Common Task Framework For a Critical Evaluation of Scientific Machine Learning Algorithms

Philippe M. Wyder, Judah Goldfeder, Alexey Yermakov, Yue Zhao, Stefano Riva, Jan Williams, David Zoro, Amy Sara Rude, Matteo Tomasetto, Joe Germany, Joseph Bakarji, Georg Maierhofer, Miles Cranmer, J. Nathan Kutz

DYNAMICS AI
NSF | AI INSTITUTE

COLUMBIA UNIVERSITY

SURF

UNIVERSITY OF CAMBRIDGE

POLITECNICO MILANO 1863

NEURAL INFORMATION PROCESSING SYSTEMS

No Common Standard

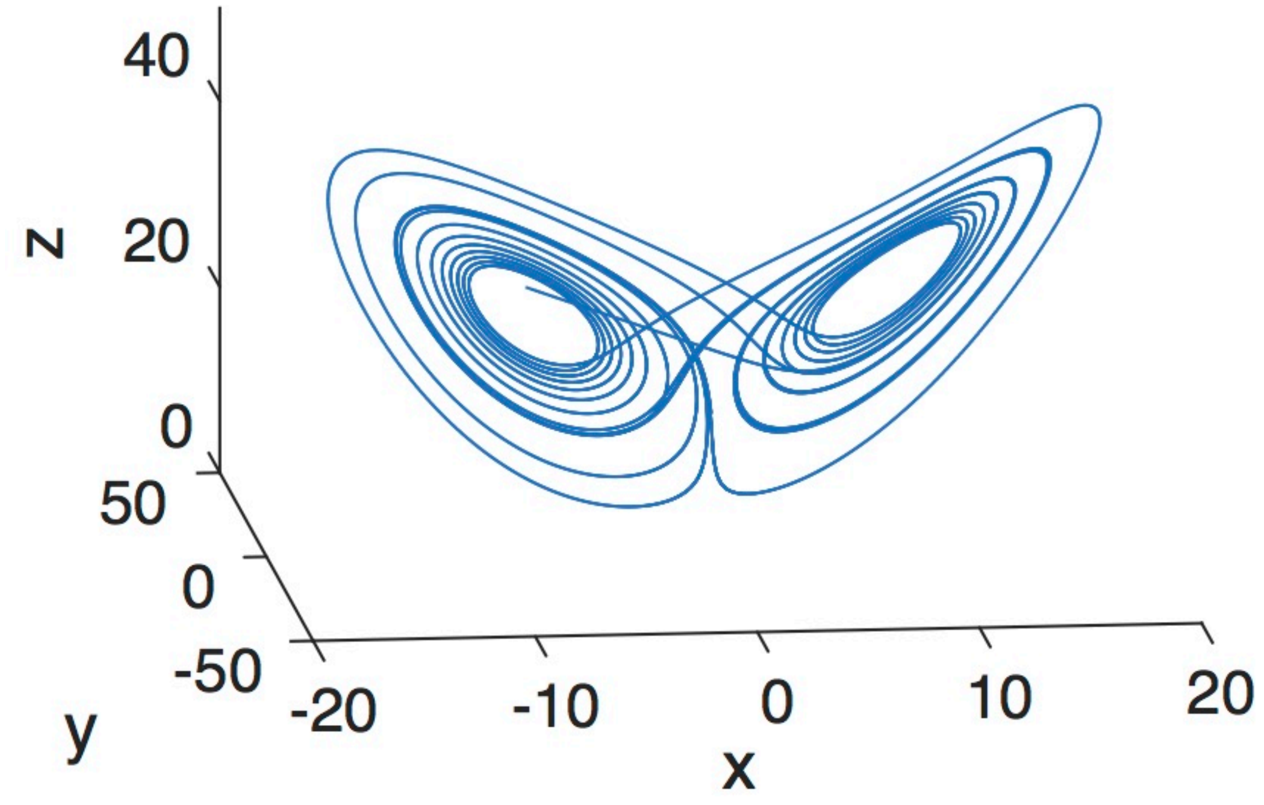
- The field is suffering from:
    - Weak Baselines
    - Reporting Bias
    - Inconsistent Evaluations
    - Lack of *hidden* test sets

Common Task Framework for Science

- We provide:
  - Curated set of datasets
  - Task specific metrics
  - Standardized Evaluation
  - Hidden test set

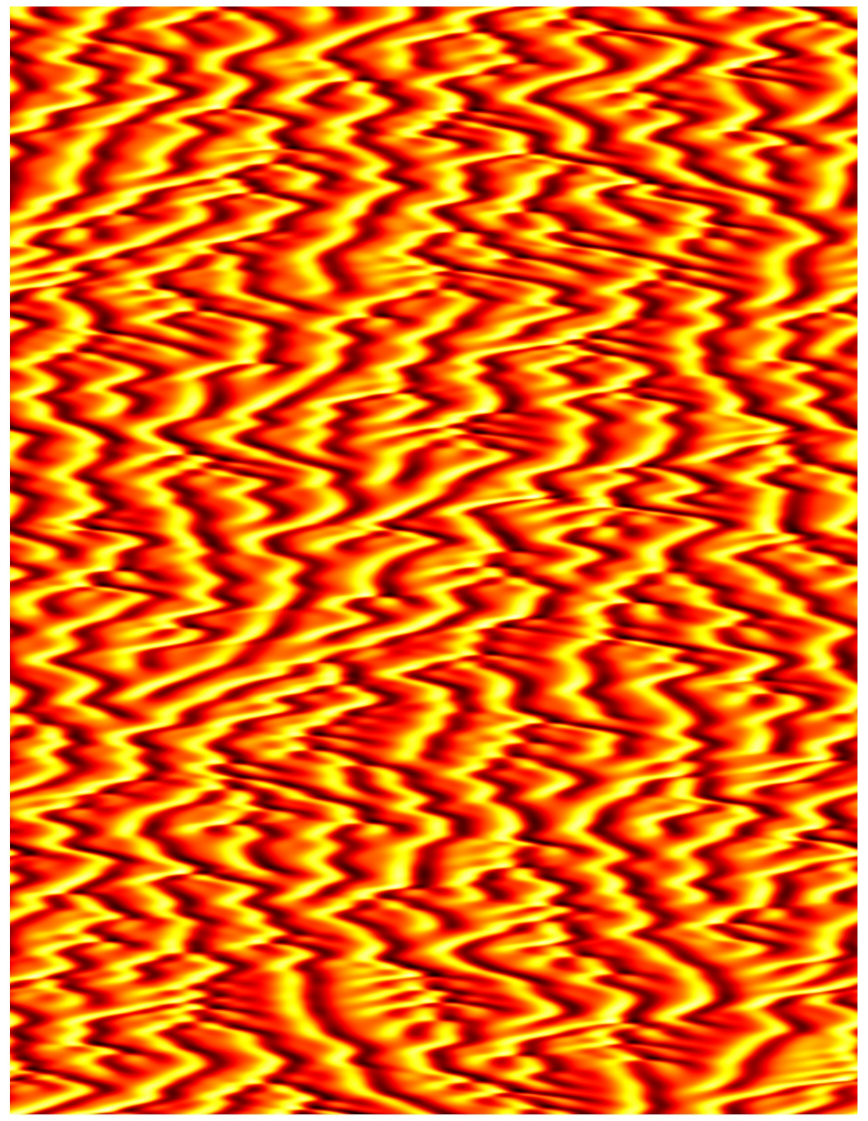
Datasets

Lorenz



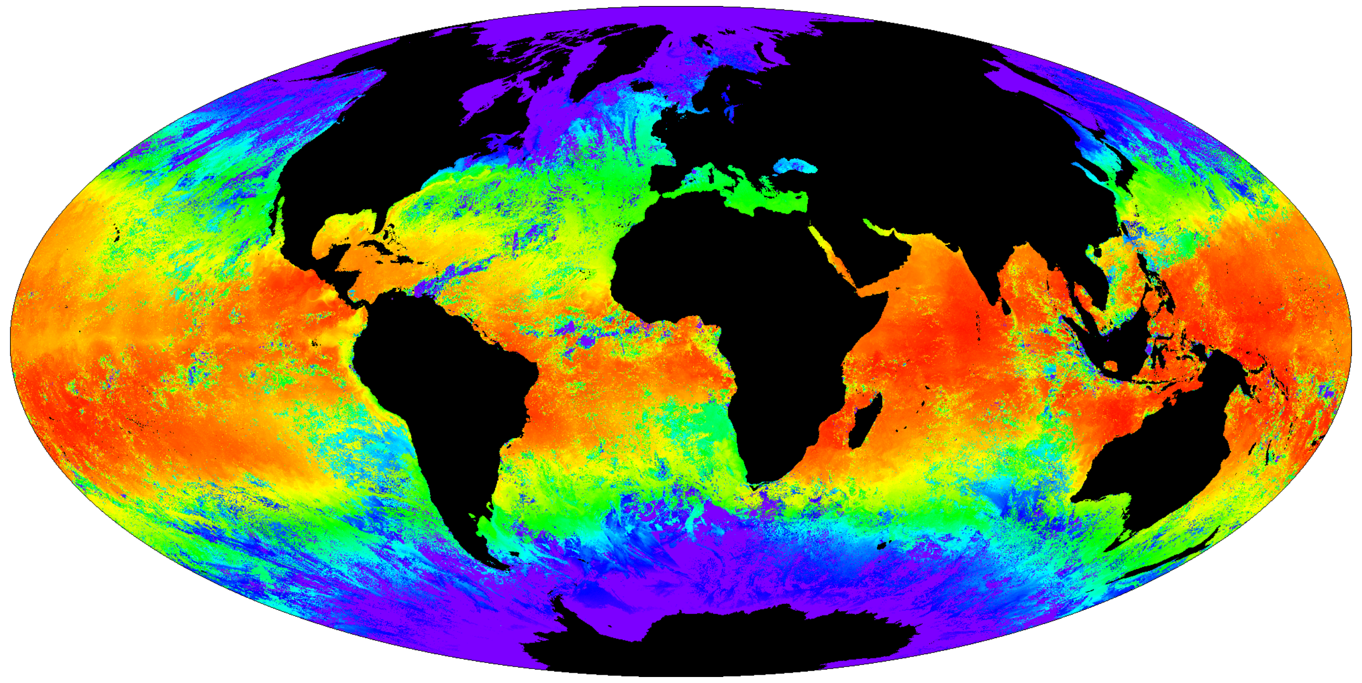
$$\mathbf{x}_j = \mathbf{x}(t_j) = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

Kuramoto-Sivashinsky



$$\mathbf{x}_j = \mathbf{x}(t_j) = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

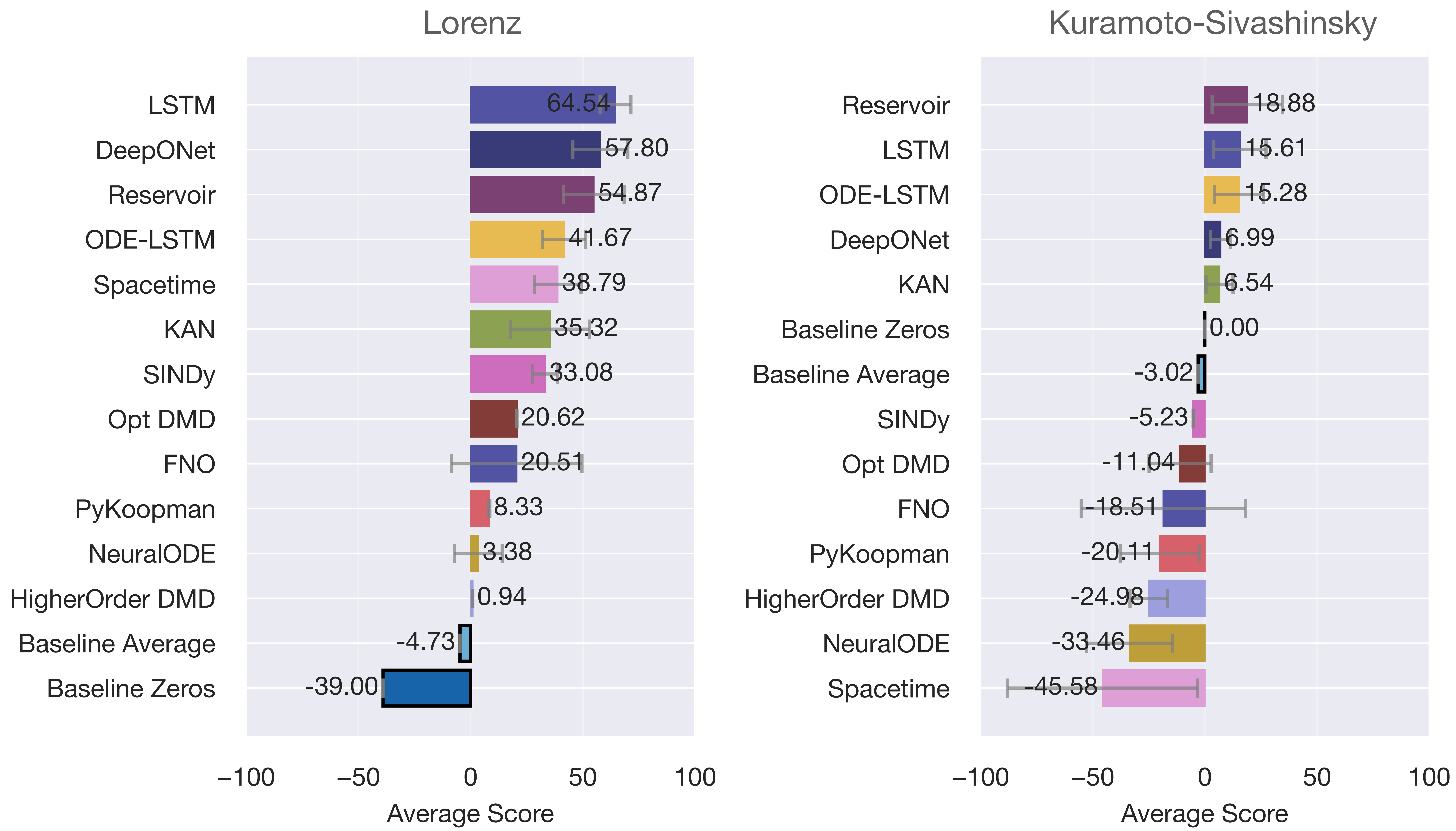
Sea Surface Temperature



More to come!

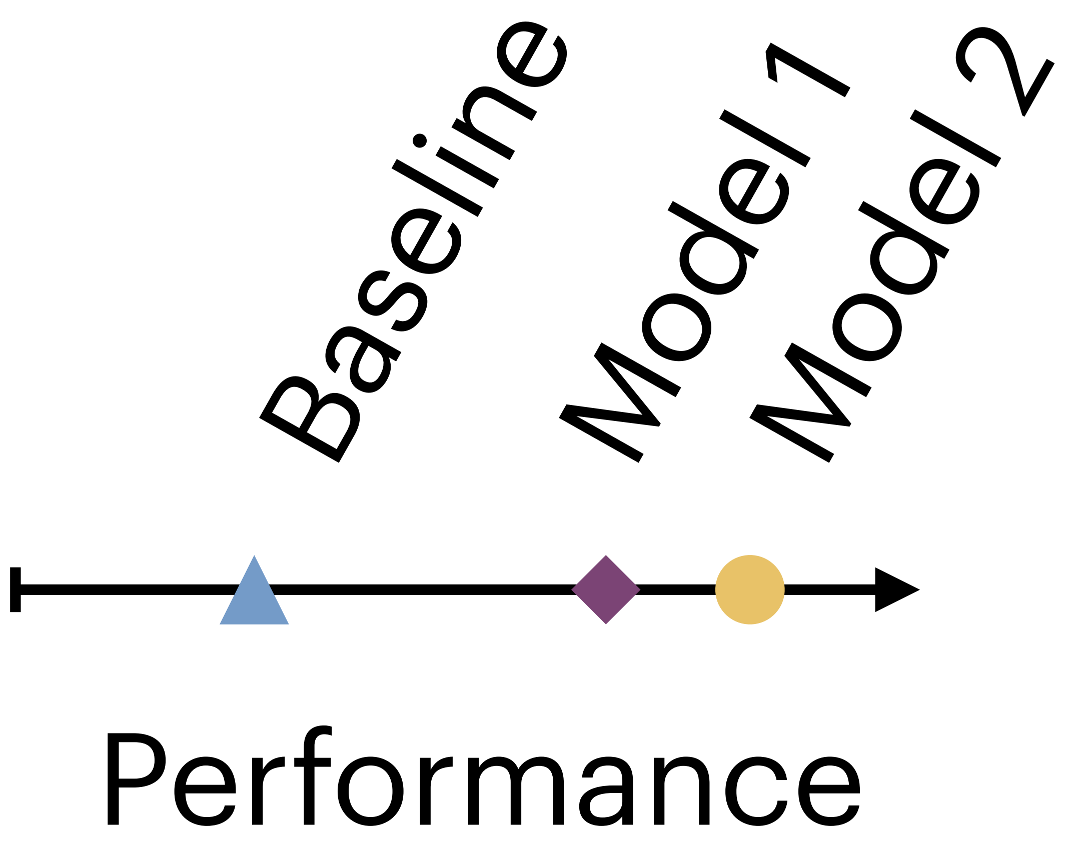
● ● ●

Average Scores

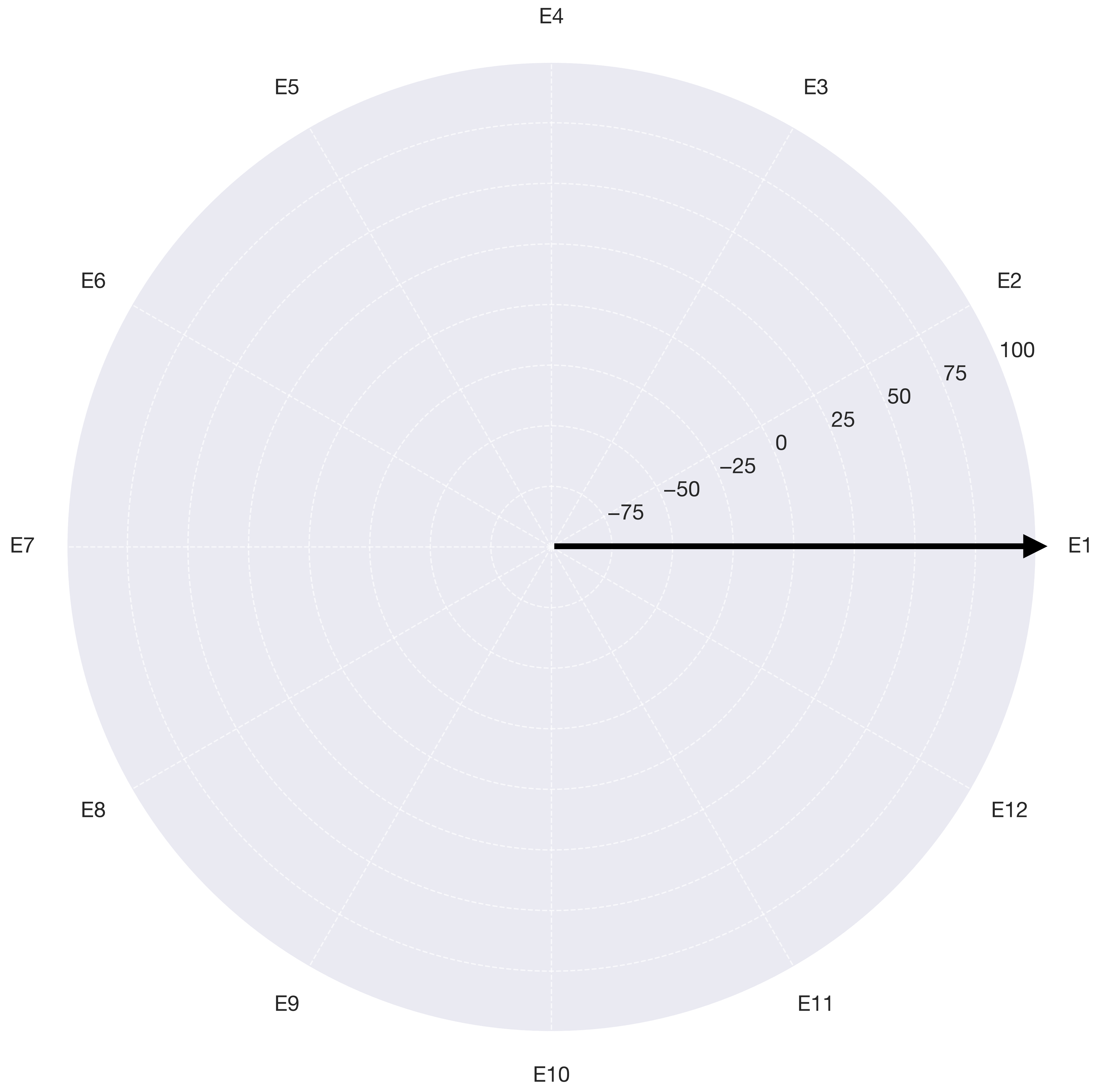


NEURAL INFORMATION PROCESSING SYSTEMS

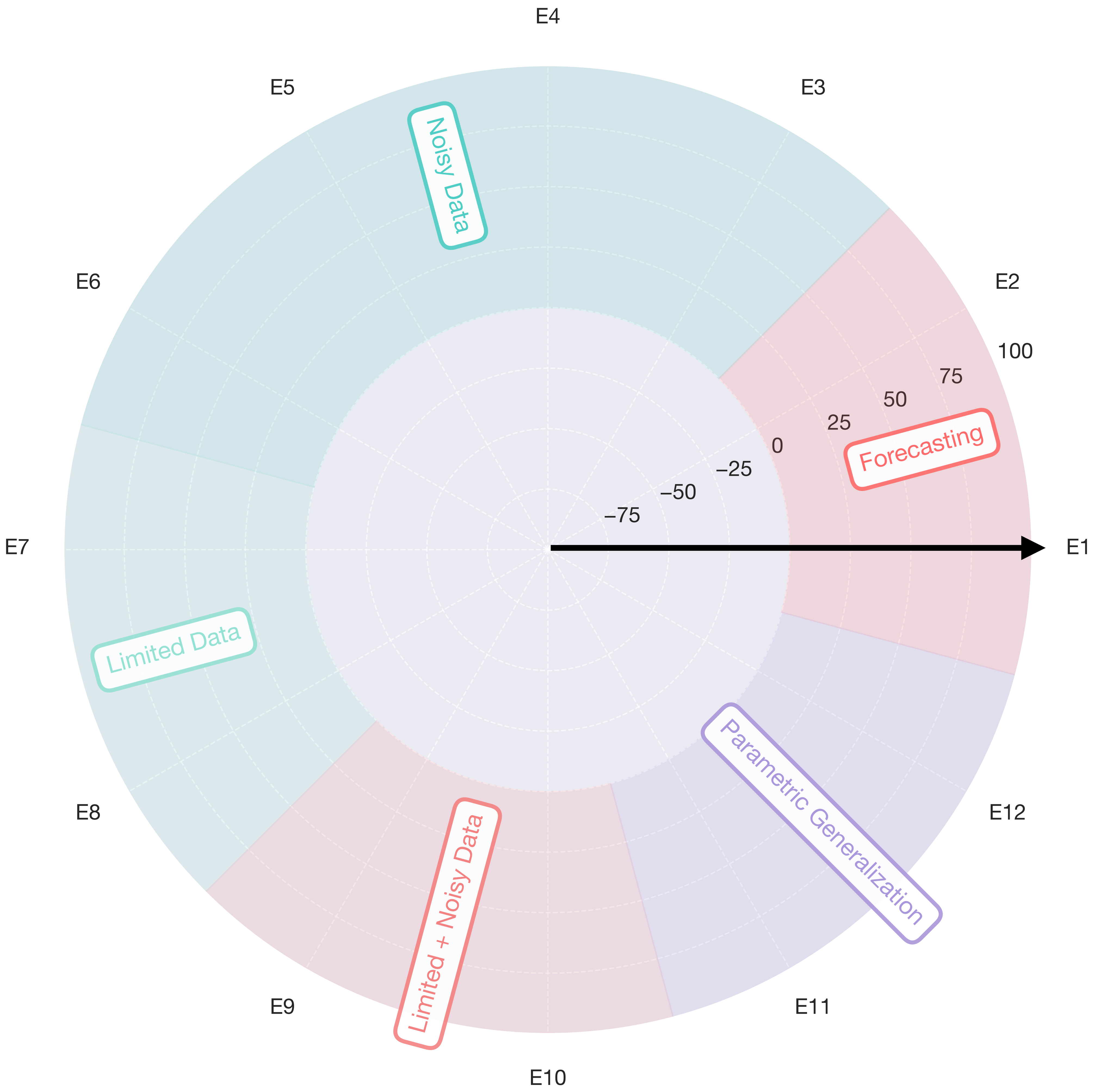
Characterizing Model Performance



Characterizing Model Performance



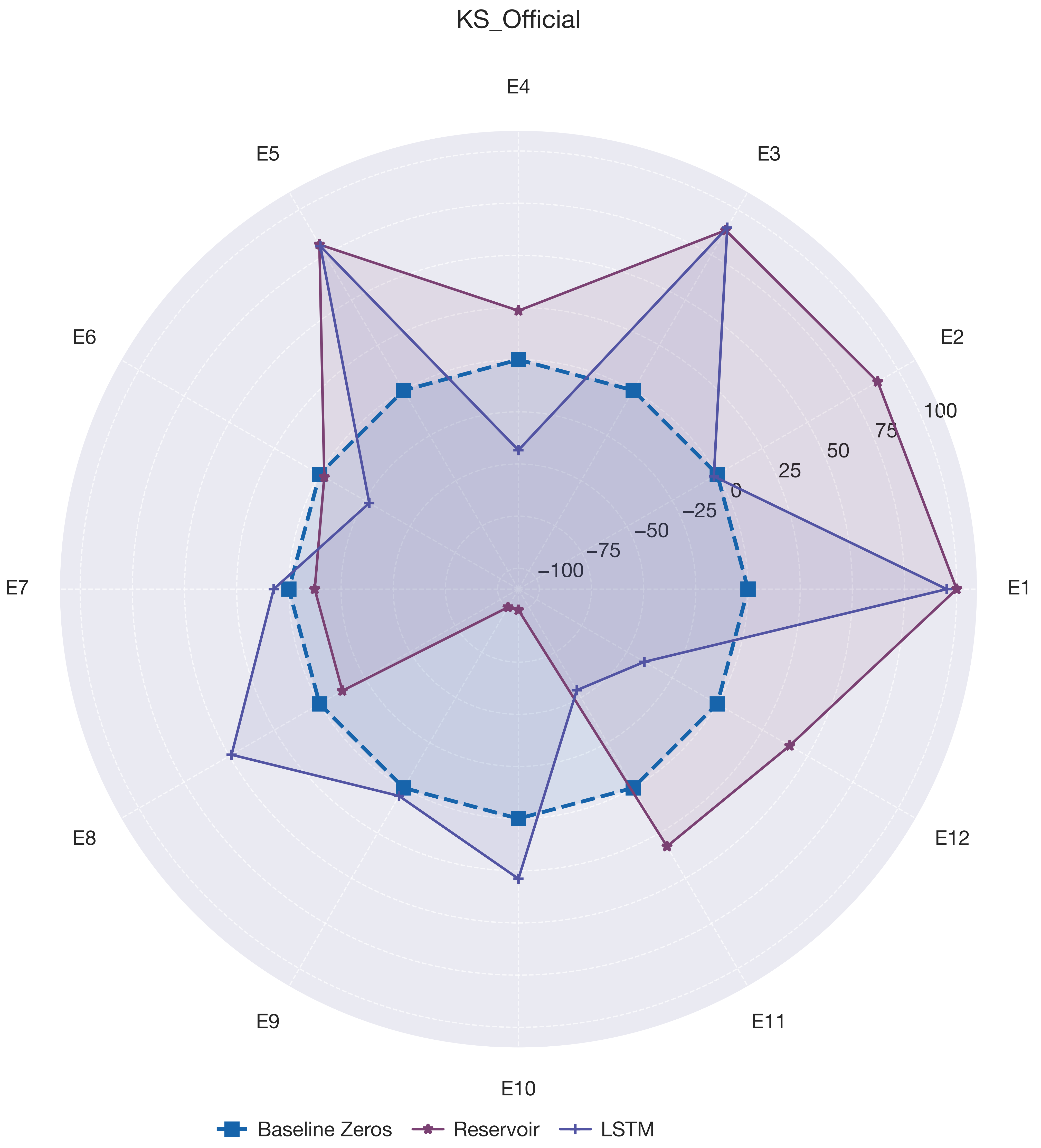
Characterizing Model Performance



Scoring Dimensions

Score	Test	Task
E1	Forecasting	Short-time
E2	Forecasting	Long-time
E3	Noisy (medium)	Reconstruction (denoising)
E4	Noisy (medium)	Forecast (long-time)
E5	Noisy (high)	Reconstruction (denoising)
E6	Noisy (high)	Forecast (long-time)
E7	Limited Data (clean)	Forecast (short-time)
E8	Limited Data (clean)	Forecast (long-time)
E9	Limited Data (noisy)	Forecast (short-time)
E10	Limited Data (noisy)	Forecast (long-time)
E11	Parametric Generalization	Interpolation forecast
E12	Parametric Generalization	Extrapolation forecast

Model Fingerprint



NEURAL INFORMATION PROCESSING SYSTEMS

CTF4Science on GitHub

- Modular code with diverse functionality
  - Data loading and model evaluation API
  - Easy hyper-parameter tuning
- Pre-configured models
  - 18 models implemented, tuned, and evaluated
- Visit https://github.com/CTF-for-Science/ctf4science

CTF for Science on Kaggle

- Let the broader community score their models
- Truly withheld test-set for fair comparison
- Display submissions per team to prevent p-hacking
- Custom website to show spider plots and detailed metrics

UNIVERSITY · OF · WASHINGTON · 1861 ·
LVX · SIT

COLUMBIA
UNIVERSITY

SURF

UT VITAM HABEANT ABUNDANTIUS HABEANT
1866

NSF

UNIVERSITY OF
CAMBRIDGE

POLITECNICO
MILANO 1863

NEURAL INFORMATION
PROCESSING SYSTEMS

AI INSTITUTE IN DYNAMIC SYSTEMS