Machine learning for data-driven discovery in solid Earth geoscience

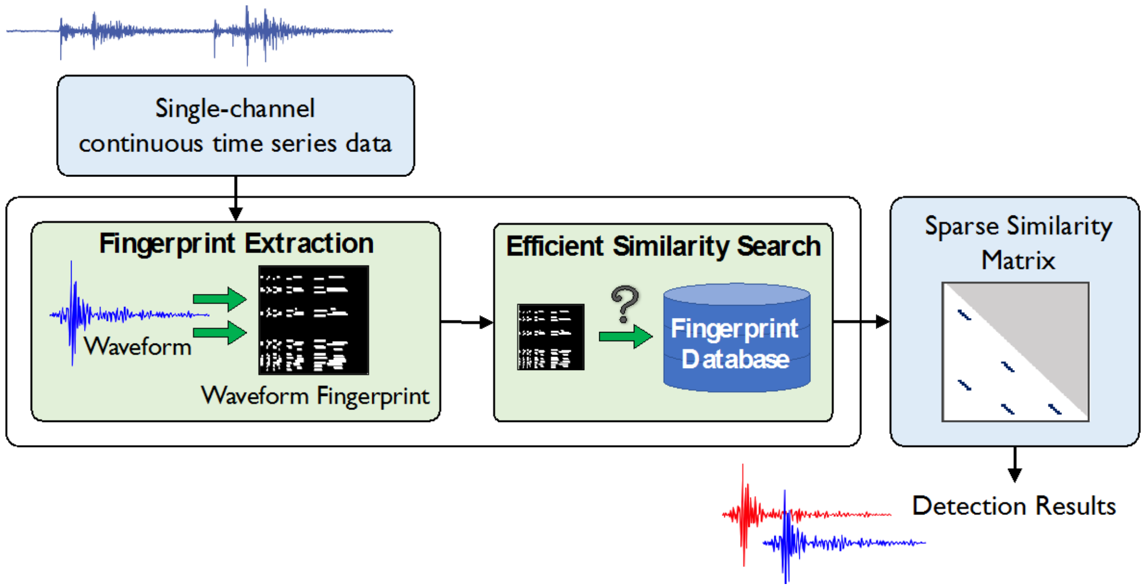
Karianne J. Bergen, Ph.D.

Harvard Data Science Initiative Postdoctoral Fellow

Harvard John A. Paulson **School of Engineering** and Applied Sciences

Beyond the Black Box: The Future of Machine Learning and Data-Intensive Computing in the Solid Earth Geosciences

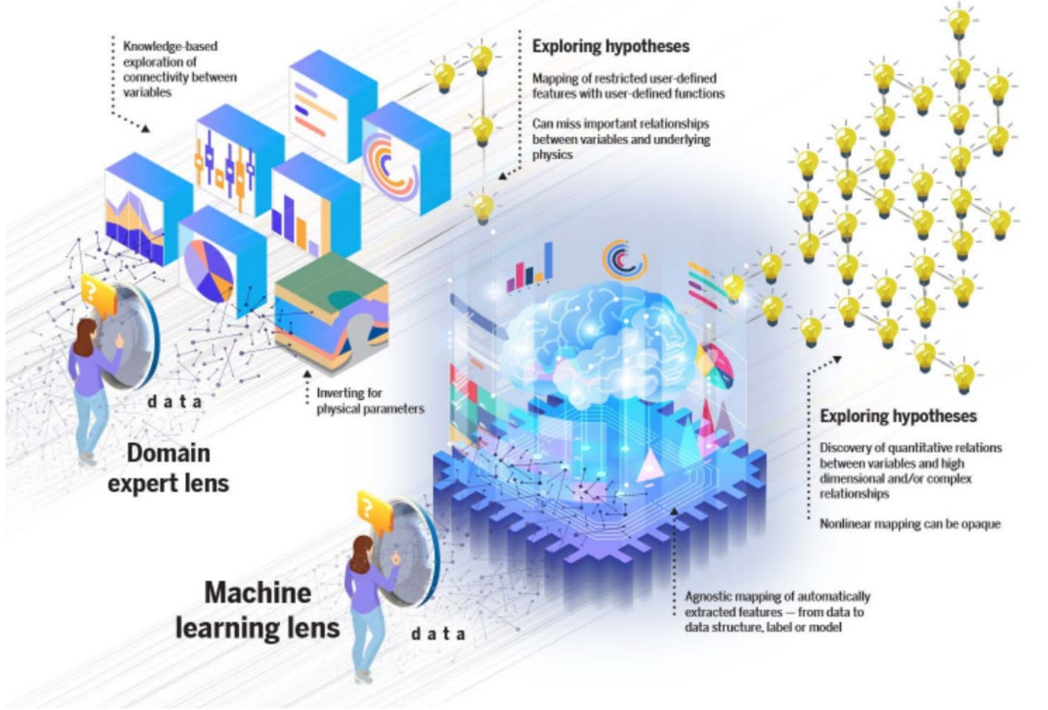
FAST: Earthquake detection by efficient similarity search



GEOPHYSICS

Machine learning for data-driven discovery in solid Earth geoscience

Karianne J. Bergen[1,2], Paul A. Johnson[3], Maarten V. de Hoop[4], Gregory C. Beroza[5]*



What is Machine Learning?

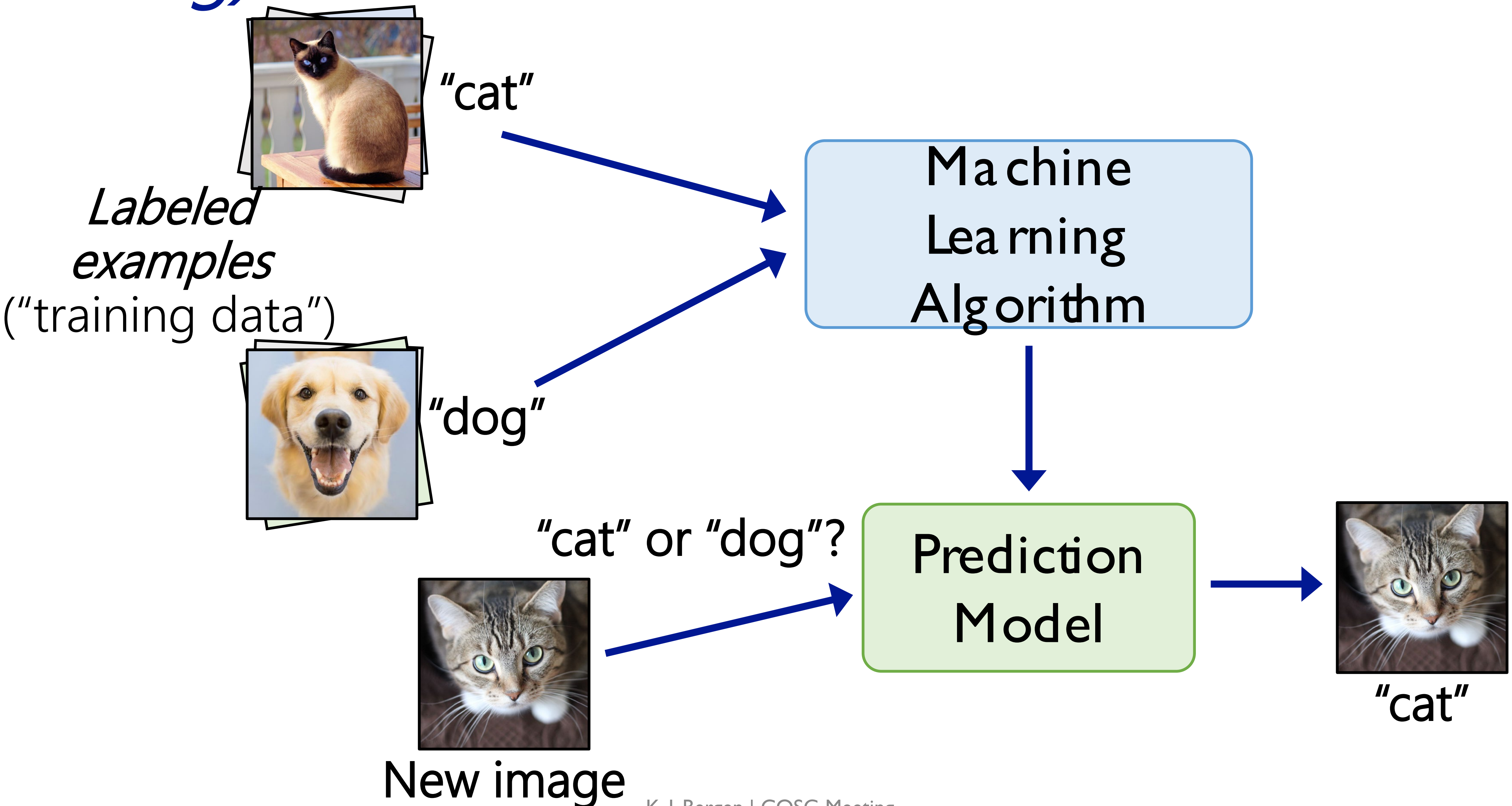
Machine learning (ML)

a set of tools for extracting patterns and building predictive models from data.

Data mining

tools for extracting unknown patterns or information from large data sets

Building models from examples *(Supervised learning)*



Finding patterns in data *(Unsupervised learning)*

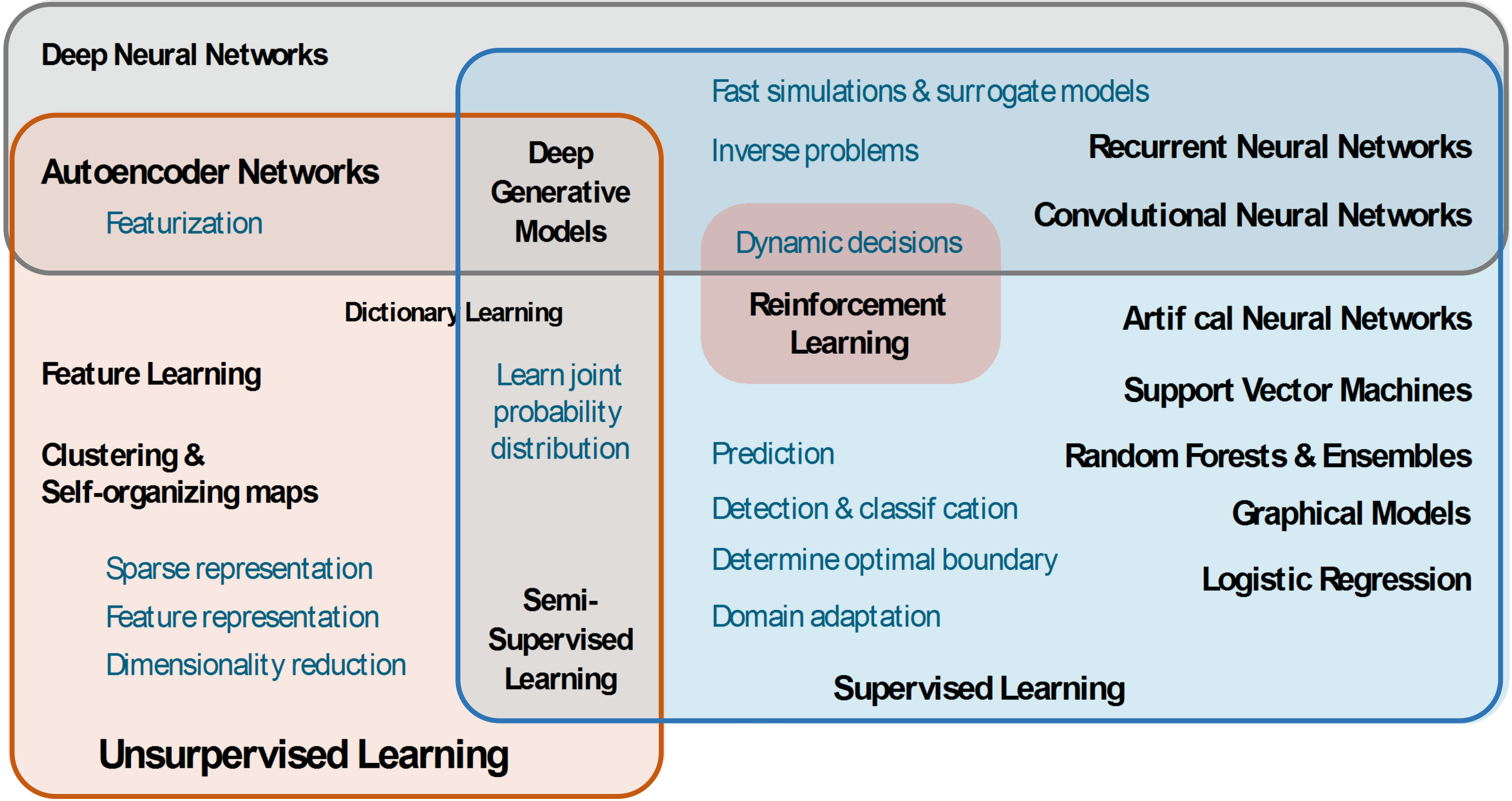
Images only (no labels)

Machine Learning Algorithm

Structure in the data

Groups of similar images

Common patterns in images



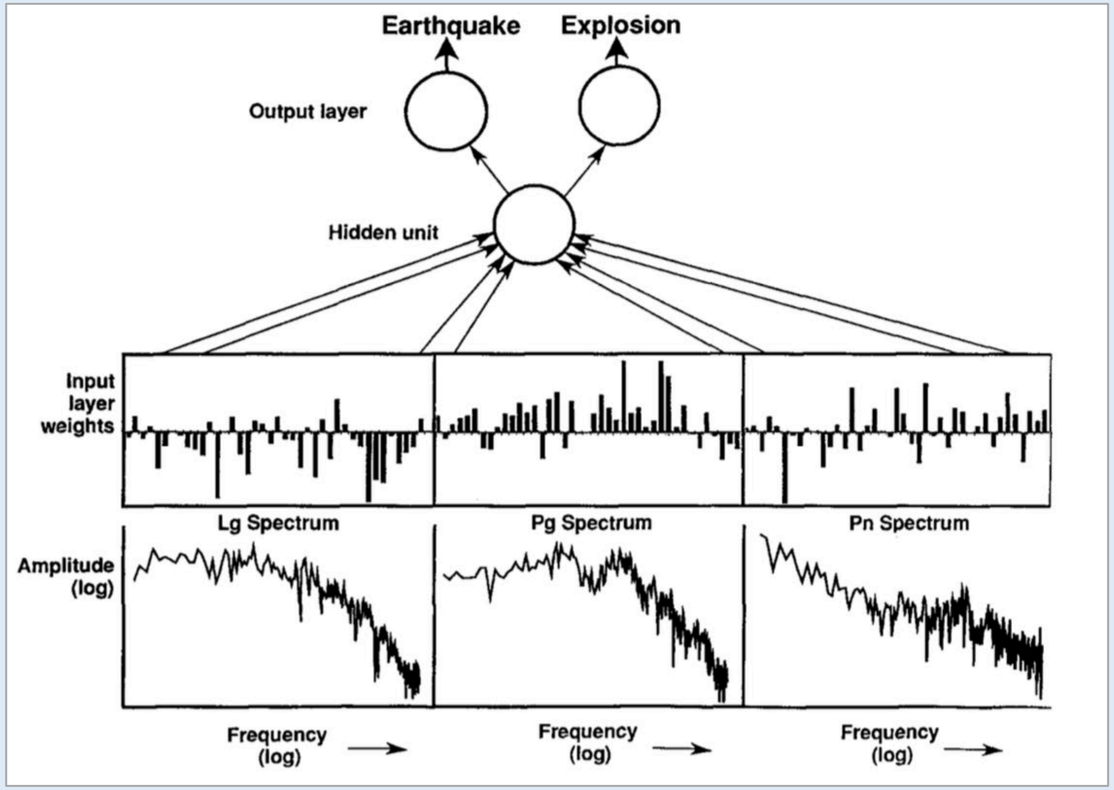
[Bergen *et al.* (2019), *Science*]

Machine learning can help geoscientists extract more knowledge & insights from larger data sets than ever before.

Geoscientists have been using ML for decades.

Artificial Neural Networks

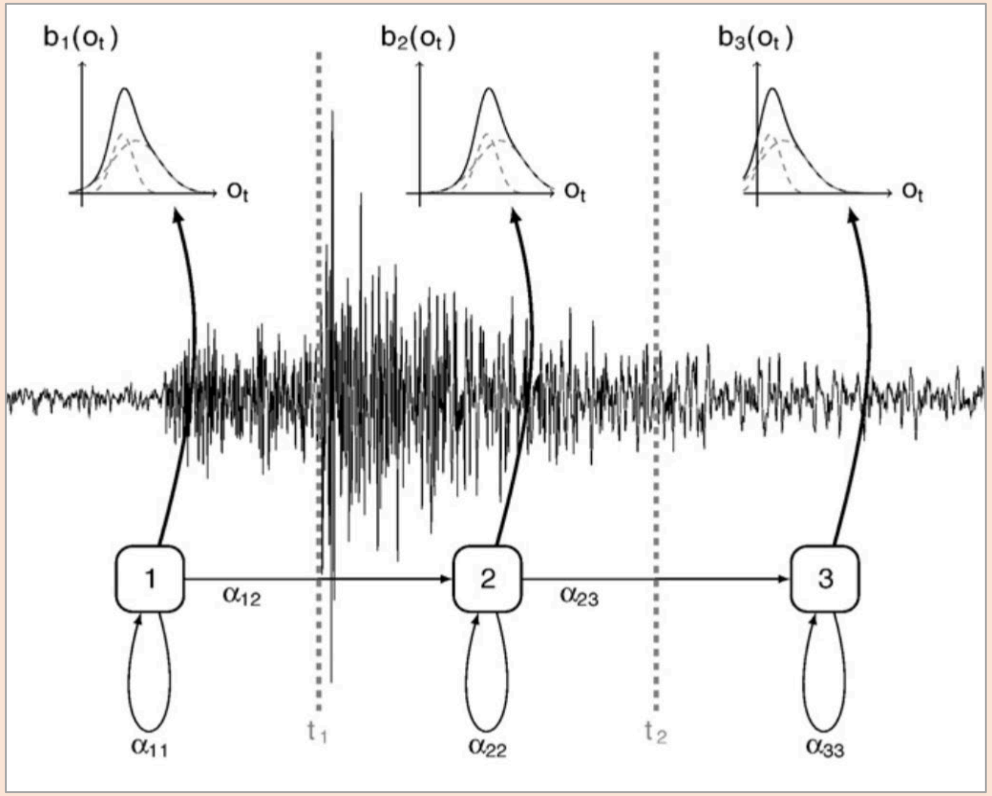
(e.g. Dowla *et al.*, 1990; Dysart & Pulli, 1990)



[Dowla *et al.*, 1990]

Hidden Markov Models

(e.g. Ohrnberger, 2001; Beyreuther *et al.*, 2008)

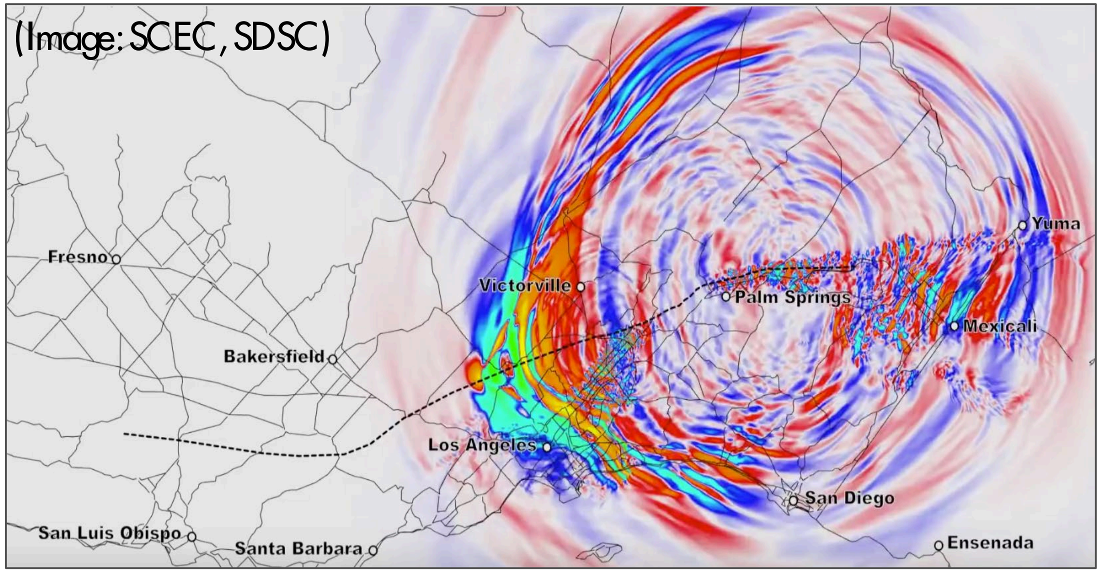


[Beyreuther *et al.*, 2008]

Recent developments have created new opportunities for scientific discovery with ML in solid Earth geoscience.

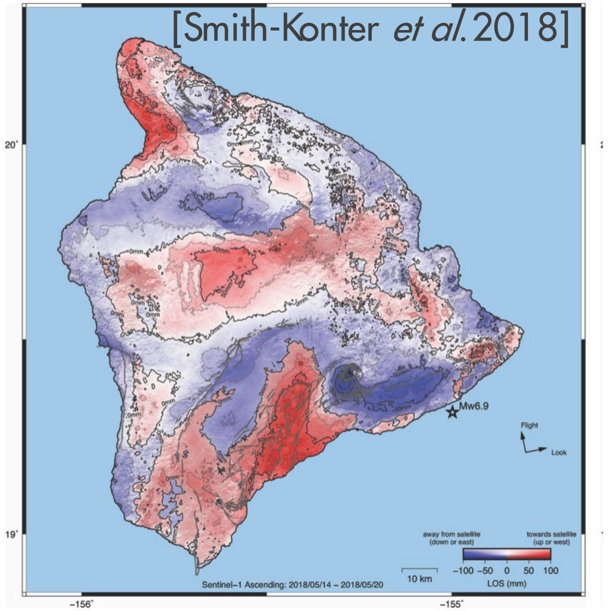
1) Massive geoscience data sets
2) New ML algorithms and models
3) Improvements in computing technology & tools

Massive geoscience data sets

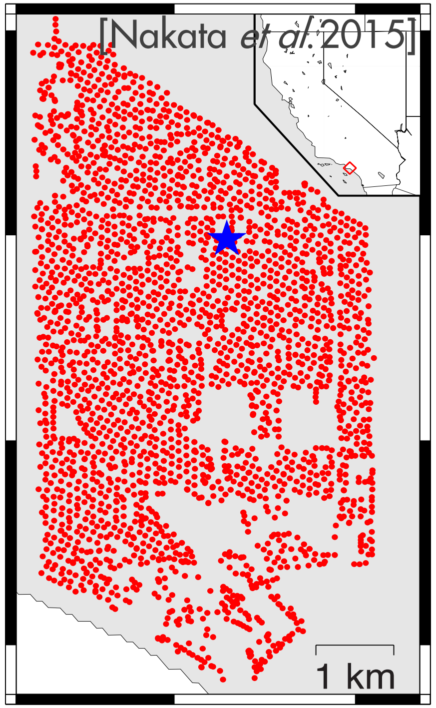


Large-scale simulations

Long-duration continuous observations

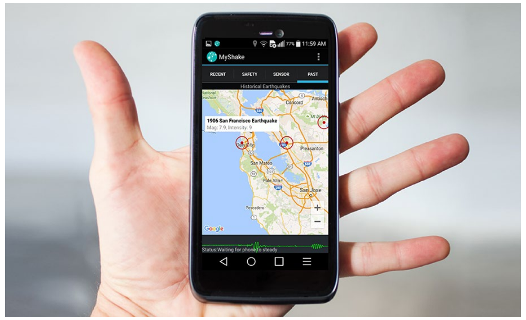


Spatiotemporal & remote sensing data



Dense sensor arrays

Crowdsourced data

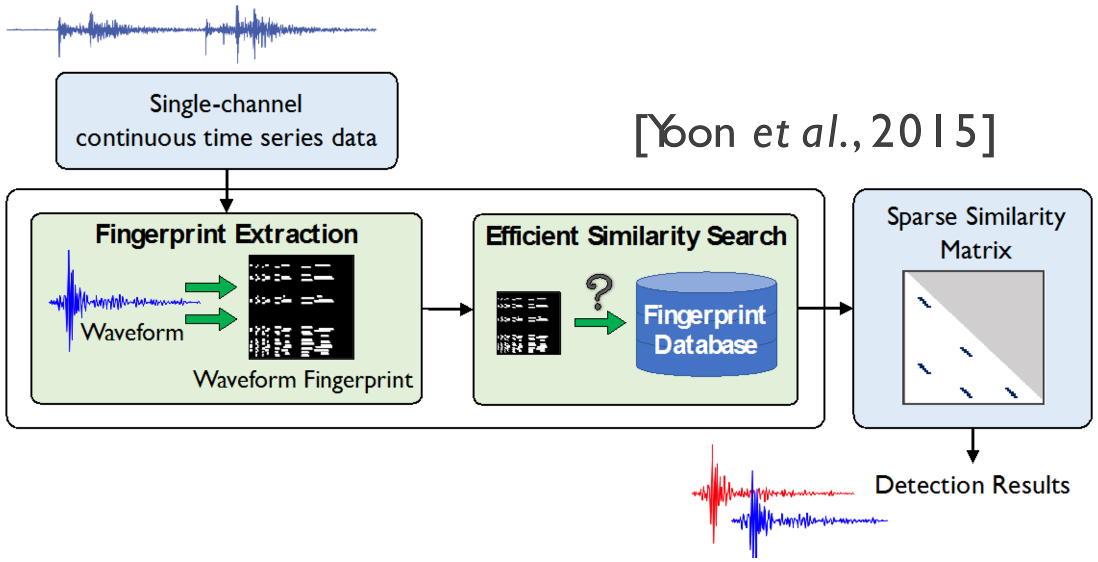


MyShake™

New ML algorithms and models

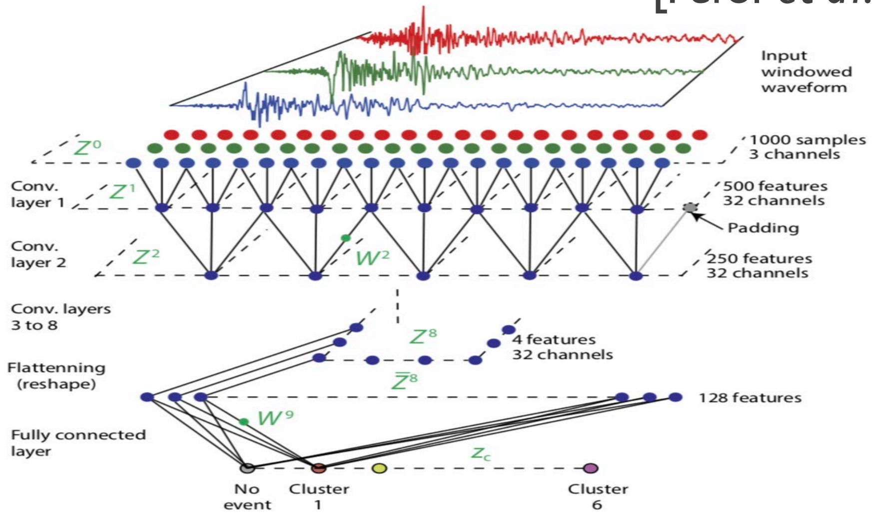
Large-scale Data Mining (FAST)

[Yoon *et al.*, 2015]

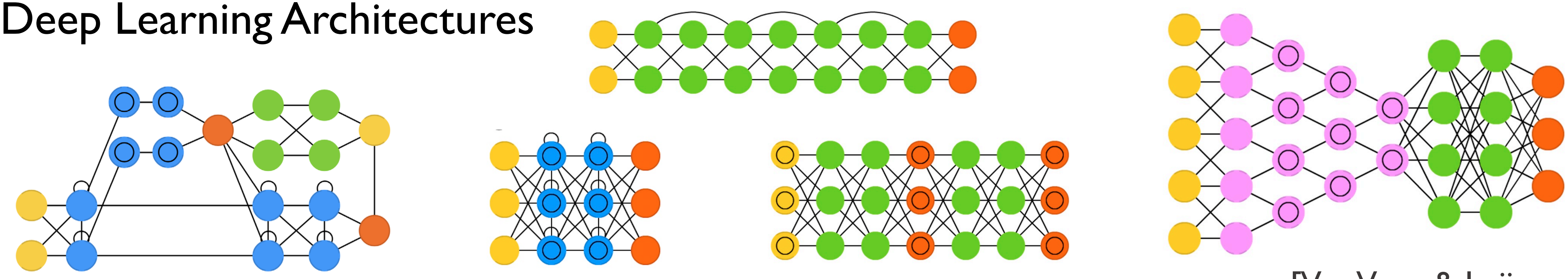


Convolutional Neural Network (ConvNetQuake)

[Perol *et al.*, 2018]



Deep Learning Architectures



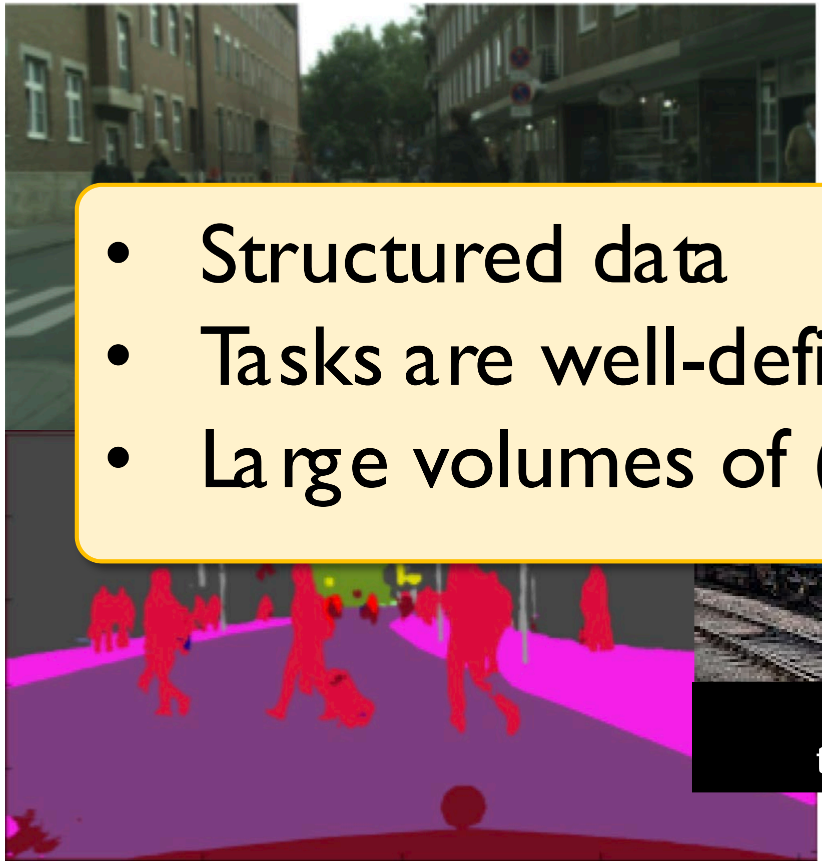
[Van Veen & Leijnen, 2019]

Improvements in computing technology & tools

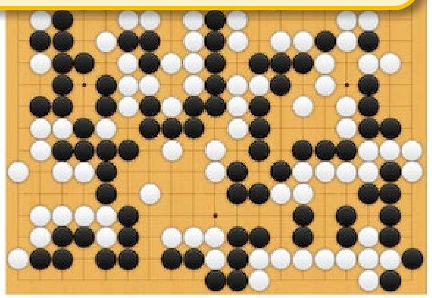
Where has ML been particularly successful?

[Treml *et al.* 2016]

[DeepMind]



[Google AI blog]



- Structured data
- Tasks are well-defined and/or easy for humans to perform
- Large volumes of (high-quality) labeled data

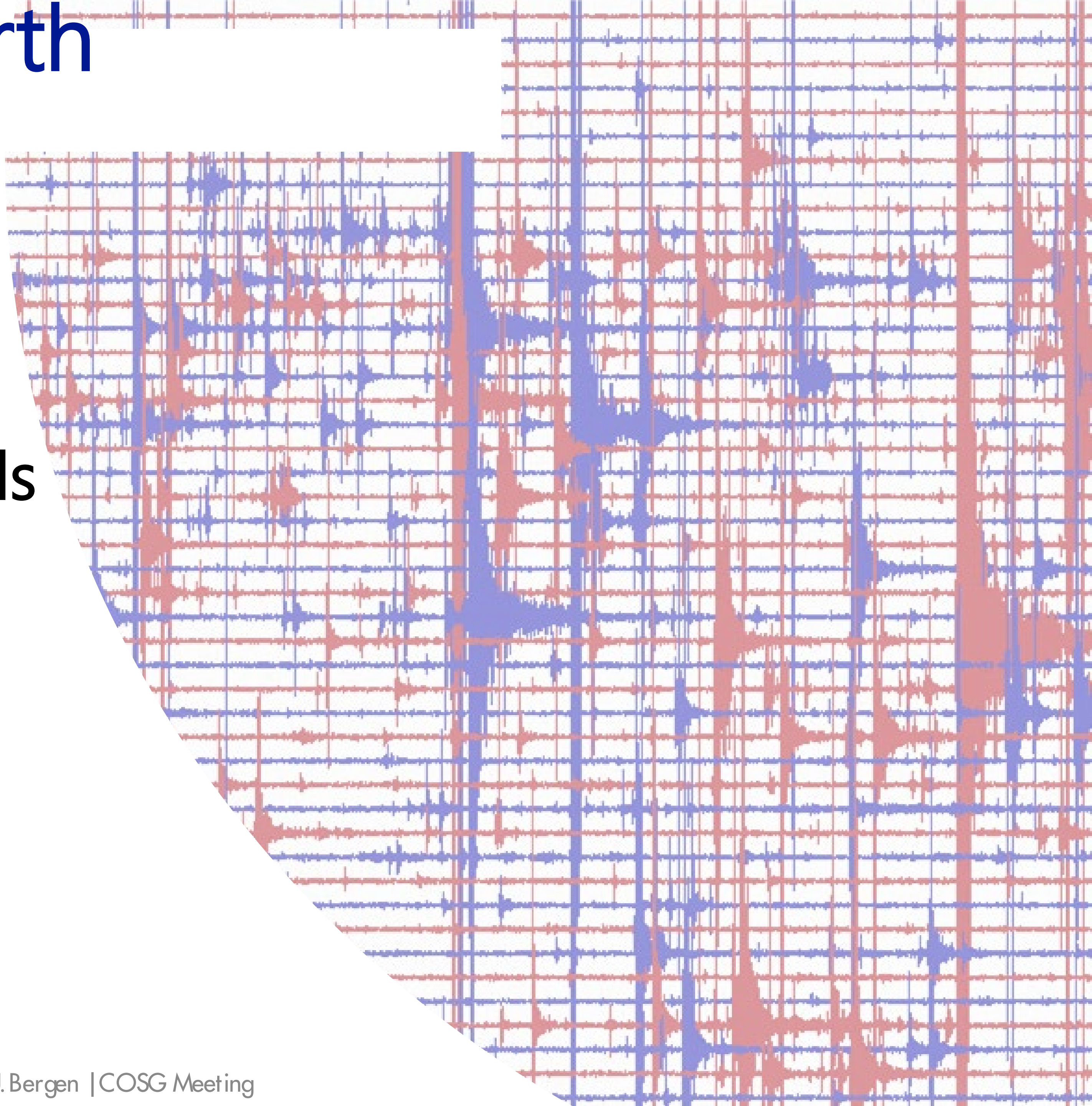
Image Processing
data on a grid: images

Language Processing
sequential data: text, audio

Game Play

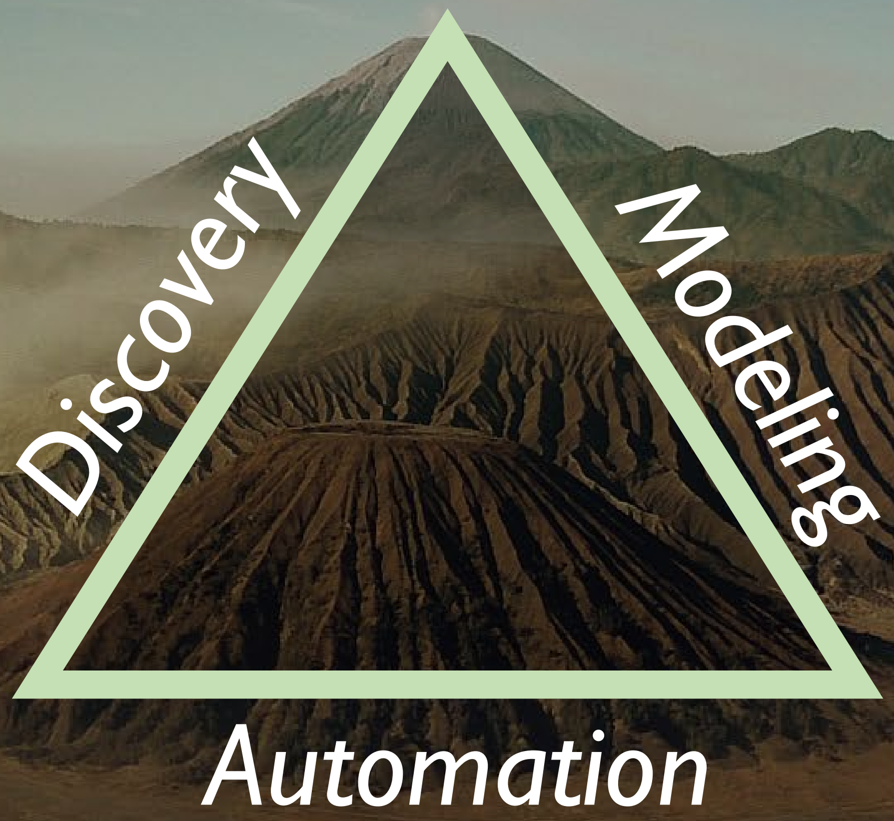
Data sets in solid Earth geoscience

- Low signal-to-noise data
- Heterogenous noise
- Limited or low-quality labels
- Ground truth often unavailable
- Massive data sets require scalable methods
- Modeling across multiple scales (spatial, temporal)



How is machine learning being used in solid Earth geoscience today?

How might it be used in the near future?

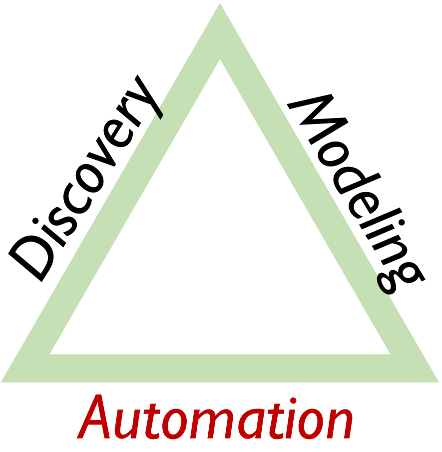


Automated prediction & analysis

Goal: Perform a complex or repetitive task

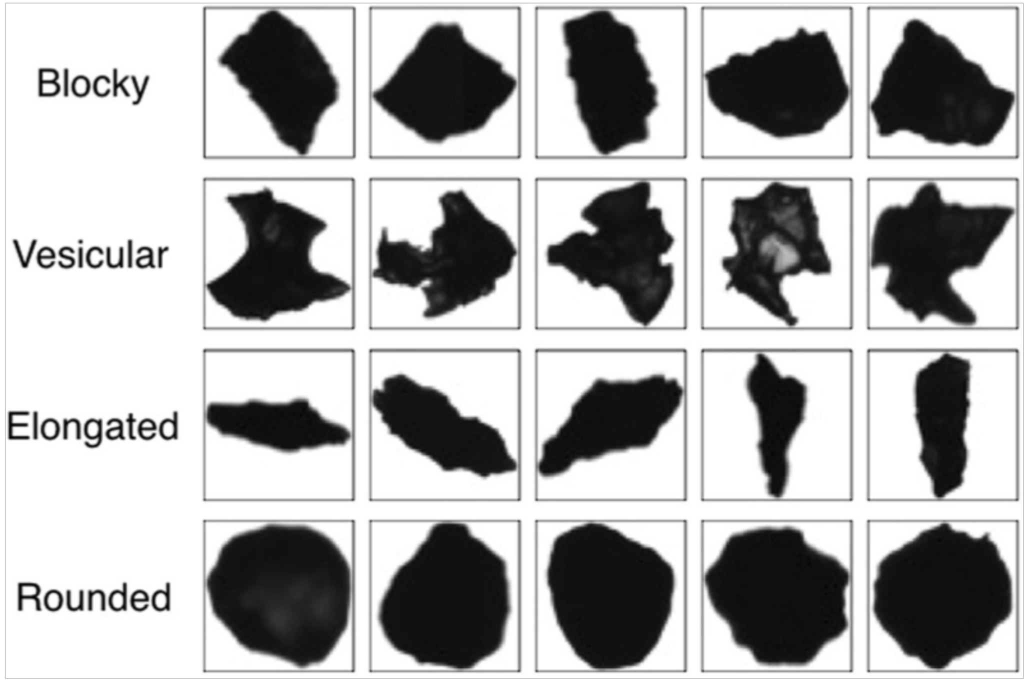
a) challenging, tedious or infeasible for an analyst to perform, OR
b) difficult to express as a set of explicit commands

- ML as a tool for high accuracy predictions



Automated data analysis

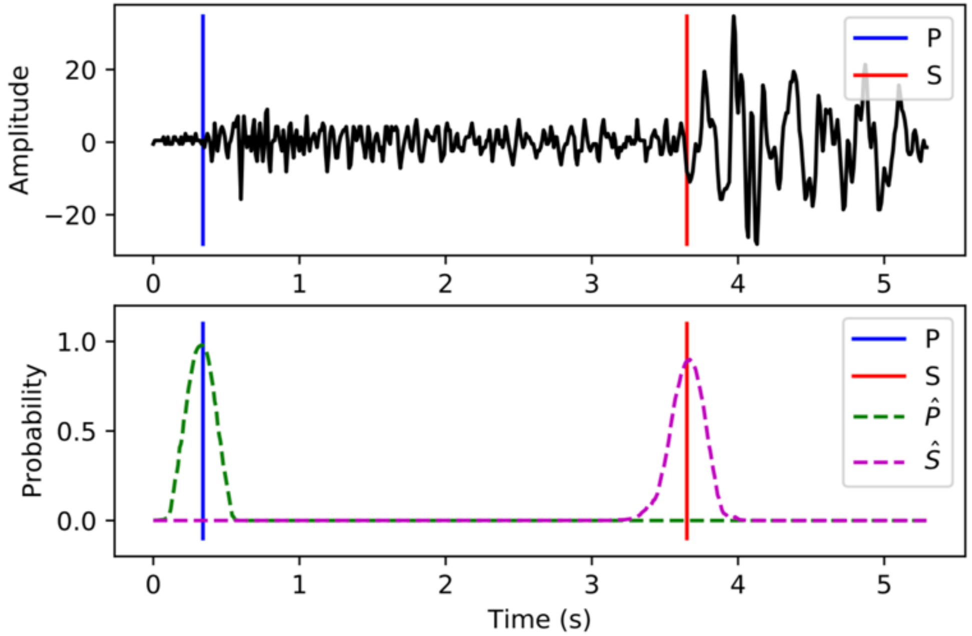
Classifying volcanic ash particles



Tedious task for an analyst, well-suited to automation with ML

[Shoji *et al.* (2018), *Scientific*

Phase-picking in seismic data

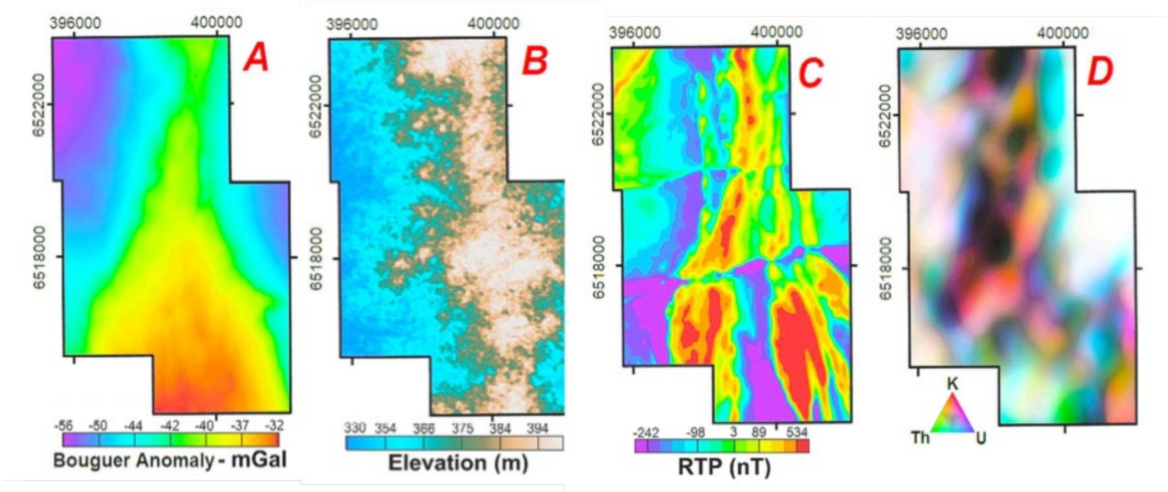


Automating, improving accuracy of existing analysis pipeline

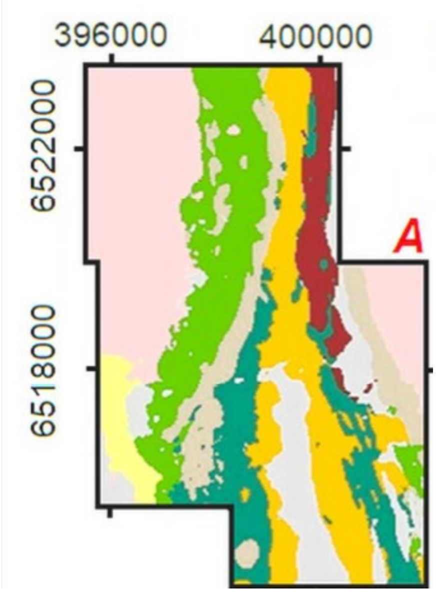
[Zhu & Beroza (2018), *Geophys J.*

Automated prediction

Lithological mapping



Geophysical & remote sensing data



Lithology

Learning to predict lithology from sparse ground-truth measurements

[Kuhn *et al.*, (2018), *Geophysics*.]

Machine learning challenges in solid Earth geoscience

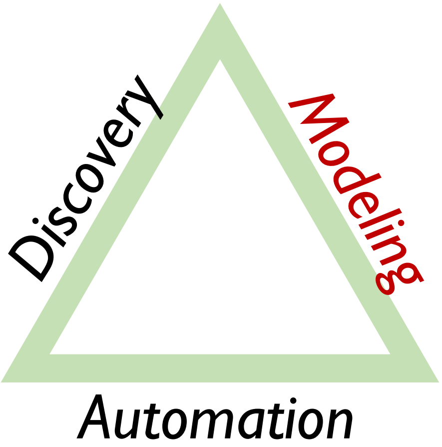
- Dataset shift, covariate shift
- Biases in data collection, labeling
- Evaluating performance

Automation

Modeling & inversion

Goal: Create a representation that captures relationships or structure in a data set.

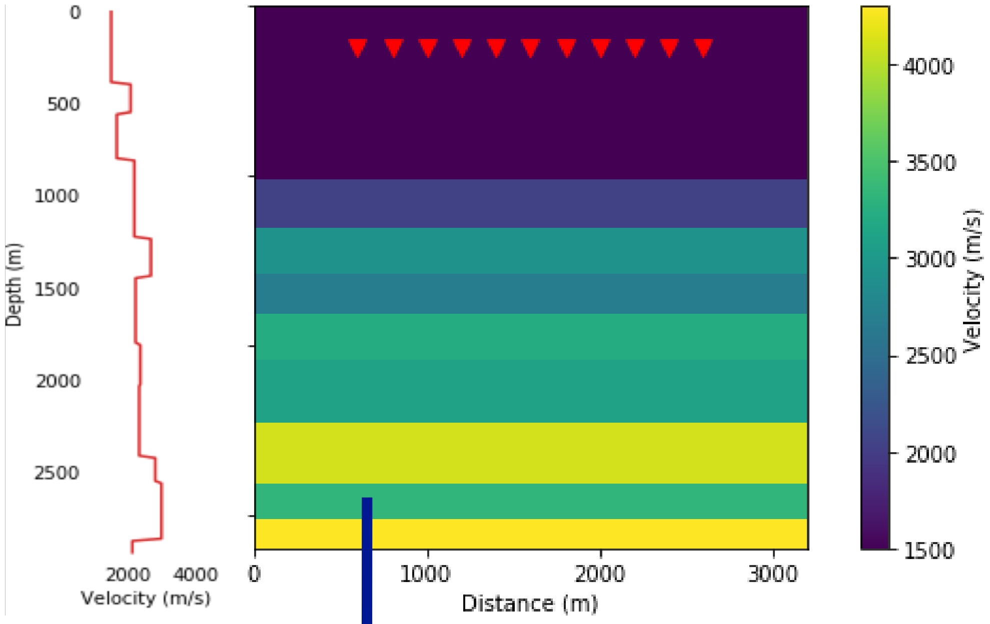
- Learning surrogate models
- Model reduction & coarse-graining
- Intersection of ML & numerical simulations



Modeling: *Simulating wavefields*

[Moseley *et al.* (2018), *arXiv*]

1D velocity model



data

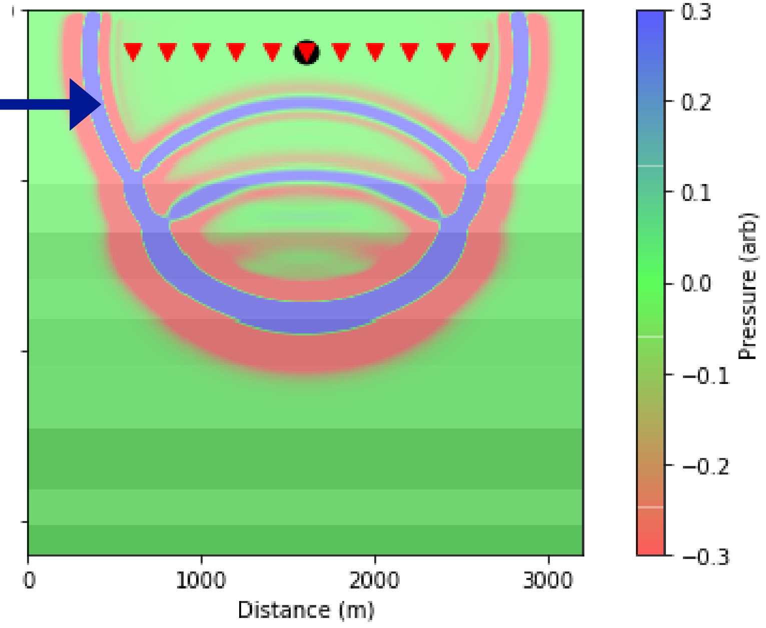
targets

Deep Neural Network

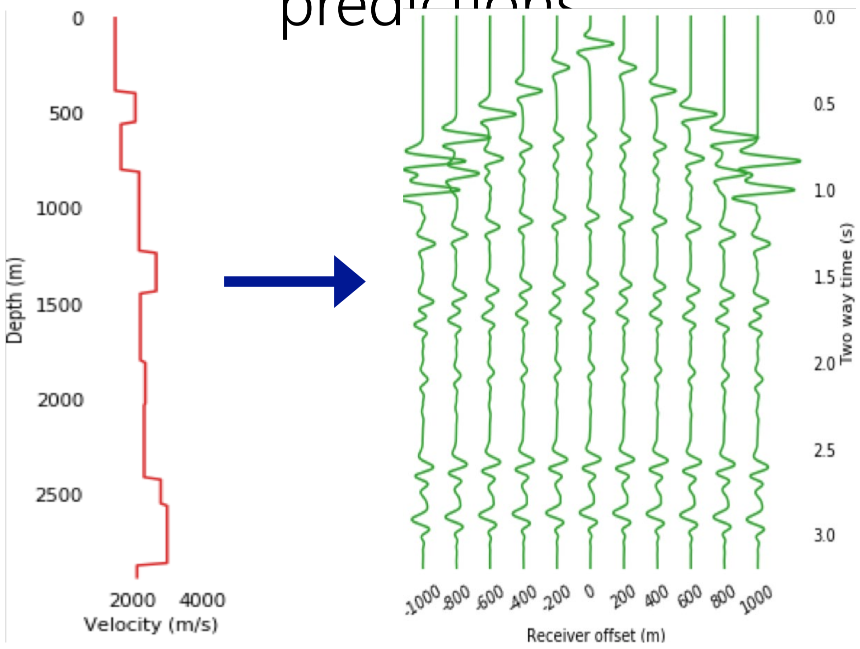
Prediction model

Directly maps 1D velocity structure to receiver pressure predictions

Acoustic finite difference model

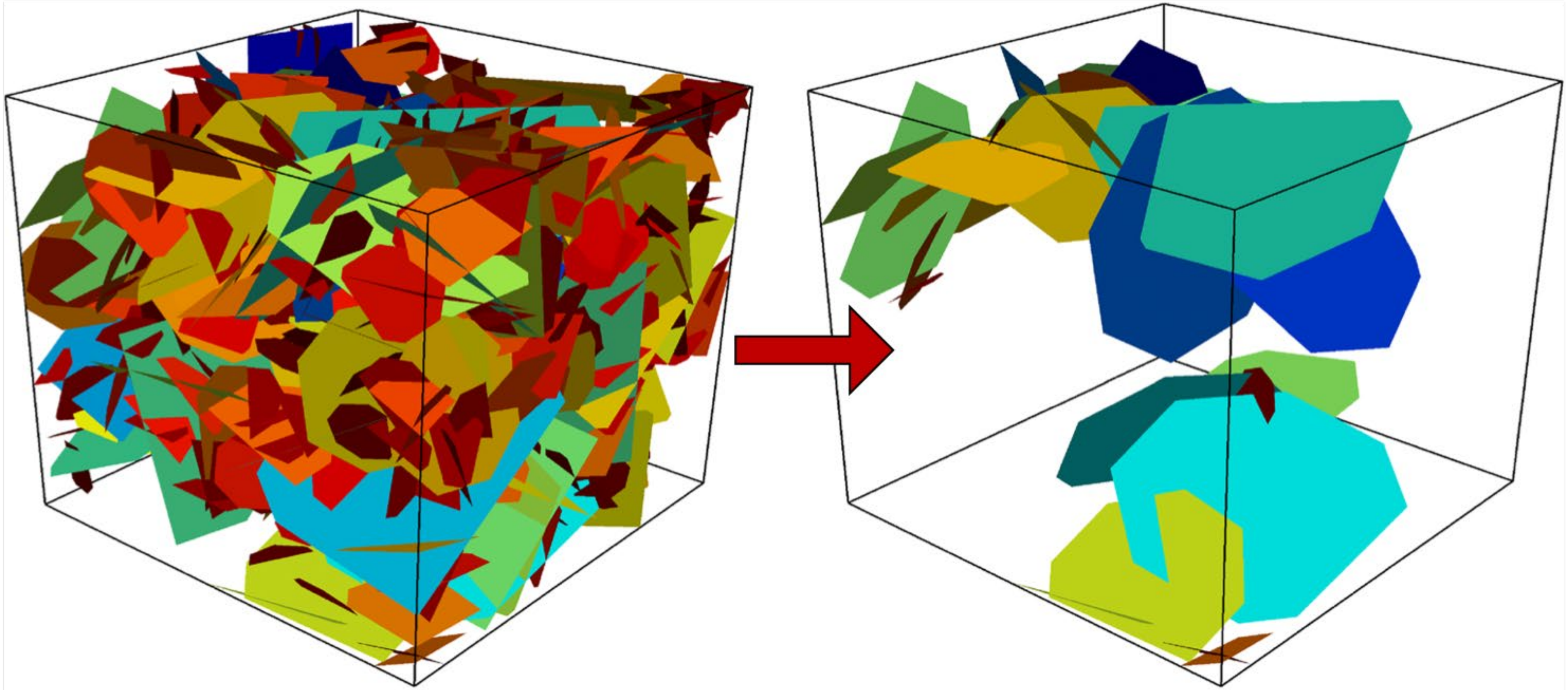


Wavefield Pressure (at 1s)



Modeling: *Model reduction for flow in porous media*

[Valera *et al.* (2018), *Comput. Geosci.*]



Full network model
Discrete fracture network

"Backbone" (reduced network model)
Subnetwork capturing flow pattern of full network

Machine learning challenges in solid Earth geoscience

Automation

- Dataset shift, covariate shift
- Biases in data collection, labeling
- Evaluating performance

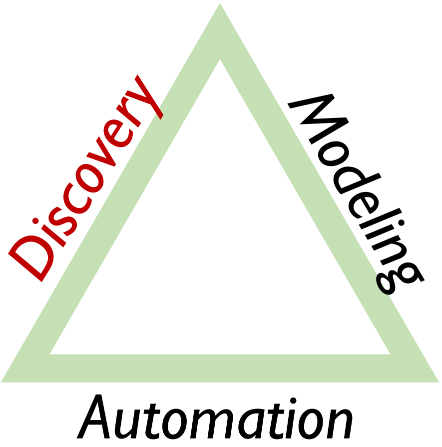
Modeling

- Quantifying model uncertainty
- Physical constraints, domain knowledge
- Expense of collecting training data (from simulations)

Discovering patterns & insights

Goal: Extract new information (complex patterns, structure, or relationships) from scientific data sets,

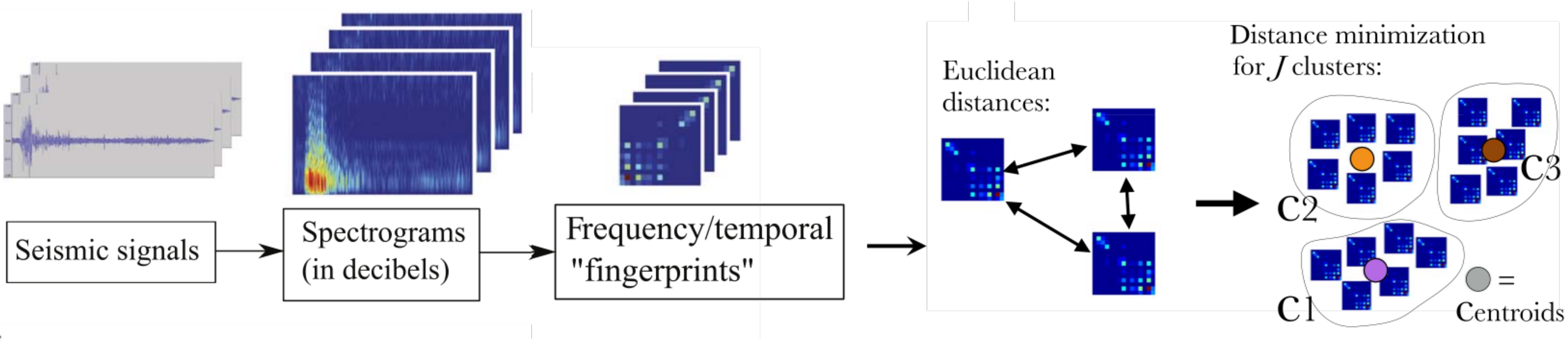
especially patterns *not easily revealed by conventional analysis techniques.*

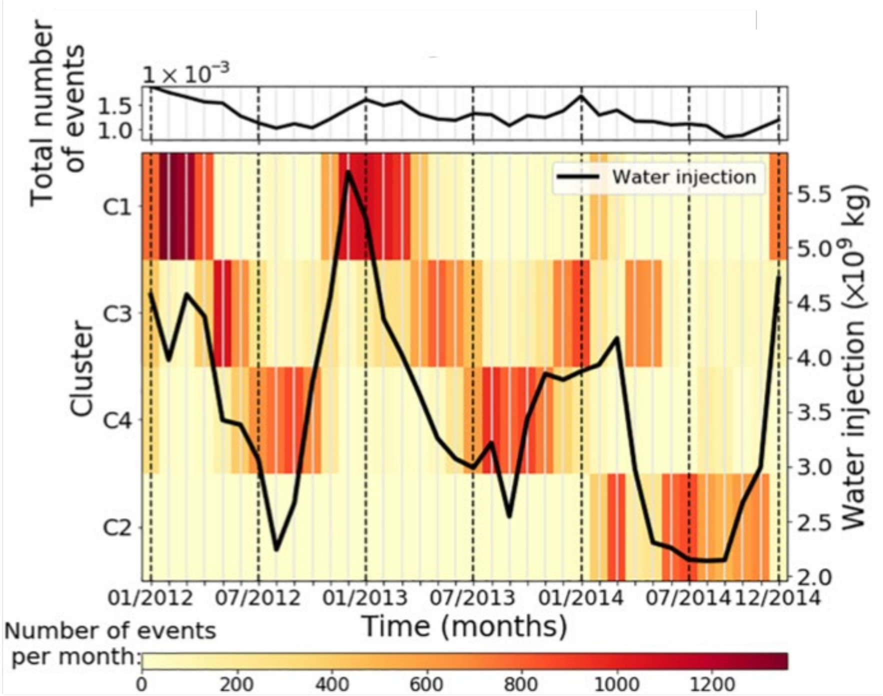


- Unsupervised learning
- Generative models

Discovery: *finding temporal patterns among seismic signals*

[Holtzman *et al.* (2018), *Sci. Adv.*]



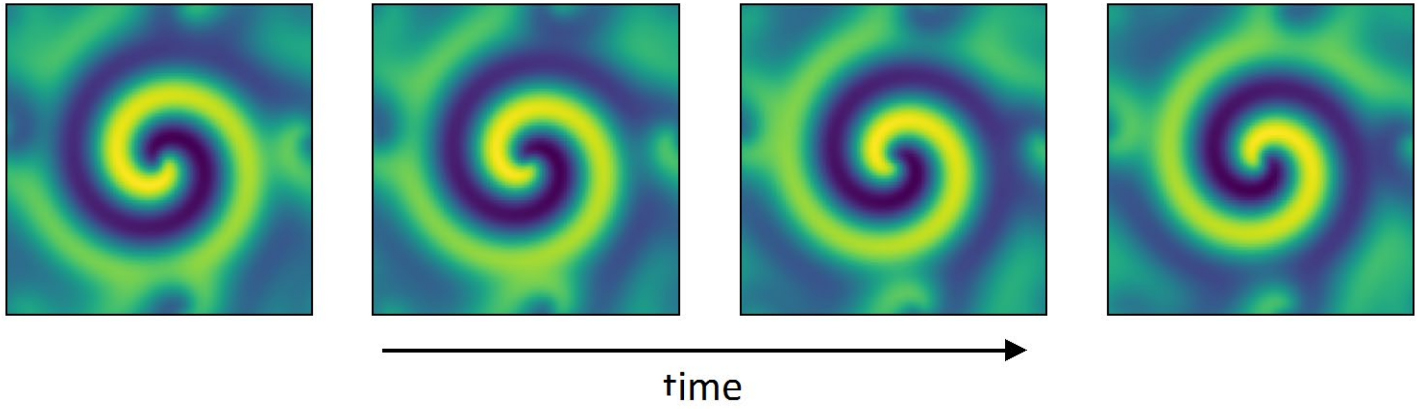


ML identifies spectra-temporal patterns, too subtle for standard methods, that reflect changes in faulting process.

Discovery: *learning governing equations from data*

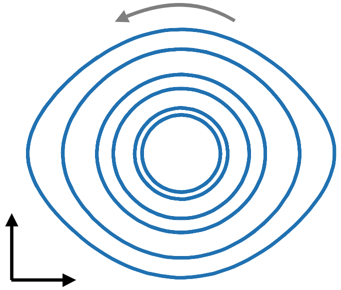
[Champion *et al.* (2019), *PNAS*]

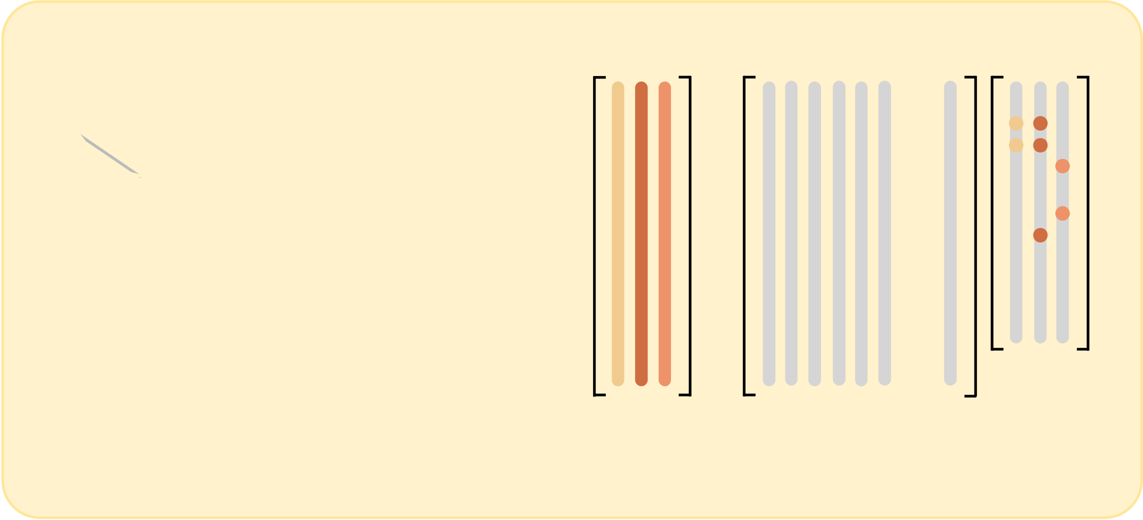
Reaction-diffusion system



$$\dot{z}_1 = -0.85 z_2$$
$$\dot{z}_2 = 0.97 z_1$$

Nonlinear pendulum





Machine learning challenges in solid Earth geoscience

Automation

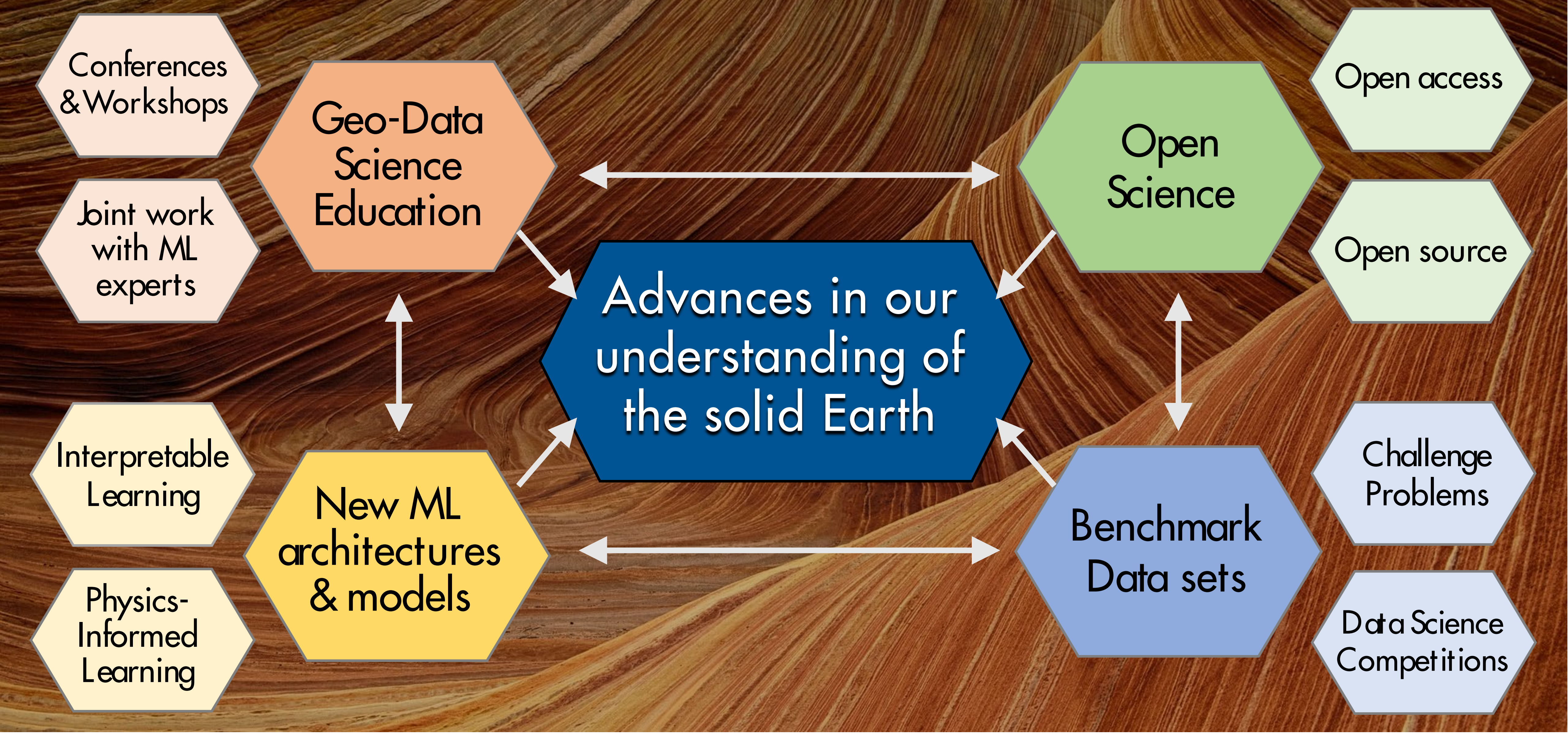
- Dataset shift, covariate shift
- Biases in data collection, labeling
- Evaluating performance

Modeling

- Quantifying model uncertainty
- Physical constraints, domain knowledge
- Expense of collecting training data (from simulations)

Discovery

- Interested in outliers, infrequent events, unexpected patterns
- Interpolation vs. extrapolation
- Discovery often requires interpreting "black box" models



Questions?

karianne_bergen@fas.harvard.edu

HDSI | Harvard Data Science Initiative

Modeling: Representing seamount bathymetry

[Valentine *et al.* (2013)]