Overview of the Role of Machine Learning in Atmospheric Research

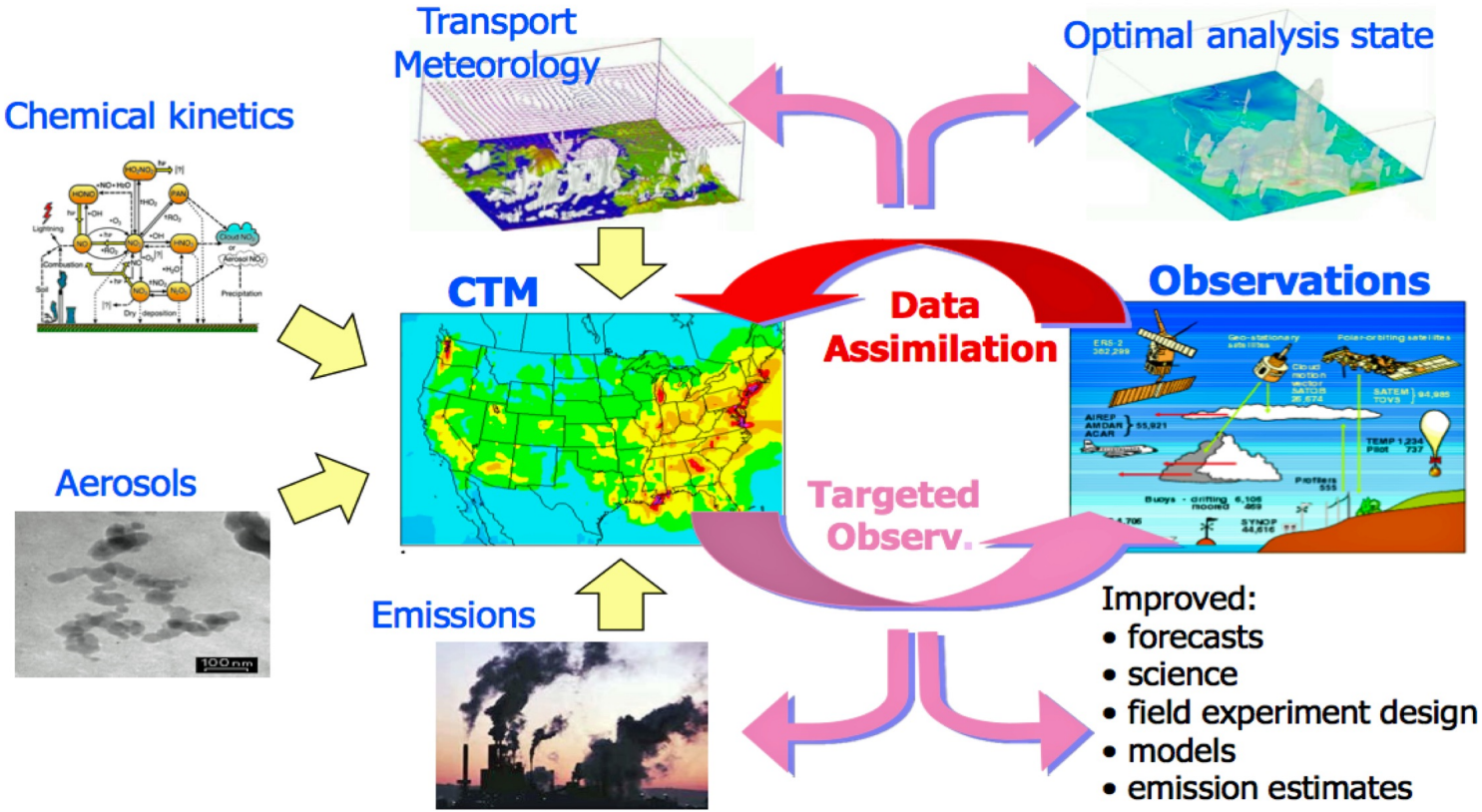
Haiyan Cheng

Willamette University

BIRS workshop
Mathematical Approaches of Atmospheric Constituents Data Assimilation and Inverse Modeling

March 22, 2023

Chemical Transport Model



Uncertainty Quantification

- Better air quality model forecast ➔ better informed decision making.
- Scientists who build the models are not the decision makers.
- Error bars are not enough for confidence information

Uncertainties

- Aleatoric Uncertainties

  due to the intrinsic randomness of a system or phenomenon, cannot be reduced even with complete knowledge and understanding

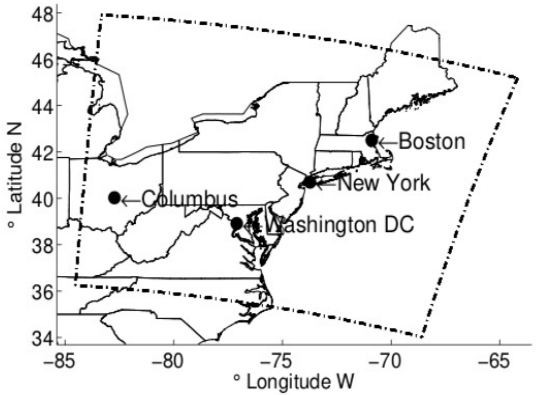
- Epistemic Uncertainty

  due to lack of knowledge or understanding of a system, can be reduced by obtaining more information or improving the understanding of the system

Uncertainty Quantification with Polynomial Chaos Method for a 3D Air Quality Model

Sulfur Transport Eulerian Model (STEM):

$$
\begin{aligned}
\frac{\partial c_i}{\partial t} &= -u \cdot \nabla c_i + \frac{1}{\rho} \nabla \cdot (\rho K \nabla c_i) + \frac{1}{\rho} f_i(\rho c) + E_i \qquad t^0 \le t \le T, \\
c_i(t^0, x) &= c_i^0(x), \\
c_i(t, x) &= c_i^{IN}(t, x) \qquad \text{for} \quad x \in \Gamma^{IN}, \\
K \frac{\partial c_i}{\partial n} &= 0 \quad \text{for} \quad x \in \Gamma^{OUT}, \\
K \frac{\partial c_i}{\partial n} &= V_i^{dep} c_i - Q \quad \text{for all} \quad 1 \le i \le n.
\end{aligned}
$$



models 93 chemical species

involves 213 chemical reactions

use KPP as the chemical time integrator

Represent second order processes as series expansion in the basis of orthogonal polynomials:

$$X(\theta) = \sum_{i=0}^{\infty} c^i \Phi^i(\xi(\theta))$$

In practice, number of terms in PC expansion:

$$S = \frac{(n+p)!}{n!\,p!}$$

n – number of uncertainties. p – maximum degree of polynomials.

Random variables ξ	Orthogonal polynomials $\Phi(\xi)$	Support
Gaussian	Hermite	$(-\infty, +\infty)$
Gamma	Laguerre	$[0, \infty)$
Beta	Jacobi	$[a, b]$
Uniform	Legendre	$[a, b]$

Source of Uncertainties

$\xi_1 \rightarrow NO_x$ ground emission (NO,NO2) (-20%--+20%)

$\xi_2 \rightarrow AVOC$ ground emission (HCHO,ALK,OLE,ARO) (-50%--+50%)

$\xi_3 \rightarrow BVOC$ ground emission (ISOPRENE, TERPENE, ETHENE) (-40%--+40%)

$\xi_4 \rightarrow$ Deposition velocity for Ozone (-50%--+50%)

$\xi_5 \rightarrow$ Deposition velocity for NO2 (-50%--+50%)

$\xi_6 \rightarrow$ West boundary condition for ozone (-5%--+5%)

$\xi_7 \rightarrow$ West boundary condition for PAN (-5%--+5%)

- 7 independent Beta distributed random variables (Jacobi polynomial of order 2 is used).
- Uncertainties in emissions, B.C., and deposition velocity .

Uncertainty Quantification with PC

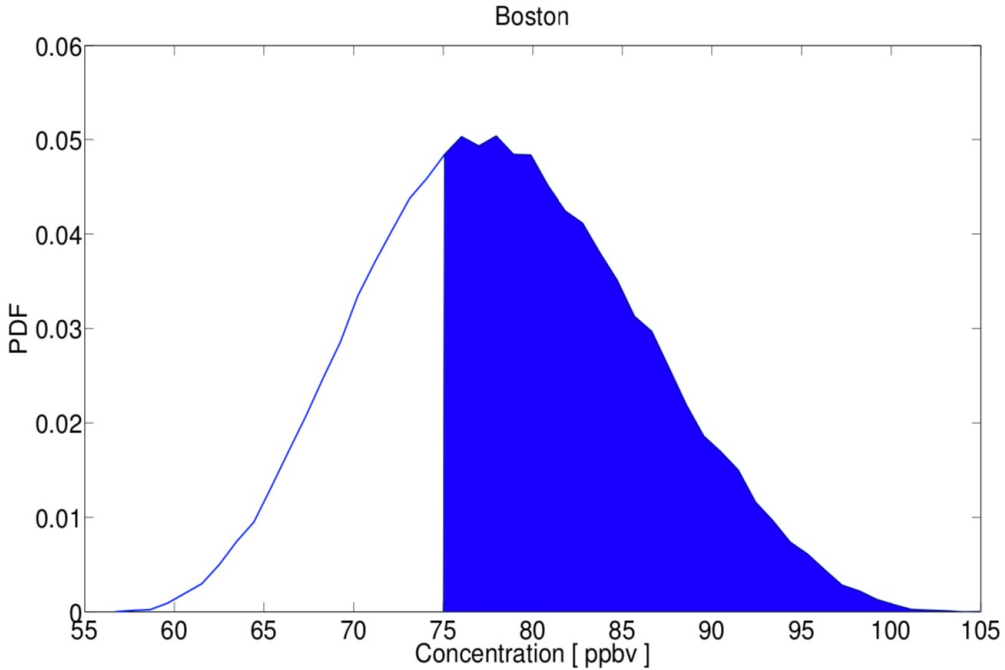


Figure: Boston average ozone PDF (68% of exceeding 75 ppbv).

Uncertainty Apportionment with PC

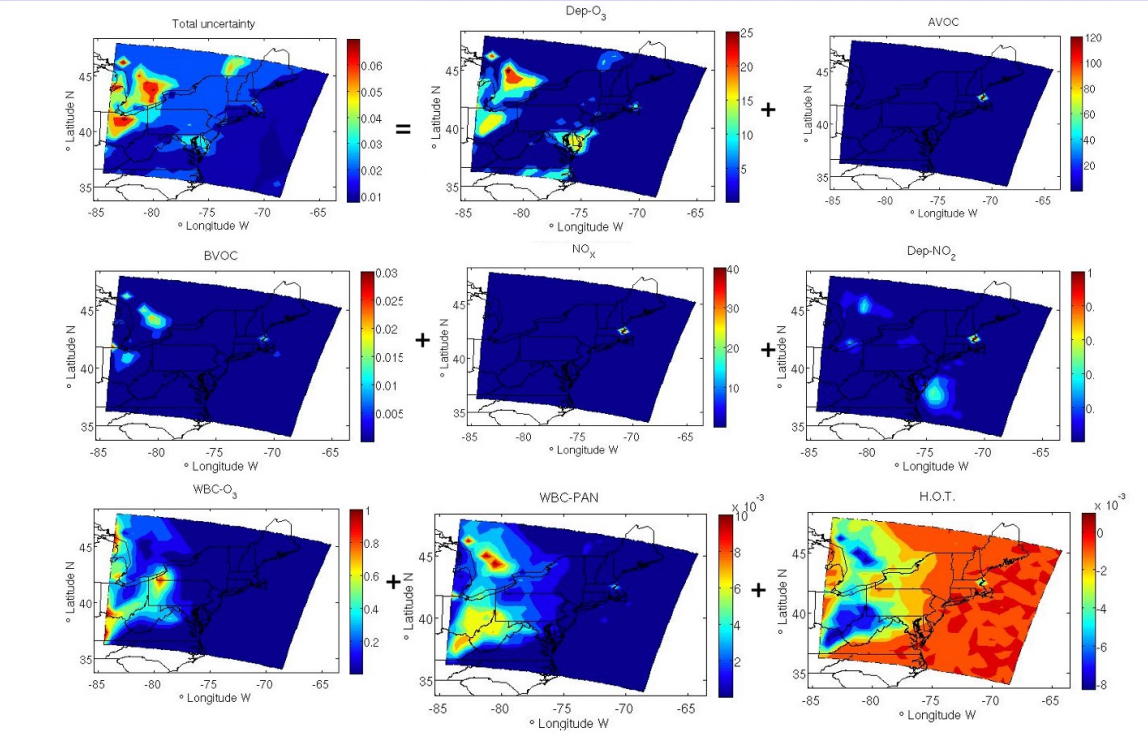
$$y_\ell(t) = \underbrace{a_\ell^0(t)}_{\text{0th order term}} + \underbrace{a_\ell^1(t)\Phi^1(\xi) + a_\ell^2(t)\Phi^2(\xi) + \ldots + a_\ell^d(t)\Phi^d(\xi)}_{\text{linear order terms}} + H.O.T. \quad 1 \leq \ell \leq n$$

Separate the terms corresponding to the linear terms from the higher order terms:

$$s_\ell^2 = \sum_{i=1}^{d} (a_\ell^i)^2 \langle \Phi^i, \Phi^i \rangle + \sum_{i=d+1}^{S} (a_\ell^i)^2 \langle \Phi^i, \Phi^i \rangle.$$

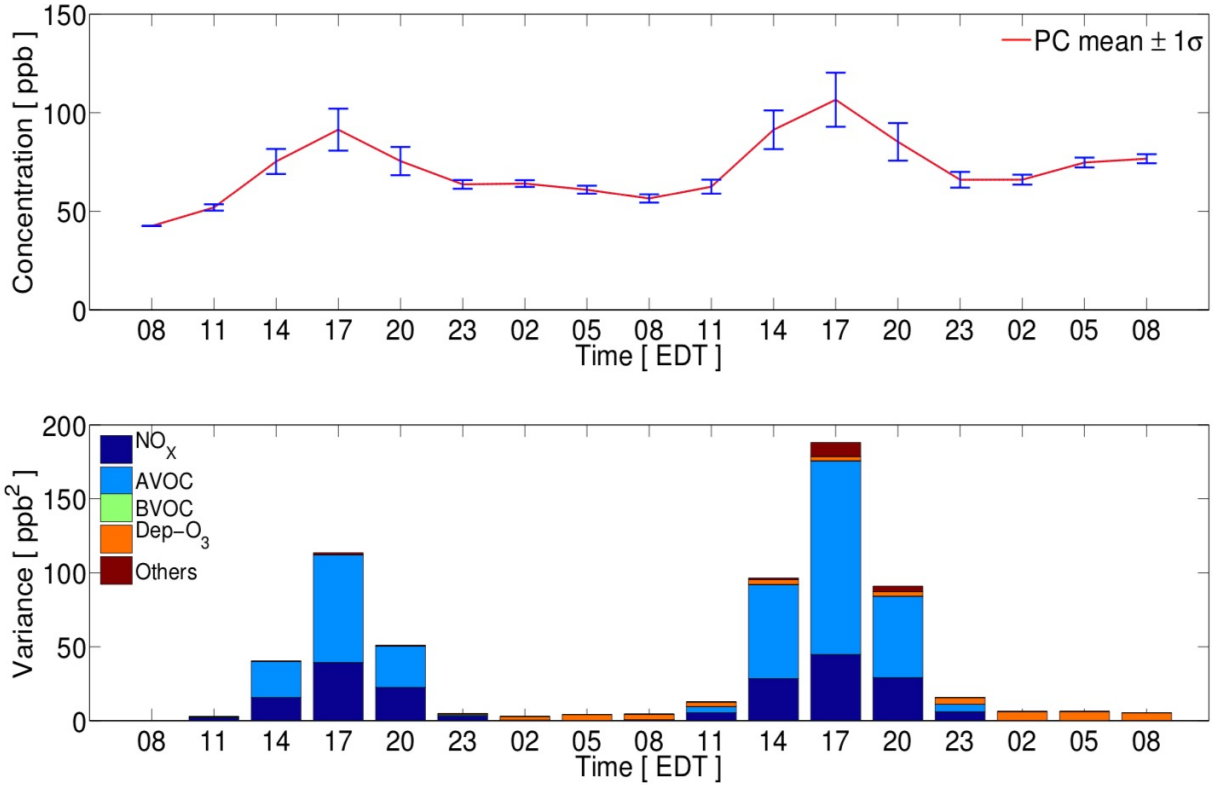
$(a_\ell^i)^2 \langle \Phi^i, \Phi^i \rangle$ in the linear portion is the part of the total variance s_ℓ^2 that can be attributed to the i-th source of uncertainty (modeled by variable ξ_i). The higher order terms represent the mixed contribution resulting from the interaction of multiple sources.

Uncertainty Apportionment with PC

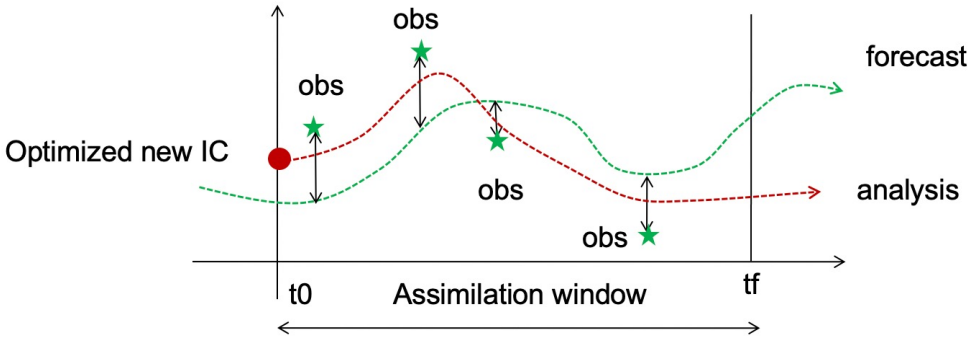


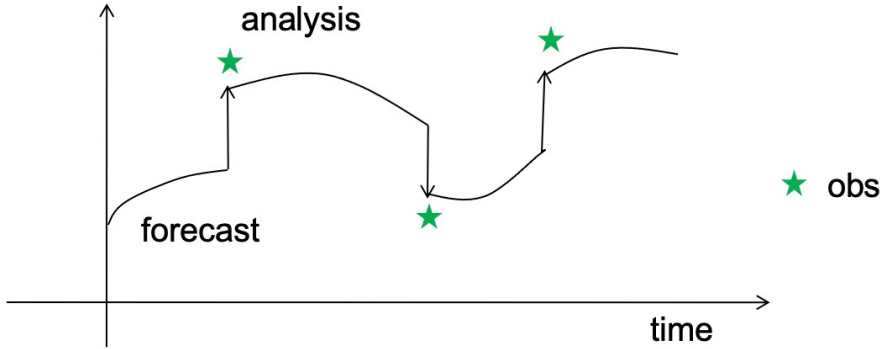
Cheng, H., & Sandu, A. (2009)

Uncertainty Apportionment with PC



Uncertainty Reduction with Data Assimilation (variational or sequential?)





Assimilation frequency is controlled by data availability

How about Variational + Sequential?

- Update and correct error covariance matrix at the end of each assimilation window by investigating 4D-Var error reduction directions

- Run a short window 4D-Var and use that information to initialize EnKF

Cheng, H., Jardak, M., Alexe, M., & Sandu, A. (2010). Sandu, A., & Cheng, H. (2015).

Hybrid Particle Filter

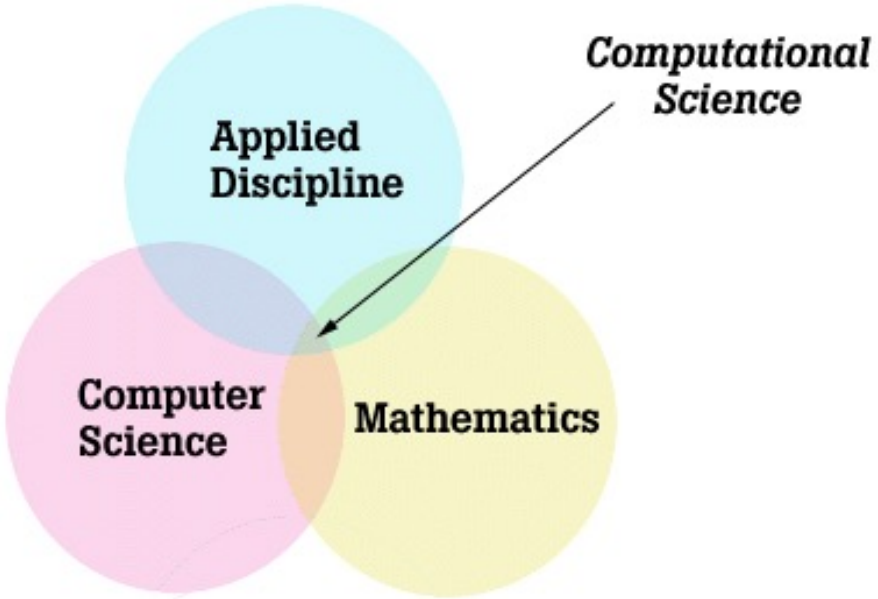
- Introduce Particle Swarm Optimization (PSO) as an auxiliary procedure to alleviate the “particle degeneration” problem.

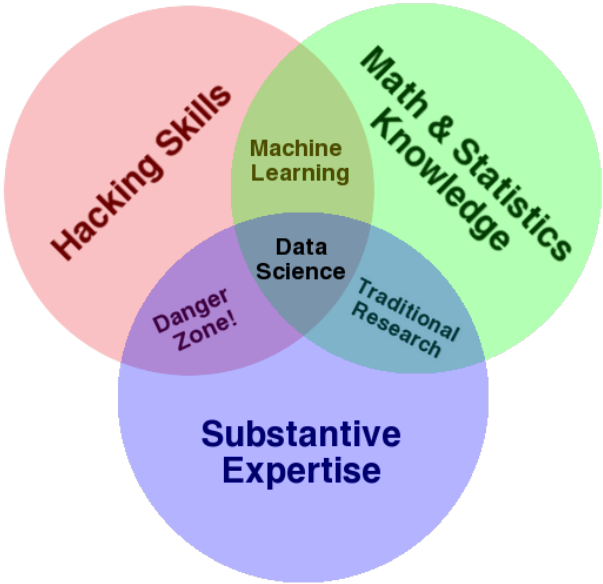
Schmal, K., & Cheng, H. (2015

I thought I had a better understanding of somewhat complex mathematical models until …..

Building a Model from Data only?

Computational Science → Data Science

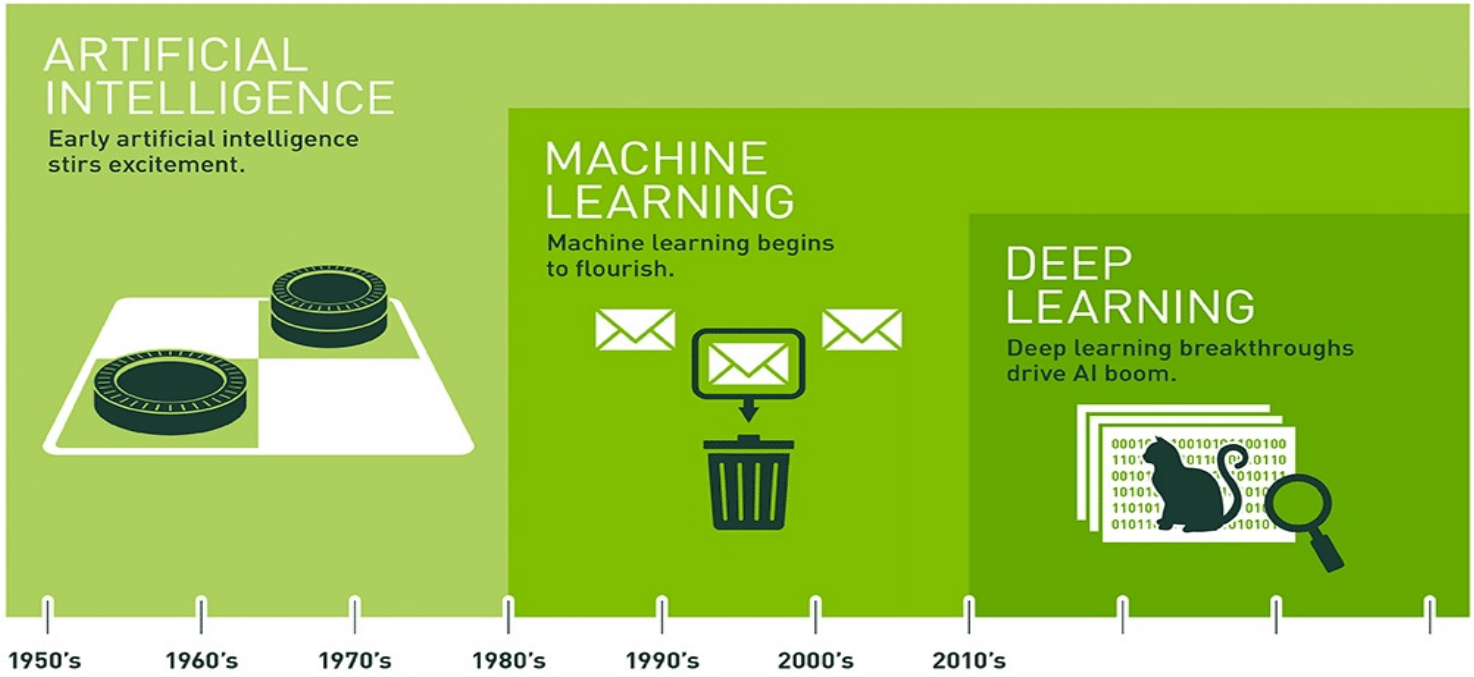




Machine Learning

- A subset of Artificial Intelligence (AI).
- The core of Data Science (DS)
- Algorithms that use data to learn, then use what was learned to make predictions.

Machine Learning Timeline

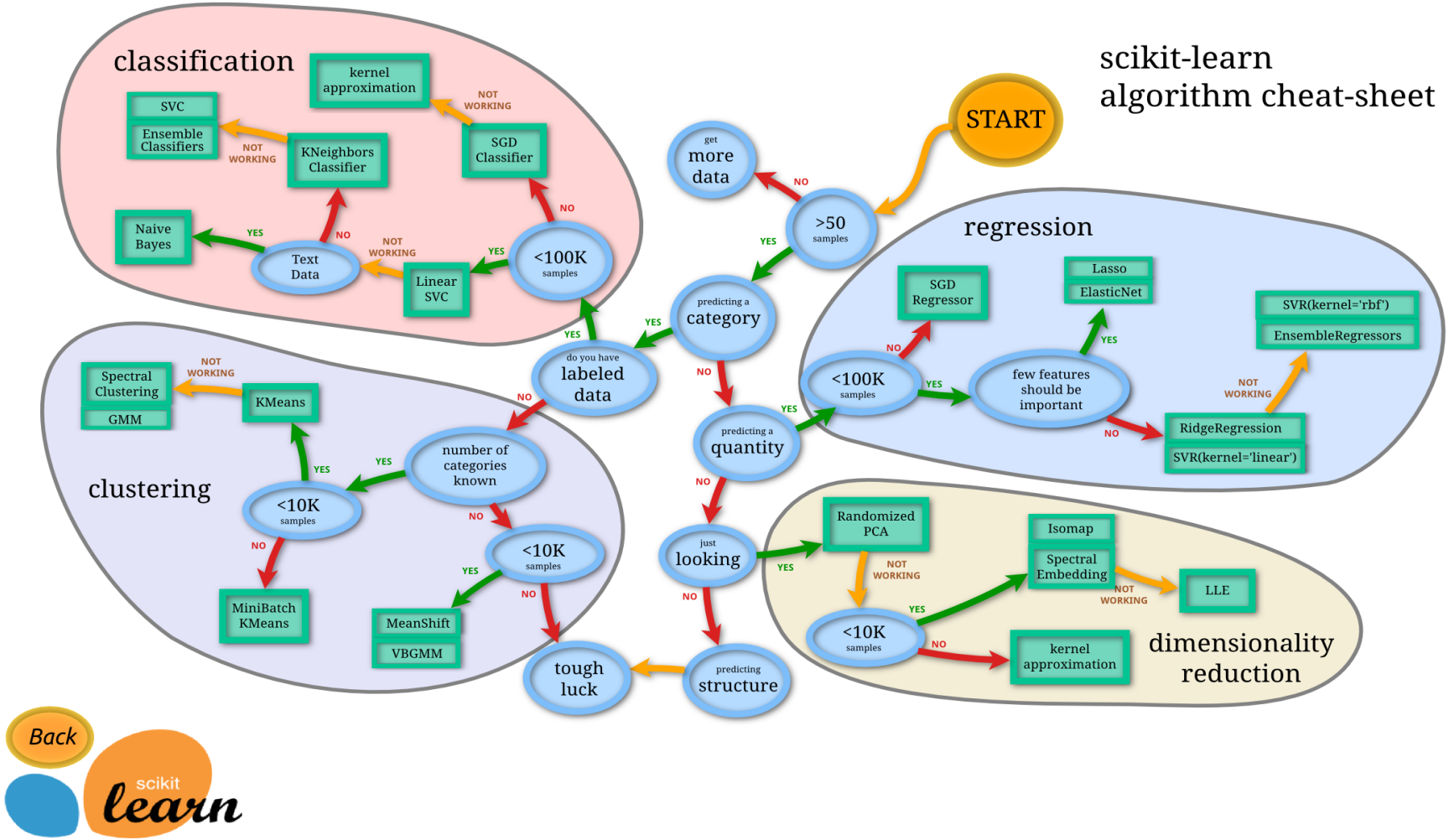


Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

Machine Learning Algorithms

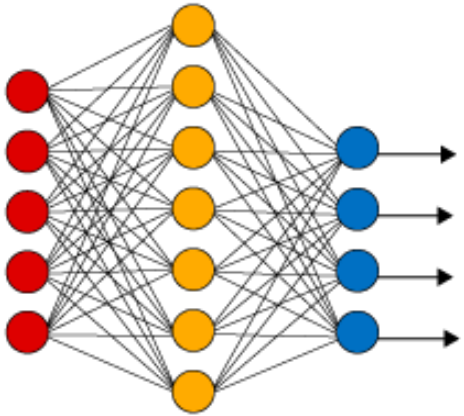
- Supervised – data is labeled
- Unsupervised – data is unlabeled, discover hidden structure
- Semi-supervised – data is partially labeled
- Reinforced – interact with environment, use reward feedback to learn best action


scikit-learn
algorithm cheat-sheet
START
get more data
NO
>50 samples
YES
classification
kernel approximation
NOT WORKING
SVC
Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
SGD Classifier
NO
Naive Bayes
YES
Text Data
NO
NOT WORKING
Linear SVC
YES
<100K samples
NO
predicting a category
YES
regression
SGD Regressor
Lasso
ElasticNet
SVR(kernel='rbf')
EnsembleRegressors
NO
<100K samples
YES
few features should be important
YES
NOT WORKING
RidgeRegression
SVR(kernel='linear')
NO
do you have labeled data
YES
NO
clustering
Spectral Clustering
GMM
NOT WORKING
KMeans
YES
number of categories known
YES
<10K samples
NO
MiniBatch KMeans
NO
<10K samples
YES
MeanShift
VBGMM
NO
predicting a quantity
YES
NO
just looking
YES
Randomized PCA
NOT WORKING
Isomap
Spectral Embedding
NOT WORKING
LLE
YES
<10K samples
NO
kernel approximation
dimensionality reduction
NO
predicting structure
tough luck
Back
scikit learn

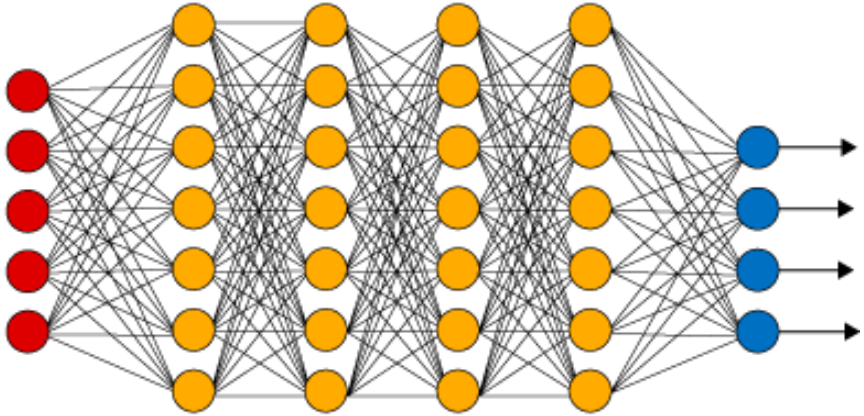
Deep Learning

- An advancement/subclass of machine learning
- Extracts features automatically using multiple-layer hidden neural networks
- Requires powerful hardware and long time to train models

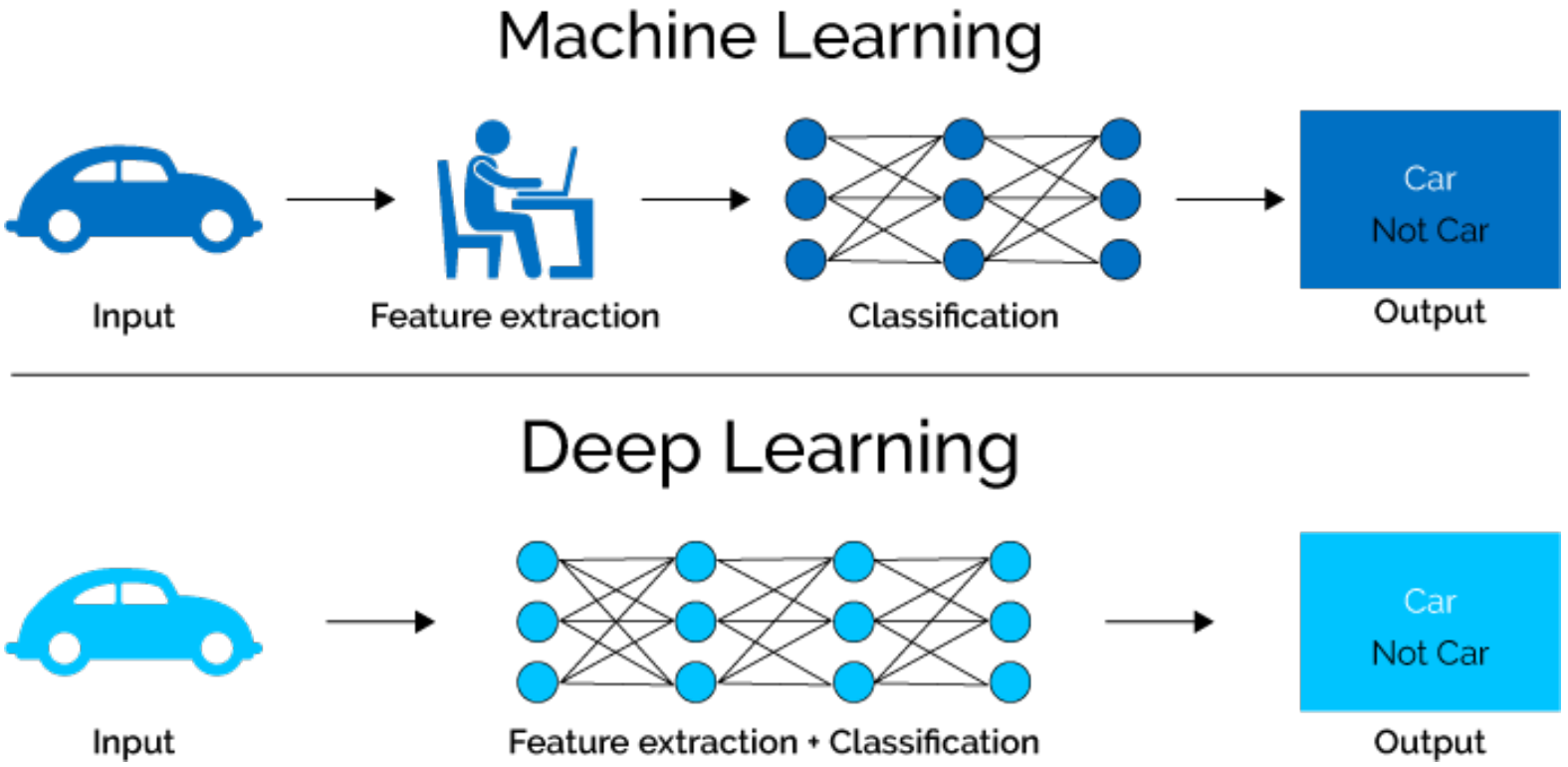
Simple Neural Network



Deep Learning Neural Network



Input Layer Hidden Layer Output Layer



https://www.xenonstack.com/blog/log-analytics-deep-machine-learning/

Computational Scientists:

Build mathematical models to represent underlying physics or chemistry

(Data Sparse -> Data Abundance)

Statistician / Data Scientists

Build statistical models or ML models from data

Use models to predict (make inferences or decisions)

Physical model vs. Data Model

Physical model + Data Model

Large Amount of Historic Data in Atmospheric Research

- weather forecast
- climate model
- air quality model
- remote sensing

Observation data + Simulation Data

➔ learn patterns and correlations➔ improve short-term forecast

Traditional ML Process

1. Data collection and preparation (clean, normalize, preprocessing, missing data handling, transformation, smoothing)
2. Exploratory Data Analysis (EDA)
3. Feature extraction and selection (e.g. PCA)
4. Model selection (regression, decision trees, random forests, neural network)
5. Model training (parameter optimization)
6. Model evaluation
7. Model tuning
8. Document and launch

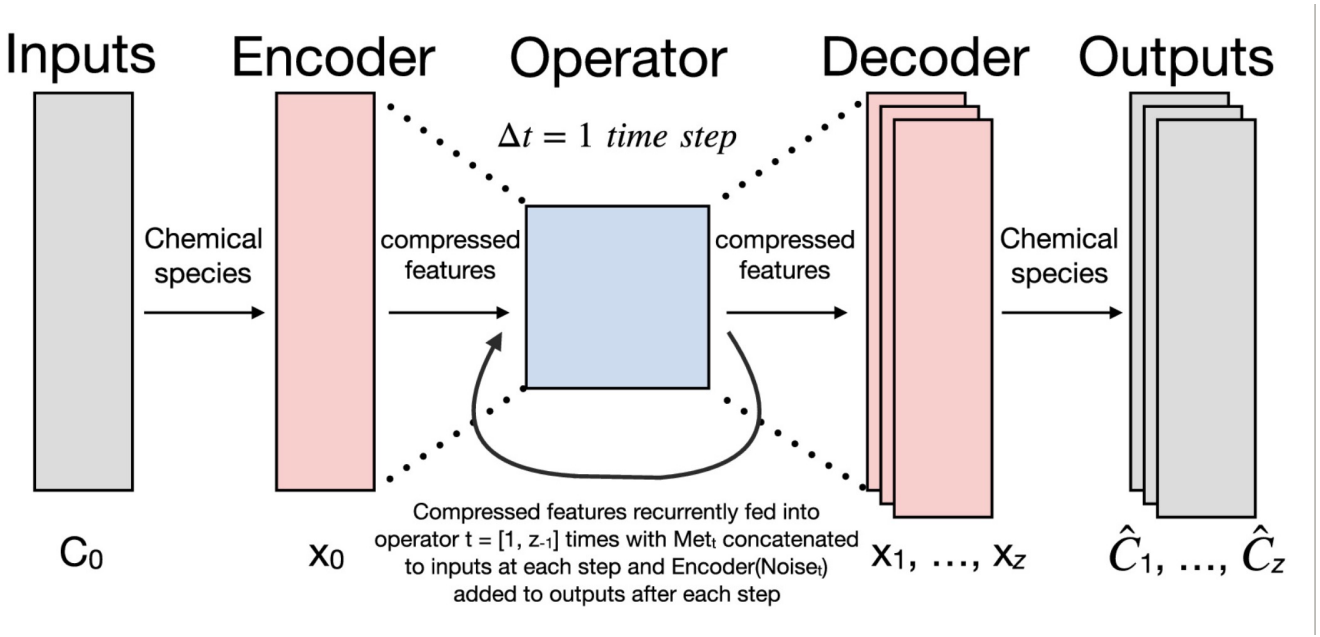
Scientific Machine Learning for Complex Models

- Use ML for parameter estimation
- Use ML to speed up scientific applications
- Use ML to help optimal design
- Automatic hypothesis generation
- Accelerated scientific simulation

Physics Informed Machine Learning

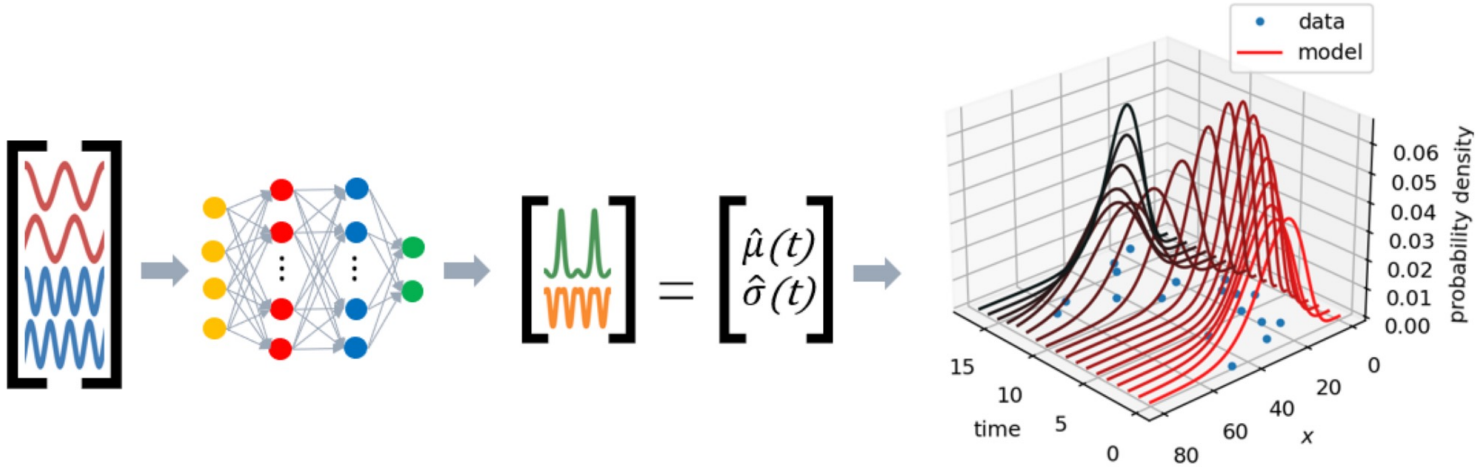
- Data-driven solutions of nonlinear PDEs
- Physics informed neural network
- Data-efficient universal function approximators that encode any underlying physical laws as prior information
- Physics-informed surrogate models that are fully differentiable with respect to all input coordinates and free parameters

Encoder-operator-decoder neural network



Kelp, Makoto M., et al. "Toward stable, general machine-learned models of the atmospheric chemical system." *Journal of Geophysical Research: Atmospheres* 125.23 (2020): e2020JD032759.

Deep Probabilistic Koopman



Mallen, Alex, Henning Lange, and J. Nathan Kutz. "Deep probabilistic Koopman: long-term time-series forecasting under periodic uncertainties." *arXiv preprint arXiv:2106.06033* (2021).

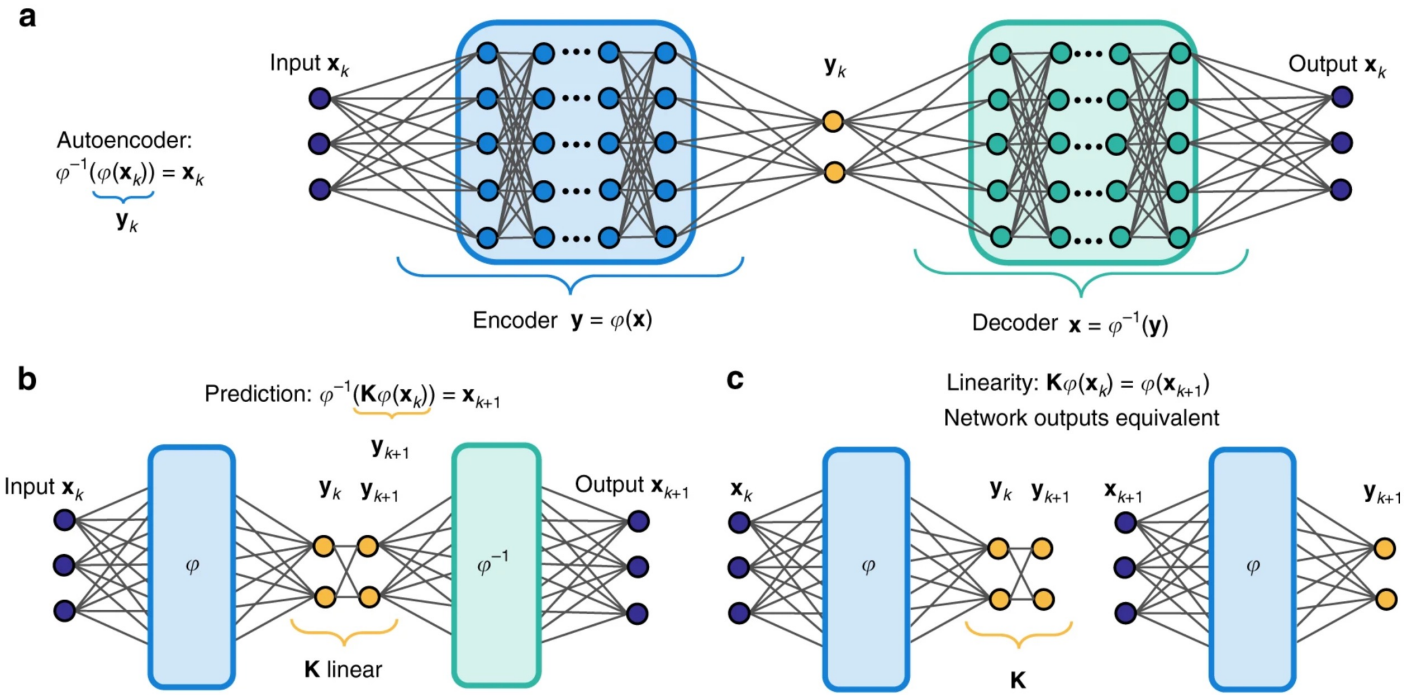
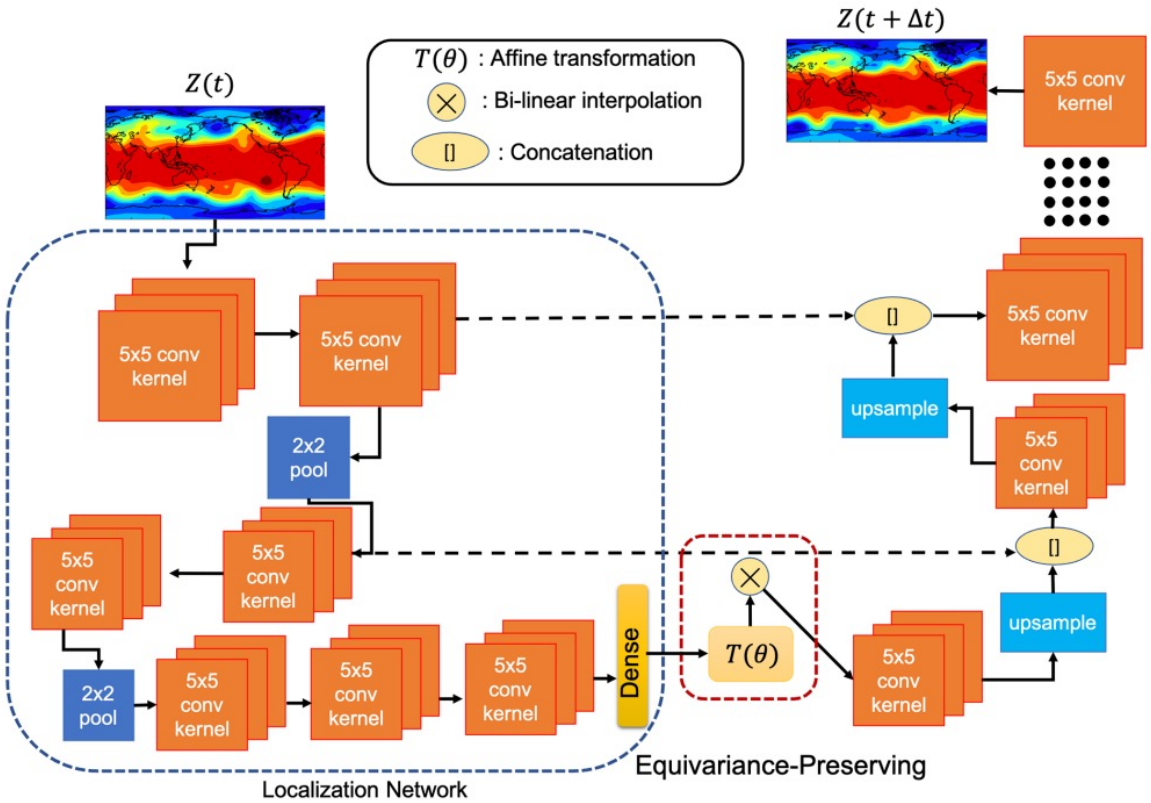
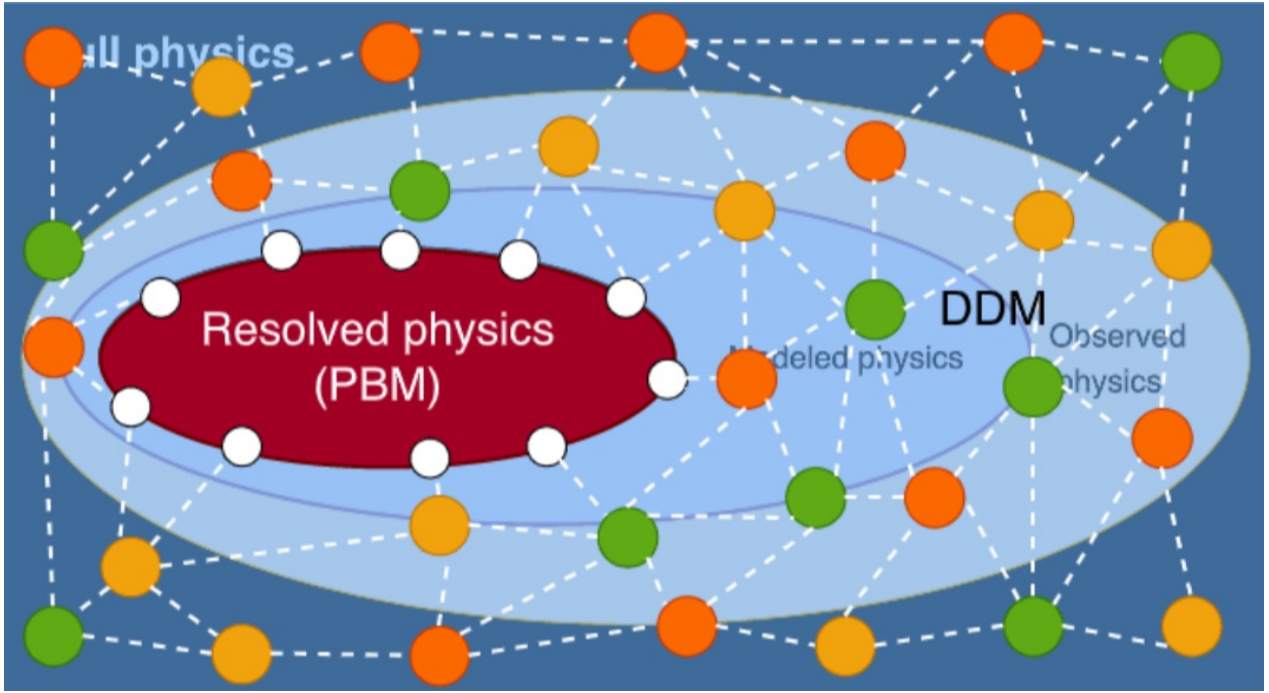


Diagram of our deep learning schema to identify Koopman eigenfunctions $\varphi(\mathbf{x})$. **a** Our network is based on a deep auto-encoder, which is able to identify intrinsic coordinates $\mathbf{y} = \varphi(\mathbf{x})$ and decode these coordinates to recover $\mathbf{x} = \varphi^{-1}(\mathbf{y})$. **b**, **c** We add an additional loss function to identify a linear Koopman model $\mathbf{K}$ that advances the intrinsic variables $\mathbf{y}$ forward in time. In practice, we enforce agreement with the trajectory data for several iterations through the dynamics, i.e. $\mathbf{K}^m$. In **b**, the loss function is evaluated on the state variable $\mathbf{x}$ and in **c** it is evaluated on y

Lusch, Bethany, J. Nathan Kutz, and Steven L. Brunton. "Deep learning for universal linear embeddings of nonlinear dynamics." *Nature communications* 9.1 (2018): 4950.



Chattopadhyay, Ashesh, et al. "Towards physics-inspired data-driven weather forecasting: integrating data assimilation with a deep spatial-transformer-based U-NET in a case study with ERA5." *Geoscientific Model Development* 15.5 (2022): 2221-2237.



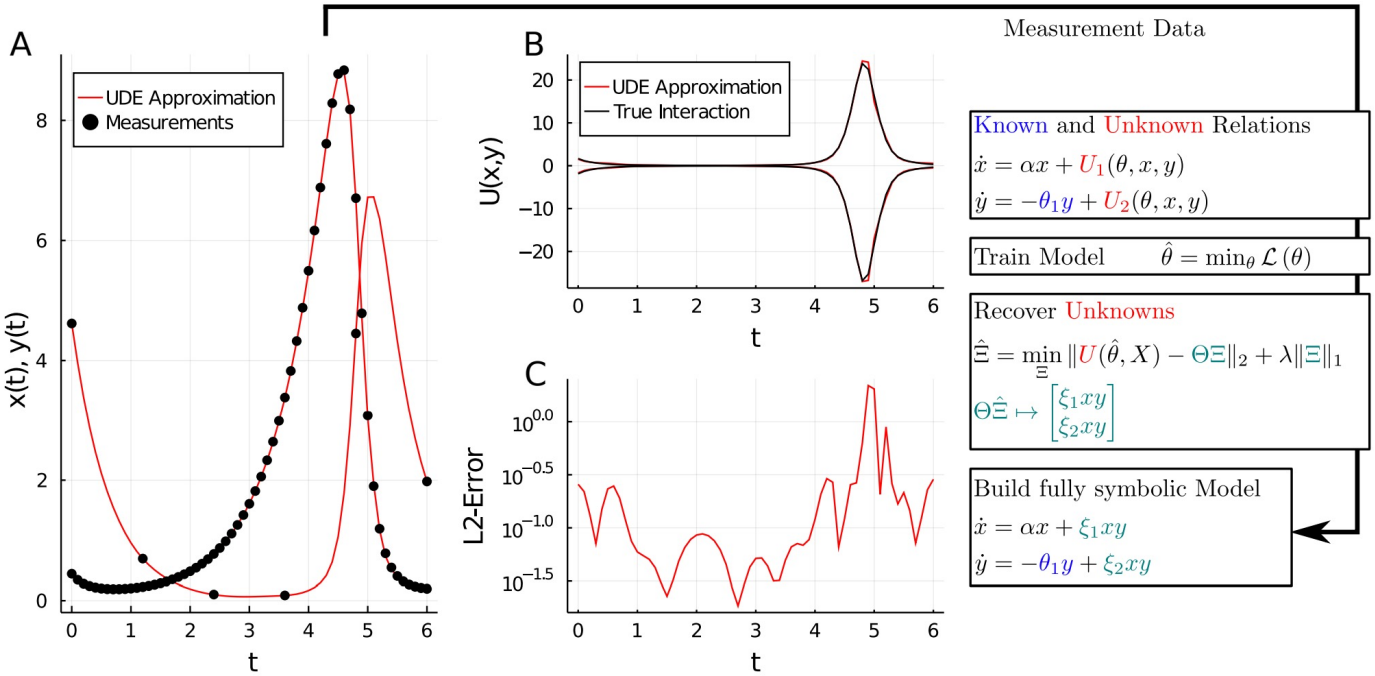
PBM DDM

$$\widetilde{\mathcal{N}}\,\hat{\tilde{u}} = \tilde{f} + \hat{\sigma}_{\mathrm{NN}}$$

CoSTA

Blakseth, Sindre Stenen, et al. "Combining physics-based and data-driven techniques for reliable hybrid analysis and modeling using the corrective source term approach." *Applied Soft Computing* 128 (2022): 109533.

Universal differential equation



Rackauckas, Christopher, et al. "Universal differential equations for scientific machine learning." *arXiv preprint arXiv:2001.04385* (2020).

- Multidisciplinary
- Interdisciplinary
- Transdisciplinary

Collaboration

Thank you!