Is data quality still the ‘elephant in the room’? Working with imperfect data

Dr Gareth Hagger-Johnson

Speaking in a personal capacity

- Director of Data Strategy, Nottingham Building Society
- Affiliate Academic, Consumer Data Research Centre (CDRC), UCL
- Honorary Research Fellow, Norwich Medical School

February 2025

Introduction and overview

- ‘Is data quality the elephant in room?’
  - A cliché, nothing new
- Many organisations still:
  - Eager to use artificial intelligence (AI)
  - Hesitate to invest in data quality
- The importance of data quality in decision making
  - Ethical AI and data-driven decisions rely on good data management, and data quality in particular
- Practical strategies for handling imperfect data

Data merely reflect processes or human interactions

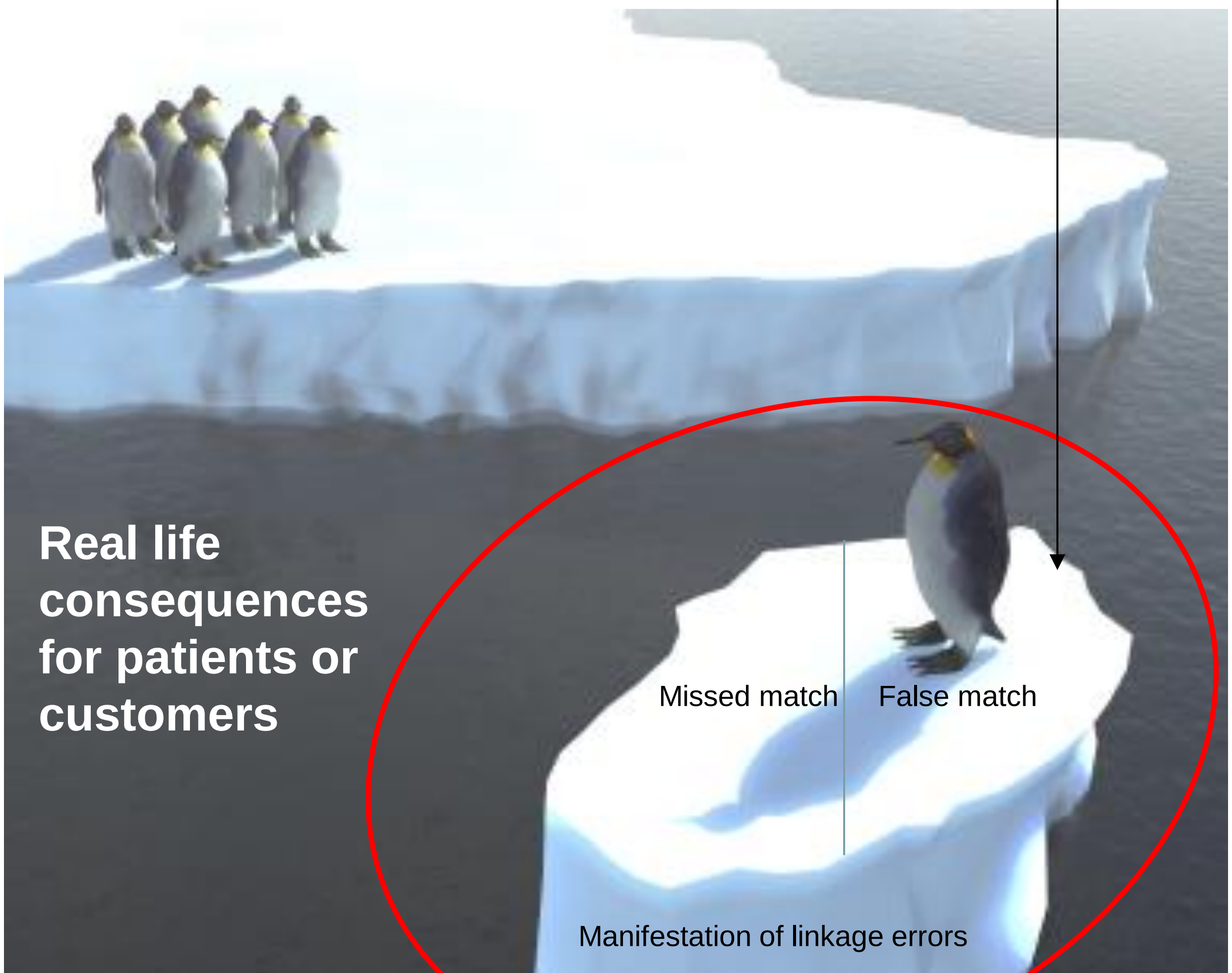
- Where there are data quality issues, these are ‘symptoms’ with causes
  - Systems
  - Database design
    - First name, surname
  - Human error
  - Policy
  - Process
  - Adherence to protocol

Data quality matters – for any algorithm

- Deterministic
- Probabilistic
- AI
- Graph based

Errors in identifiers → Algorithm → Linkage errors

Data quality



Example: Hospital Episode Statistics (HES) algorithm for admissions and readmissions

- False matches
- Clinically implausible scenarios
  - multiple births sharing ID
  - ID readmitted after death
  - ID born, then born again
  - same ID admitted to >1 hospital
  - mother and child sharing ID
  - ID of child gives birth
- More common for males, younger patients, Mixed ethnic group, in frequent readmissions

Hagger-Johnson et al. (2014) Identifying possible false matches in anonymized hospital administrative data without patient identifiers. *Health Services Research.*

- Missed matches
  - Missed matches more common for younger, Asian/Black/Other/Missing (vs. White)
  - Readmission underestimated
  - Minority group readmission underestimated

Hagger-Johnson et al. (2015) Data linkage errors in hospital administrative data when applying a pseudonymisation algorithm to paediatric intensive care records. *BMJ Open.*

Data is a representation

Imperfect map

- ‘We should not confuse the map (**a model of reality**) for the territory (actual reality)…Coming to a good map (model) of the territory (the part of reality that we are interested in) requires careful thought and discussion with key stakeholders’

Van Gils (2023). *Data in context*. Milton Keynes: Springer.

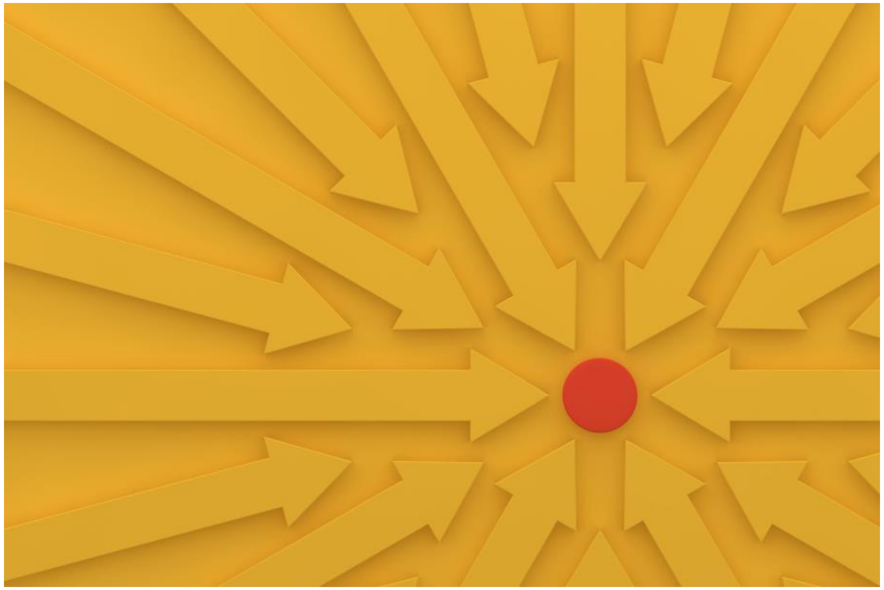
Poor data quality can reverse even the *direction* of effects, not just size

- Long known in public health research that data quality creates biased effects, driven by data linkage quality
  - Relative risk of mortality US Hispanic (vs. White) = 0.97
  - Stricter algorithm = 0.78 (Hispanic at increased risk)
  - Relaxed algorithm = 1.24 (Hispanic at decreased risk)
- Why? Healthier subgroups tend to link more successfully
  - Missed matches among groups with poorer quality data
- Apparent 'Hispanic mortality advantage' is reported in official reports and surveys
  - Important public health policy implications
  - Used in resource allocation, decision making and service delivery

Lariscy (2011) *Journal of Aging and Health*.

Defining ‘bias’ in financial services context

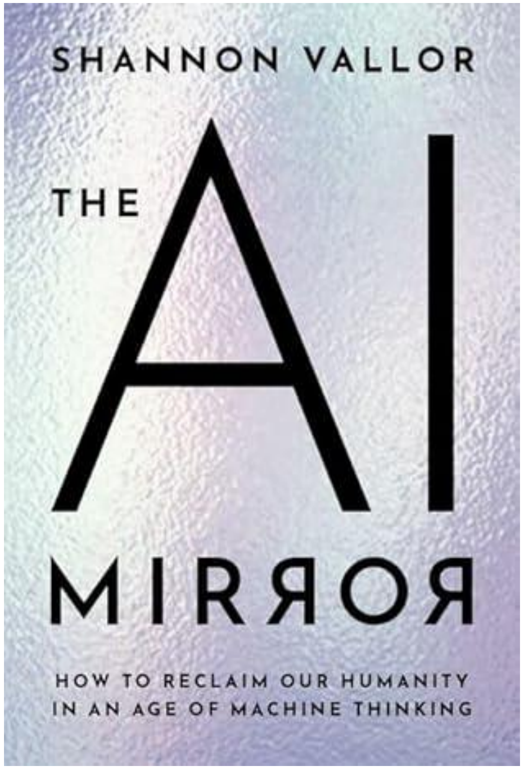
- ‘**unjustified differences** by demographic **characteristics** or characteristics of **vulnerability** in **predictions** and subsequently in **decisions** made in either an **automated** or an **algorithm-assisted** way’



Bogiatzis-Gibbons et al. (2024). *A literature review on bias in supervised machine learning*. Financial Conduct Authority.

'This issue of fairness is regularly raised but rarely solved'

Yesterday’s data



- Popular assumptions
  - Historic data favours ‘majority groups’, and men
  - Protected characteristics correlate with credit risk factors, creating bias
    - Postcode, credit history, complex income, life events
- Causes difficult to identify in practice
  - Algorithms are rarely transparent or shared for evaluation
  - Firms cannot collect data on many protected characteristics to undertake retrospective or prospective ‘equality audits’
- But absence of evidence is not evidence of absence

Financial Services firms’ personal data use – is this leading to bias and detriment for consumers with protected characteristics? Financial Services Consumer Panel Evidence Review – Final Report (FCA, 2023)

We need to better understand the role of data quality and identify ‘other factors’

UK ‘Ethnic Minority Businesses and Access To Finance’ report

‘There is no evidence to indicate that disparities are due to racial discrimination per se, but variances could be accounted for by other business characteristics’

Department for Communities and Local Government. (2023). *Ethnic minority businesses and access to finance*.

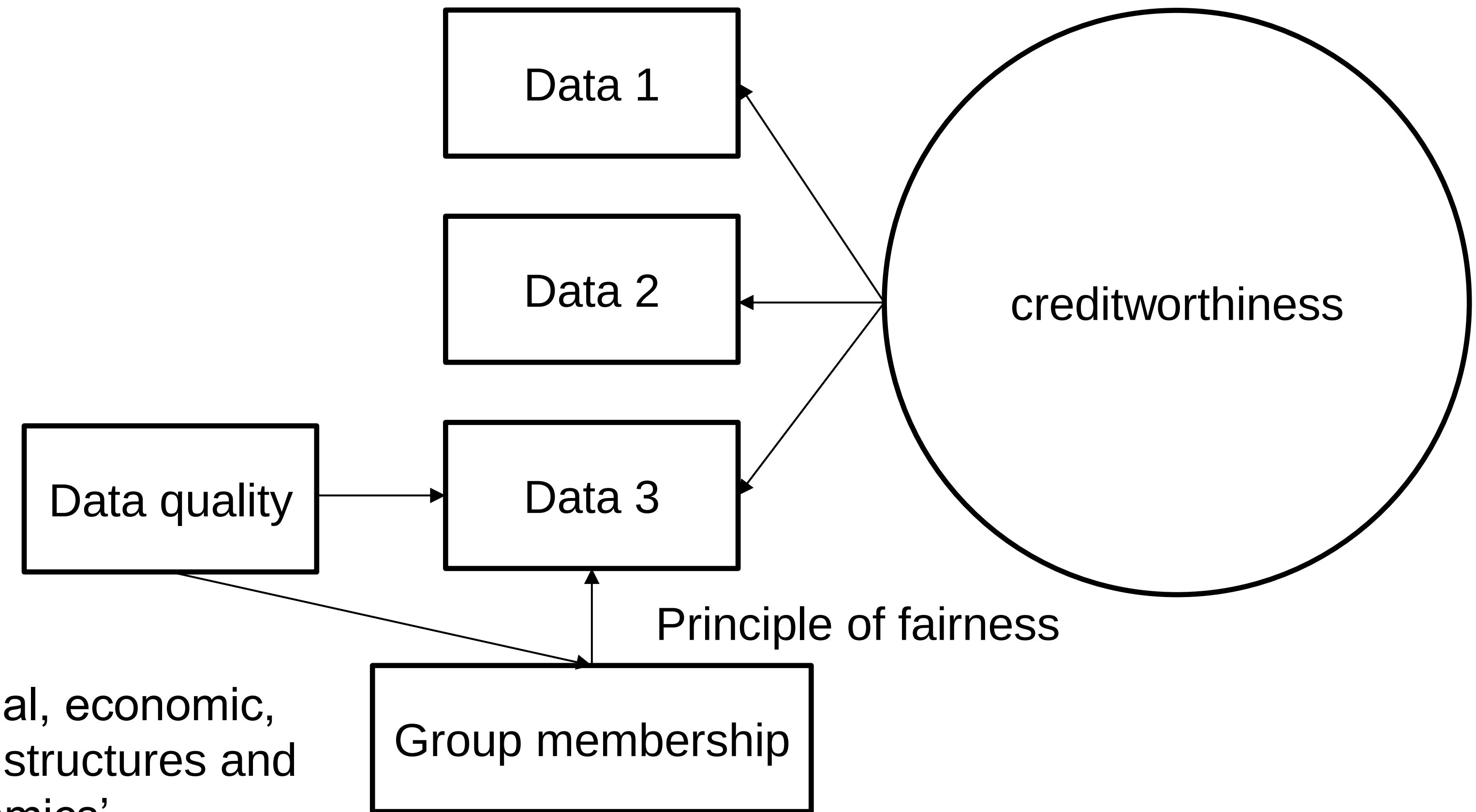
‘This is exactly how institutional racism works: the clearly discriminatory playing field is acknowledged and then rejected on the basis that there could be “**other factors**” at work…A lack of access to finance is a key cause of the staggering ethnic wealth gap, which further prevents people from owning homes and setting up businesses’

Kehinde Andrews

Practical suggestions are available

- Breeden & Leonova (2021). Creating unbiased machine learning models by design. *Journal of Risk and Financial Management*
- Ethnicity estimator software (UCL CDRC) for possible ‘retrospective equality audits’
- Differential item functioning (DIF) drawing on psychometric methods for identifying statistical bias
- Financial Conduct Authority (2024). *A literature review on bias in supervised machine learning*.

Testing for statistical bias



‘existing social, economic, and historical structures and dynamics’

Ostmann & Dorobantu (2021)

What is current best practice?

Further reading

- The High-Level Expert Group on AI (AI HLEG) (2020). *The assessment list for trustworthy artificial intelligence (AI) for self-assessment*.
- European Banking Authority (2021). *EBA discussion paper on machine learning for IRB models*.
- Ostmann & Dorobantu (2021). *AI in financial services.* The Alan Turing Institute.

Key points

- Constantly evolving
- Ethical, explainable, trustworthy
  - Poor data quality undermines all three
- Bias could undermine public trust
- FCA Consumer Duty: Fair outcomes
- Firms restricted in processing ‘protected characteristics’
  - Industry / academic collaboration?

European Banking Authority (2021)

Four pillars

data management and data quality

technological infrastructure

organisation and governance

analytics methodology

Trust elements

- Ethics
- Explainability and interpretability
- Traceability and auditability
- Fairness and bias prevention/detection
- Data protection and quality
- Consumer protection aspects and security

AI could help reduce bias: opportunities

- Risk differentiation
- Risk quantification
- Improving data collection and preparation including data quality checks, outlier detection
- Improving credit risk mitigation techniques
- Robust validation and monitoring
- Automating bias prevention

European Banking Authority (2021). *EBA discussion paper on machine learning for IRB models*.

Bias at the human manual review stage

Human decisions

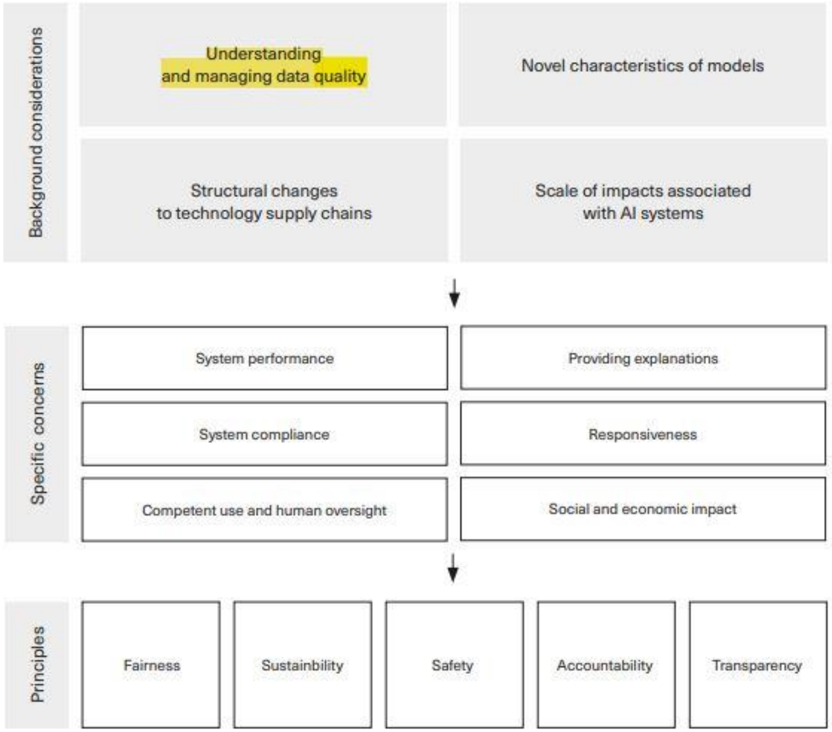
"the increased reliance on models, data, and responsible forms of automation can help to reduce the occurrence of unwanted forms of differential treatment that result from biased human judgments"

e.g. differential treatment at the manual review stage of lending applications

Ostmann & Dorobantu (2021). *AI in financial services*. The Alan Turing Institute.

Data quality is necessary, but not sufficient

Ethical AI



- Understanding and managing data quality is a fundamental consideration in ethical use of AI
- First and foremost (top left)
- Investing in data quality will pay dividends later
  - Ethical AI builds trust
  - Poor data quality decreases revenue and diverts resources
- Get specific!

Ostmann & Dorobantu (2021). *AI in financial services*. The Alan Turing Institute.

Thank you

g.hagger-johnson@ucl.ac.uk

@hssghj

@gh-j.bsky.social