Theory-guided Data Science: A New Paradigm for Scientific Discovery from Data

Theory-guided Data Science: A New Paradigm for Scientific Discovery from Data
Data science models, although successful in a number of commercial domains, have had limited applicability in scientific problems involving complex physical phenomena. Theory-guided data science (TGDS) is an emerging paradigm that aims to leverage the wealth of scientific knowledge for improving the
https://arxiv.org/abs/1612.08544

Abstract

- Data Science models have limited applicability in scientific problems involving complex physical phenomena
- TGDS aims at using scientific knowledge for improving the effectiveness of data science models in scientific discovery
- Vision of TGDS
  1. Introduce scientific consistency in models
  2. Advance scientific understanding by discovering novel domain insights

---

Introduction

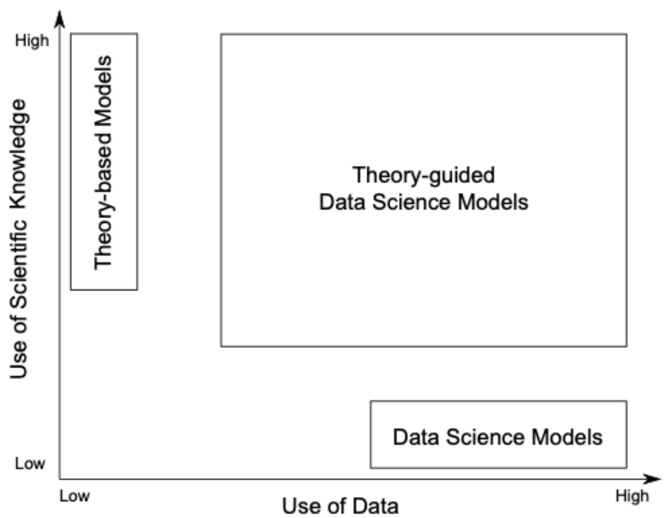
- Google Flu Trends: model overestimated the flu propensity since the data used for training the model was not representative of the trends in subsequent years
- two primary characteristics of knowledge discovery in scientific disciplines
  - **scientific problems are often under-constrained**: high number of variables with complex and non-stationary patterns, small training set, risk of learning spurious relationships

- **basic nature of scientific discovery:** translation of learned patterns and relationships to *interpretable* theories and hypotheses
    - black-box models lack the ability to deliver a mechanistic understanding of the underlying process
    - interpretable model that is grounded by explainable theories is more robust against the learning of spurious patterns

---

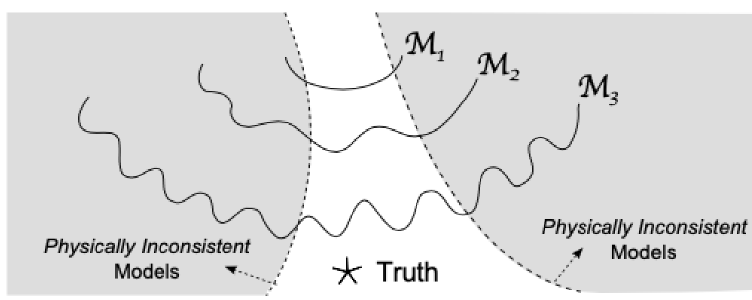
Theory-Guided Data Science

- problem: represent relationships among physical variables



- the two models and their drawbacks
    - **theory-based models**: represent relationships based on models from scientific knowledge
        - in complex, unknown settings these models need no make simplifying assumptions about the physical processes which leads to poor performance and worse interpretability of the model
    - **data-based models**: use a set of training examples for learning a model that can automatically extract relationships between the variables
        - available data inadequately represents the complex spaces of hypothesis

        - since these models only capture associative relationships between variables, they do not serve the goal of understanding causative relationships in scientific problems
- TGDS
    - **physically consistent models**
        - lean dependencies that have a sufficient grounding in physical principles and thus have a better chance to represent causative relationships
        - better generalizability since the models are consistent with scientific principles
    - **principle of bias-variance trade-off**
        - scientific knowledge can help in reducing the model variance by removing physically inconsistent solutions without likely affecting their bias



    - **overarching vision**
        - include physical consistency as a critical component of model performance along with training accuracy and model complexity

$$\text{Performance} \propto \text{Accuracy} + \text{Simplicity} + Consistency$$

I. Theory-guided Design of Data Science Models

→ restrict the space of models to physically consistent solutions

- **Theory-guided Specification of Response**

  → use synergistic combinations of response and loss functions

    - simplify optimization → low training errors
    - consistent with our physical understanding → generalizable solutions

  - generalized linear model (GLM): g(mu) = w^Tx + b
    - important to choose an appropriate link function g that matches with domain understanding

- **Theory-guided Design of Model Architecture**

  → design compliant with scientific knowledge

  1. decompose the overall problem into modular sub-problems each representing a different physical sub-process
  2. specify node connections that capture theory-guided dependencies among variables (e.g. time dependency in RNN)

II. Theory-guided Learning of Data Science Models

→ guide a learning algorithm to focus on physically consistent solutions

- **Theory-guided Initialization**

  → initializing the model with physically meaningful parameters

  - matrix completion by using the species mean
  - ANN pretraining by initializing the model with computational simulations

- **Theory-guided Probabilistic Models**

  → encoding scientific knowledge as probabilistic relationships among variables

  - graph Lasso: automated graph estimation techniques to find relationships
    - limit search to physically consistent models
  - introduce priors in model space

- **Theory-guided Constrained Optimization**

  → use constrains to ensure self-consistent models

- **Theory-guided Regularization**

  → introduce regularization terms inspired by physical understanding

  - use variants of Lasso to incorporate domain specific structure among parameters
  - multitask learning when facing heterogeneity in data sub-populations
    - treat learning at every sub-population as a different taks
    - share the learning at related tasks

III. Theory-guided Refinement of Data Science Outputs

→ refine output of models using explicit or implicit knowledge

- **Using Explicit Domain Knowledge**
  - reduce the effect of noise and missing values
  - refine outputs to improve quality measures e.g. through pruning
- **Using Implicit Domain Knowledge**
  - domain structure among the output may not be known through explicit equations
  - jointly solve: inferring the domain constraints & using the learned constraints to refine model outputs

IV. Learning Hybrid Models of Theory And Data Science

→ construct hybrid models where some aspects of the problem are modeled using theory-based components wile other aspects are modeled using data science components

- **two-component model**
  - outputs of the theory-based component are used as inputs in the data science component

    - these outputs can also be used to supervise the training of data science models
- **predict intermediate quantities**

  → use data science methods to predict intermediate quantities in theory-based models that are currently being missed or inaccurately estimated

    - amend the deficiencies in theory-based models
    - use theory-based outputs as training samples in data-science components

V. Augmenting Theory-based Models using Data Science

→ make effective use of observational data

- **Data Assimilation in Theory-based Models**
    - infer the most likely sequence of states such that the model outputs are in agreement with the observations available at every time step
    - values of the current state are constrained to depend on previous state values as well as the current observation
- **Calibrating Theory-based Models using Data**
    - calibrating model parameters with the help of observational data
    - e.g. model parameter combination uncertainty using Monte Carlo approaches