Pathway toward prior knowledge-integrated machine learning in engineering

Xia Chen and Philipp Geyer

Leibniz University Hannover, Institute for Design and Construction,
Sustainable Building Systems Group, Hannover, Germany

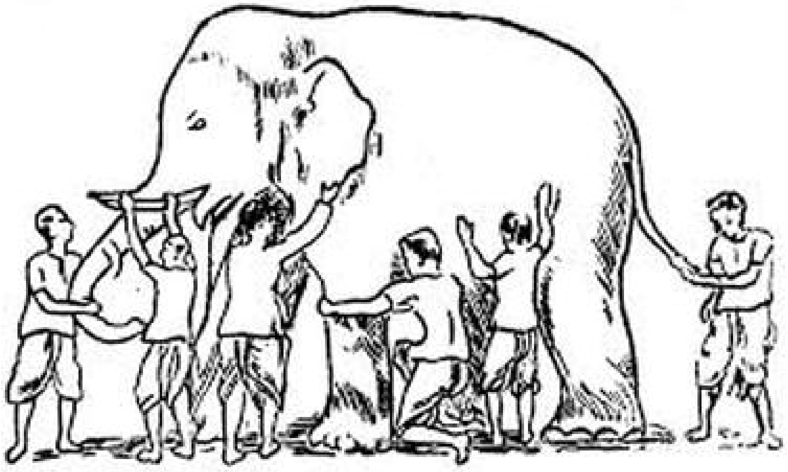
xia.chen@iek.uni-hannover.de

Nachhaltige
Gebäudesysteme

To understand what is an elephant…

Deconstruction perspective: tusks, tail, legs, ears, and their connections

Reductionism, Symbolism



Entity perspective: movement, behavior, and interactions with its environment

Holism, Connectionism

There is no one best way to formalize information for problems.

First-principles modeling
Knowledge, logic-based

Prior knowledge

Data

?

Data-driven/ML methods
Experience, heuristic

1010
1010

Methodology Framework

1. **Uncertainty analysis (Where is the gap?)**
   - Data
   - Prior knowledge
   - Data-driven model
2. **Knowledge-based decomposition (What information/knowledge we can use?)**
   - Domain know-how
   - Scientific Method
   - Complexity/Scale
3. **Ladder of knowledge-integrated ML (What advantages we can achieve by the integration?)**
   - Interpolation
   - Extrapolation
   - Representation

1. Uncertainty analysis (Where is the gap?)

(1) Uncertainty due to the available data/measurement/collection

Gap comes from:

1. **First-principles simulation** and **measurements**;
2. **ML** and **measurements**,
3. **Measurements** from **different sources**.

Key idea:

They are complementary!

General uncertainty
Performance gap between actual and predicted values.

Epistemic
limitation because of biased or lack of understanding.

Aleatoric
the natural inherent noise.

Parametric
Limitations under the current model specification.
(Implicit factors, information hidden in the data)

Structural
Whether model specification is sufficient.
(Decomposition patterns explained by knowledge)

Data-driven ML methods

First-principles modeling

knowledge-integrated machine learning

1. Uncertainty analysis

(2) Uncertainty due to physics (domain knowledge), first-principles model, symbolism

Gaps	Description	Case	Reference
Model over-simplification	Unable to capture synergistic or non-linear effect from **hidden factors**	Structure engineering in extreme condition	(Stochino 2016)
Context constraints in model development	Symbol-based rules derived from a strict logical deduction process limit the ability to accommodate **exceptional conditions** and implicit interactions	Transitioning from experimental modeling or simulation in lab environments to real-world projects	(Tang et al. 2019, Durdyev et al. 2021)
Confirmation bias in modelling	The **reliance on informative priors** does not guarantee inferential perfection or even consistency in problem-solving	Energy system optimization modeling regardless of spatiotemporal boundaries	(DeCarolis et al. 2017)

→ Integration of implicit patterns learned from data

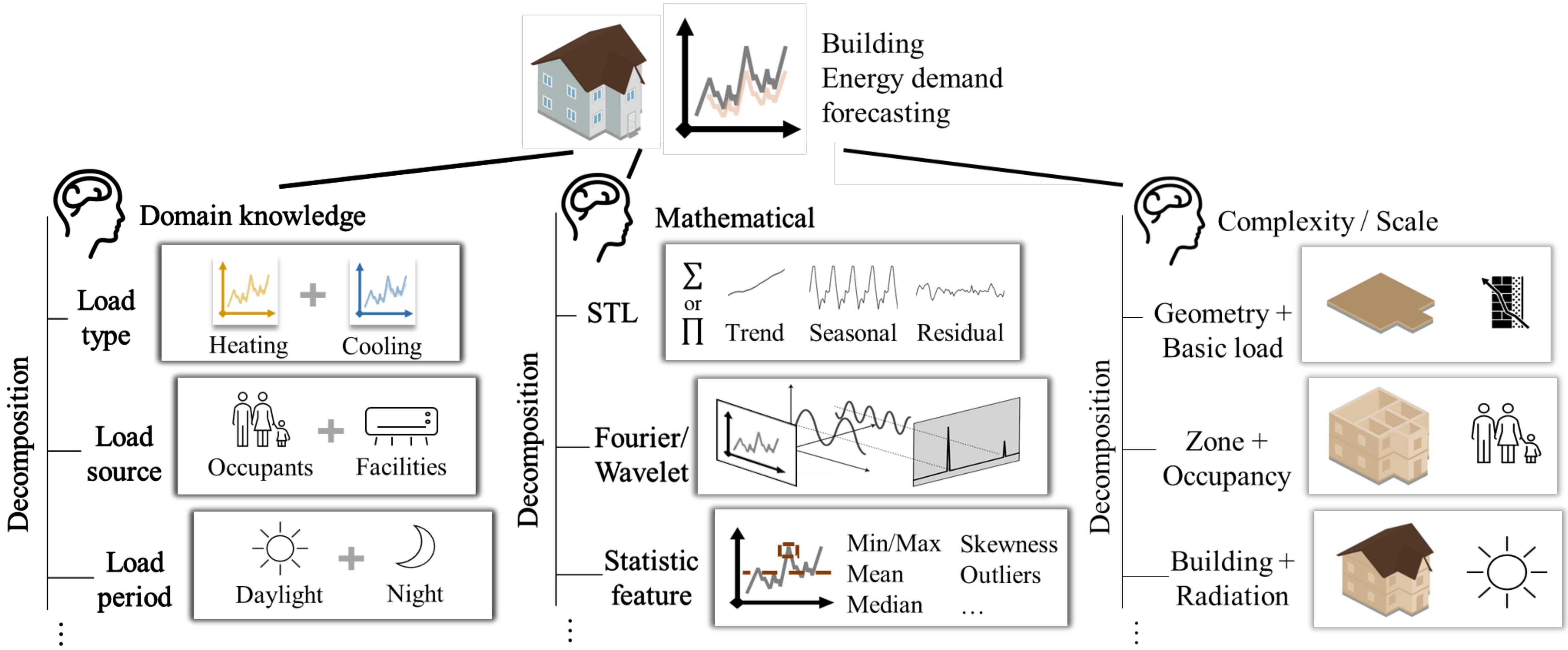
1. Uncertainty analysis

(3) Uncertainty due to the learning models (ML), data-driven model, connectionism

Gaps	Description	Case
Approximation error Model architecture (how is it organized?)	Whether the ML **model organization** (e.g., the design of the model structure, depth of model) approximates a solution to accurately describe complex system behavior	CNN/RNN/Tree whether a model is designed to capture the autocorrelation
Optimization error Learning rules (how does it learn?)	**Choice of learning rules** cause difficulty in finding or result in convergence to a suboptimal solution	Over-/underfitting issues
Generalization error Objective functions (what does it learn?)	Whether training **error minimization** to approaching the defined indicator leads to a more accurate prediction for the solution	Mean squared error / cross-entropy

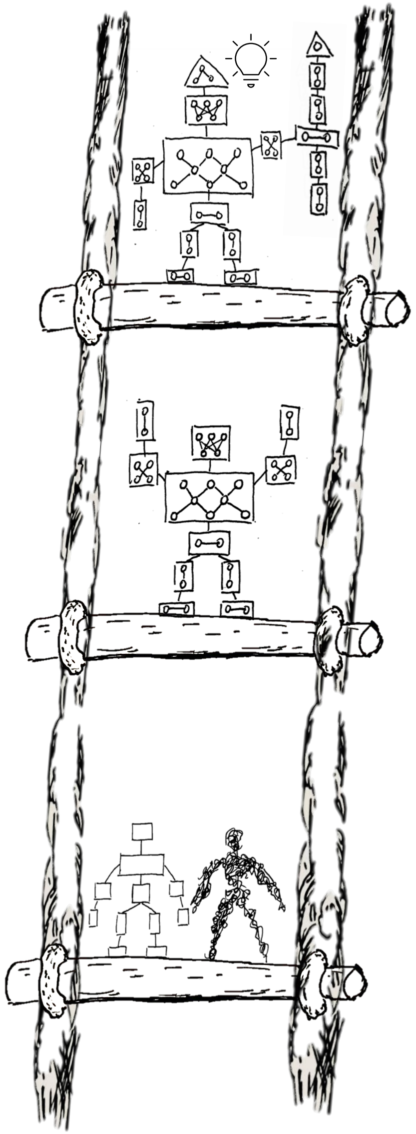
→ **Integration of explicit prior domain knowledge**

2. Knowledge-based decomposition (What knowledge we can use?)



3. The Ladder of knowledge-integrated machine learning

Transfer information into machine-learnable information to achieve better



- **Level 3** ***Representation***

Typical methods: knowledge discovery, representation learning

- **Level 2 -** ***Extrapolation***

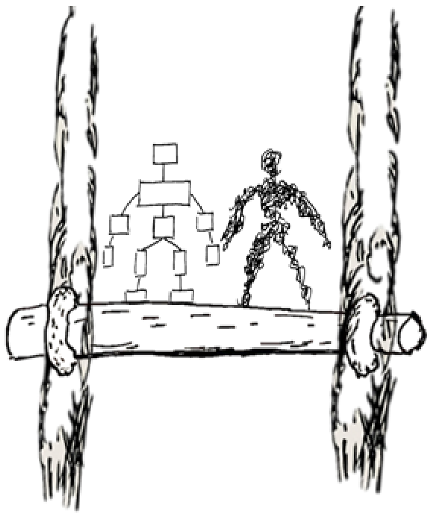
Typical methods: transfer domain knowledge into modeling process

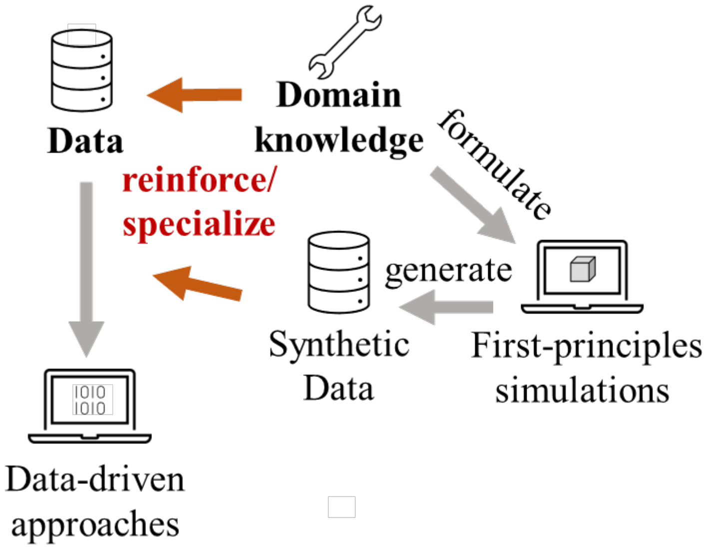
- **Level 1 -** ***Interpolation***

Typical methods: data argumentation; feature engineering

3. The Ladder of knowledge-integrated machine learning

Level 1 - *Interpolation: data argumentation*

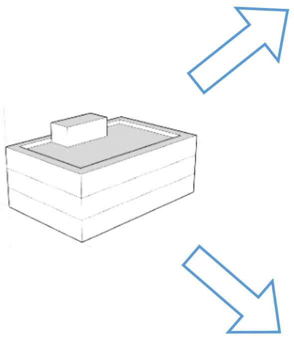




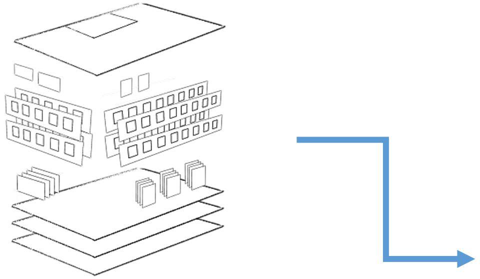
Incorporate prior understanding into *data*: better generalization; more efficient training; reduce overfitting; and compensate for sparse data **within observed range**

3. The Ladder of knowledge-integrated machine learning

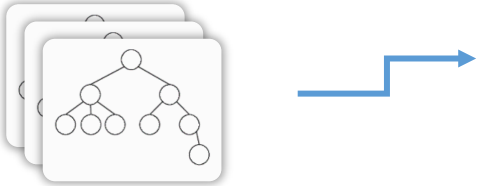
Level 1 - *Interpolation: data argumentation*



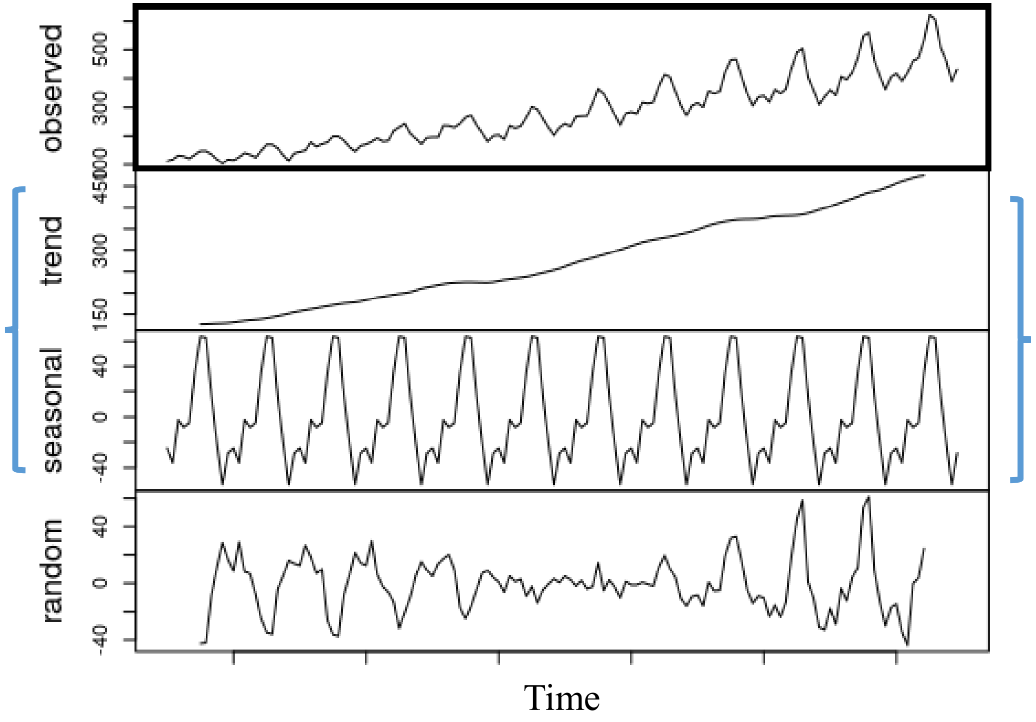
First-principles model:
Good explainability



Machine learning model:
Flexible



Decomposition of additive time-series



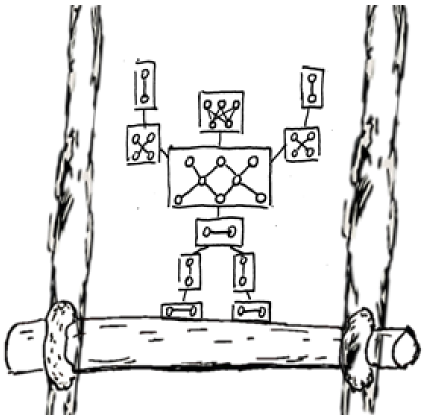
Linear patterns obtained by domain knowledge (e.g., physics or thermal equations)

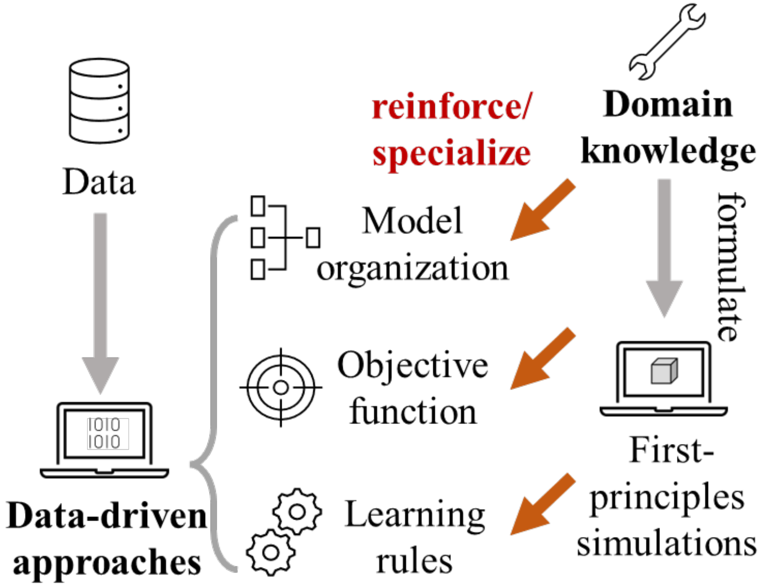
Nonlinear pattern from unknown factors (e.g., environmental noise)

- *Chen, X., Guo, T., Kriegel, M., & Geyer, P. (2022). A hybrid-model forecasting framework for reducing the building energy performance gap. Advanced Engineering Informatics, 52, 101627.*

3. The Ladder of knowledge-integrated machine learning

Level 2 - *Extrapolation: Physical-informed*

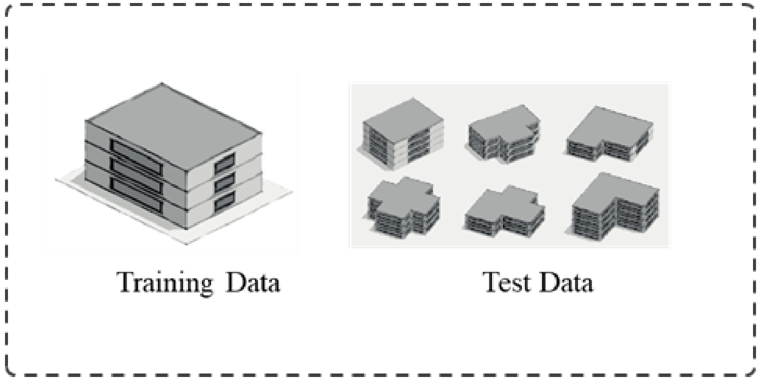


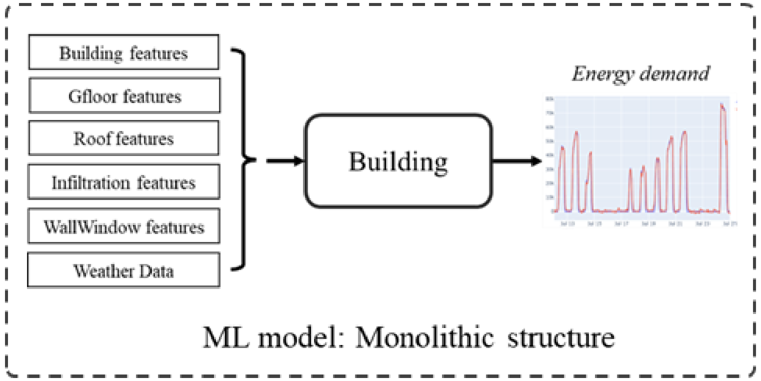


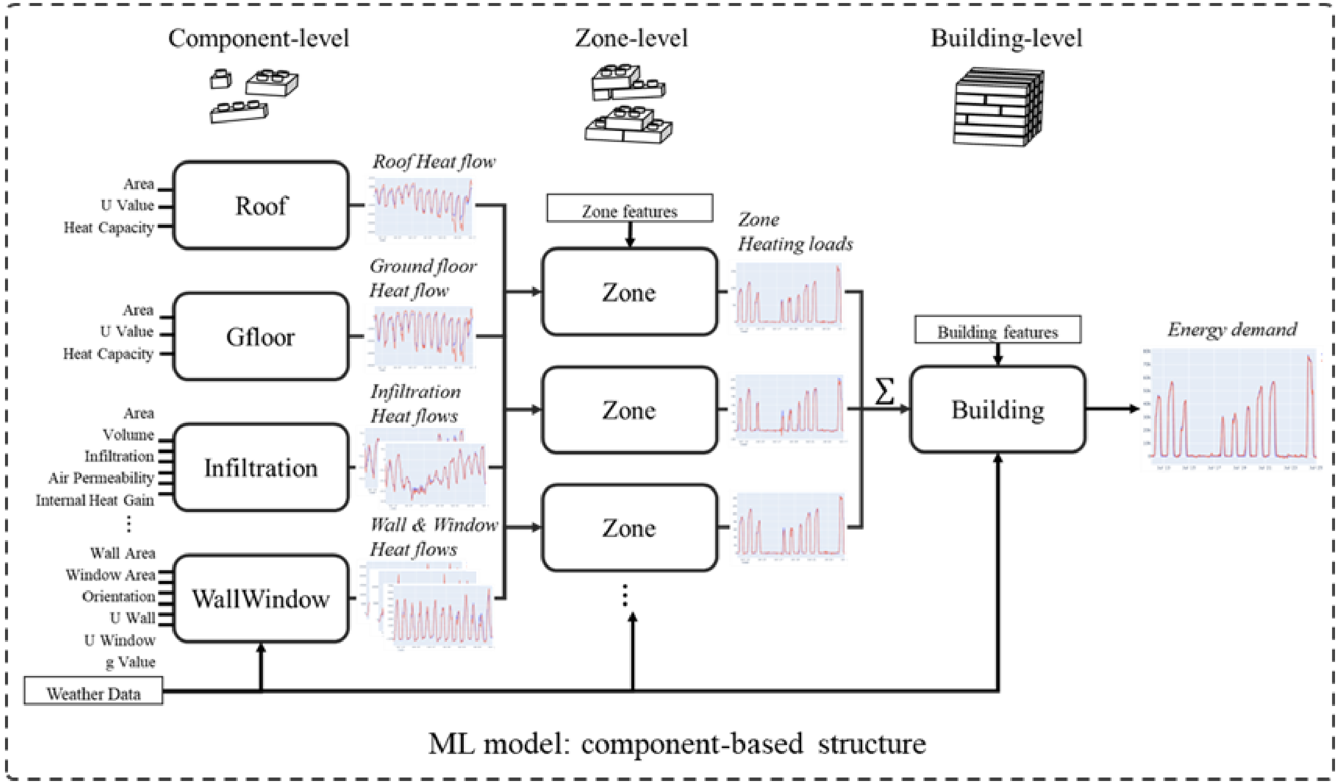
Incorporate prior understanding into *model*:
better generalization, regularization; more efficient training; contextual understanding, informed predictions;
outside the observed range

3. The Ladder of knowledge-integrated machine learning

Level 2 - *Extrapolation: Physical-informed*



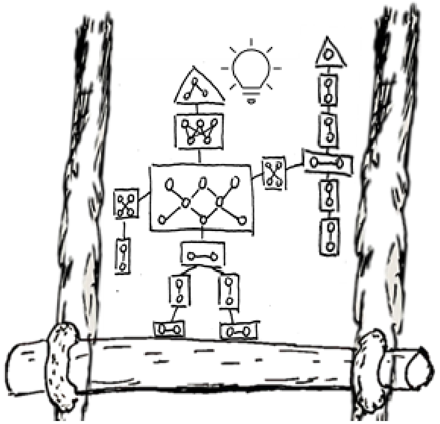


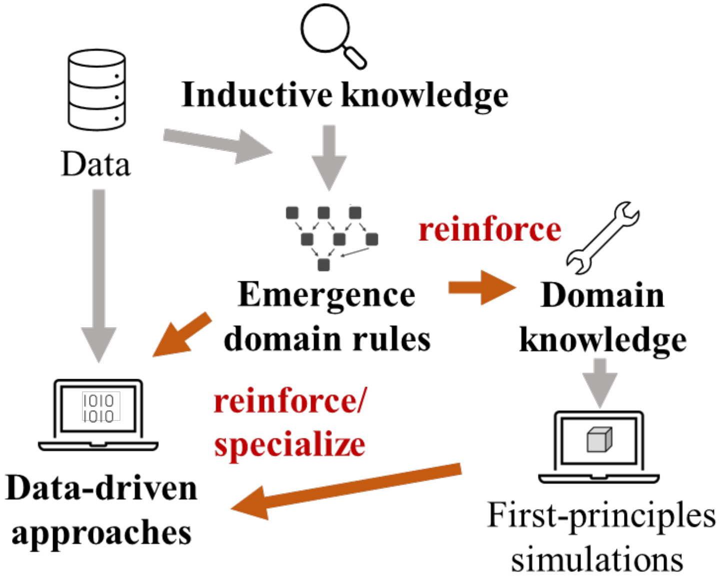


- *Chen, X., Singh, M.M., & Geyer, P. (2022). Utilizing domain knowledge: robust machine learning for building energy performance prediction with small, inconsistent datasets. arXiv preprint arXiv:2302.10784.*
- *Chen, X., Singh, M.M. & Geyer, P. (2021). Component-based machine learning for predicting representative time-series of energy performance in building design. In 28th International Workshop on Intelligent Computing in Engineering, EG-ICE 2021. Berlin, Germany.*

3. The Ladder of knowledge-integrated machine learning

Level 3 - *Representation: Knowledge discovery*

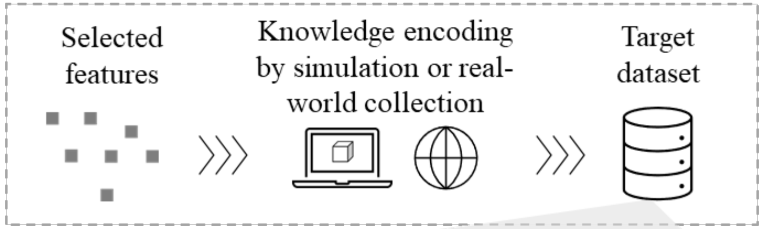


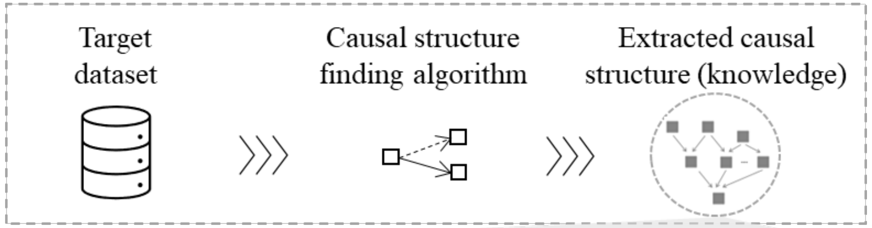


Incorporate knowledge discovery mechanism into *model*:
reducing prior knowledge biases;
encoding, representing, and transforming effective information concisely and self-continuously, reasoning
from domain data

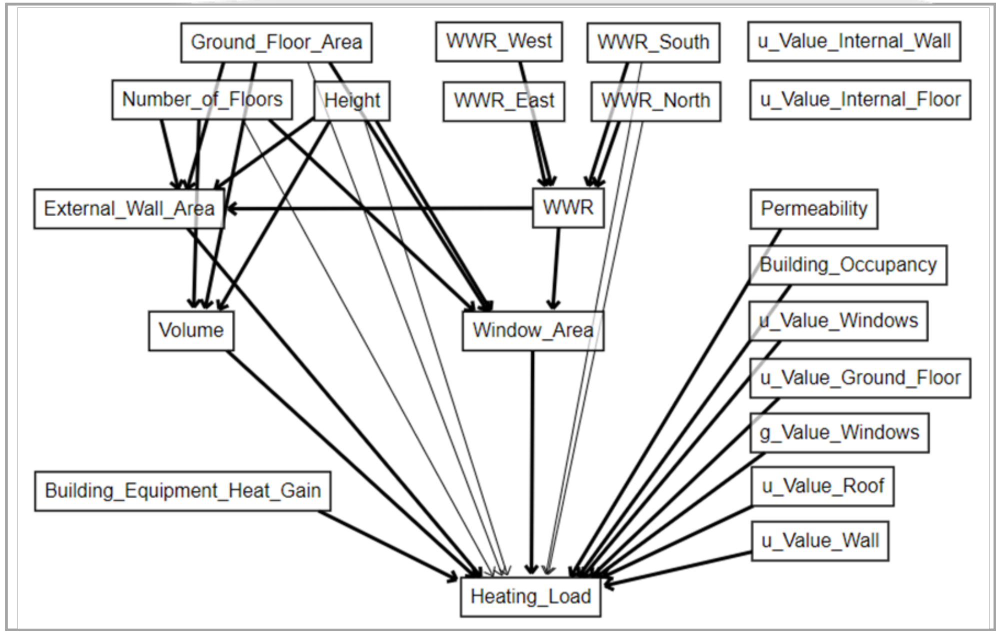
3. The Ladder of knowledge-integrated machine learning

Level 3 - *Representation: Knowledge discovery*





Height	Volume	Number of Floors	External Wall	Ground Floor	Window Area	u-Value (Wall)	u-Value (Ground	u-Value (Roof)
3.74219	8039.57	4	1051.36	537.09	357.575	0.23828	0.21797	0.20234
3.24219	5150.12	3	610.043	529.49	305.469	0.18828	0.16797	0.15234
3.82813	11041.8	4	1050.95	721.1	597.062	0.22031	0.15156	0.24531
3.46875	2524.66	3	467.647	242.61	195.751	0.23438	0.15313	0.21563
3.65625	7635.9	5	1018.12	417.69	476.369	0.20313	0.23438	0.19688
3	864	2	259.2	144	28.8	0.15	0.15	0.15
3.64063	6369.49	4	1039.34	437.39	200.297	0.22656	0.24531	0.15156
3.14063	2683.73	2	341.977	427.26	192.714	0.17656	0.19531	0.20156
3.96875	9691.53	4	1119.71	610.49	463.823	0.18438	0.20313	0.16563
3.15625	8205.33	3	871.051	866.57	243.894	0.15313	0.18438	0.24688
3.75	7315.31	3	803.25	650.25	344.25	0.225	0.175	0.175
3.80469	7637.68	4	938.138	501.86	425.843	0.20703	0.18672	0.19609
3.30469	1186.71	2	243.495	179.55	110.933	0.15703	0.23672	0.24609
3.89063	4455.16	3	691.053	381.7	302.516	0.20156	0.17031	0.22656
3.39063	6138.12	4	790.975	452.58	363.533	0.15156	0.22031	0.17656
3.04688	3689.52	3	503.973	403.64	253.556	0.17969	0.16719	0.19844
3.90625	2976.56	3	583.649	254	163.422	0.22813	0.15938	0.22188
3.5625	2896.1	3	533.64	270.98	171.735	0.15625	0.15625	0.15625



To correctly estimate the direct causal effect between ***Window Area*** and ***Heating Load***,

- *Ground Floor Area*
- *Floor Height*
- *Number of Floor*
- *WWR*

should be controlled.

- *Chen, X., Abualdenien, J., Singh, M. M., Borrmann, A., & Geyer, P. (2022). Introducing causal inference in the energy-efficient building design process. Energy and Buildings, 277, 112583. https://doi.org/10.1016/j.enbuild.2022.112583*

Key takeaways

- A systematic review of performance gaps and uncertainties in problem formalization in the field of engineering.
- Knowledge decomposition paves the path toward knowledge-integrated machine learning - a three-level ladder of integration paradigms.
- Reconciling first-principles simulation and data-driven methods contributes to effective engineering solutions.

Thank you!
Questions?

Nachhaltige Gebäudesysteme

Wechat

Personal page