Universal Approach to Science Time Series: Deep Learning on Hydrology

Junyang He
Dr. Geoffrey Fox

Background

AI for Science

- Traditional science is driven by theories and differential equations. Predictions are made based on these established formulas.
- New approach to science is driven by data. Predictions are made based on models that rely entirely or heavily on data. DL is used to learn the "hidden variables" and to find the complex formula behind a phenomenon.

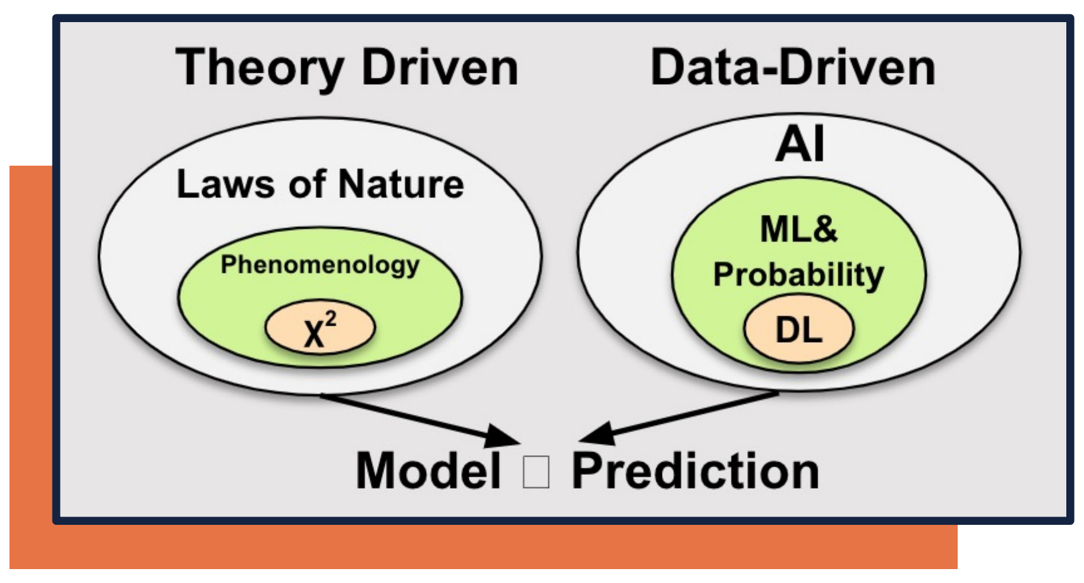

Figure #1. Theory driven vs. data-driven science

Science Time Series (Spatial Bag)

- Science time series involves groups of time series data collected at different geographic locations with different static data.
- In Hydrology, data includes a collection of 671 time series collected at 671 different catchments that vary by geological, soil, and climate characteristics.
- The Spatial Bag allows future forecasting in the y-direction and sequence-to-sequence mapping in the x-direction.

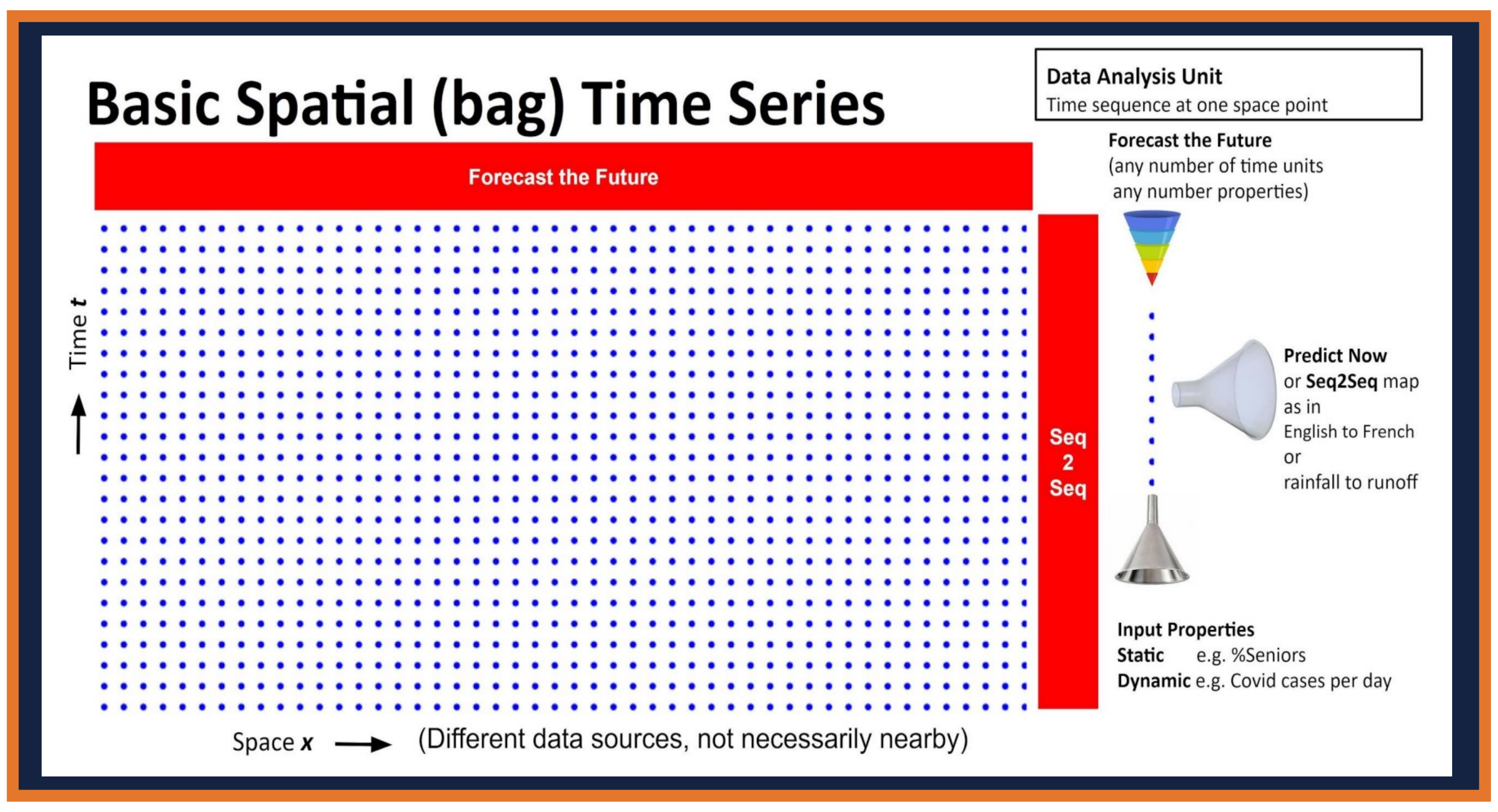

Figure #2. Spatial bag structure.

Future Work

Science Transformer and TFT Model

- Set up Science Transformer and Google's TFT model with the same Hydrology training data from US, UK, and Chile.
- Compare model efficiencies between LSTM and the two Transformer models widely used to deal with sequence-to-sequence mapping in NLP.

Transfer Learning

- Utilize structure of the science time series to perform sequence-to-sequence learning.
- Train model with CAMELS data from US and predict Hydrological characteristics of UK and Chile with local static variables.

References

- G. Fox, J. Rundle, A. Donnellan, and B. Feng, "Earthquake nowcasting with Deep Learning," *arXiv.org*, 18-Dec-2021, https://arxiv.org/abs/2201.01869.
- Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: The CAMELS data set: catchment attributes and meteorology for large-sample studies, Hydrol. Earth Syst. Sci., 21, 5293–5313, https://doi.org/10.5194/hess-21-5293-2017.
- D. Feng, K. Fang, and C. Shen, "Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales," arXiv [cs.LG], 18-Dec-2019, http://arxiv.org/abs/1912.08949.
- Kratzert, Frederik, "CAMELS Extended Maurer Forcing Data." https://www.hydroshare.org/resource/17c896843cf940339c3c3496d0c1c077/.

Goals

Overall Goal: Universal Approach to Science Time Series

- Establish the standard procedure to predicting science/spatial bag time series with Hydrology, Earthquake Nowcasting, and COVID models.

Current Goal: Build Hydrology Model for US, UK, and Chile

- Process input data from US, UK, and Chile to contain the same static and dynamic variables.
- Run LSTM and Science Transformer models for predictions in US, UK, and Chile.
- Compare LSTM vs. Science Transformer in predictions and model fit.

Current Work

Data Collection and Preprocessing

- Found Hydrology data for UK and Chile based the standards set by the CAMELS-US hydrometeorological forcing data taken from 671 catchments in the US.
- Removed static and dynamic variables that are not present in all three datasets and unified the format of the filtered data.

LSTM Model Training

- Normalized input time series, numerical and categorical static data.
- Set up LSTM model with Mean Square Error (MSE) as loss function and Normalized Nash-Sutcliffe Model Efficiency (NNSE) coefficient as model performance monitor.
- Trained the model with a sequence length of 21 days and an epoch size of 50.
- The LSTM network consisted of an input layer, a dense/MLP encoder layer, two LSTM layers, a dense/MLP decoder layer, and an output layer.

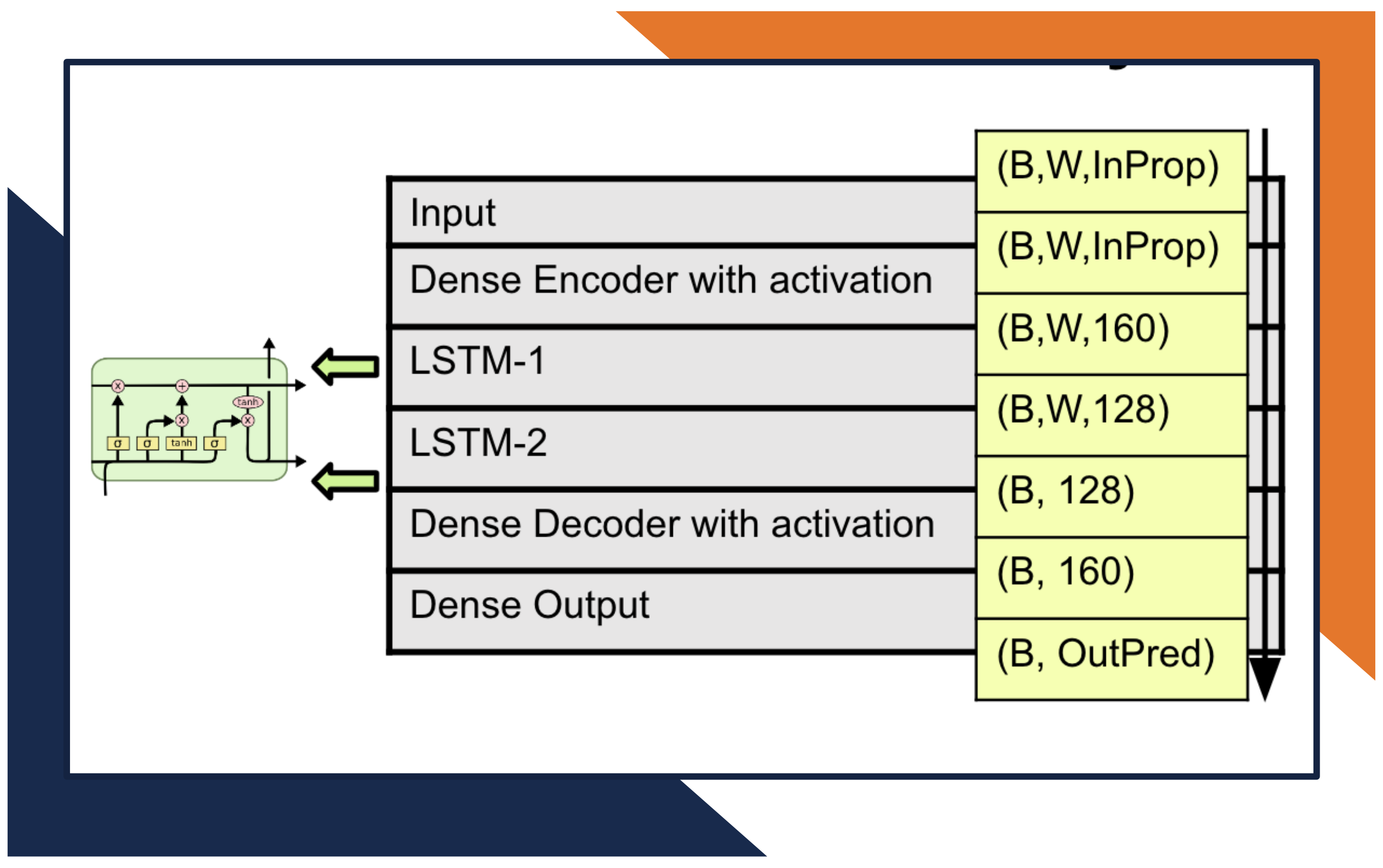

Figure #3. LSTM network architecture

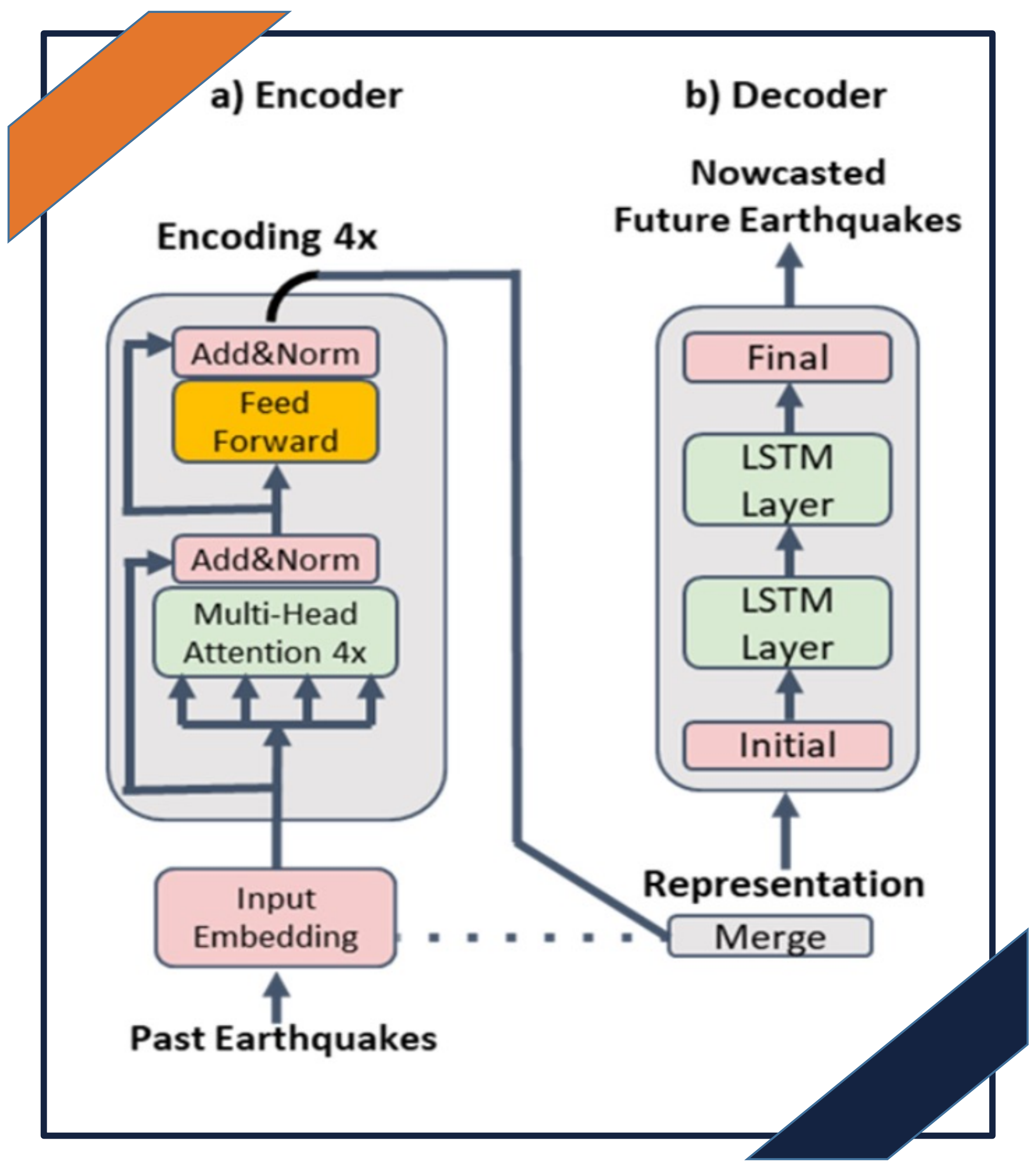

Figure #4. Science transformer architecture for Earthquake

Computing for Global Challenges

UNIVERSITY of VIRGINIA
BIOCOMPLEXITY INSTITUTE