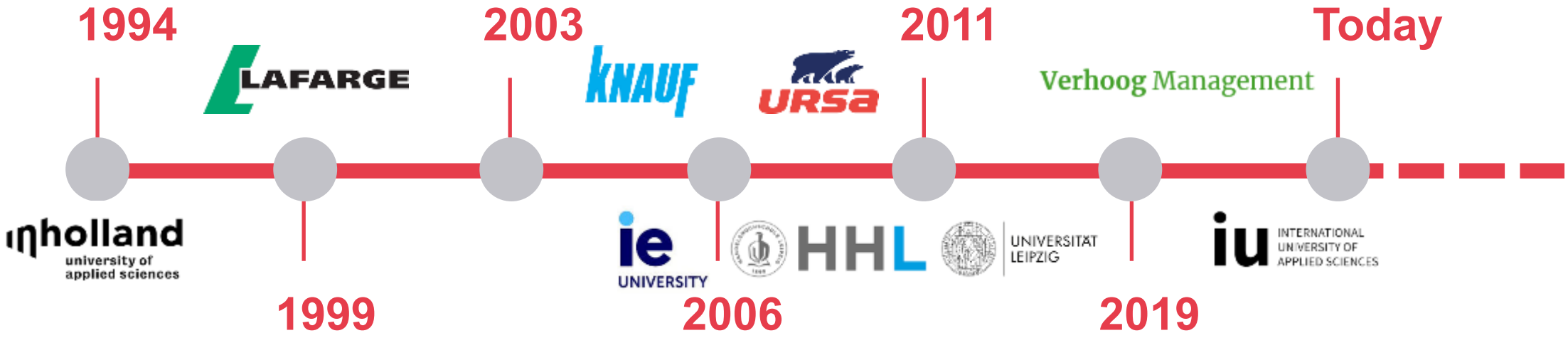
PROF. DR. MART VERHOOG

Traditional Statistics and Machine Learning in SNA: A Comparative Reanalysis of Empirical Data

The 17th International Conference on Advances in System Modeling and Simulation SIMUL 2025

PROF. DR. MART VERHOOG

- Worked for over 15 years as a marketing consultant and held various roles in the building materials industry across the Netherlands, Germany, and Belgium.
- Holds a Bachelor in Economics, an MBA in General Management, and a PhD in Marketing from HHL Leipzig; completed Executive Education at IE Madrid.
- Led the Research Centre for Municipal Energy Economics (FKE) at Leipzig University before taking up a professorship at IU International University of Applied Sciences
- Continues to work as an independent consultant and lives near Bonn, Germany with his wife, two children, and a dog.



"Prediction vs. explanation is a long-standing unresolved tension (Breiman, 2001; Radford & Joseph, 2020; Raeini et al., 2020)."

"Prediction vs. explanation is a long-standing unresolved tension (Breiman, 2001; Radford & Joseph, 2020; Raeini et al., 2020)."

"In social network research, the tension between prediction and explanation remains largely unresolved. We attempt a head-to-head in the domain of energy transition networks."

MOTIVATION AND PROBLEM STATEMENT

- Many social scientists rely on regression (e.g. OLS, logistic regression) because of their interpretability, theory alignment, and familiarity.
- However, real-world networks often display **nonlinearities**, interactions, threshold effects, and structural dependencies that traditional models may miss.
- Machine Learning (ML) methods (decision trees, random forests, gradient boosting, neural nets) may better capture complex patterns *but* at the cost of opacity.
- **Gap:** In applied social network research (especially in domains like energy transitions), there is little systematic comparative work showing when and how ML "wins" over regression, or vice versa.

RESEARCH QUESTIONS

1. Network conditions – when do ML models achieve higher predictive accuracy than regression models?

2. 2. Feature relevance – which network features emerge as most influential in ML compared to regression?

3. 3. Interpretability – can tools, such as SHAP reconcile predictive accuracy with explanatory clarity in applied SNA for energy transition research?

EMPIRICAL FOUNDATION

"This project builds on my earlier doctoral research, published as the book Controlling Actors and Decisions in Construction Networks (Springer, 2017).

Original study: ~700 household decisions on energy-efficient refurbishment analyzed with traditional statistical methods (regression, factor analysis). Provided insights into stakeholder roles, household attitudes, and network influence

Now:
We re-examine the same empirical data using Machine Learning methods
Goal: uncover additional structural patterns that may have remained hidden to traditional inference.

DATA AND EMPIRICAL CONTEXT

- We use the dataset from Verhoog (2017) with ~700 household decisions regarding energy-efficient refurbishment.

- Variables include:
  - Household socio-demographic features (income, education, dwelling type)
  - Physical building features (insulation, age, heating systems)
  - Social network measures: degree, betweenness, closeness, tie strength, density, interaction intensity
  - Institutional variables: involvement of experts, consultants, banks, subsidy programs

- The goal: predict whether a given household undertakes energy retrofit, or the magnitude of retrofit, conditional on network embedding.

COMPARATIVE METHODOLOGICAL PIPELINE

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

1	2	3	4	5	6
Preprocessing	**Feature Engineering**	**Regression Models**	**ML Models**	**Validation**	**Evaluation**
Impute missing values, normalize, encode categorical variables	derive network metrics (degree, betweenness, closeness, density, tie strength)	linear and logistic regression as transparent baselines	Random Forests, Gradient Boosting, Support Vector Machines, optional Neural Nets	cross-validation with consistent train/test splits; hyperparameter tuning via grid search	predictive performance (R^2, ROC-AUC) and interpretability with SHAP values

EXPECTED CONTRIBUTIONS & IMPLICATIONS

- A systematic benchmark of regression vs ML in social network–embedded decisions (energy transitions)
- Insights into when ML is worthwhile vs when regression suffices
- Demonstrating the power of interpretability tools to “open the black box”
- Toward a hybrid approach: theory-driven models enriched by data-driven methods
- Practical guidance for social scientists or policy modelers: *“If your network is sparse and variables are few, regression is likely enough. But when complexity arises, ML + SHAP bridges both worlds.”*
- Broader relevance: transferable to domains like diffusion, adoption, organizational networks, etc.

CHALLANGES, LIMITATIONS AND NEXT STEPS

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Challenges:

- Overfitting risk in ML (especially with n ~700)
- Multicollinearity and feature redundancy
- Dependence across observations (network autocorrelation)
- Interpretability limits (some interactions remain obscure)

Limitations:

- Single cross-sectional dataset
- Sample size vs. model complexity

Next steps:

- Prepare dataset
- Build comparative pipeline (regression vs ML)
- Benchmark performance & interpretability
- Report results & submit full paper

SUMMARY & TAKE-HOME MESSAGE

- Regression and ML offer complementary strengths: explanation vs. Prediction
- Interpretability tools (e.g., SHAP) can help bridge the gap
- Our contribution: an empirical roadmap for researchers facing this methodological dilemma
- Within the limits of a single, cross-sectional dataset