“ML Oops”: How data simulation can help your quants avoid modeling errors

Michel Debiche
Director of Analytics Research, STAC

michel.debiche@STACresearch.com

Start with an example from this study

"Why Should I Trust You?" Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro
University of Washington
Seattle, WA 98105, USA
marcotcr@cs.uw.edu

Sameer Singh
University of Washington
Seattle, WA 98105, USA
sameer@cs.uw.edu

Carlos Guestrin
University of Washington
Seattle, WA 98105, USA
guestrin@cs.uw.edu

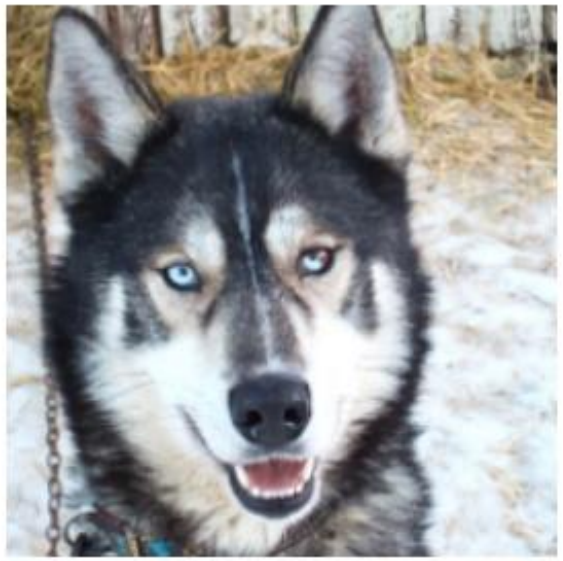
arXiv:1602.04938v3 [cs.LG] 9 Aug 2016

Husky or wolf?

Source of this slide: Alexiei Dingly, https://becominghuman.ai/its-magic-i-owe-you-no-explanation-explainableai-43e798273a08

STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

Model explanation: snow!



(a) Husky classified as wolf

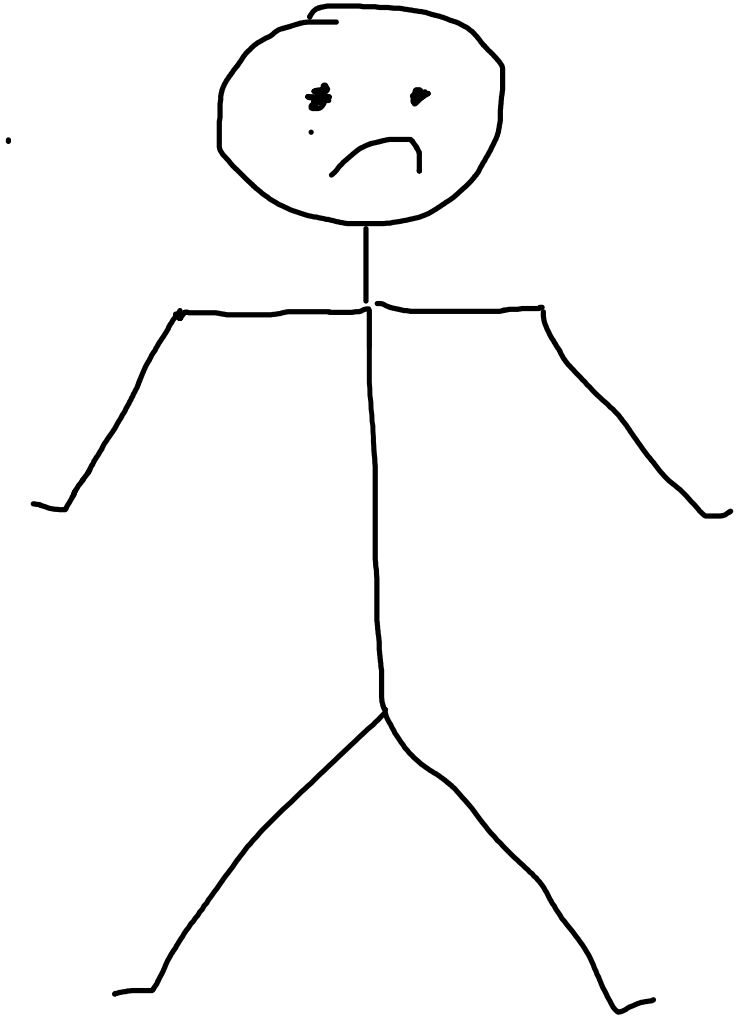
(b) Explanation

Symmetric errors vs. asymmetric risk

Husky or Wolf?

Says wolf instead of Husky → Opportunity cost (avoids Husky)

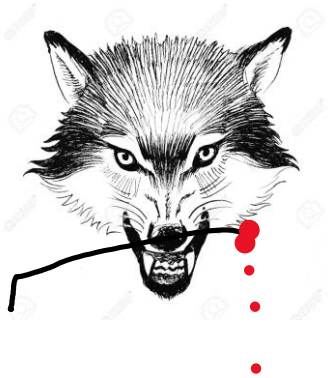
Sad!

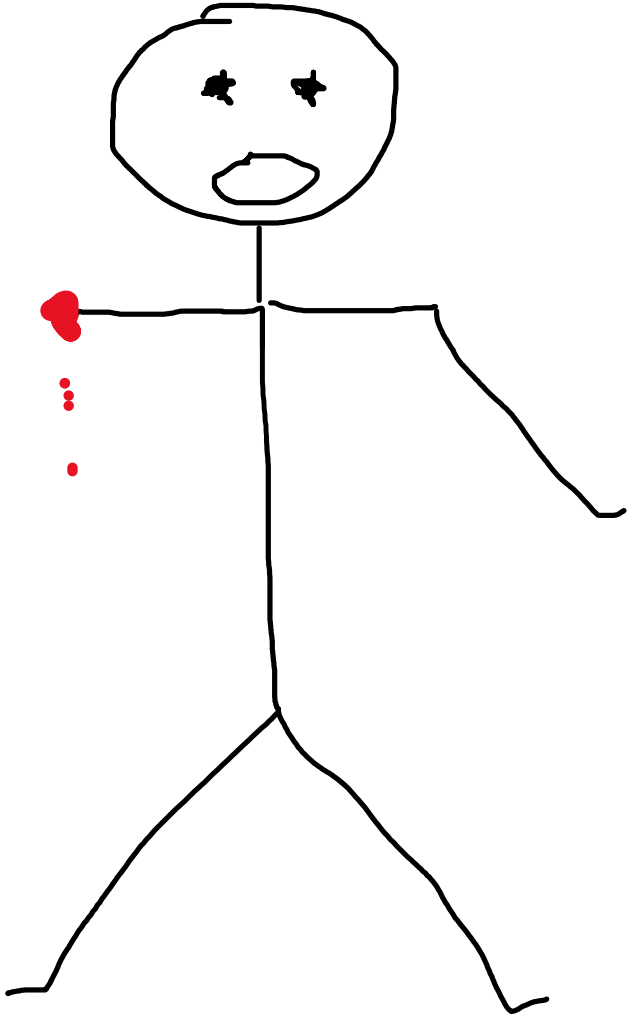


STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

Says Husky instead of wolf → Realized loss (tries to pet wolf)

Bad! Loses 25% of limb portfolio.

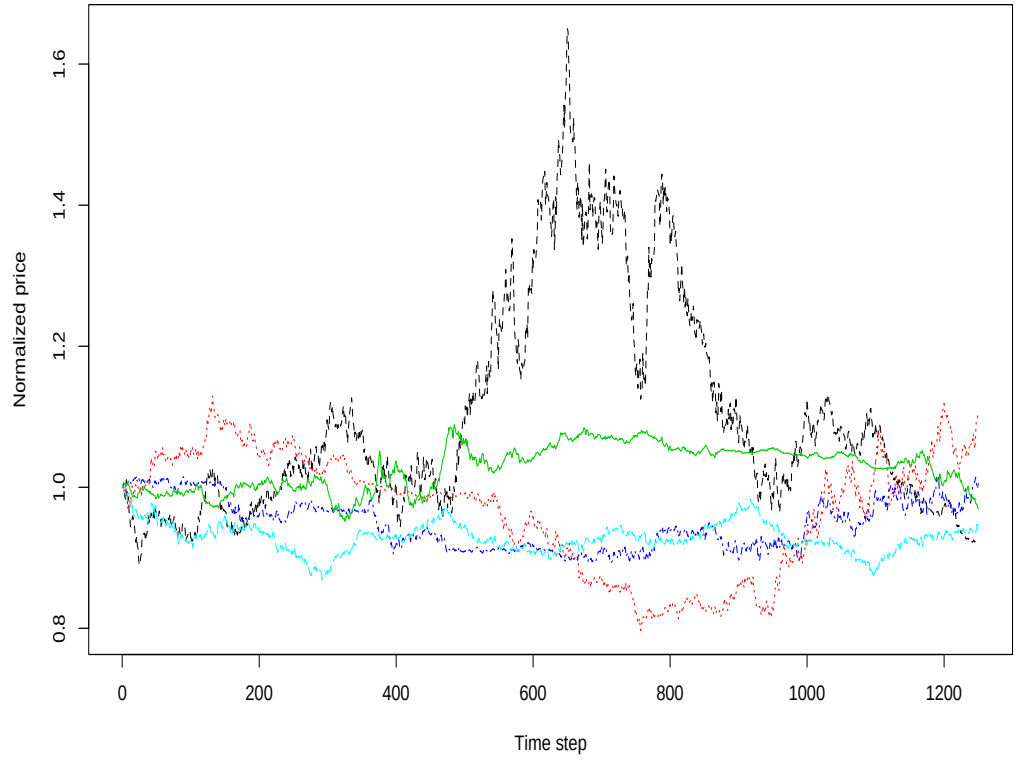




STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

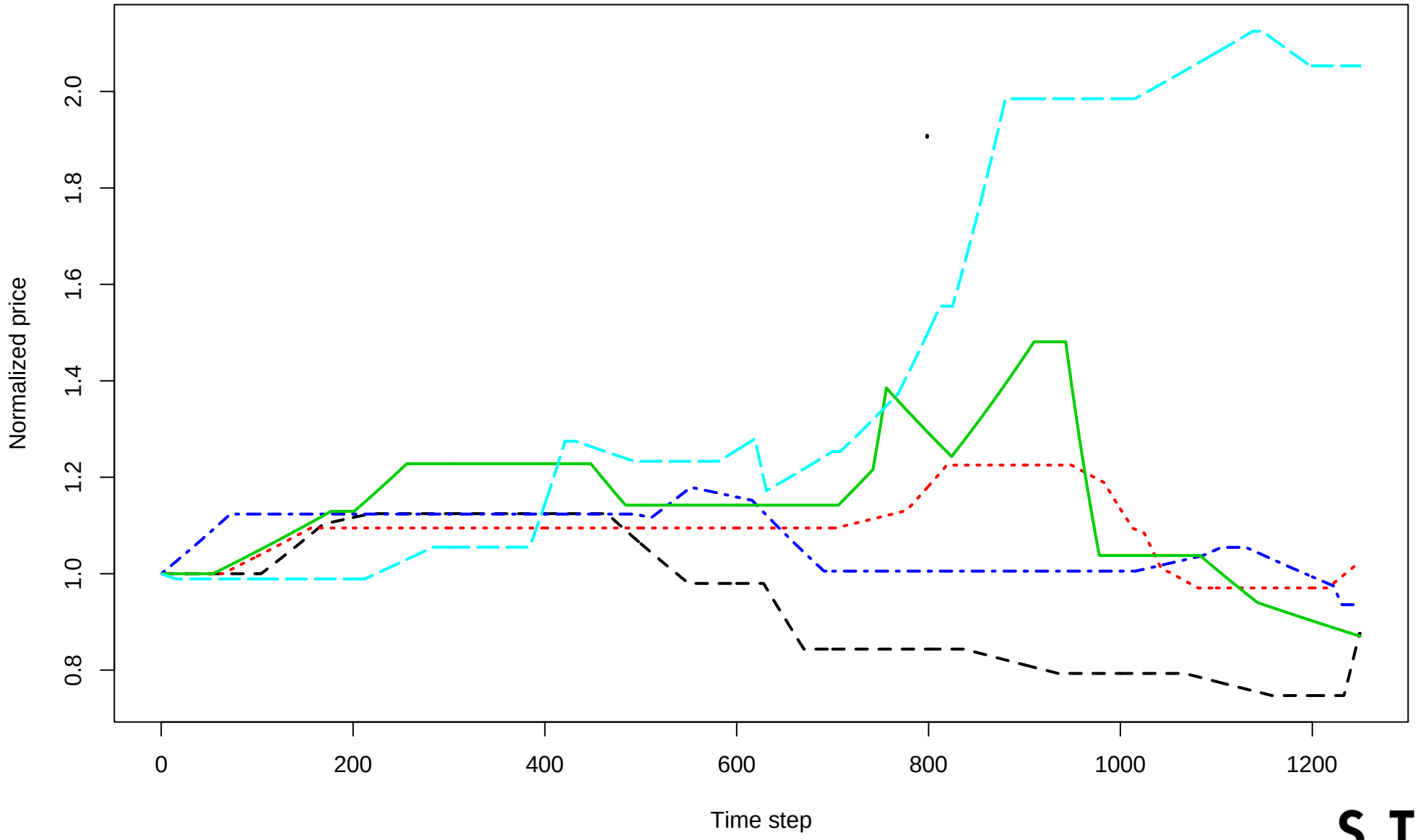
Exploring multivariate time series models

- What would be the equivalent to “seeing” only the snow in multivariate time series?
- We will explore multivariate time series modeling using simulated data
- Goals:
  - Appreciate what is involved in trying to “explain” such models
  - Understand the potential of using simulated data to understand and test models of all kinds



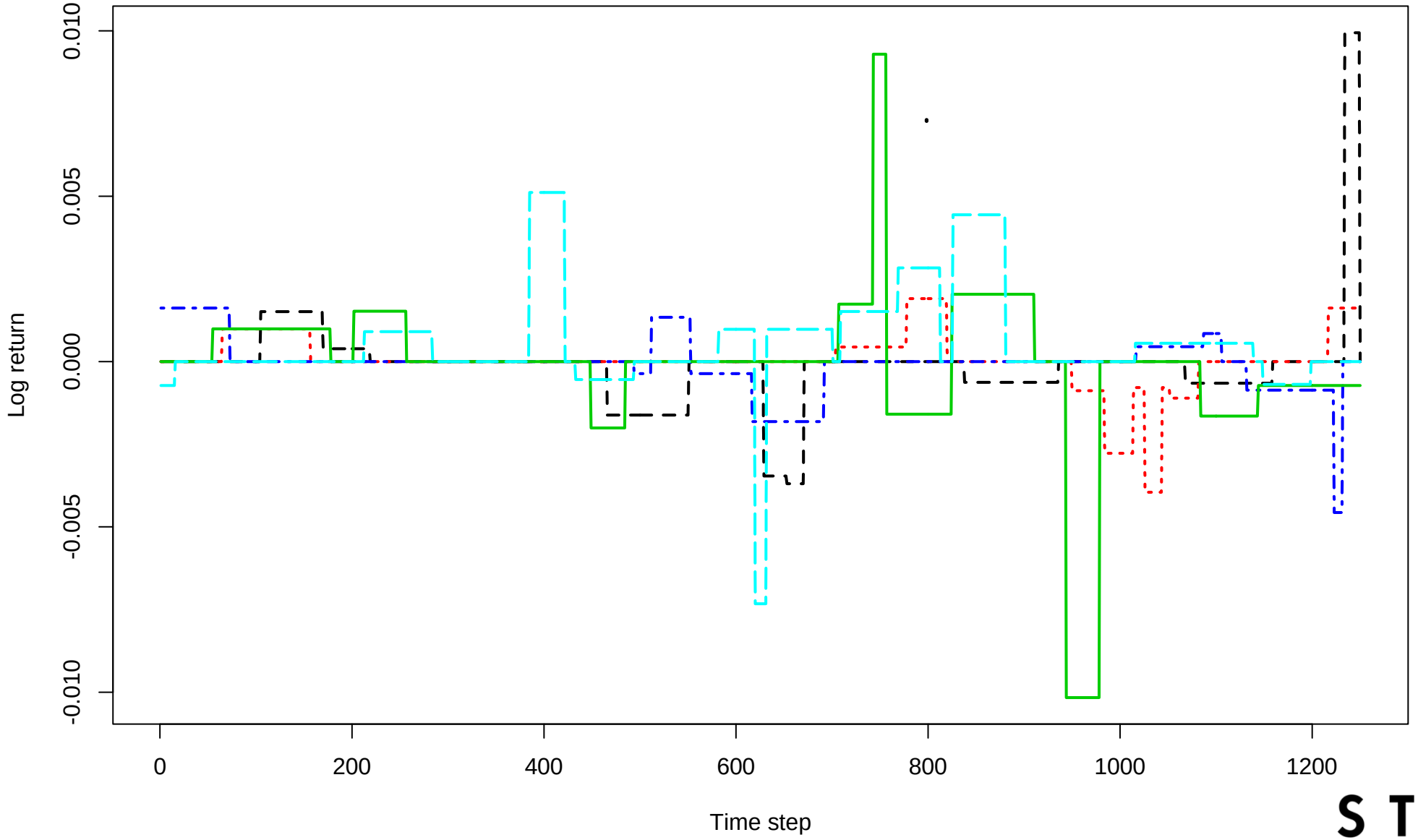
STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

Generated signals (normalized price paths)



STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

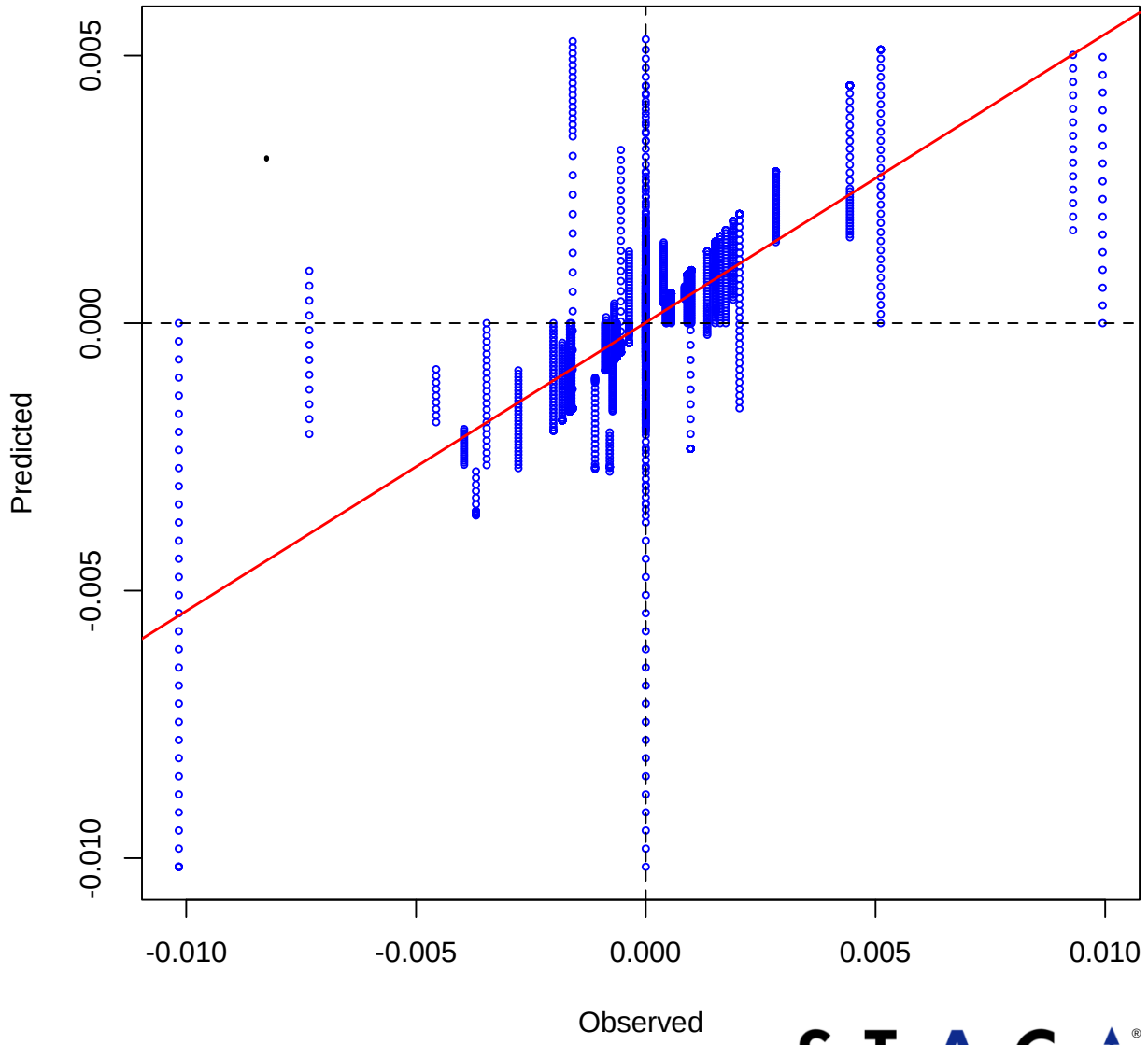
Generated signals (log returns)



STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

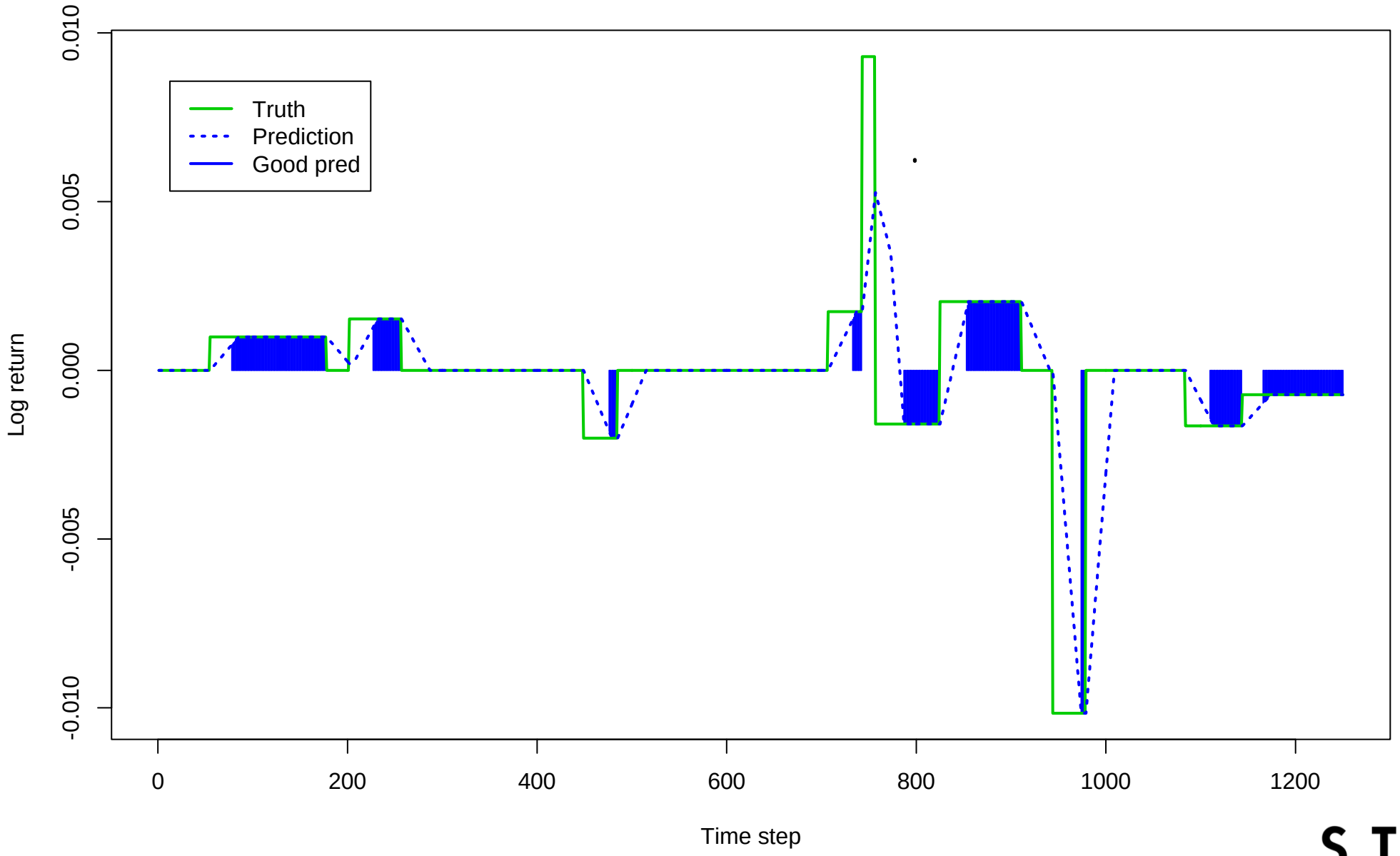
A very simple model

- Predicted value = average of last 30 values
- Regress predicted value vs. observed value at next time step
- R-squared for these generated signals is 46%



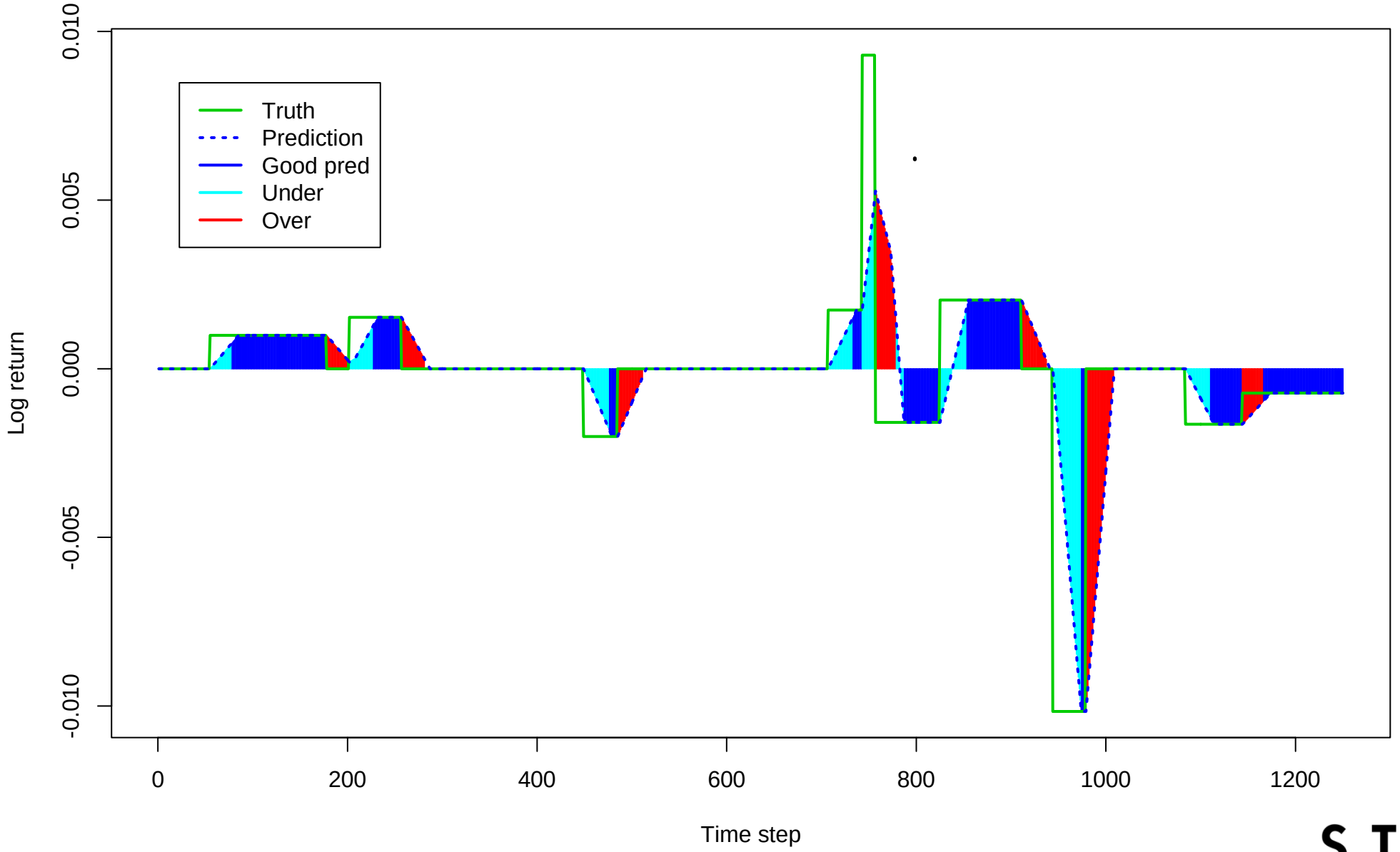
STAC
SECURITIES TECHNOLOGY ANALYSIS CENTER

Generated signal vs. model predictions



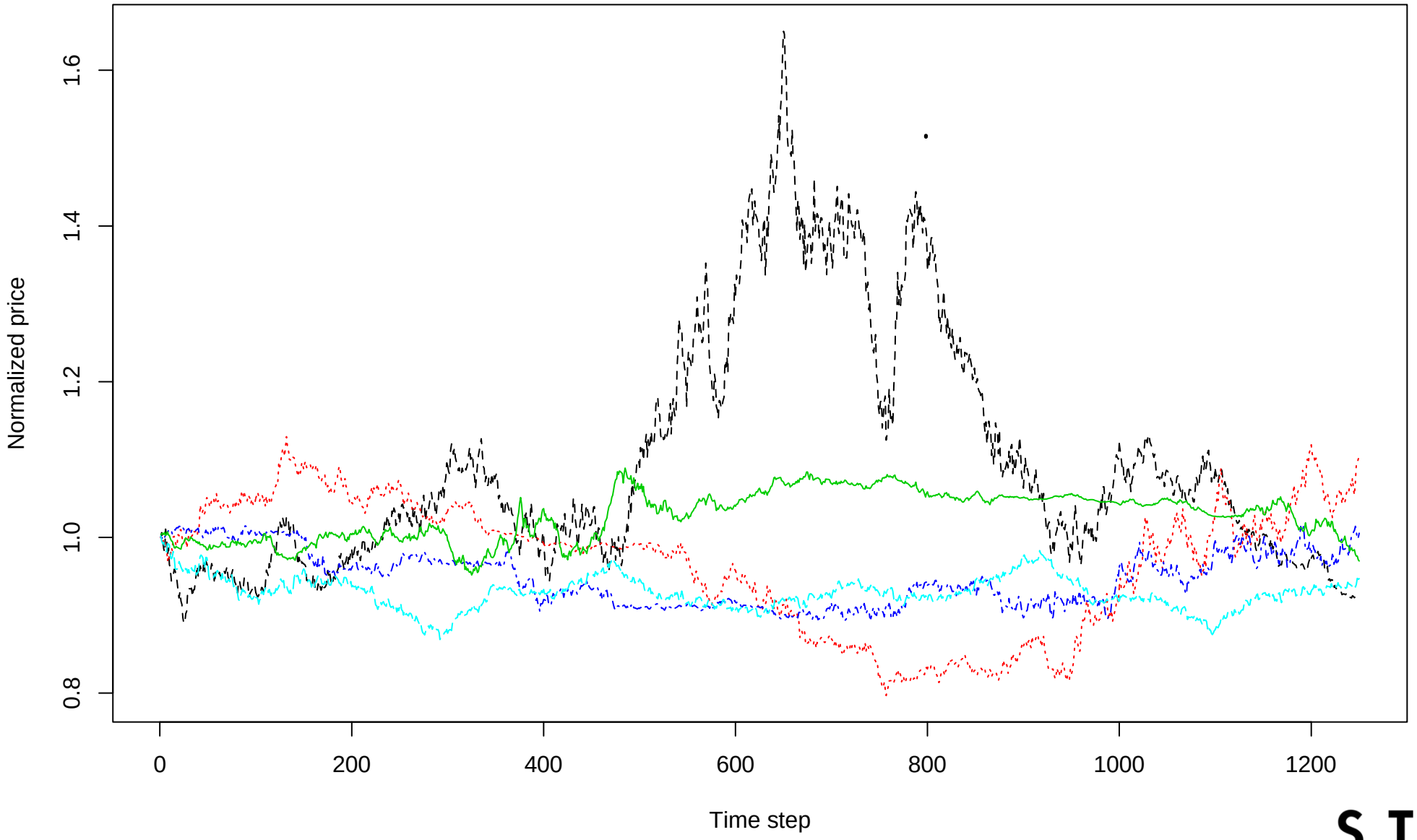
STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

Generated signal vs. model predictions (cont'd)



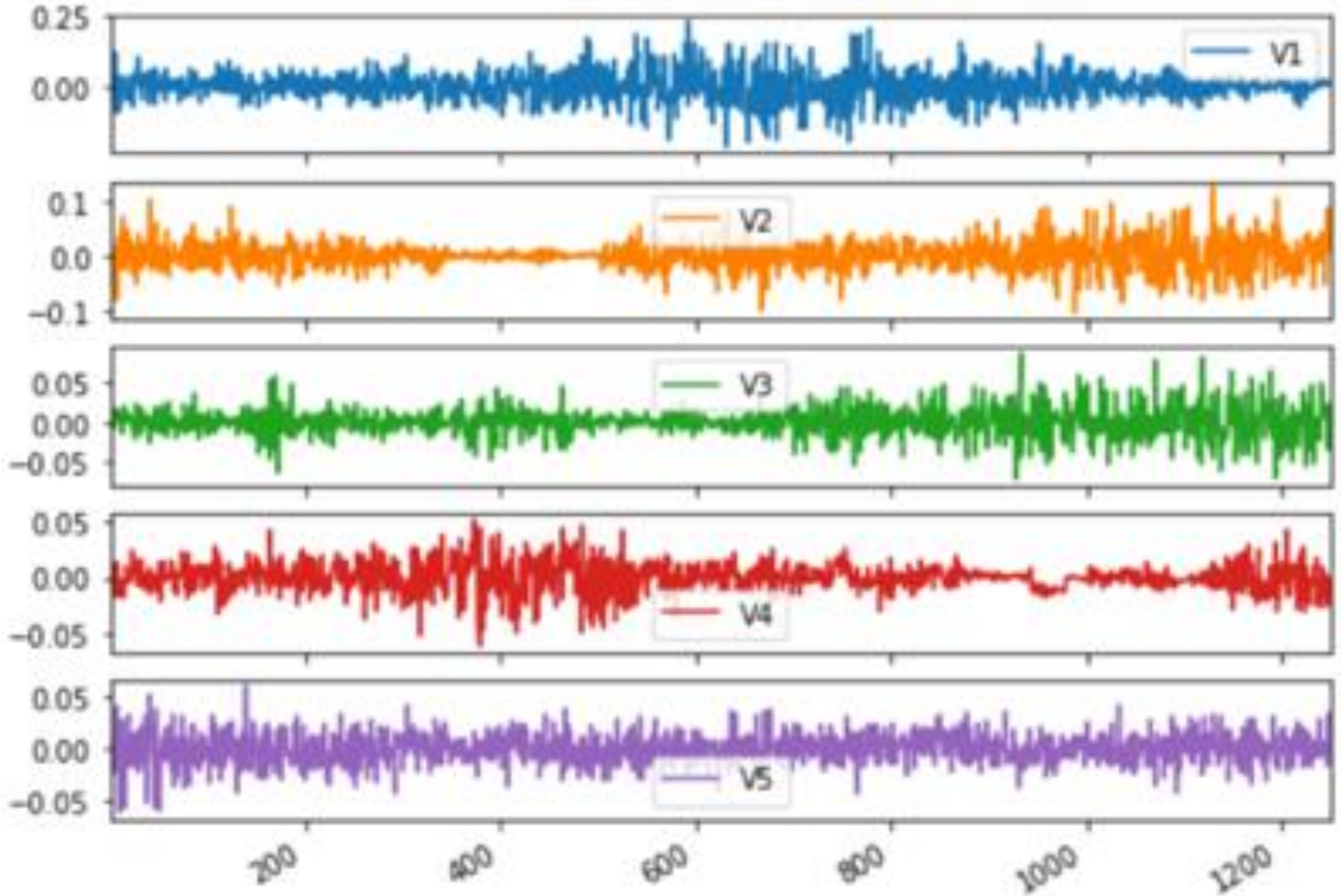
STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

Generated noise (normalized price paths)



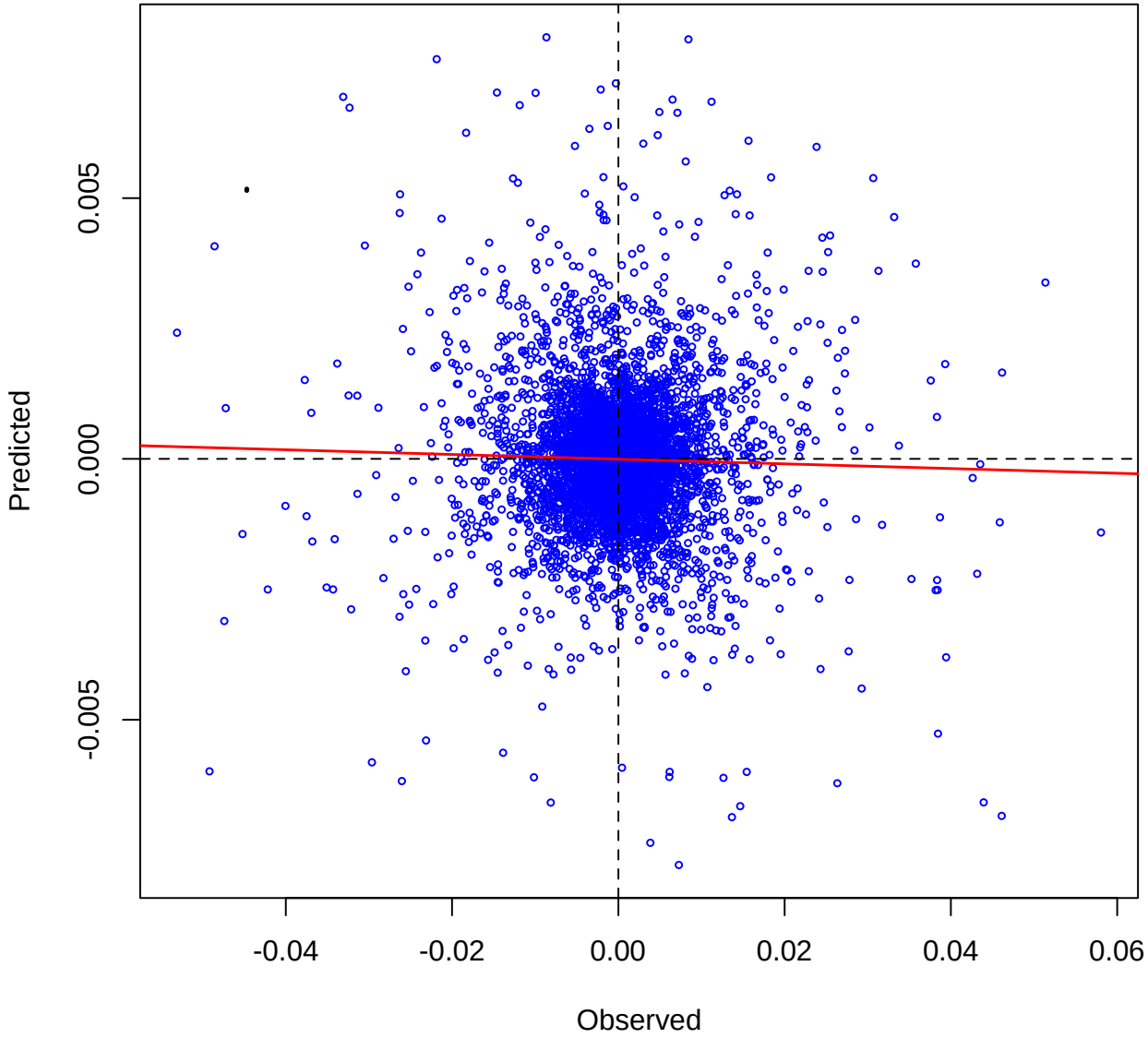
STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

Generated noise (log returns)



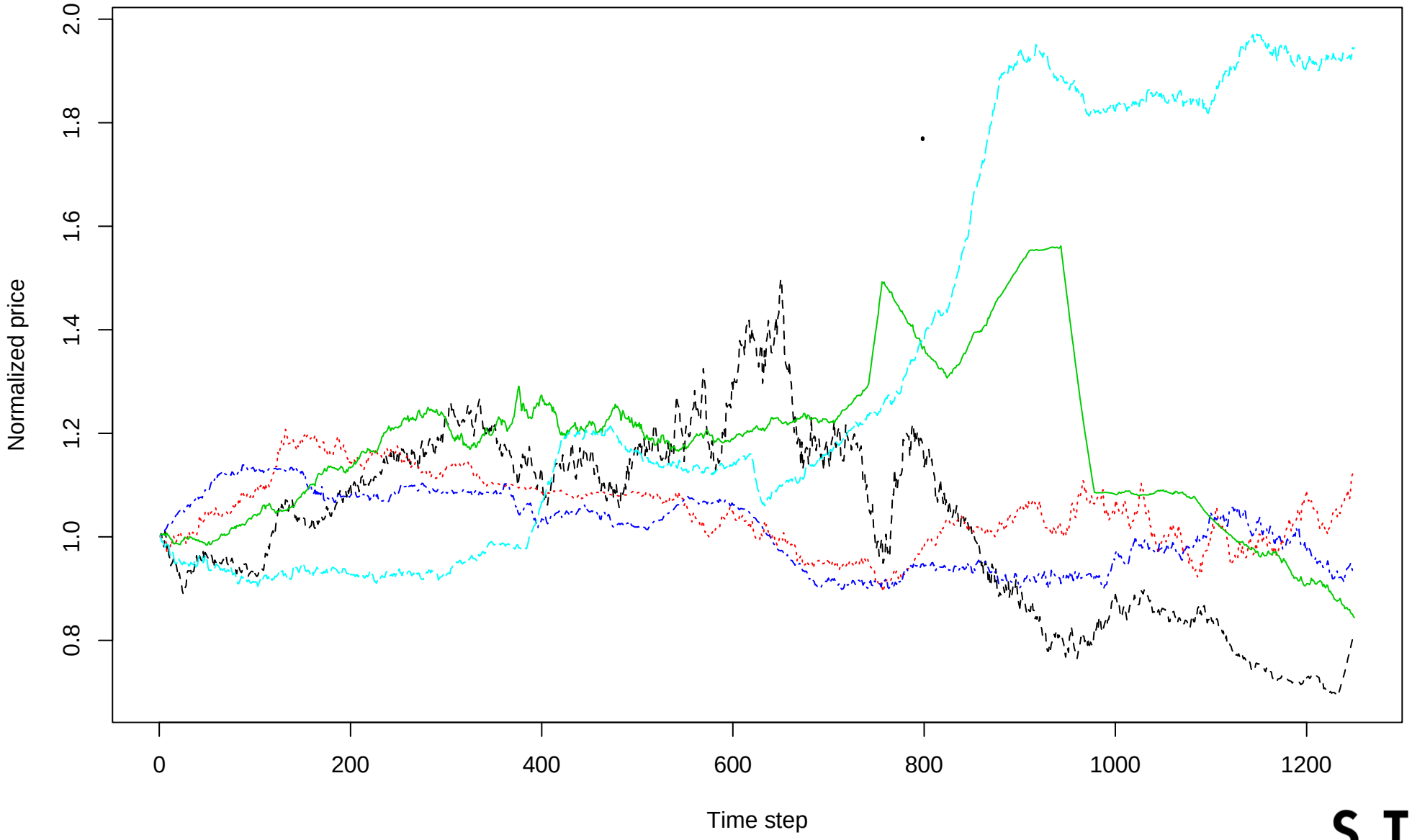
Try to model the noise

- R-squared is only 0.07%
- Yes, it's noise!



STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

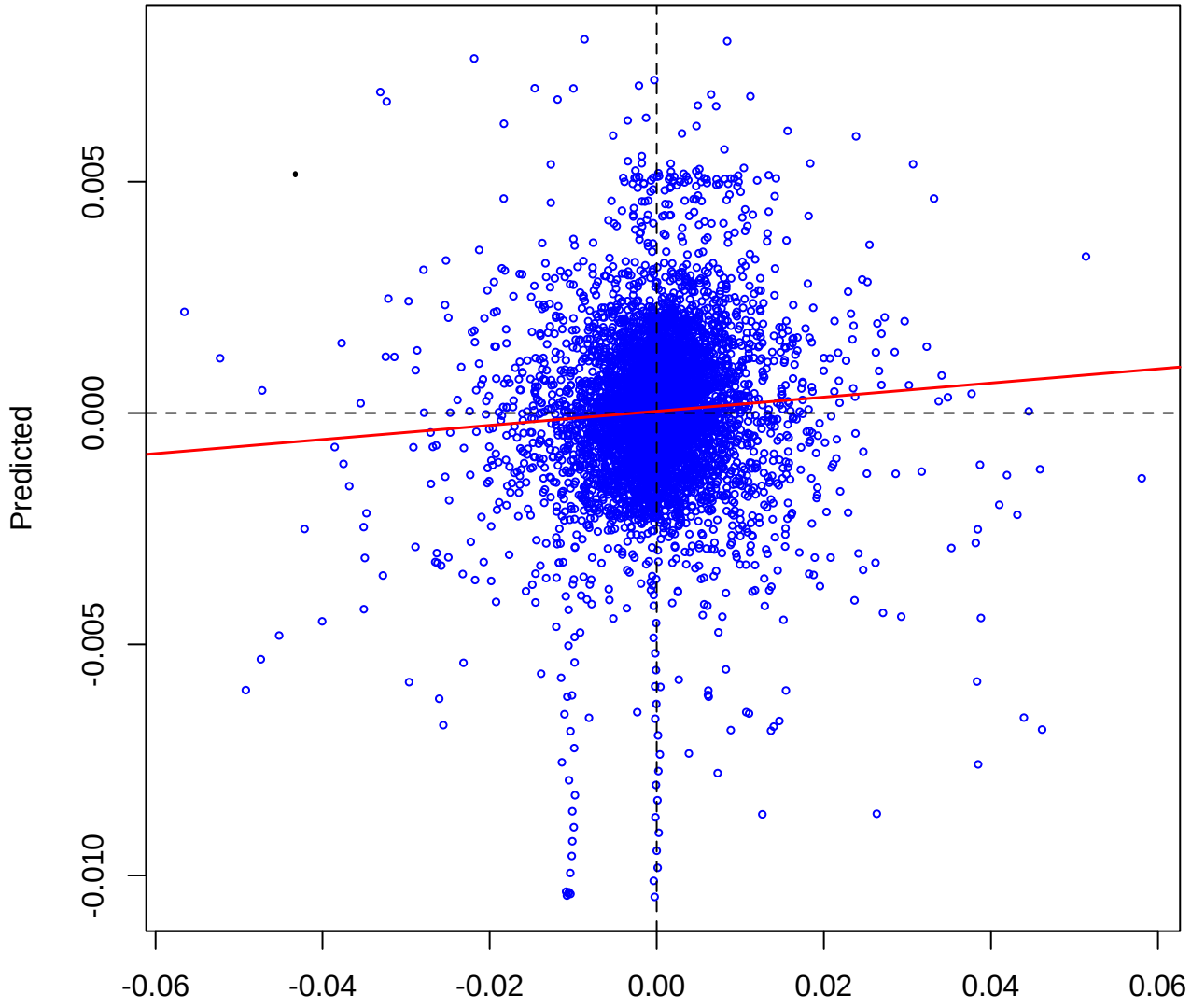
Simulated data (signal + noise)



STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

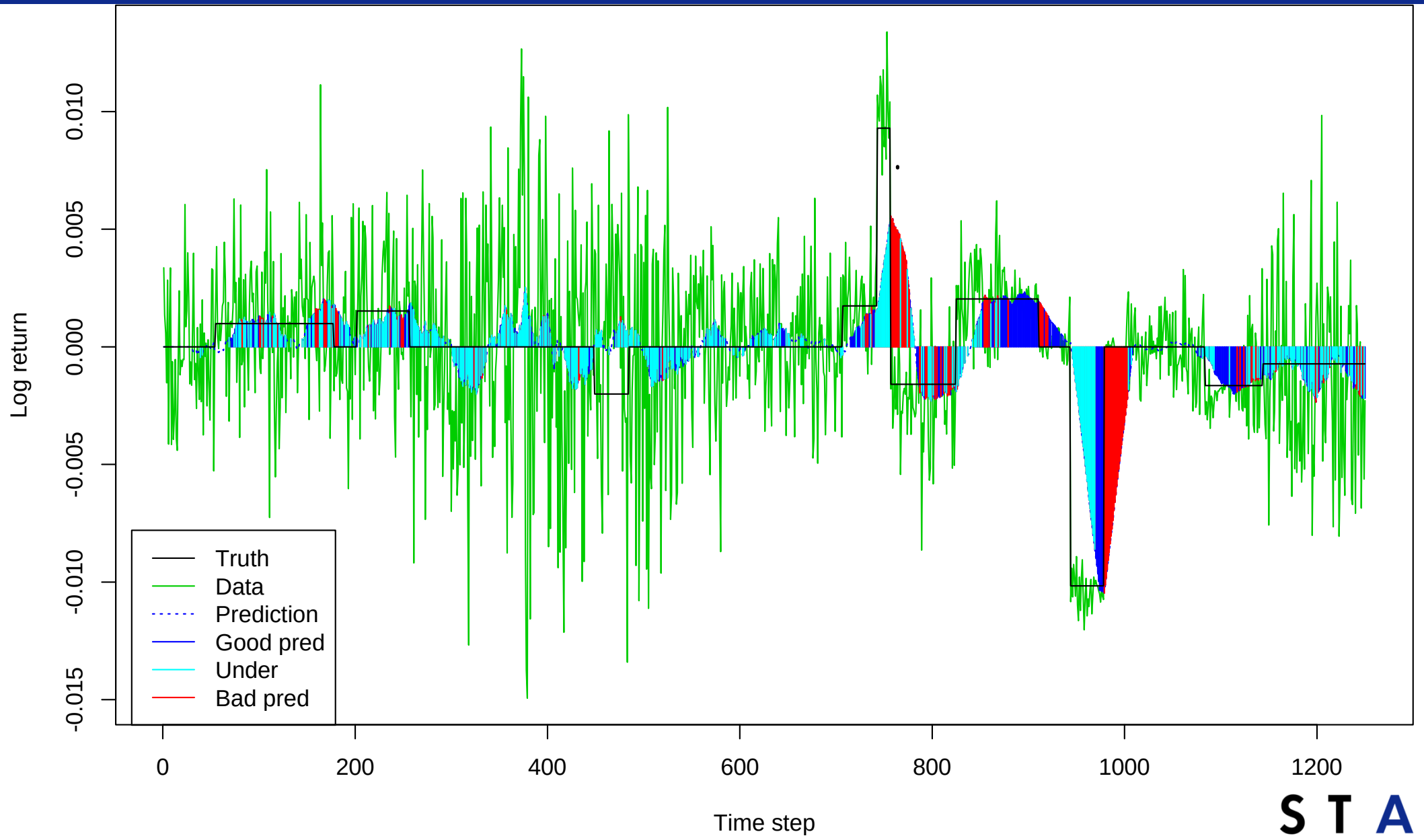
Model the simulated data

- R-squared is 0.45%
- The fit is statistically significant



STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

Truth, noisy data and predictions



STAC®
SECURITIES TECHNOLOGY ANALYSIS CENTER

Some ways to tempt or torture models with simulated data

- Bull market -> overfitting to random patterns (superstition)
- Single dominant symbol
  - E.g. Signal in one symbol, only noise in others -> how many false positives?
- Time window with perfect correlation across variables (proxy for market crash)
- Symbols with scaling errors
- Symbols with all zeros
- Time windows with all zeros
- All variables driven by correlated multifactor model + noise
- Non-linear signals (e.g. jump up or down + rebound)

Challenges in explaining multivariate time series models

- Features typically include functions of sliding windows
- Features from overlapping windows are not independent
- Features may be correlated
- For dense data, number of features rises rapidly
- General methods exist for trying to assess the importance of features in models
- These require extensive computation or extensive manipulation of data or both
- The explanation methods themselves have to be tested (for example, with simulated data)
- Results may be hard to display or visualize

Conclusions

- Simulating data provides insights into both data and models
  - Variations on signal type, distribution, density, strength and continuity highlight sensitivities and vulnerabilities of the model
  - Likewise for attributes of background market “noise”
- Models should be *routinely* tested against signal and noise patterns known to be challenging
  - This should be built into operational architectures
- Explaining opaque models such as Deep Learning is active research, but:
  - Business, compliance, regulators will require it
  - Will need to be built into operational architectures as well
  - Will most likely require enormous resources (compute, memory and/or I/O)
  - These workloads may behave differently from both training and inference
  - May require data simulators to test the explainers
- Data can be generated in interesting ways; e.g. multiple agents