Theory-guided Data Science:

A new paradigm for accelerating scientific discovery

Vipin Kumar

University of Minnesota

kumar@cs.umn.edu
www.cs.umn.edu/~kumar

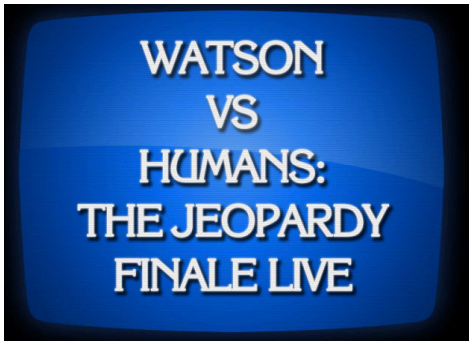
AAAI Fall Symposium on Accelerating Science:
A Grand Challenge for AI

EXPERT OPINION

Contact Editor: **Brian Brannon,** bbrannon@computer.org

The Unreasonable Effectiveness of Data

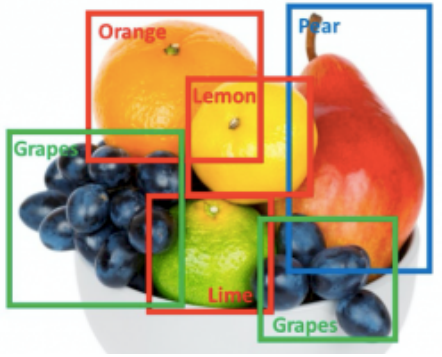
-Halevy, Norvig, and Pereira
IEEE Intelligent Systems (2009)



NETFLIX

Google DeepMind

facebook



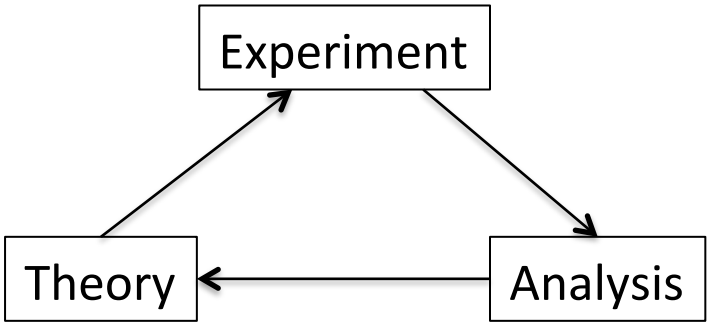
Google Ads

Google's AI won the game Go by defying millennia of basic human instinct -- Quartz, 2016

The “end of theory” in scientific domains?

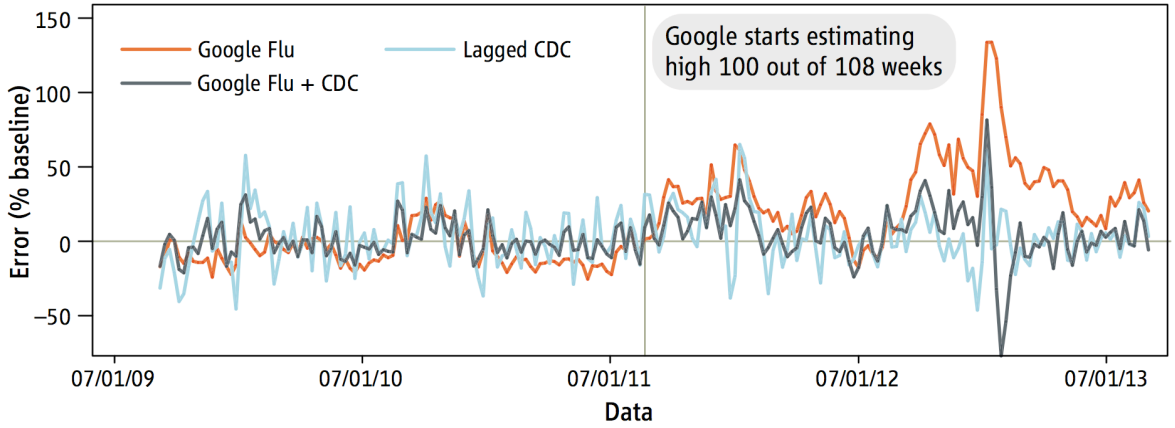
- Will the rapidly growing area of “black-box” data science models (that completely rely on the data) make existing theory-based models obsolete?

- Standard scientific paradigm:



Limits of “black-box” data science approaches

- The rise and fall of Google Flu Trends
  - Predicted flu occurrences solely on the basis of Google search queries



  - Lazer, David, et al. "The Parable of Google Flu: Traps in Big Data Analysis." **Science** 343 (2014): 1203-1205.

- Similar observations in other scientific domains:

Climate Science:

Caldwell et al. "Statistical significance of climate sensitivity predictors obtained by data mining.” **Geophysical Research Letters** (2014)

The New York Times

The Opinion Pages | OP-ED CONTRIBUTORS

Eight (No, Nine!) Problems With Big Data

By GARY MARCUS and ERNEST DAVIS APRIL 6, 2014

" . . . you will always need to start with an analysis that relies on an understanding of physics and biochemistry."

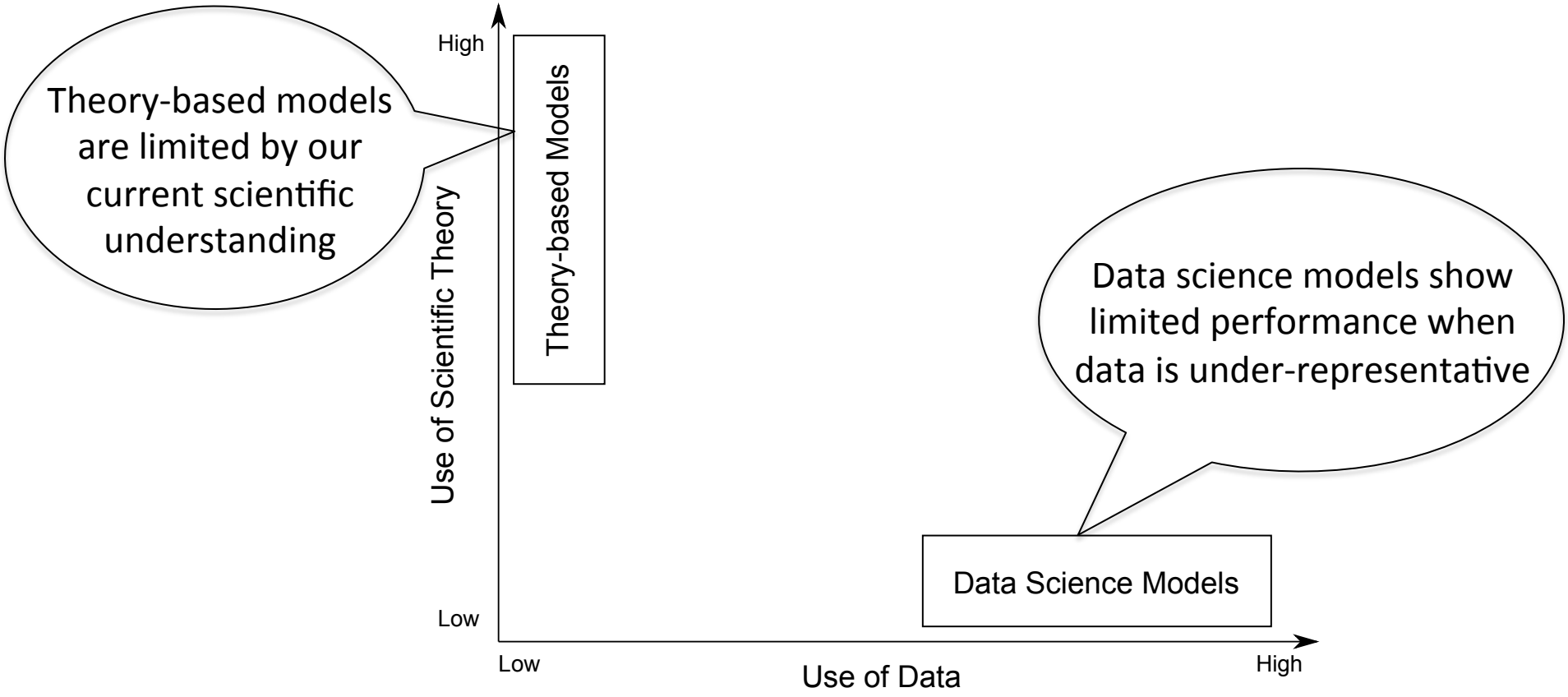
Why does "black-box" data science fail in scientific domains (1)

- *Poor predictive performance*
  - Many scientific problems are complex and under-constrained
  - May involve "non-stationary" relationships
  - Training and test sets only provide limited representations of overall phenomena
  - Relying solely on data can lead to *spurious* relationships that …
    - Look deceptively good on the training and test sets
    - Lead to poor generalization outside the available data

Why does "black-box" data science fail in scientific domains (2)

- *Poor scientific understanding*
  - Basic nature of goal is different in scientific domains
    - Solely improving predictive performance is not useful unless it **advances our scientific knowledge**
  - Non-interpretablity can be seen as *"lack of trustworthiness"*
    - A model for medical diagnosis is more reliable if it can be understood by a medical practitioner
  - Easy translation leads to faster adoption as scientific knowledge
    - Models can serve as building blocks of science only if they are easy to communicate to other scientists
  - End-goal is not the output but their resulting scientific insight
    - Detecting bio-markers for a certain disease is useful only if they are accompanied by a biological pathway explaining its causal relations

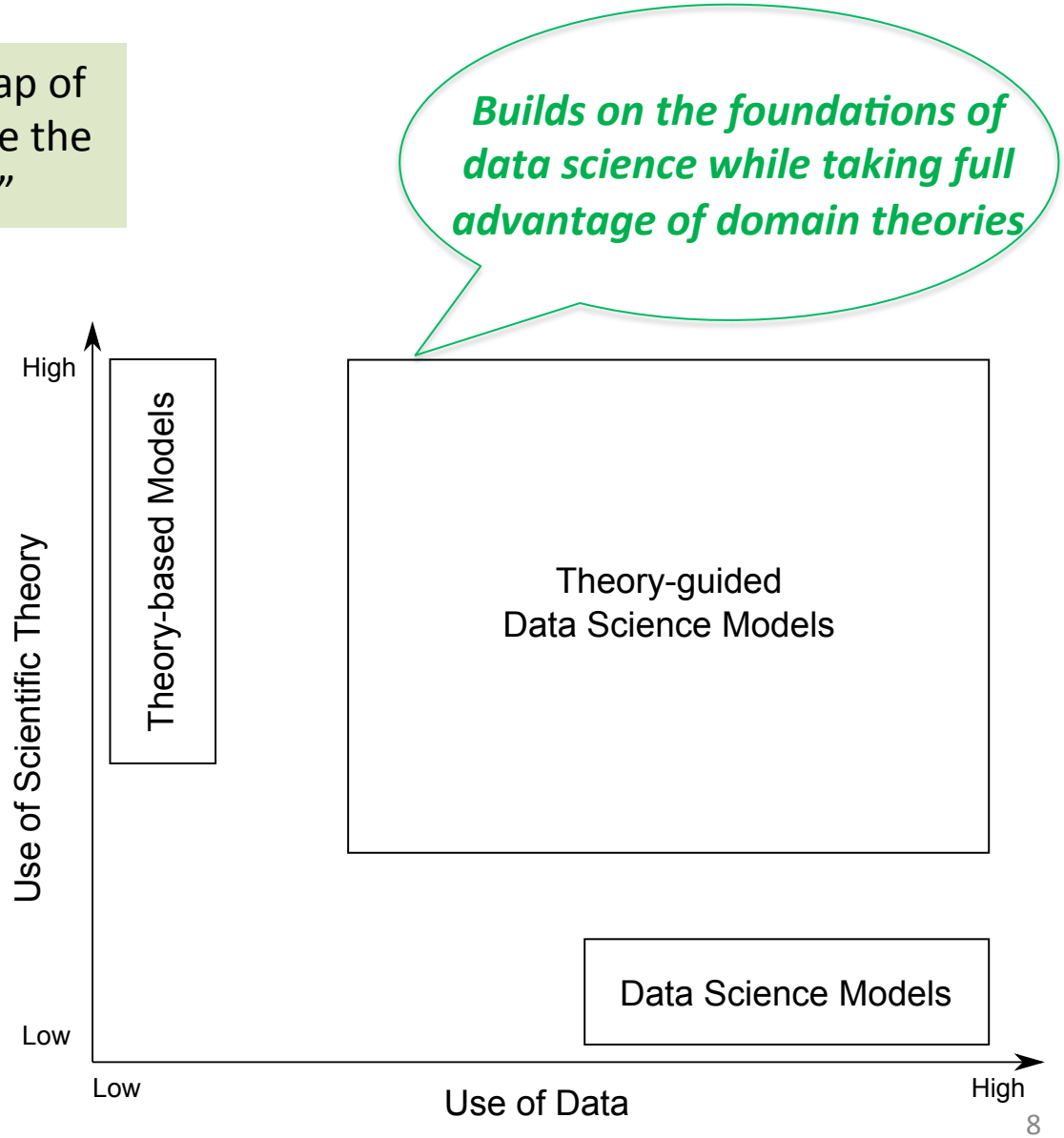
Dichotomy b/w Theory-based and Data Science Models



Both use incomplete sources of information about the two key components of knowledge discovery: *scientific theory* and *data*

Theory-guided Data Science (TGDS)

TGDS attempts to take the next leap of knowledge discovery to fully realize the vision of the "fourth paradigm"



Theory-guided Data Science (TGDS)

IEEE Computer

"*Theory-guided data science for climate change*"
Faghmous et al. 2014

IBM Research

"Physical Analytics" Research Division

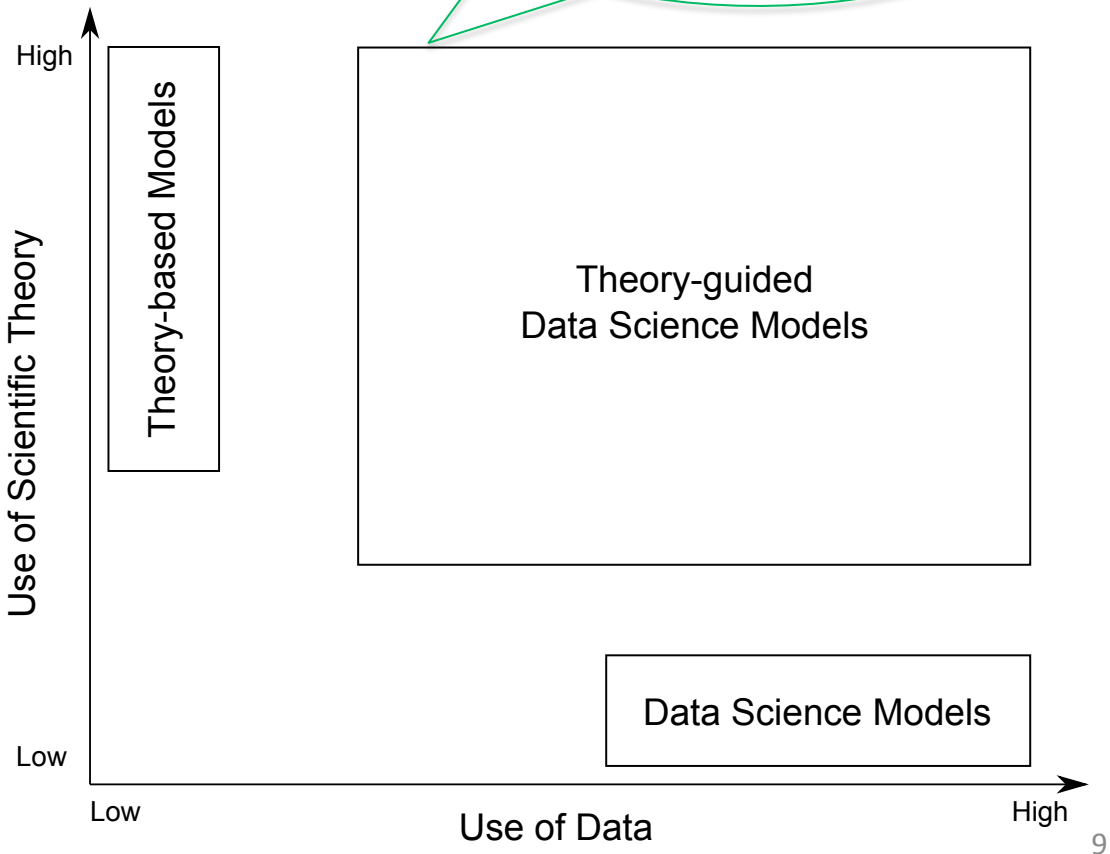
PHYSICS INFORMED MACHINE LEARNING

Symposium by Los Alamos National Laboratory, 2016

Builds on the foundations of data science while taking full advantage of domain theories

Emerging Applications of TGDS:

1. *Climate science and hydrology* [Faghmous et al. 2014 a,b; Khandelwal et al. 2015]
2. *Turbulence modeling* [Duraisamy et al. 2016; Xiao et al. 2016]
3. *Bio-medical science* [Wong et al. 2009; Xu et al. 2015]
4. *Bio-marker discovery* [Liu et al. 2013]
5. *Material discovery* [Hautier et al. 2010; Curtarolo et al. 2013]
6. Computational chemistry [Li et al. 2015]



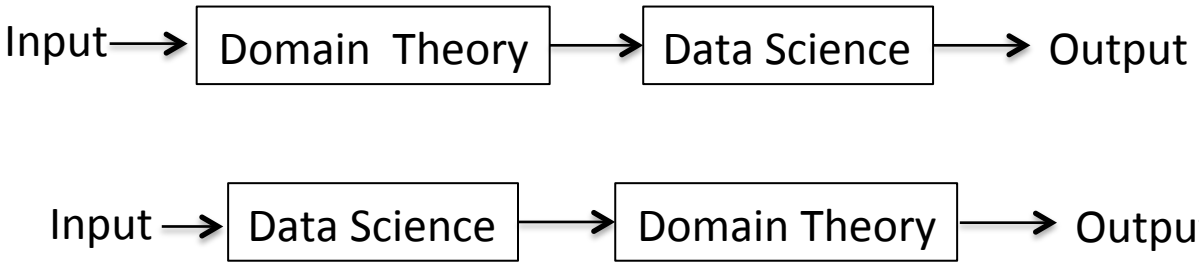
Ensuring Scientific Interpretability:
An Overarching Objective of TGDS

- Traditional measures of model performance motivated by …
  - Empirical estimates of model performance
    - Examples: training error, cross-validation error
  - Statistical Notions of Generalization Performance
    - Examples: MDL, VC-dimensions, AIC/BIC Criterion
- Empirical and statistical notions not sufficient to capture *"physical meaningfulness"* of the model
- Need to introduce scientific interpretability as a key component of model performance

$$\text{Performance} \propto \text{Error} + \text{Simplicity} + \textit{Interpretability}$$

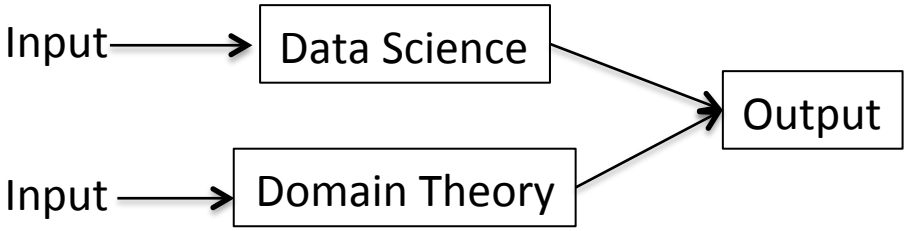
Research Themes in TGDS: An Overview

1. *Creating Composite Hybrid Models (output of one feeds into another)*



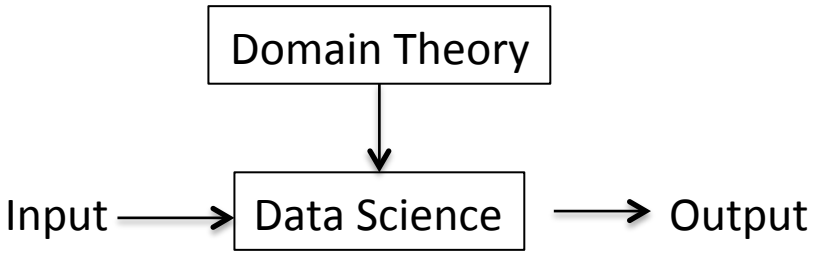
- Theory-based Feature Selection
- Feeding data science outputs in theory-based models

2. *Ensuring Compliance between Model Outputs*



- Data Assimilation
- Theory-based Supervision
- Theory-based Post-processing
- Theory-based Pruning

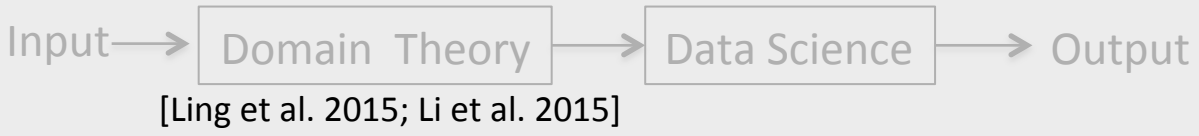
3. *Theory-guided Learning of Data Science Models*



- Theory-based Selection of Model Family
- Theory-based Parameter Initialization
- Theory-based Probabilistic Models
- Theory-based Constrained Optimization
- Theory-based Regularization

Research Themes in TGDS: An Overview

1. *Creating Composite Hybrid Models (output of one feeds into another)*

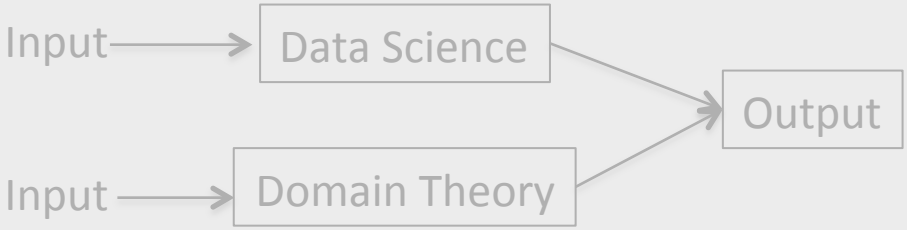


[Ling et al. 2015; Li et al. 2015]

[Duraisamy et al. 2016; Xiao et al. 2016]

- Theory-based Feature Selection
- Feeding data science outputs in theory-based models

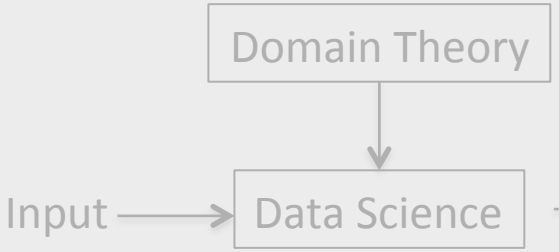
2. *Ensuring Compliance between Model Outputs*



- Data Assimilation
- Theory-based Supervision
- Theory-based Post-processing
- Theory-based Pruning

[Guo et al. 2016; Khandelwal et al. 2015; Hautier et al. 2010; Curtarolo et al. 2013]

3. *Theory-guided Learning of Data Science Models*



[Wong et al. 2009; Xu et al. 2015; Li et al. 2015; Liu et al. 2013]

- Theory-based Selection of Model Family
- Theory-based Parameter Initialization
- Theory-based Probabilistic Models
- Theory-based Constrained Optimization
- Theory-based Regularization

1. Creating Composite Hybrid Models (1)

- *Theory-guided Feature Selection*

  - Domain expertise commonly used for manually extracting features
  - Need for automated approaches to encode domain theories:
    - Variants of Lasso that encode domain structure (e.g. group-Lasso)
    - Principal components that not only capture the variance but can also be explained using domain theories
    - Manifold learning approaches that incorporate domain characteristics:
      - Scaling, translation, and rotational invariance in ***turbulence modeling*** [Ling et al. 2015]
      - Symmetry and near-symmetry of electron density in ***computational chemistry*** [Rupp et al. 2012; Li et al. 2015]

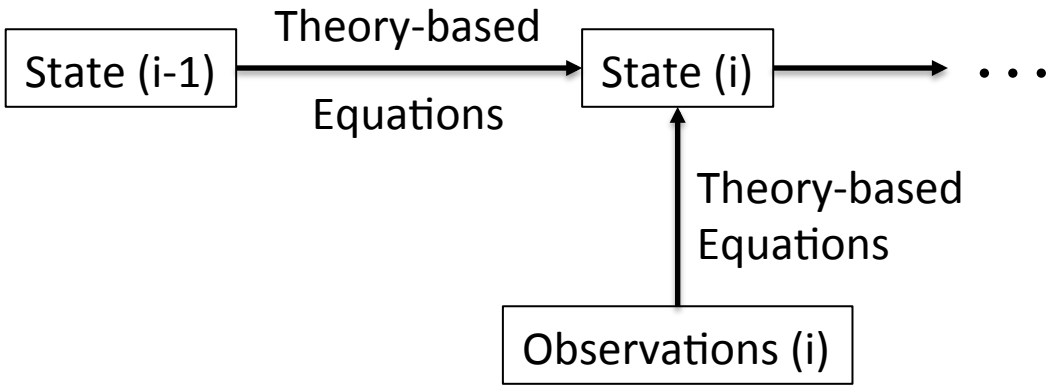
1. Creating Composite Hybrid Models (2)

- *Feeding Data Science Outputs in Theory-based Models*

  - Data science can predict intermediate quantities in theory-based models
  - Can help fill knowledge gaps where current scientific theories are *weak*
  - *Illustrative example:* TGDS models in ***turbulence modeling***
    - Reynolds-averaged Navier-Stokes (RANS) models are standard tools for modeling turbulent flow
    - Provides poor approximation in modeling complex flows with separation, curvature, or swirling
    - Recent approaches explored data science methods to improve RANS models
      - Random forests to predict the discrepancy of RANS models [Xiao et al. 2016]
      - ANN models to augment RANS approximations [Duraisamy et al., 2016]

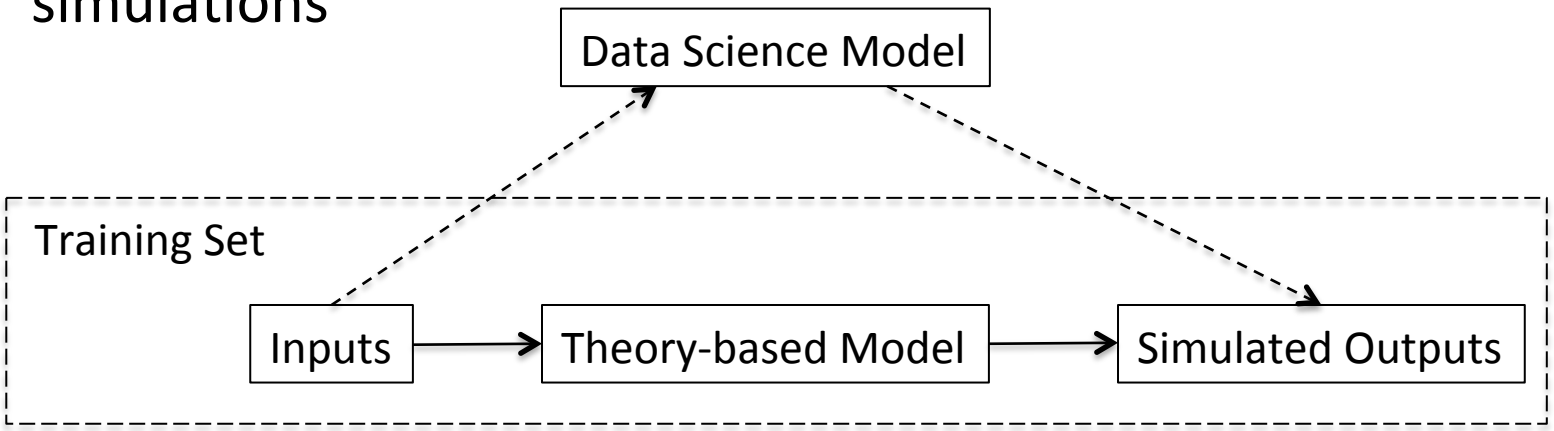
2. Ensuring Compliance between Outputs (1)

- *Data Assimilation*
  - Common approach for calibrating theory-based models using observational data in ***climate science*** and ***hydrology***
  - Use state-based representation of dynamic physical processes
  - State depends on previous states and current observations
    - Dependencies are modeled using theory-based equations
  - Model outputs are constrained to comply with observational data



2. Ensuring Compliance between Outputs (2)

- *Theory-based Supervision*
  - Training data science models using theory-based model simulations



  - Minimizing the training error ensures compliance between theory and data science
  - Types of theory-based supervision:
    - Accurate but costly [***Computational Fluid Dynamics:*** Guo et al. 2016]
    - Approximate but inexpensive (*weak supervision*), e.g. RANS model simulations

2. Ensuring Compliance between Outputs (3)

- *Theory-based Post-processing*
  - Data science outputs can be refined using domain theory
  - Post-processing commonly used with spatio-temporal outputs
    - Spatial smoothing using majority filters or MRF-based methods
    - Temporal smoothing using HMM-based and frequency-based methods
  - Domain constraints among outputs may be *latent* in nature and not explicitly known
    - Requires simultaneous estimation of outputs and latent constraints
    - *Illustrative example* in ***surface water mapping***:
      - Classification of water bodies using implicit elevation-based constraints [Khandelwal et al. 2015]

2. Ensuring Compliance between Outputs (4)

- *Theory-based Pruning*
  - Data science outputs can be pruned to discover patterns that are consistent with scientific understanding
  - *Illustrative example*: TGDS models for ***material discovery***
    - Objective is to find novel materials that show a desirable property
    - Existing methods using quantum chemical models are expensive and cannot be scaled to the exponential number of candidate materials
    - Probabilistic models can predict candidate structures that can be further pruned using theory-based models

      [Fischer et al. 2006; Hautier et al. 2010; Curtarolo et al. 2013]
    - Helped in the discovery of a *hundred new ternary oxide compounds* that were previously unknown [Hautier et al. 2010]

3. Theory-guided Learning of Data Science Models (1)

- *Choosing Family of Data Science Models*
  - Use same form of relationship b/w input and output as that suggested by domain knowledge
  - Examples:
    - Selecting suitable combinations of loss and response functions
    - Use of appropriate link functions in generalized linear models (GLM)
    - Design of ANN architecture can be inspired by domain knowledge
      - Number of hidden layers
      - Nature of connections among layers
      - Sharing of model parameters among nodes
      - Decomposing ANN model into modular components representing sub-processes of the physical system

3. Theory-guided Learning of Data Science Models (2)

- *Constraining Model Parameters*
  - Theory-based constraints can help reduce the model search space and learn generalizable models that are physically meaningful
  - Approaches:
    - Theory-based Parameter Initialization
      - ANN models can be *pretrained* using theory-based simulations
    - Theory-based Probabilistic Models
      - Use of theory-based graphical models and priors

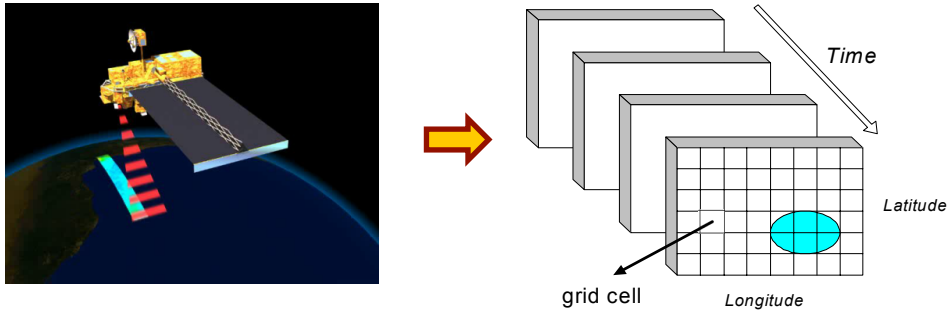
      [***Electrophysiological imaging of the heart***: Wong et al. 2009; Xu et al. 2015]
    - Theory-based Constrained Optimization Methods
      - Using Theory-based equations as constraints in data science models

      [ML-based density functionals in ***computational chemistry***: [Li et al. 2015]

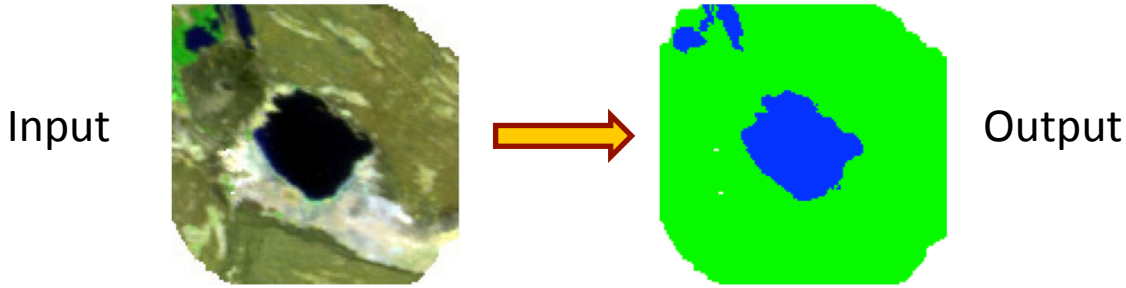
      [***Surface water mapping*** using elevation constraints]
    - Theory-based Regularization Techniques
      - [***Bio-marker Discovery***: Liu et al. 2013]

Case Study: Mapping Surface Water Dynamics

- **Binary Classification Problem**
  - Attributes: Multi-spectral data
    - MODIS (at 500m spatial resolution, daily since 2000)
    - Landsat (at 30m spatial resolution, every 16 days since 1970s)



  - Binary Labels (at every location and time): Water or Land

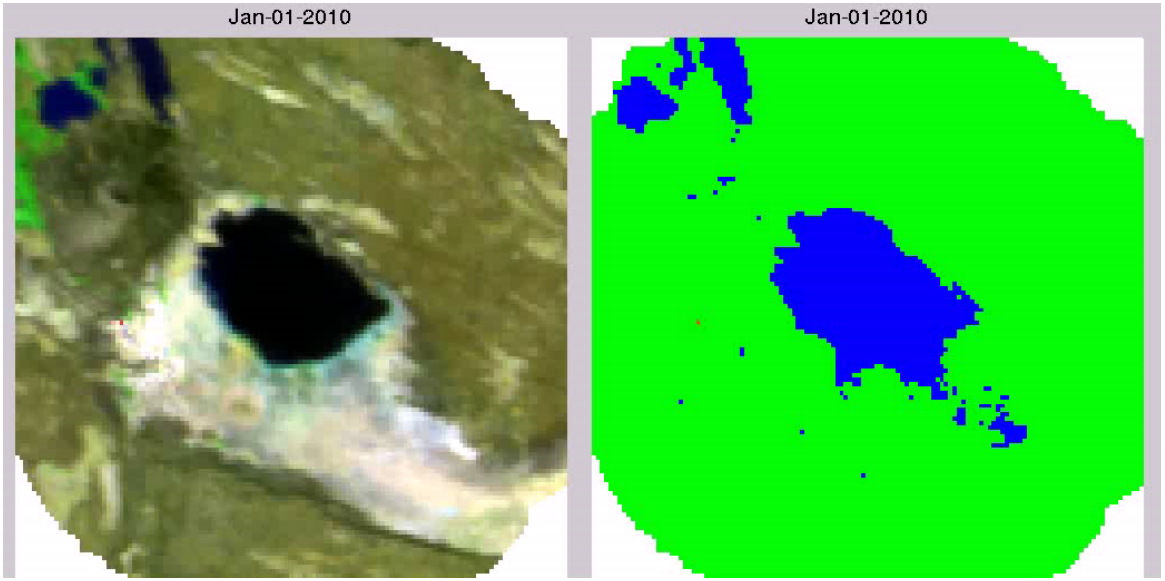


  - Ground-truth labels:
    - Available on specific dates from various sources (GLWD, SWBD)

Limitations for Existing Supervised Learning Methods

Even after training with ample labeled samples, classification performance is poor due to heterogeneity and poor data quality [Karpatne et al. 2014]

Lake Abhe (on the border of Ethiopia and Djibouti)



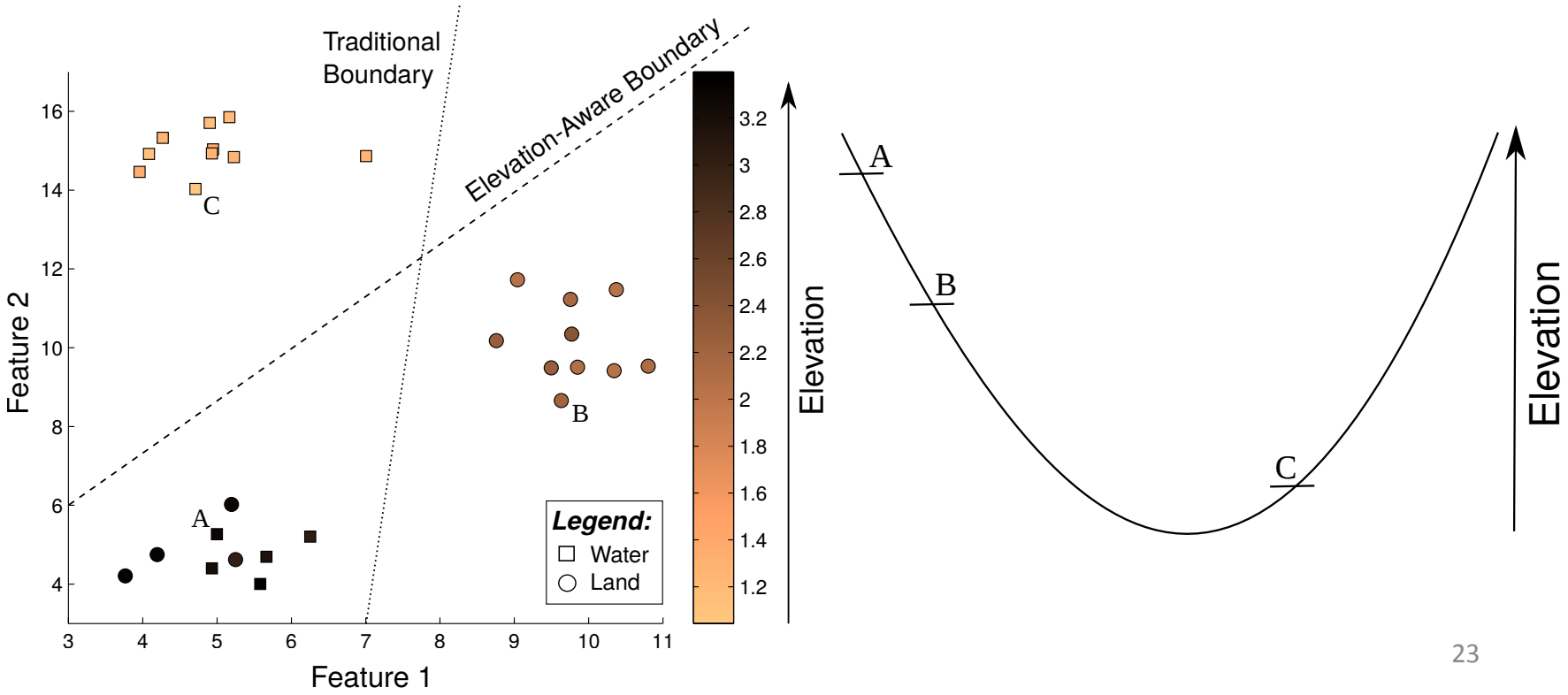
Multispectral Composites

Classification Results

TGDS Approach 1:

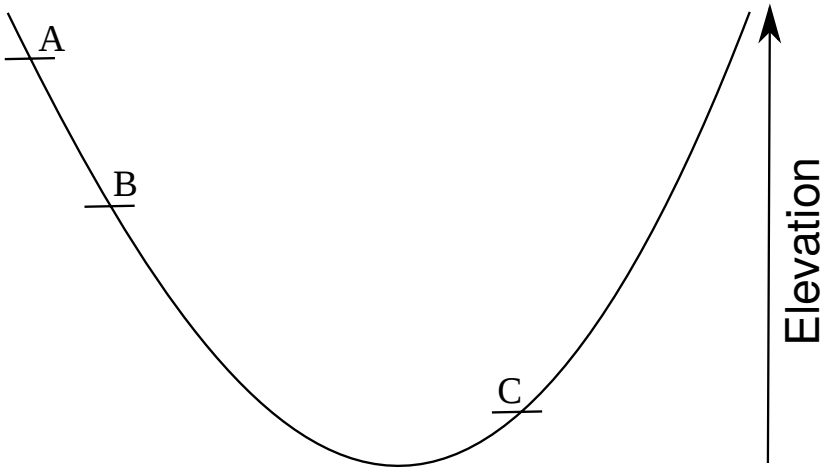
Learning using Elevation-based Constraints

- Knowledge about the elevation structure of lakes can help in the learning of classification boundary
- *Theory-guided constrained optimization*
  - **Constraint**: Locations at a lower elevation should be labeled as *water* if higher-elevation locations are labeled as *water*



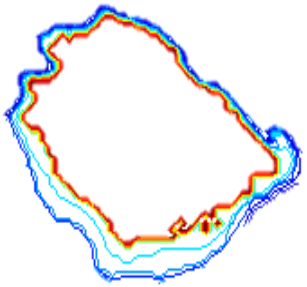
TGDS Approach 2:
Post-processing Outputs using *Latent* Elevation-ordering

- Knowledge about elevation structure is scarcely available
  - Requires use of expensive sonar-based instruments
- Elevation-based ordering is thus *hidden* and needs to be inferred
- *Potential approach* [Khandelwal et al. 2015; Mithal et al. 2016]:
  - Use the history of imperfect labels to estimate *relative* elevation of locations



The SELPh Approach:
Simultaneous Estimation of Labels and Physics

Initial classification map



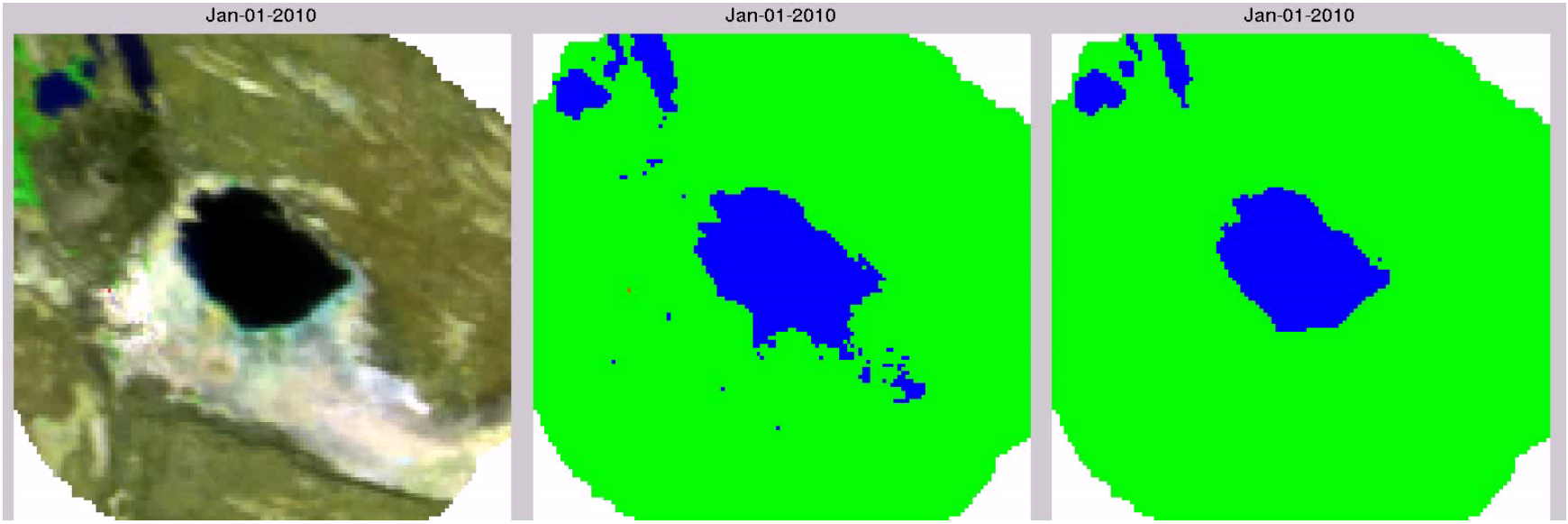
Depth contours

Improved Classification map

Step 1: Estimate Depth Contours
Step 2: Improve Classification Labels
Iterate until convergence

Illustration of the SELPh Approach

Lake Abhe



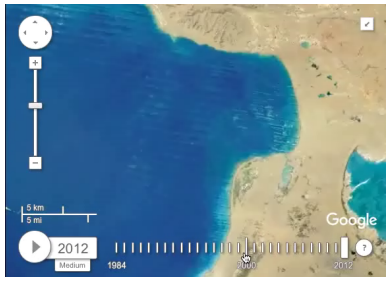
Multispectral composites

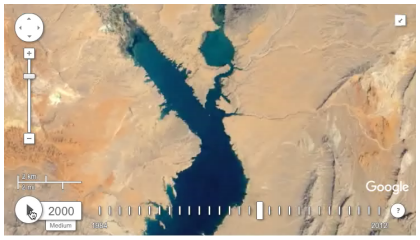
Classification results using traditional framework

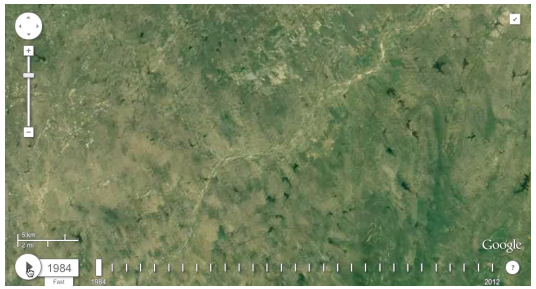
Classification results using SELPh

A Global Surface Water Monitoring System
http://z.umn.edu/monitoringwater

- Maps the dynamics of all major surface water bodies (surface area > 2.5 km^2) shown as *blue dots*







Key Highlights:

- Detects melting of glacial lakes
- Maps changes in river morphology
- Identifies reservoir constructions
- Finds relationships b/w surface water and precipitation/groundwater

Concluding Remarks

- "Black-box" data science models not sufficient for knowledge discovery in scientific domains
- Need to deeply integrate theory and data science for accelerating knowledge discovery using the *fourth paradigm*
- The paradigm of "Theory-guided data science" (TGDS) holds great promise for knowledge discovery in scientific applications
- TGDS is already starting to gain attention in several disciplines:
  - Climate science and hydrology
  - Turbulence modeling
  - Bio-medical science
  - Bio-marker discovery
  - Material discovery
  - Computational chemistry

Thank You! Questions?

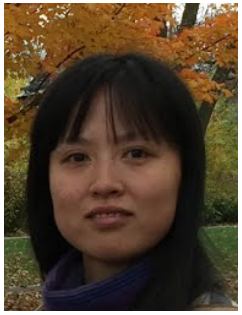
UMN Graduate Students

Anuj Karpatne

Ankush Khandelwal

Varun Mithal

Guruprasad Nayak


Xi Chen

Xiaowei Jia

Saurabh Agrawal

UMN Team Members
Arindam Banerjee, Snigdhansu Chatterjee, Stefan Liess, Shashi Shekhar, Michael Steinbach

NSF Expeditions Collaborators
NCSU: Nagiza Samatova, Fredrick Semazzi; Northeastern: Auroop Ganguly; North Carolina A&T: Abdollah Homaifar, Fred Semazzi

External Collaborators
NASA Ames: Rama Nemani, Nikunj Oza; IonE, UMN: Kate Braumann; UCLA: Dennis Lettenmaier, Miriam Marlier

Work supported by NASA and NSF Expeditions in Computing project on Understanding Climate Change using Data-driven Approaches