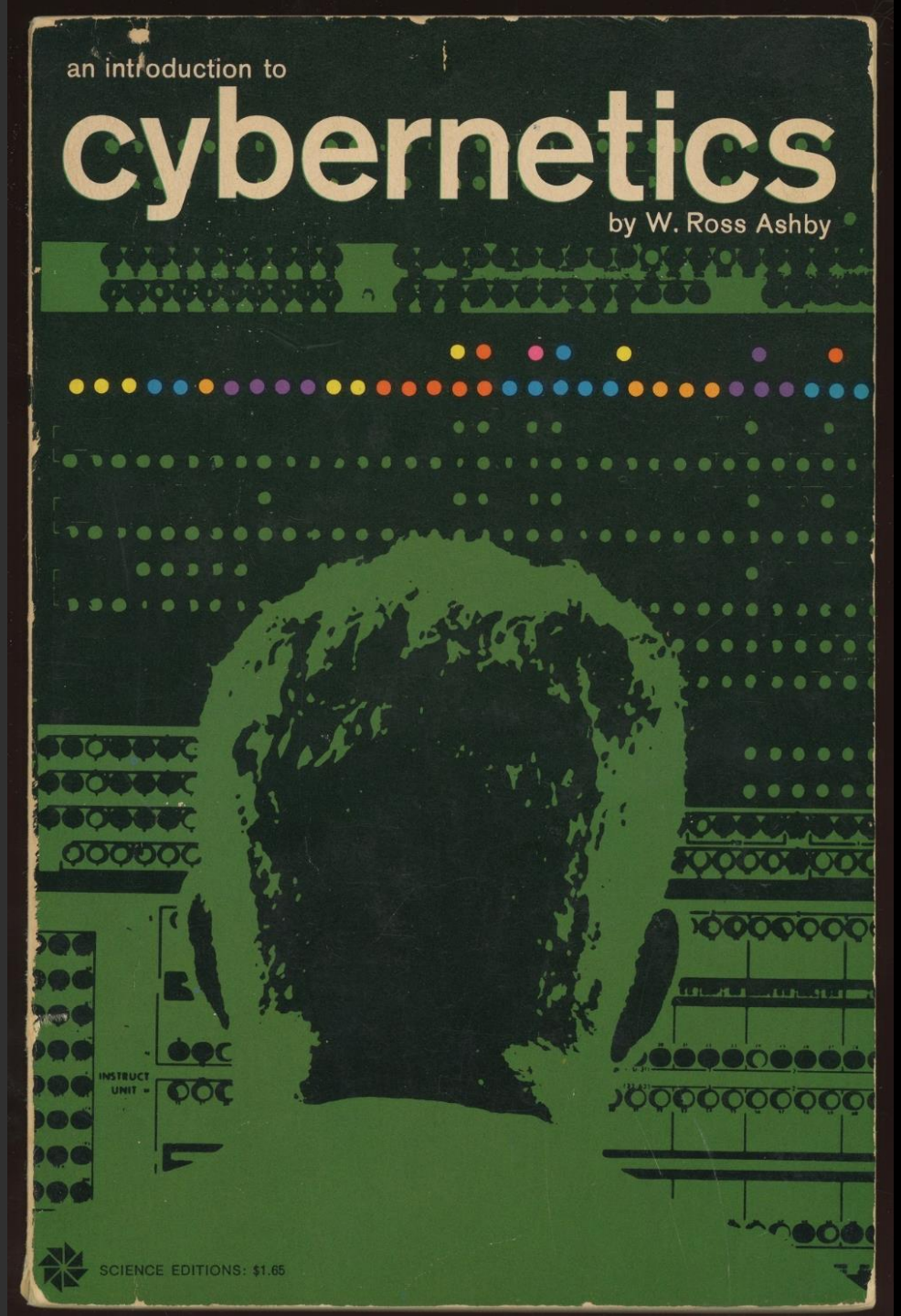
Explainable AI (XAI)



Gunning, D., & Aha, D. (2019). DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*, *40*(2), 44-58. https://doi.org/10.1609/aimag.v40i2.2850

Principles of XAI

- NIST's four principles (Phillips et al., 2021)
  1. Explanation
     1. System provides evidence or reason leading to its outputs
  2. Meaningful
     1. System's explanation is understandable for the intended audience
  3. Explanation accuracy
     1. Explanation accurately reflects the system processes
  4. Knowledge limits
     1. System acknowledges its own limitations and indicates confidence level in output



Limitations on explainability in black box systems

- Post-hoc, reconstructive explanations
- Prediction vs. explanation

Post hoc explanations

Ashby, "Introduction to Cybernetics," 1956

- "every system, fundamentally, is investigated by the collection of a long protocol, drawn out in time, showing the sequence of input and output states… It is now clear that something of the connexions within a Black Box can be obtained by deduction" (88-92).

Philips et al., NISTIR 8312, 2021

- Two categories of AI explanations:
  1. Self interpretable
     - The model itself is the explanation
     - E.g. decision trees
  2. Post hoc explanations
     - The model's logical is reconstructed after the fact by an accessory tool
     - Two subcategories:
       - Local explanations provide an explanation of part of the model or specific decisions
       - Global explanations approximate the entire model's functioning

Prediction vs. explanation

Ashby, "Introduction to Cybernetics," 1956

- "It will be seen that prediction of the [black box's] behaviour can be based on complete or on incomplete knowledge of the parts... When the knowledge of the parts is so complete, the prediction can also be complete, and no extra properties can emerge. Often, however, the knowledge is not, for whatever reason, complete. Then the prediction has to be undertaken on incomplete knowledge, and may prove mistaken" (110-111).

Andreson, A discussion frame for explaining records that are based on algorithmic output, 2021

- "For an algorithm with an uncertain outcome, the least unlikely prediction could be the closest available approximation to an explanation."

Paradata as AI processual documentation

- Paradata must document the full scope of application and context of use – not just the algorithm itself.
  - XAI: why did a given tool produce a given output from a given set of inputs?
  - Paradata: why, how, and to what effect was a given tool used in a particular context?

The National Archives (UK): “Building explainable AI is not just an algorithmic matter, but needs to consider the individuals and the environment in which it will operate” (Jaillant et al., 2020)

Relevant questions to ask for paradata

- *What records are created within AI research teams to document their process?*
- *What records are created of the decisions to procure or deploy systems utilising AI?*
- *What records are created of the decisions and impact of such systems?*
- *Are the created records sufficient to meet existing legal provisions?*
- *Do the created records meet the required standards of quality?*

(Bunn, 2020)

What might comprise relevant paradata?

Operational records documenting an ML tool itself

- Training, testing, and validation data
- Performance information
- Versioning information and the ML instrument itself

(Davet et al. 2023)

Operational records documenting organization's AI practices

- Procurement process
- Implementation
- Quality control and assessment
- Response to complaints

What might comprise relevant paradata?

Policy records documenting organization's AI policies

- AI use cases
- Risk assessment policy
- AI documentation practices
- Subjects' rights
- Avenues of recourse for harm

(See Mooradian on the AI Record, 2018; Andreson, 2020)

Open questions on paradata

- How do requirements for paradata collection increase in higher risk applications?
- What is the relationship between paradata and metadata in existing metadata schema and systems?
- Can meaningful paradata be generated when archivists must document the activity of proprietary black box systems?
- Case studies and context-specific analyses of necessary paradata in given cases, especially as AI legislation mandates algorithmic transparency and accountability

Paradata for accountability

- The introduction of AI should not eliminate transparency or undermine accountability; instead it can be an opportunity to increase the accountability of both computerized as well as human processes.
  - AI-based processes can and should be as transparent and accountable as human processes, and vice versa
  - Processual documentation in both cases can enable accountability

While AI can introduce opacity into information processing and challenge accountability, processual documentation in the form of paradata can embed accountability in AI processes into their context of use.

Cited in this presentation

- Andresen, H. (2019). A discussion frame for explaining records that are based on algorithmic output. *Records Management Journal*, *30*(2), 129–141. https://doi.org/10.1108/RMJ-04-2019-0019
- Ashby, W. R. (1956). *An introduction to cybernetics*. Chapman & Hall Ltd. http://pcp.vub.ac.be/books/IntroCyb.pdf
- Bunn, J. (2020). Working in contexts for which transparency is important: A recordkeeping view of Explainable Artificial Intelligence (XAI). *Records Management Journal*, *30*(2), Article 2.
- Cameron, S., Hamidzadeh, B. & Franks, P. (2023 submitted). Positioning paradata. *Journal on Computing and Cultural Heritage*. Currently under review.
- Davet, J., Hamidzadeh, B., & Franks, P. (2023). Archivist in the machine: Paradata for AI-based automation in the archives. *Archival Science*. https://doi.org/10.1007/s10502-023-09408-8
- Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, *4*(1), eaao5580. https://doi.org/10.1126/sciadv.aao5580
- Mooradian, N. (2019, November 12). AI, Records, and Accountability. *ARMA Magazine*, *ARMA-AIEF Special Edition 2019*, 9–13.
- Phillips, P. J., Hahn, C. A., Fontana, P. C., Yates, A. N., Greene, K., Broniatowski, D. A., & Przybocki, M. A. (2021). *Four Principles of Explainable Artificial Intelligence* (Interagency or Internal Report NISTIR 8312; p. 43). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8312