GREEN, SAFE, UNDERSTANDABLE, COMPACT, COST EFFECTIVE use of Quantitative BIG DATA:

Continuous, Interpretable, Minimalistic Machine Learning for Repeated Multi-/Hyper-spectral EO Imaging

Harald Martens[1,2]*, Frank Westad[1,2], Bharath Nagaraja[1], Christian Landaas[1]

[1]Idletechs AS, Trondheim Norway. [2]Professors in Big Data Cybernetics, Dept. Engineering Cybernetics, Norwegian U. Science and Technology NTNU, Trondheim Norway. *Corresponding author: *harald.martens@idletechs.com*

idletechs NTNU

Repeated Hyperspectral Earth Observation (EO) = "slow video" : Series of hyperspectral images, registred wrt a common, fixed reference image, comprehensively quantified by models of all relevant KNOWNS and UNKNOWNS

How to maximize the EO data's ***usefulness,*** with minimum ***risk*** and ***information loss***, minimum ***training*** data, ***energy*** use and human ***alienation*** *and* ***uncritical use?***

Multi-channel EO gives information – rich, but confusing **BIG DATA**, with both **«KNOWN»** and **«UNKNOWN»** variations:

Satellite HSI with ***multi-wavelength*** pixels: Several overlapping **spectral** and **spatial** variations sources.

Repeated satellite HSI of **same location**, spatially aligned = «fixed camera slow video»: Also **temporal** variation sources.

Each multi-channel image is affected by **several *spectral*** and ***spatial*** variation types **:**

- Δ Ground physics (surface & particles → Δ Spectral scattering),
- Δ Sun angle × Landscape 3D topology → Δ Shadows,
- Δ Atmospheric distance (Rayleigh scattering/smog) → Δ Nonlinear baseline,
- Δ Ground chemistry (constituents → Δ Spectral absorption),
- Δ Weather, clouds and haze → Δ Illumination and visibility,
- Δ Satellite position (Δ distance, angle) → Δ Displacements (size, perspective & 3D).

How to **disentangle** these partially over-lapping **variation types** mathematically, **assess** them **statistically** and inspect them **graphically**?

This requires ***Purpose-driven***, pragmatic, abductive, ***hybrid modelling,*** combining

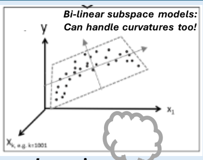
Theory-driven, *deductive* quantification of all relevant **KNOWNS** (Semi-mechanistic Meta-Modelling)

& ***Data-driven,*** *inductive* discovery and quantification of all systematic **UNKNOWNS** (Machine Learning)

Different science cultures handle BIG DATA differently

Data-driven chemometrics, etc:

*«Get **INFORMATIVE** DATA. Linearize & simplify them, using prior knowledge to quantifying **KNOWN** variations. Then, by transparent ML, discover and quantify all remaining, observed but **UNKNOWN** variation patterns. Plot them to try to understand them too. Later: Use model for classification and prediction, but **Keep Learning**»*



Continuous, Interpretable, Minimalistic Machine Learning (CIM-ML)

ML: As simple as possible, but not simpler

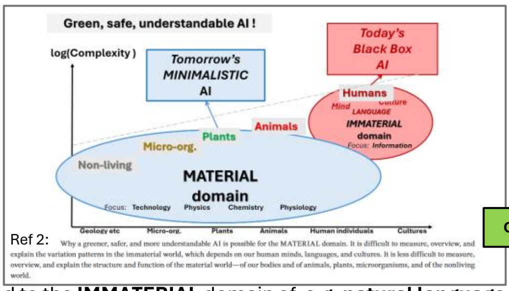
Data-driven computer science / AI:

*«The system as such is too complex to understand. Get **LOTS** of DATA. Use unprocessed raw data in Deep Learning ML, ideally to develop brain-like «intelligence». Later: Use model for classification and prediction»*

Black Box ANN/CNN based Machine Learning, (much of today's AI ?)

ML: Start with very complex model, then try to simplify by XAI

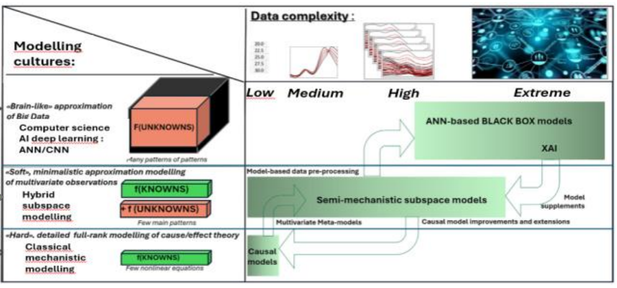
WHY CIM-ML works:



Ref 2: Why a greener, safer, and more understandable AI is possible for the MATERIAL domain. It is difficult to measure, overview, and explain the variation patterns in the immaterial world, which depends on our human minds, languages, and cultures. It is less difficult to measure, overview, and explain the structure and function of the material world—of our bodies and all plants, microorganisms, and of the nonliving world.

Compared to the **IMMATERIAL** domain of e.g. **natural language**, the **MATERIAL** domain of Science and Technology is **far simpler**. This is due to **constraints** from **Laws of Nature, Geological History, Biological Evolution** & **Accumulated Human Experience**. EO BIG DATA are **overwhelming** and **need ML**, but **not BLACK BOX**.

Modelling cultures for different levels of complexity



Soft multivariate data-modelling and meta-modelling: A bridge beteween over-simplified mechanistic modelling and overly complicated BLACK BOX ML.

HOW ? CIM-ML for BIG DATA from EO HSI: Apply well-proven, industrial self-modelling methodology:

Combine relevant **KNOWNS** + all clearly observed **UNKNOWNS**.

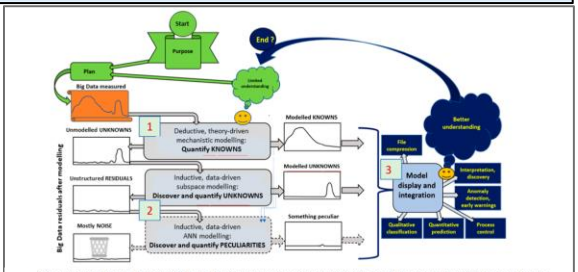
⇒ Minimalistic, compact ML (N-linear and non-linear subspace hierarchies):

Green: Less energy needed: Fast self-modelling algorithms
Less data storage & transmission: Strongly compressed results.

Safe: All observable anomalies reported
All parameters can have uncertainty estimates & statistical validation
Accessible for for human inspection and critique (and of course, no hallucination).

Understandable: Transparent hybrid models combine Human Knowledge and Big Data
Low dimensional local subspace models, easy to study graphically
Mixed additive/multiplicative spectral effects (EMSC) and motions (IDLE).

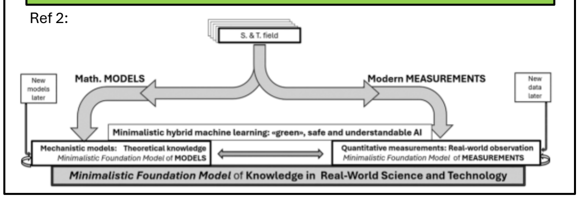
Cost-effective: Less training data: Combines prior theory, experience & new HSI data
Semi-causal models of the relevant KNOWNS and significant UNKNOWNS
Results are interpretable and usable in their compressed form
Fast reconstruction of compressed, noise-reduced input data. Lossless reconstruction also possible.



Ref 1: A framework for understandable machine learning with an eye for human knowledge: Hybrid modelling of technical BIG DATA on a small computer [3] by
1) Deductive use of expert knowledge: Causal modelling of KNOWNS.
2) Inductive machine learning: Data driven modelling of systematic and – if needed, peculiar - UNKNOWNS [4].
3) Abductive model coordination and use, as a hybrid "deep learning" [5] for technical data.

Two types of CIM-ML based Foundation Models for summarizing HSI-based Earth Observations:

Ref 2:



MECHANISTIC MODELS' combined behavioral repertoires = Summary of **theoretical KNOWNS (or BELIEFS).**

BIG DATA MEASUREMENTS' combined variation patterns = Summary of **real-world KNOWNS & valid UNKNOWNS.**

Background: The **CIM-ML** hybrid hierarchical, dual-domain, nonlinear subspace modelling for **Continuous, Interpretable, Minimalistic ML,** builds on our long R&D experience in several fields, ranging from chemometrics and cognitive science, via quantitative linguistics, psychometrics, medical imaging and mass media video modelling to continuous industrial monitoring by thermal video, and airborne/satellite EO: InSAR and thermal/vis/NIR «hyperspectral slow video».

Its mathematical and statistical basis is described in our several hundred research papers, in our books and in our patent literature, **in fields outside the computer science based AI/ML culture**.

Summary references:

1 H. Martens (2023): **Causality, machine learning and human insight**. Analytica Chimica Acta. 1277: doi.org/10.1016/j.aca.2023.341585

2 H. Martens (2025): **A Greener, Safer, and More Understandable AI for Natural Science and Technology.** J. Chemometrics; 39, 2, 27 pages: /doi/10.1002/cem.3643 . More detailed references etc at https://www.ntnu.edu/employees/harald.martens & upon request to the authors.

We seek: ´**Space-hardened partners** for developing compact, green, safe, understandable and cost-effective **HSI Digital Twins of the Earth.**