Big Data Analysis in Geoscience

Frontier of Understanding Earth's Interior and Dynamics

Tohoku University, Sendai, Japan
August 8-9th, 2022

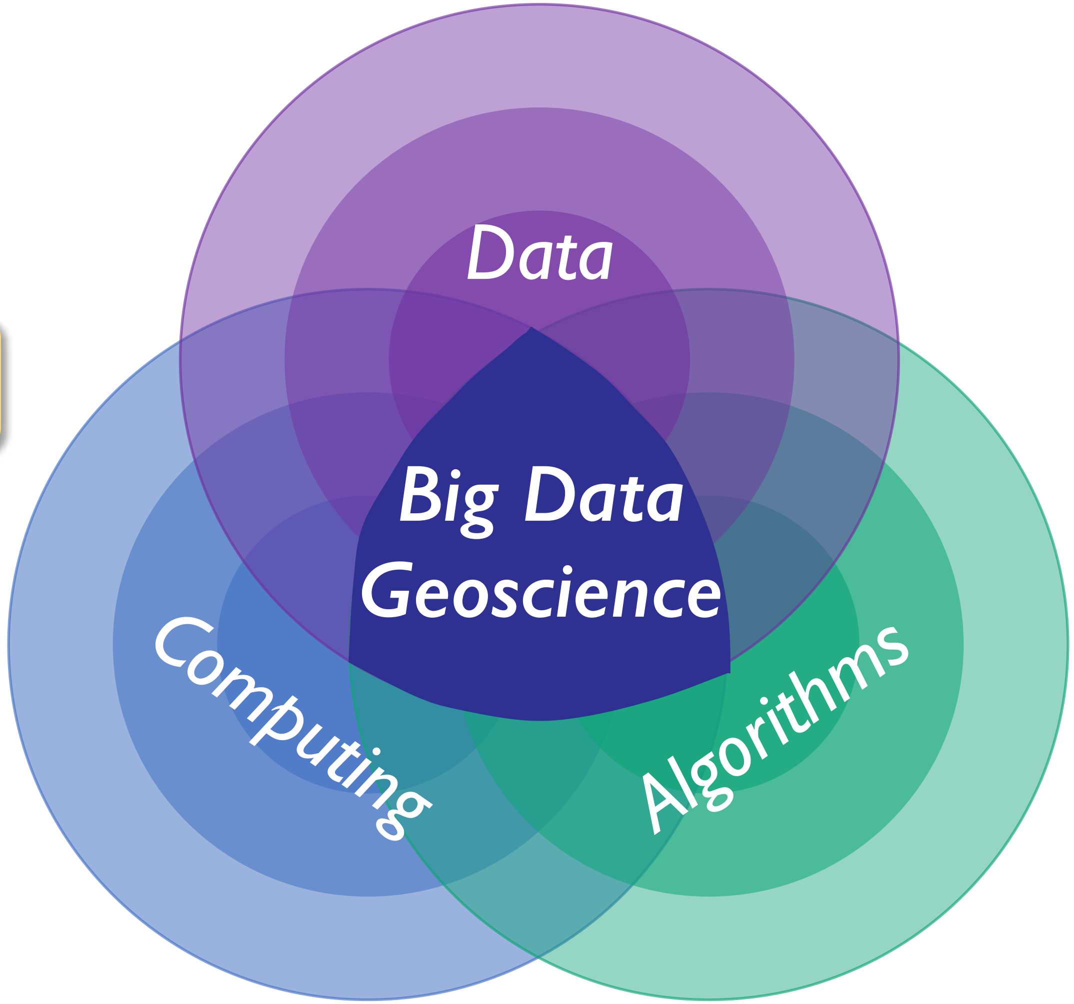
Karianne J. Bergen

Assistant Professor of Data Science
Brown University, Providence, RI, USA

Big Data Analysis *is helping Earth scientists extract more* ***knowledge & insights*** *from* ***larger, more complex data sets*** *than ever before.*

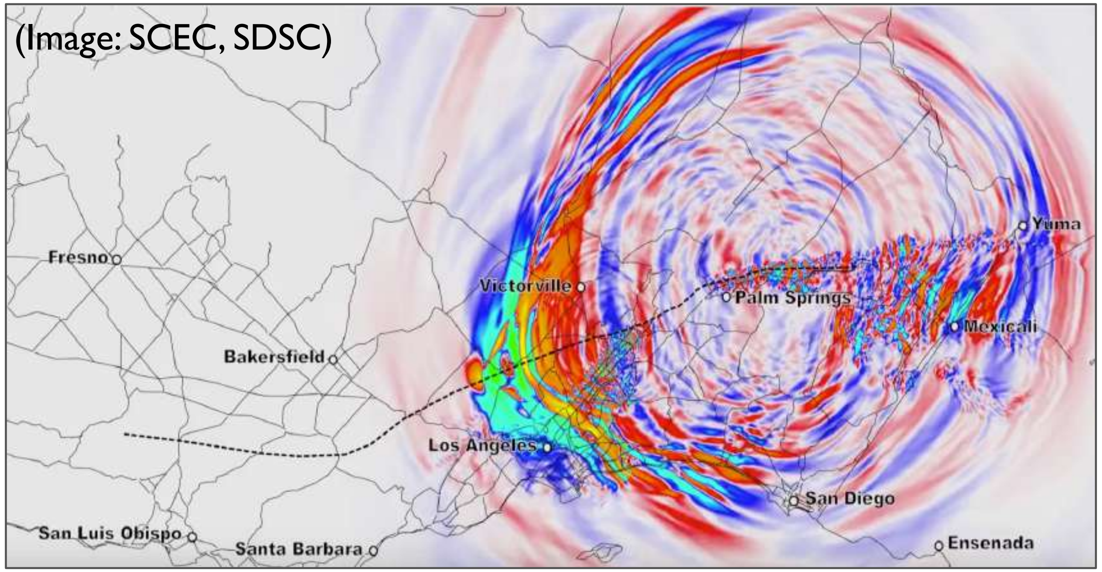
Drivers of Big Data Geoscience

1) Massive datasets
2) Advances in computing
3) New techniques and algorithms



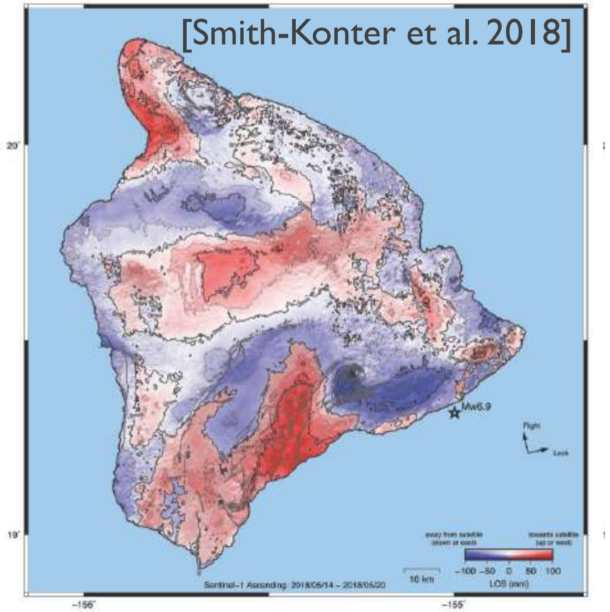
adapted from **Arrowsmith et al. (2022)**

Massive geoscience data sets

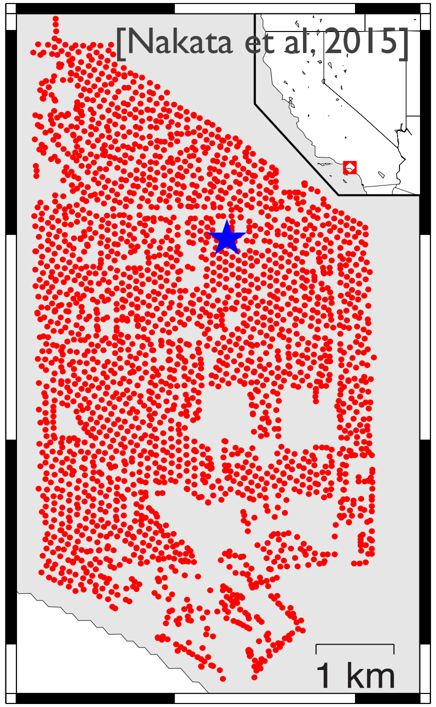


Large-scale simulations

Long-duration continuous observations



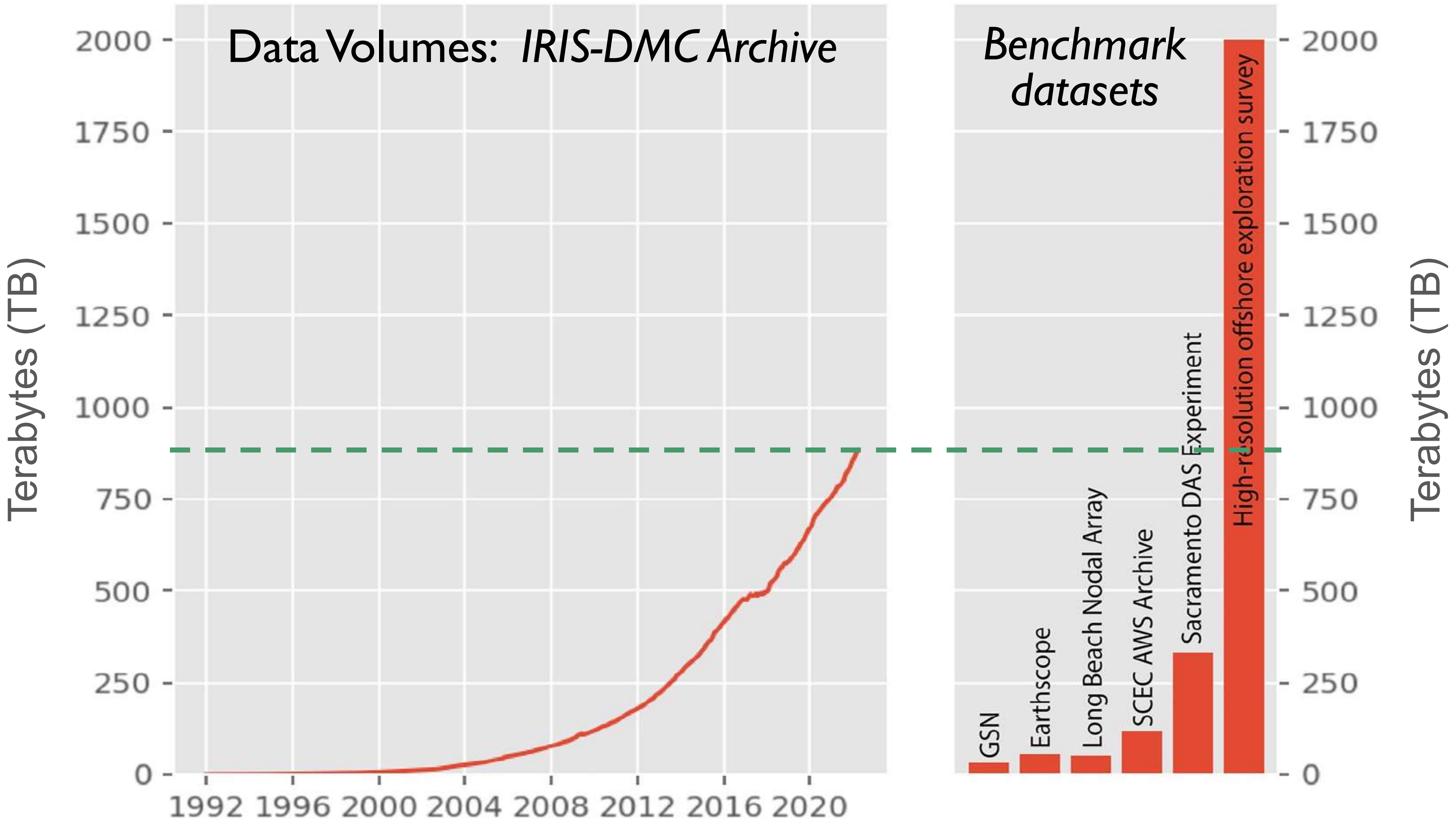
Remote sensing observations



Dense sensor arrays

Crowdsourced data

MyShake™



Arrowsmith et al. (2022)

The challenges of big geoscientific data analysis

Volume: Data-gathering capabilities – GB to TBs per day

- Extracting information – **automated analysis, scalability**

Velocity: Near real-time analysis, **e.g. for hazard assessment**

- Streaming data – **fully automated (no configuration)**

Variety: Multimodal datasets, e.g. seismometers + GNSS

- Sensor fusion – **combining multiple data sources**

Veracity: Data quality, e.g. noisy environments, instrument error

- automatic data cleaning, quality control, denoising

Algorithms for big scientific data

- **Efficient algorithms:** linear / sub-quadratic scaling with data volume
  - randomized algorithms, streaming algorithms, etc.
- Data-driven algorithms: **large-scale machine learning** (e.g. deep learning)
- Custom, task-specific algorithms
- Data reduction, data compression
- **More computation:** parallel and distributed computing, cloud computing

Can new *Big Data* algorithms detect small earthquakes and identify seismic phases?

FAST: *How can we detect more small earthquakes?*

Yoon et al., (2015)

Leverages technology for efficient audio recognition

Discovers new event waveforms (without labeled data):
10 – 100× earthquakes detected

Computationally efficient:
500× more data with reduced runtime

P. Bailis G. Beroza P. Levis O. O'Reilly K. Rong C. Yoon

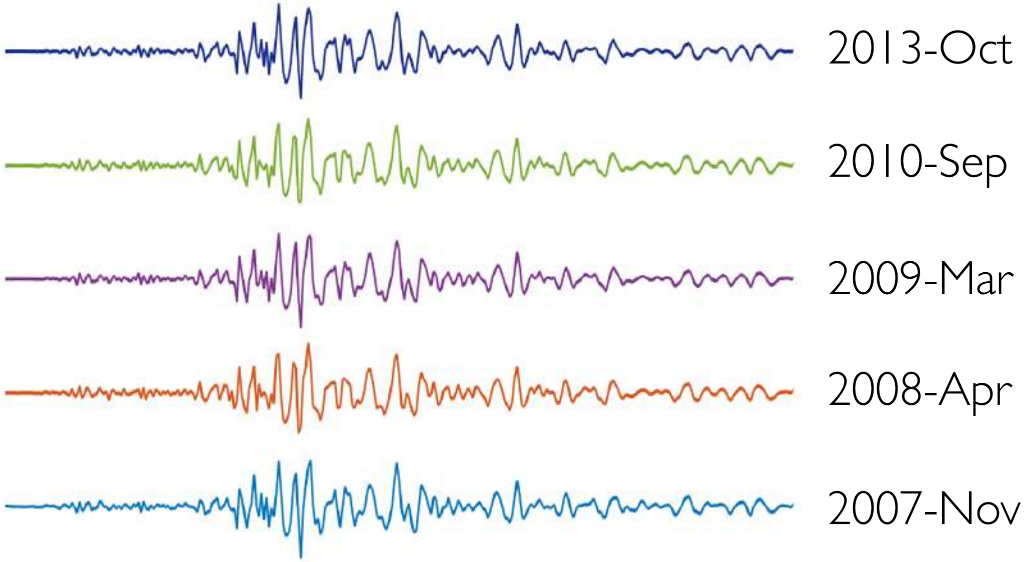
FAST: scalable "Large-T" earthquake detection

Data mining: extract similar waveforms from large datasets

Naïve (slow, exact) search: *small data*

Efficient (fast, approximate) search:

vs.



Searching a well-organized database is faster – cluster similar waveforms for quick retrieval

Sacrificing (a little) accuracy can substantially reduce runtime.

Algorithms for big scientific data

- **Efficient algorithms:** linear / sub-quadratic scaling with data volume
  - randomized algorithms, streaming algorithms, etc.
- Data-driven algorithms: **large-scale machine learning** (e.g. deep learning)
- Custom, task-specific algorithms
- Data reduction, data compression
- **More computation:** parallel and distributed computing, cloud computing

What is Machine Learning?

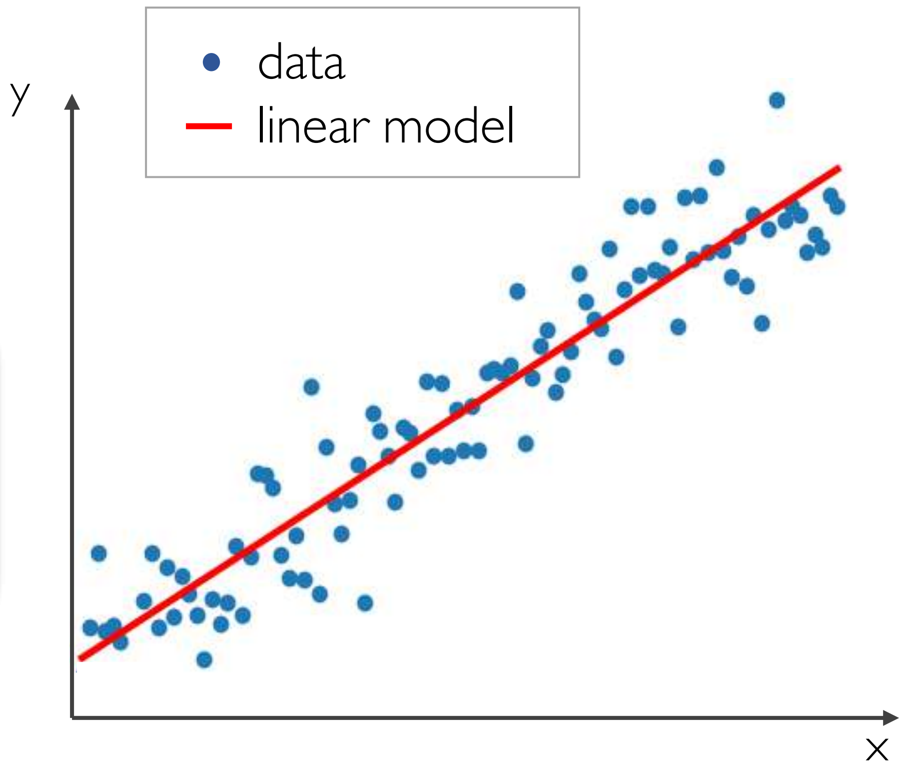
Automating and scaling data analysis

Machine learning (ML)

a set of tools for recognizing complex patterns and building predictive models *automatically from data*

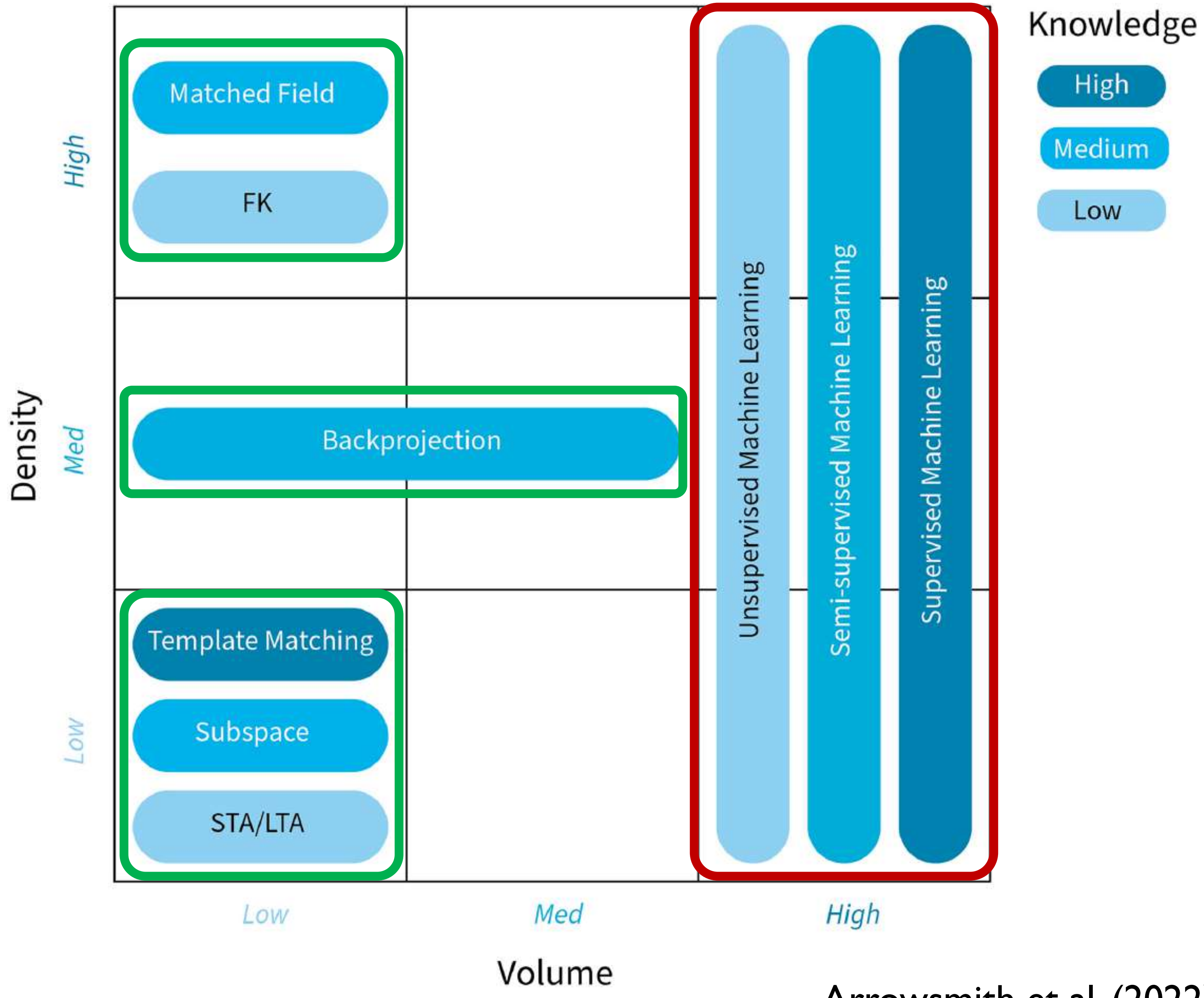
Unsupervised
Discovering structure in data

Supervised
Learning a pattern from examples

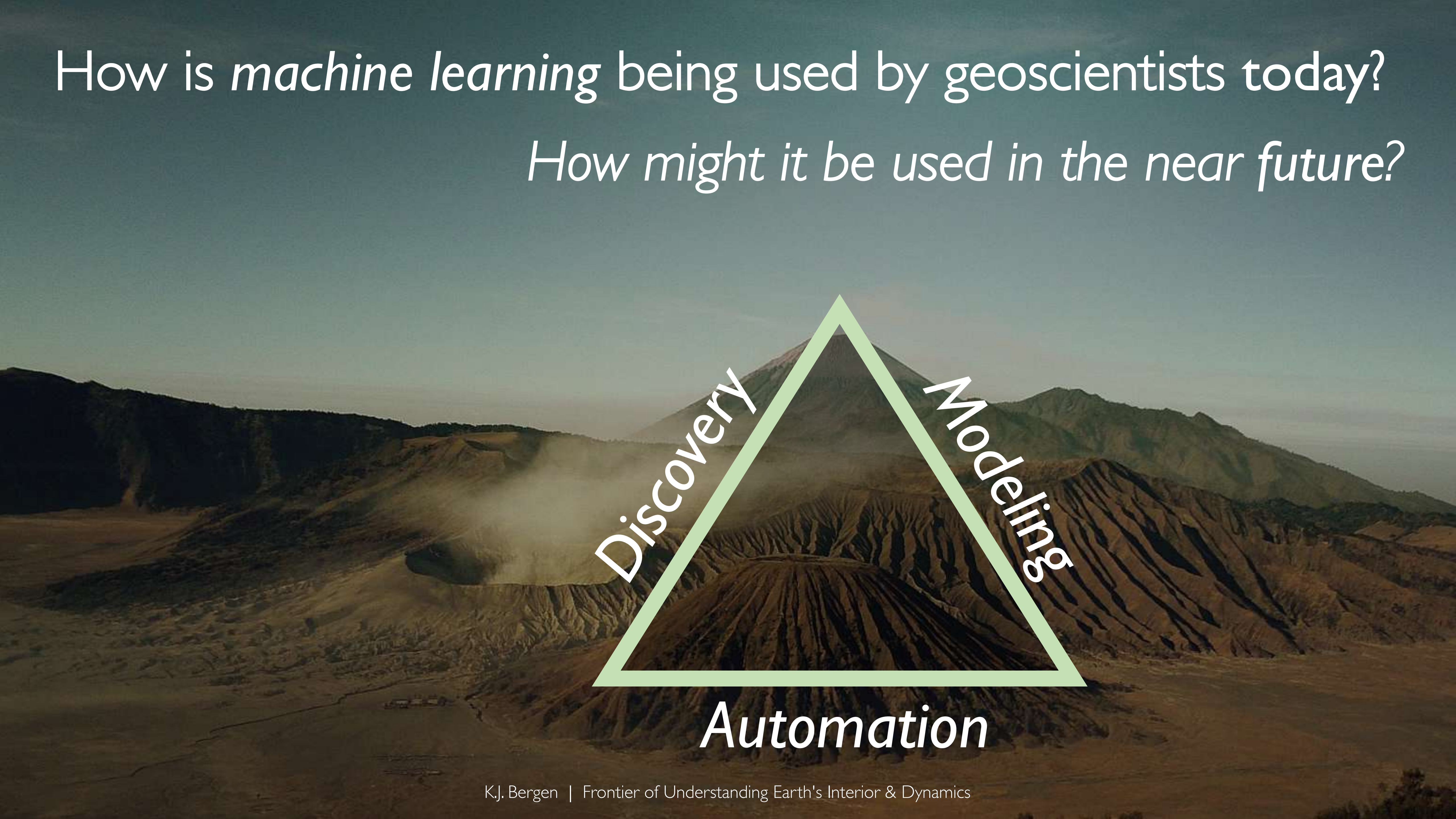


Machine Learning is a key *tool for high-volume data*

- diverse datasets & tasks
- dense & sparse data
- high- or low-knowledge



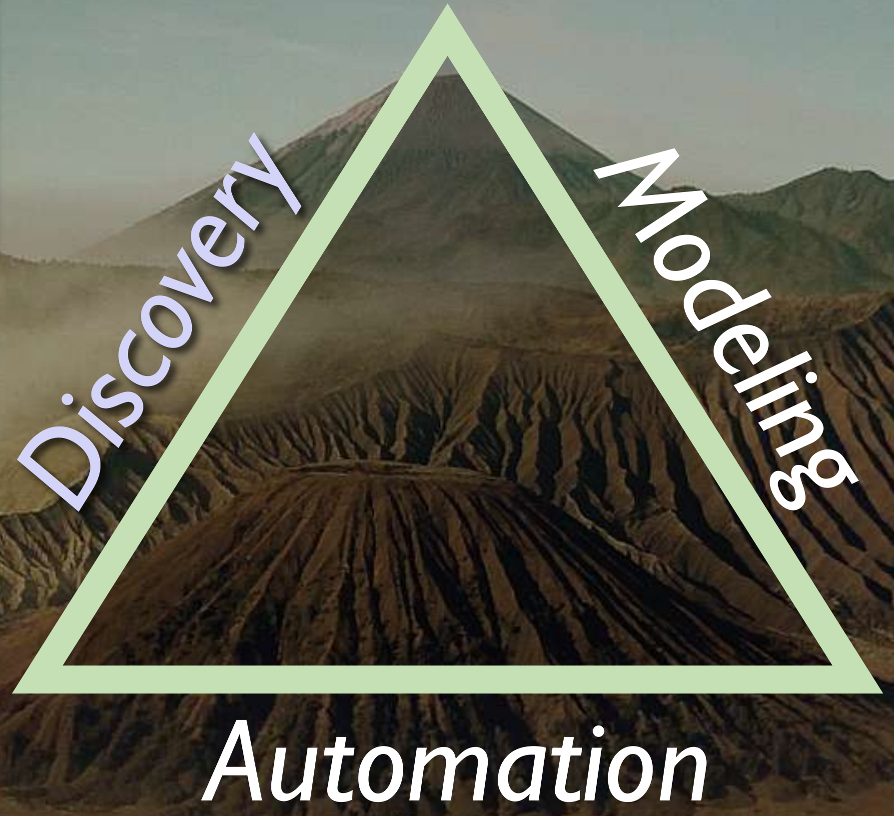
Arrowsmith et al. (2022)


How is *machine learning* being used by geoscientists today?
How might it be used in the near future?
Discovery
Modeling
Automation
K.J. Bergen | Frontier of Understanding Earth's Interior & Dynamics

How is *machine learning* being used by geoscientists today?

How might it be used in the near future?

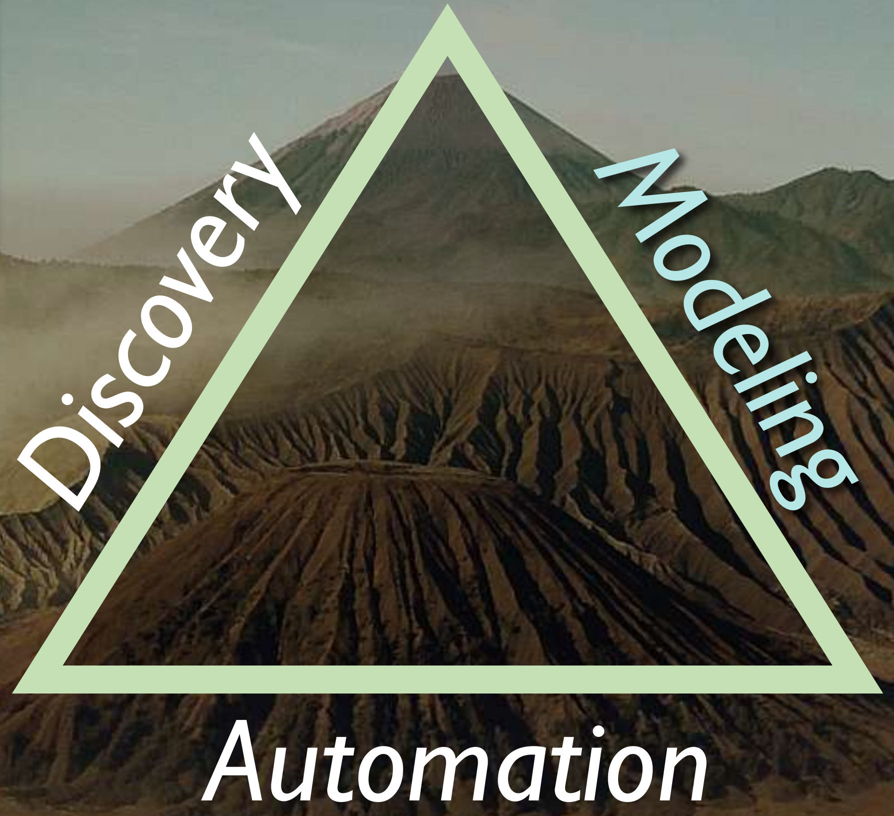
Extract new information, patterns, structure, or relationships from data



How is *machine learning* being used by geoscientists today?

How might it be used in the near future?

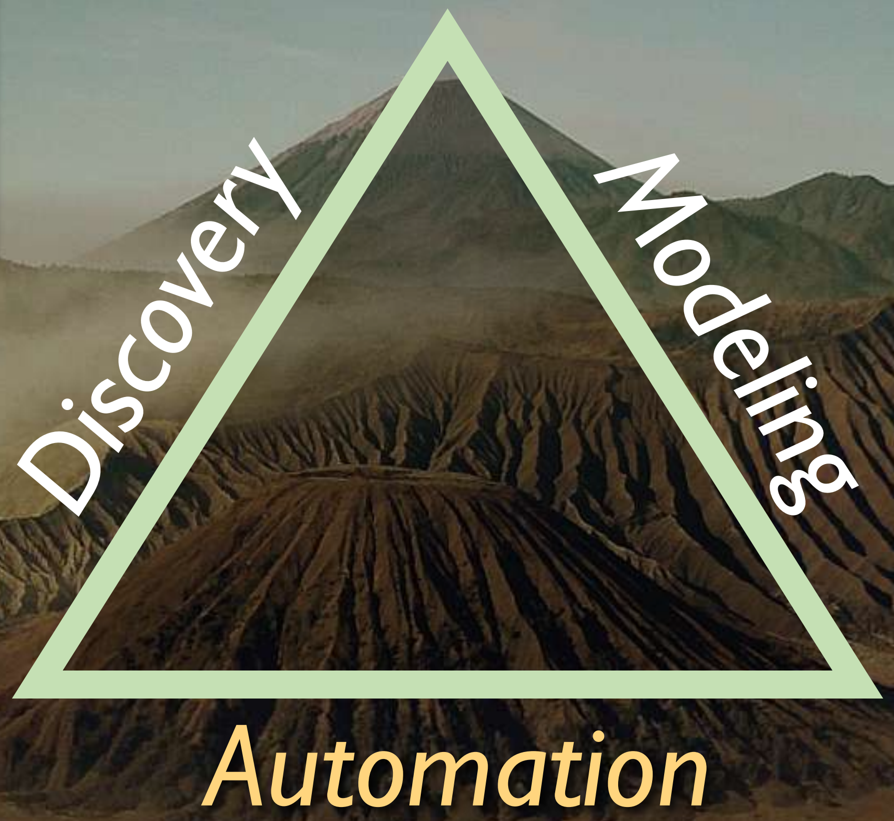
- Learn representations
- Build surrogate models
- ML + simulations



How is *machine learning* being used by geoscientists today?

How might it be used in the near future?

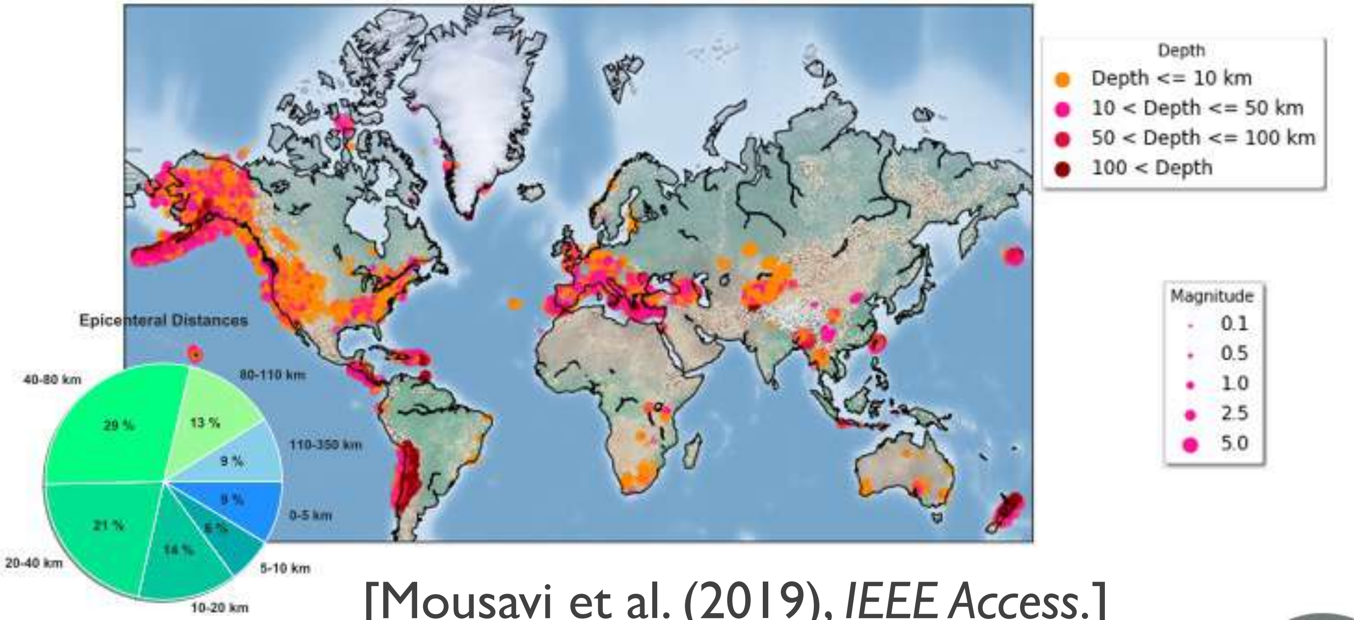
- Perform a complex or repetitive task
- High accuracy predictions



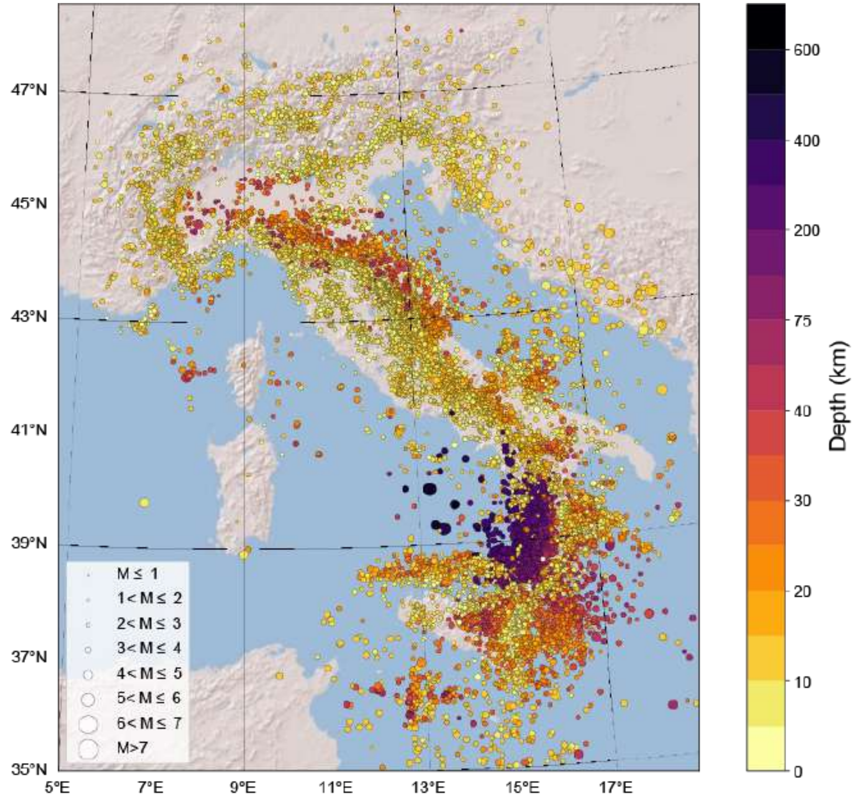
Benchmark & training datasets for supervised ML

STEAD dataset

1.2 M Labeled Waveform. **450 k** Earthquakes. **19,000** Hours of Data.



[Mousavi et al. (2019), *IEEE Access*.]

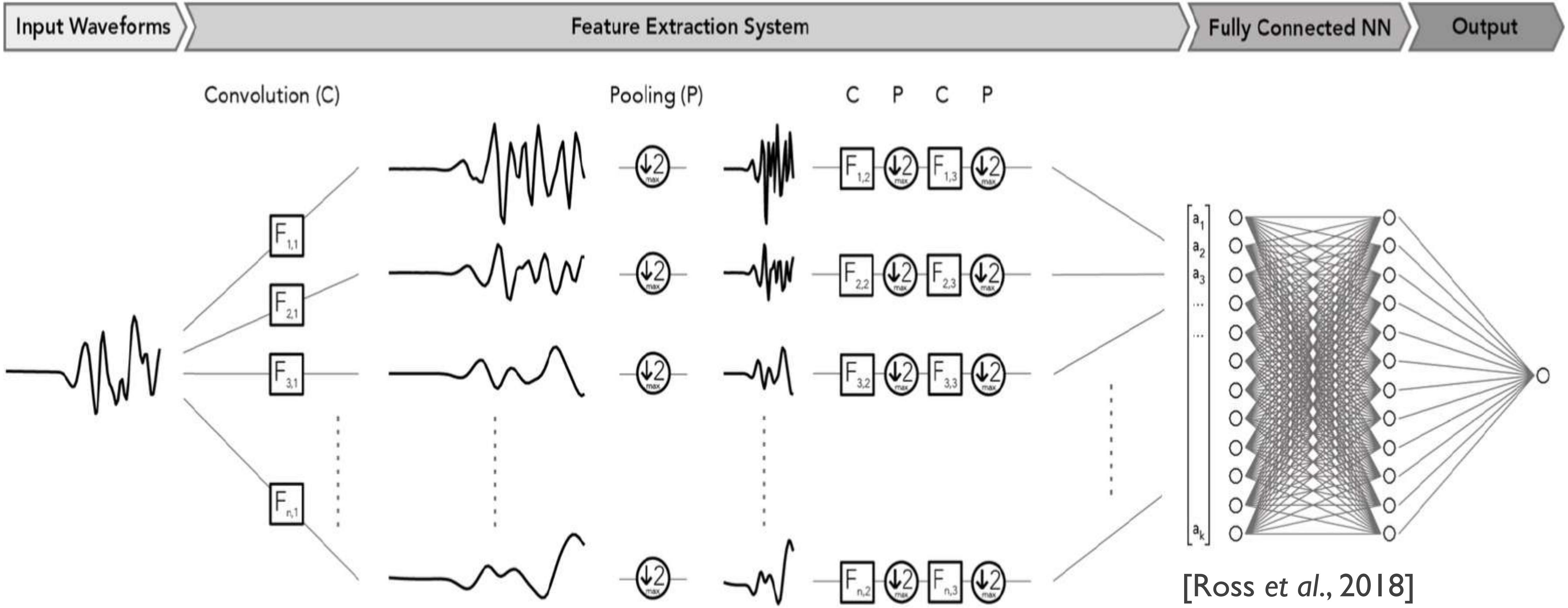


INSTANCE THE ITALIAN SEISMIC DATASET FOR MACHINE LEARNING

[Michelini et al. (2021), DOI: 10.13127/instance]

1D CNN for detection and phase-picking

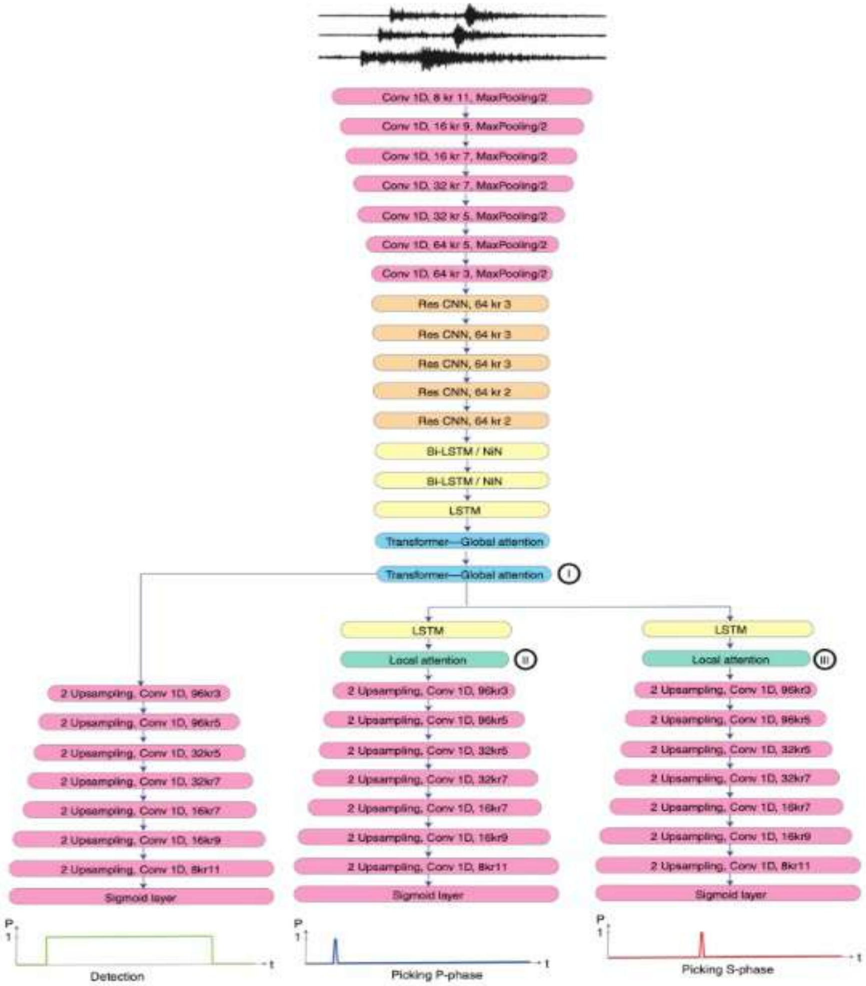
[e.g. Perol *et al.* (2018)]



[Ross *et al.*, 2018]

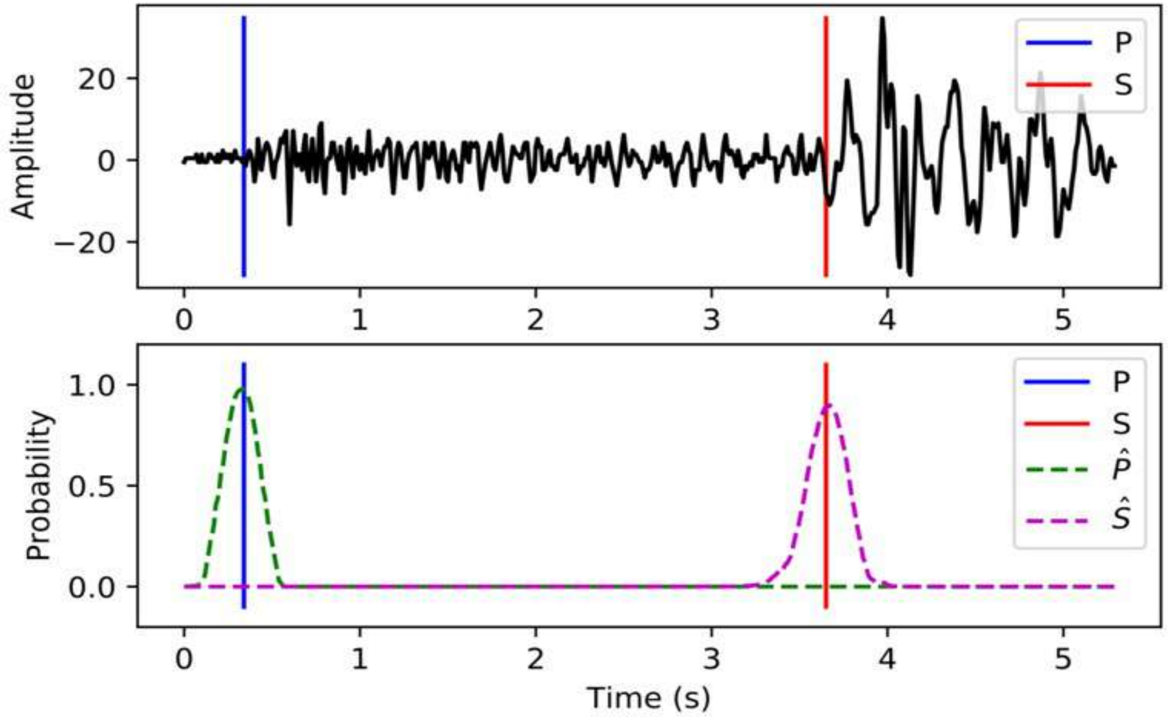
EQTransformer: Attention-based model for detection & phase picking

[Mousavi et al. (2020), *Nature Communications*]



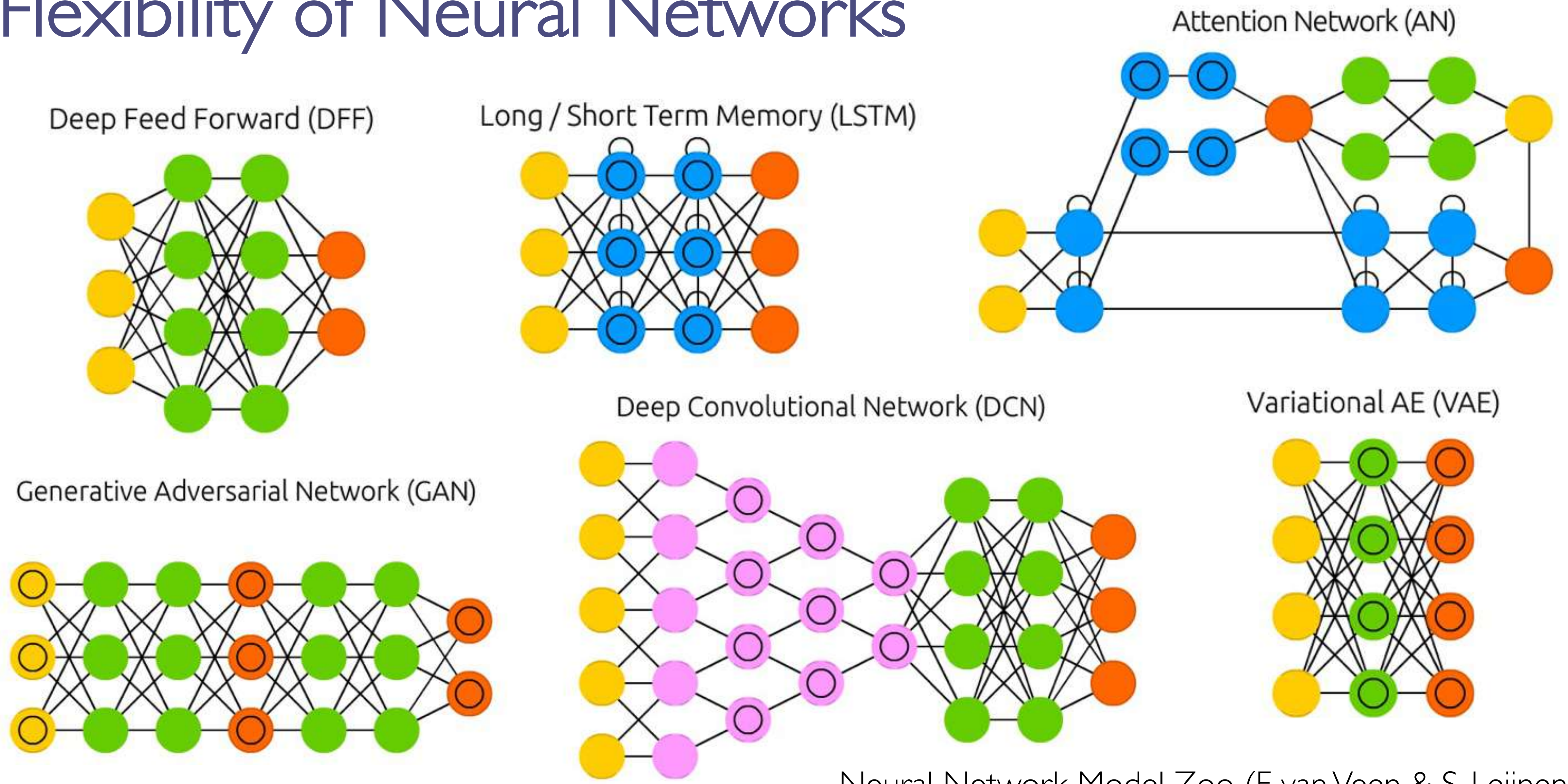
PhaseNet: U-net for phase picking

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



Bergen et al. (2019), *Science*; Kong et al. (2019), *BSSA*; Dramsch (2020), *Adv. in Geophys*; Yu & Ma (2021), *Rev. Geophys.*

Flexibility of Neural Networks



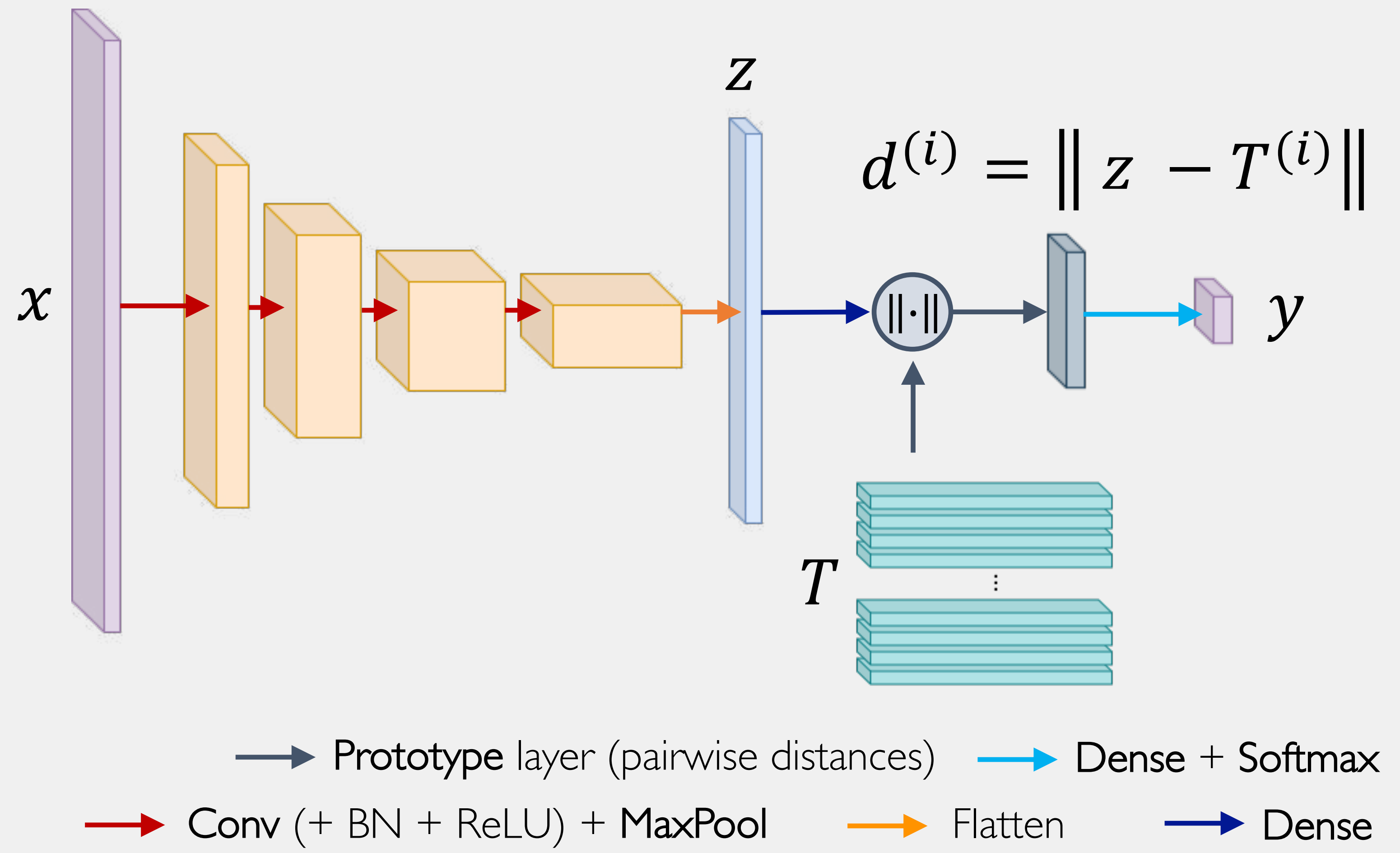
Neural Network Model Zoo (F. van Veen & S. Leijnen)

Can we build an interpretable deep NN for detection?

Domain- or task- specific NN architectures

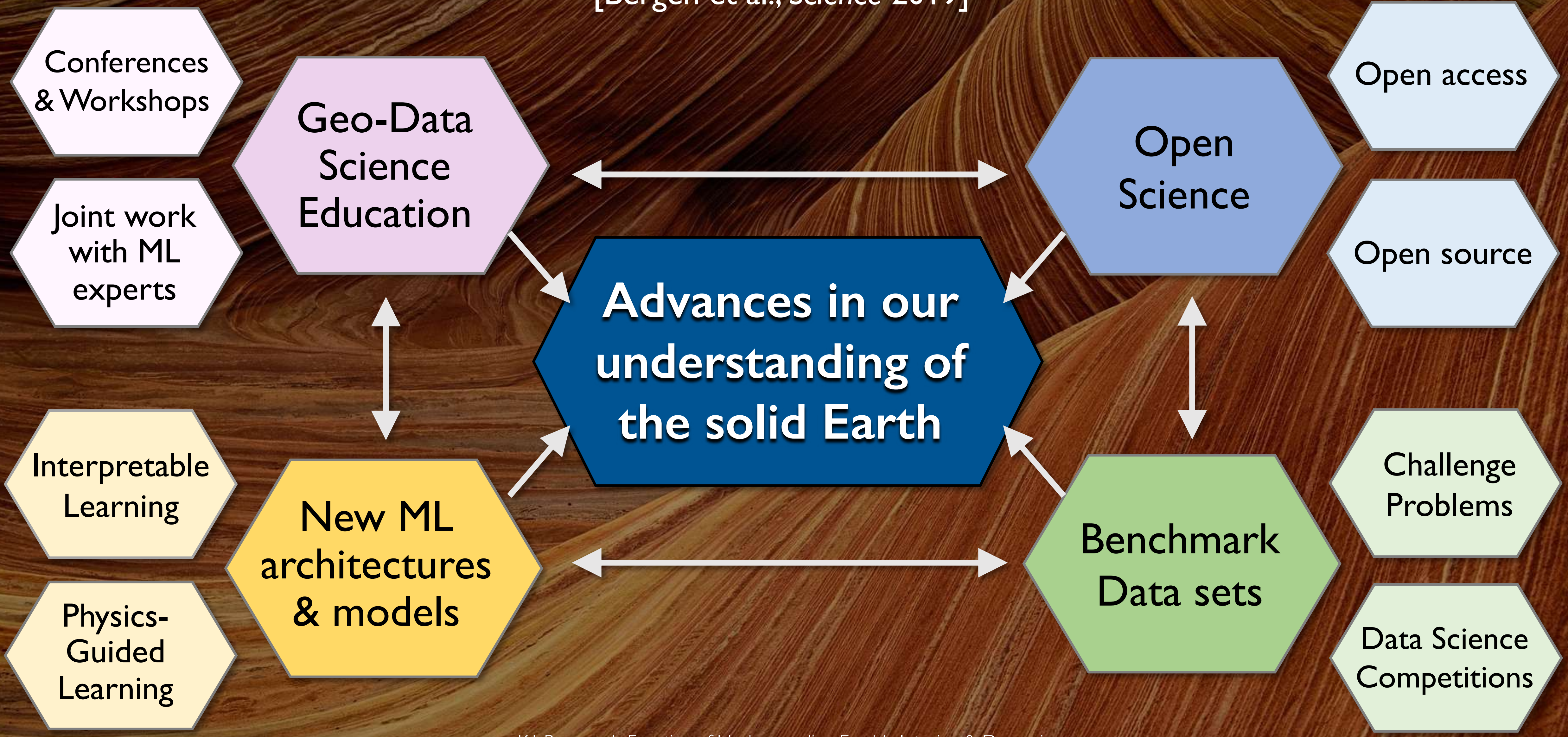
- Design human-interpretable prediction systems
- Jointly analyze data from multiple sources
- Incorporate physics into data-driven NN model

Chen et al. (2019)



Future of Machine Learning for the solid Earth

[Bergen et al., *Science* 2019]



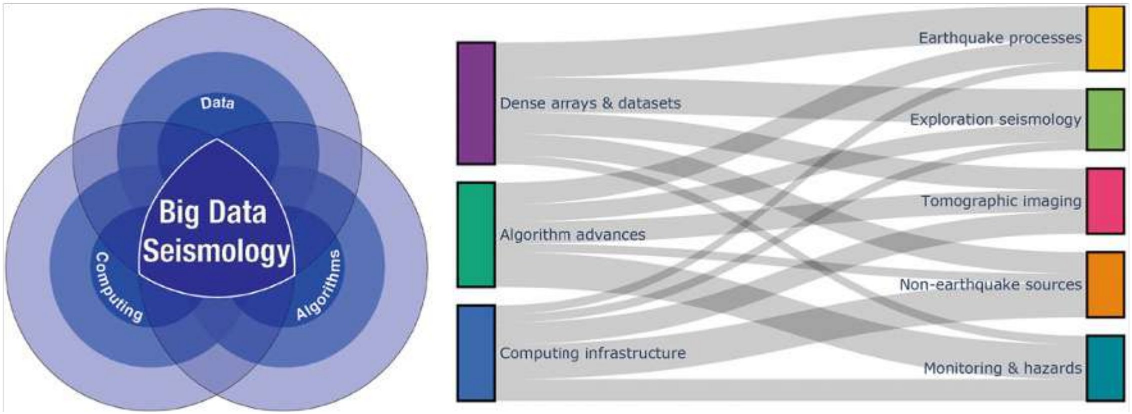
Further reading

Reviews of Geophysics

Review Article | Full Access

Big Data Seismology

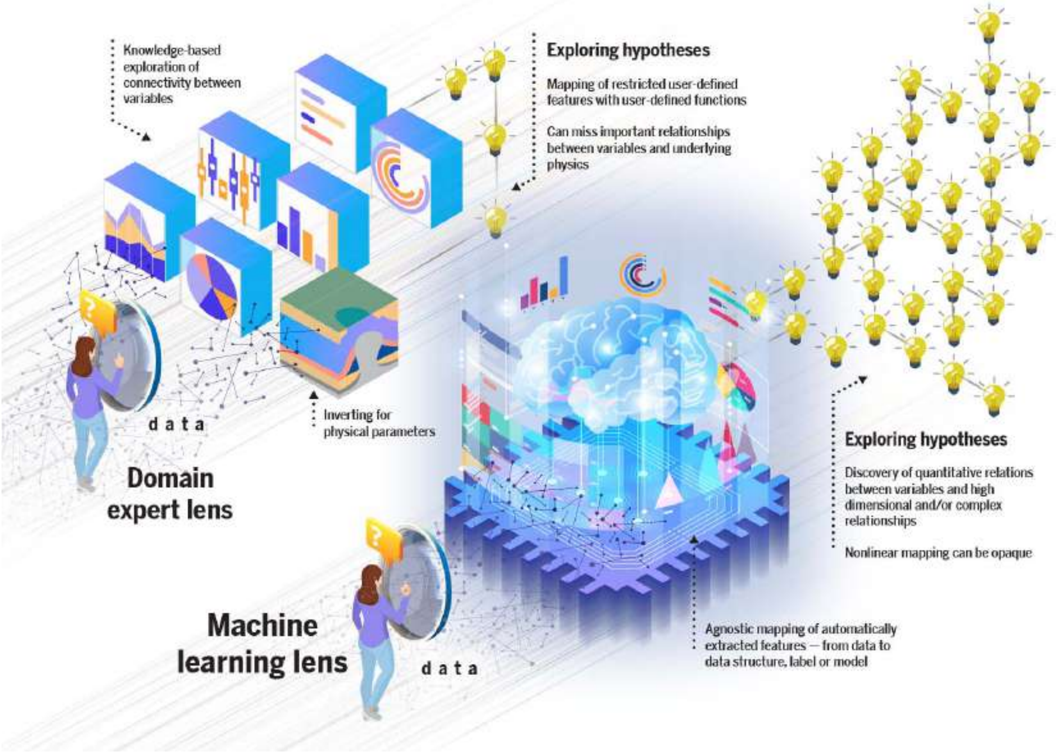
S. J. Arrowsmith, D. T. Trugman, J. MacCarthy, K. J. Bergen, D. Lumley, M. B. Magnani



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]*



Big Data Analysis *is helping Earth scientists extract more* ***knowledge & insights*** *from* ***larger, more complex data sets*** *than ever before.*

Questions?

karianne_bergen@brown.edu

@KarianneBergen

ご清聴ありがとうございました。