DAS and Big Scientific Data Analysis

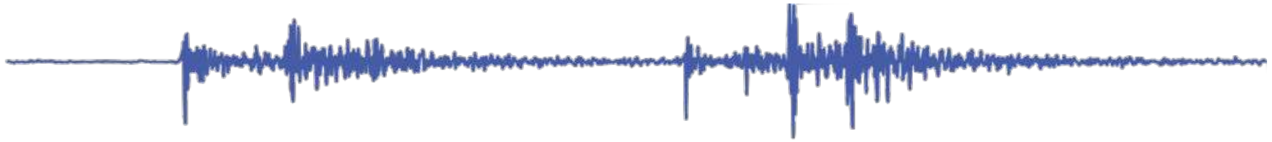
Distributed Acoustic Sensing Virtual Workshop and Tutorial
August 10, 2020

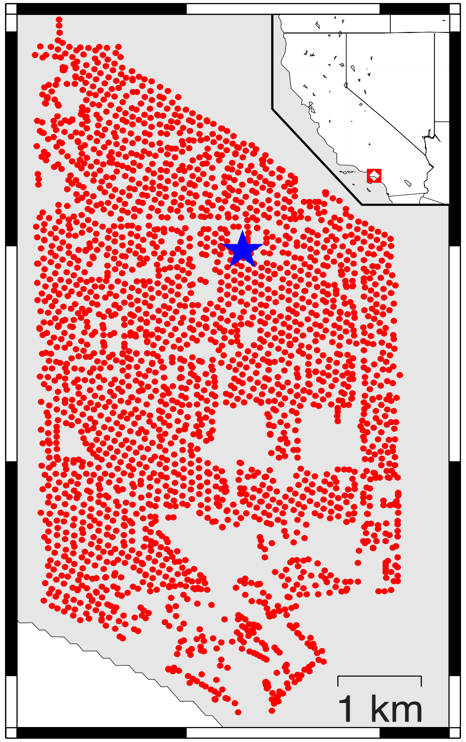
Karianne J. Bergen
Data Science Initiative Postdoctoral Fellow, Harvard University
Assistant Professor, Brown University *(January 2021)*

Harvard John A. Paulson
School of Engineering
and Applied Sciences

Massive data sets in earthquake seismology

Long Duration (Large-T)
>10 years continuous waveform data





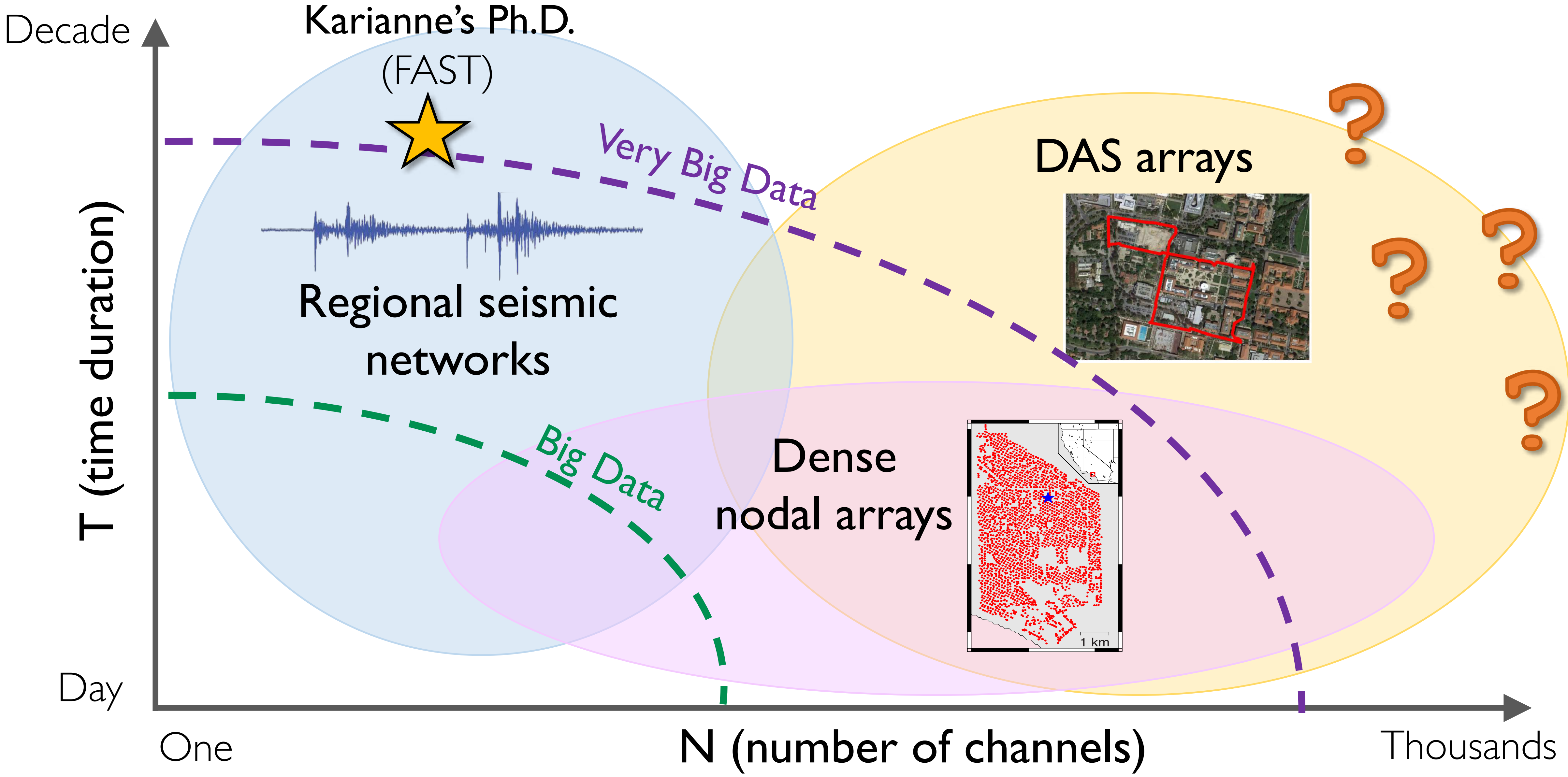
[Nakata et al., 2015]

Big Networks (Large-N)
1000's of sensors

New data sources:
DAS arrays

[Martin et al., 2017]

Massive data sets in earthquake seismology



DAS and the challenges of "big data" analysis

Volume: 100 GB to 10 TB per day for a single DAS array

- Extracting information – automated analysis, scalability

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

- Streaming data – fully automated (no configuration)

Variety: Variation in DAS systems; seismometers, geophones

- Sensor fusion – combining DAS with other sources

Veracity: Data quality: noisy environments, poor coupling

- Automatic data cleaning, quality control, denoising

DAS and the opportunities of "big data" analysis

Novelty: New data source requires new algorithms

- Existing methods for seismic analysis may not account for:
  - Data volume
  - **Data quality** (vs. purpose-built scientific sensors)
  - **Measurements** – distributed (vs. point), directional sensitivity
  - **Flexible array geometry**
- **Insufficient labeled data** – unsupervised methods, transfer learning / domain adaption

Algorithms for big scientific data

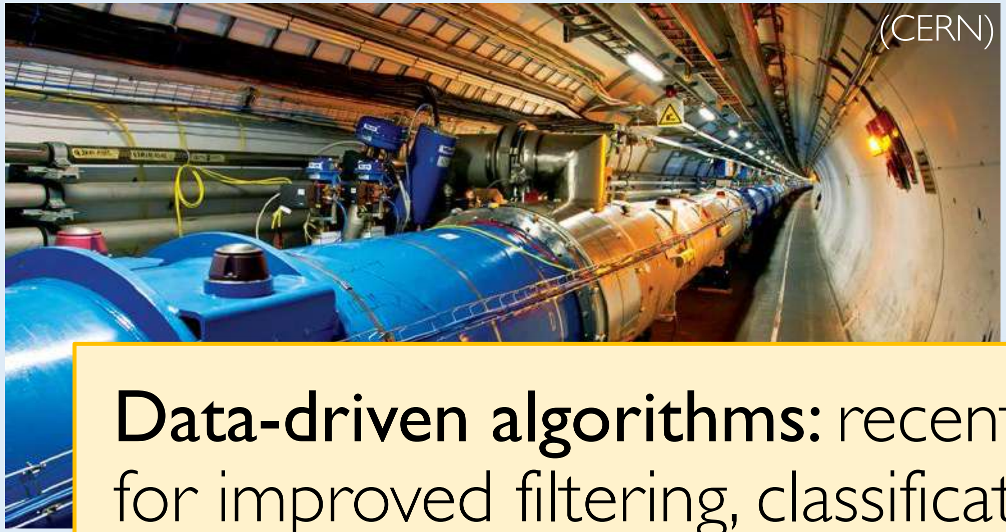
- **Efficient algorithms:** linearly, sub-quadratically scaling with data volume
    - randomized algorithms, streaming algorithms, etc.
- Data-driven algorithms: **large-scale machine learning** (e.g. deep learning)
- More computation: parallel, distributed computing
- Data reduction, data compression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

Extremely Big Scientific Data (outside geophysics)

Particle Physics

Large Hadron Collider

- Generates too much data to store: 1 billion collisions/s → 1 petabyte/s
- **Data filtering:** "interesting" events only (~1000 events/s) stored for analysis


(CERN)

Astronomy

Square Kilometer Array (Phase 1)

- Processing **reduces data size**, increases information content.
- **Real-time** processing of raw data, **supercomputers** generate science data products (600 PB/year)


(SKA Project)

Data-driven algorithms: recent work explores **machine learning** for improved filtering, classification of events at LHC

[Nguyen et al. (2019); Radovic et al. (2018)]

Algorithms for big scientific data

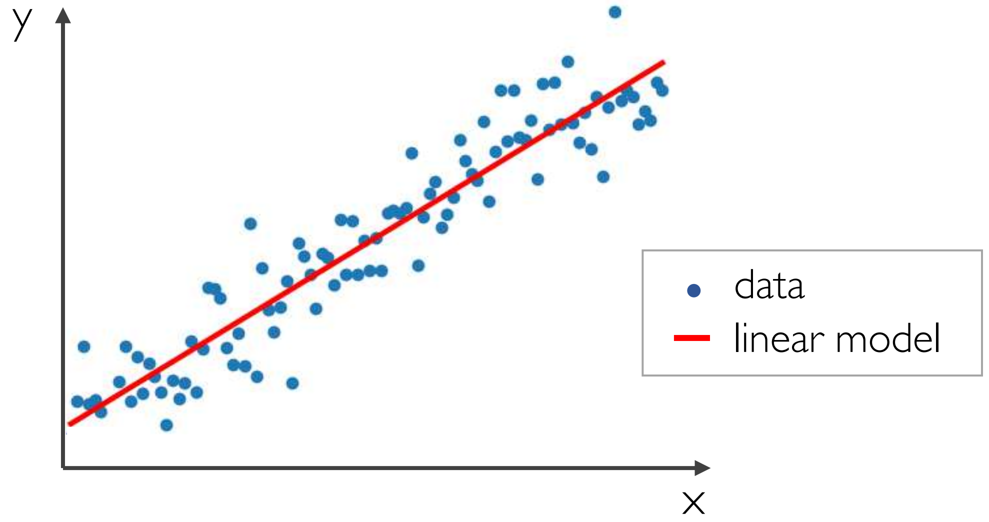
- **Efficient algorithms:** linearly, sub-quadratically scaling with data volume
  - randomized algorithms, streaming algorithms, etc.
- Data-driven algorithms: **large-scale machine learning** (e.g. deep learning)
- More computation: parallel, distributed computing
- Data reduction, data compression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

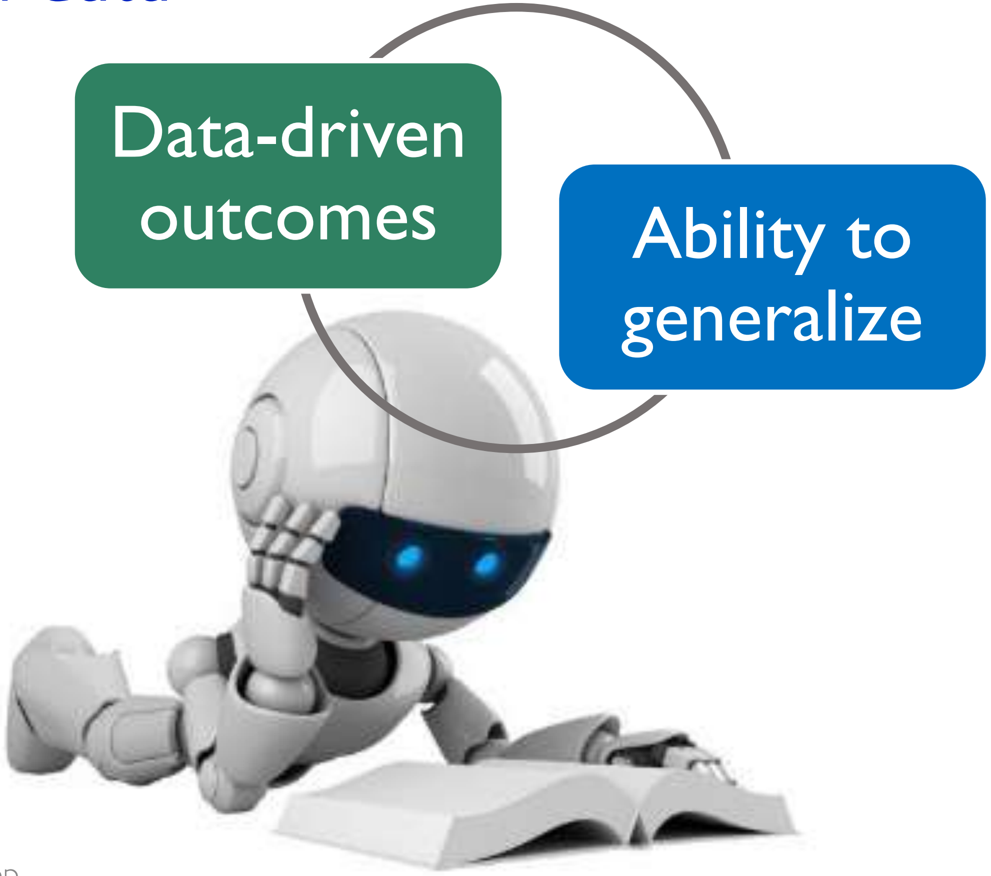
What is Machine Learning?

Machine learning (ML)

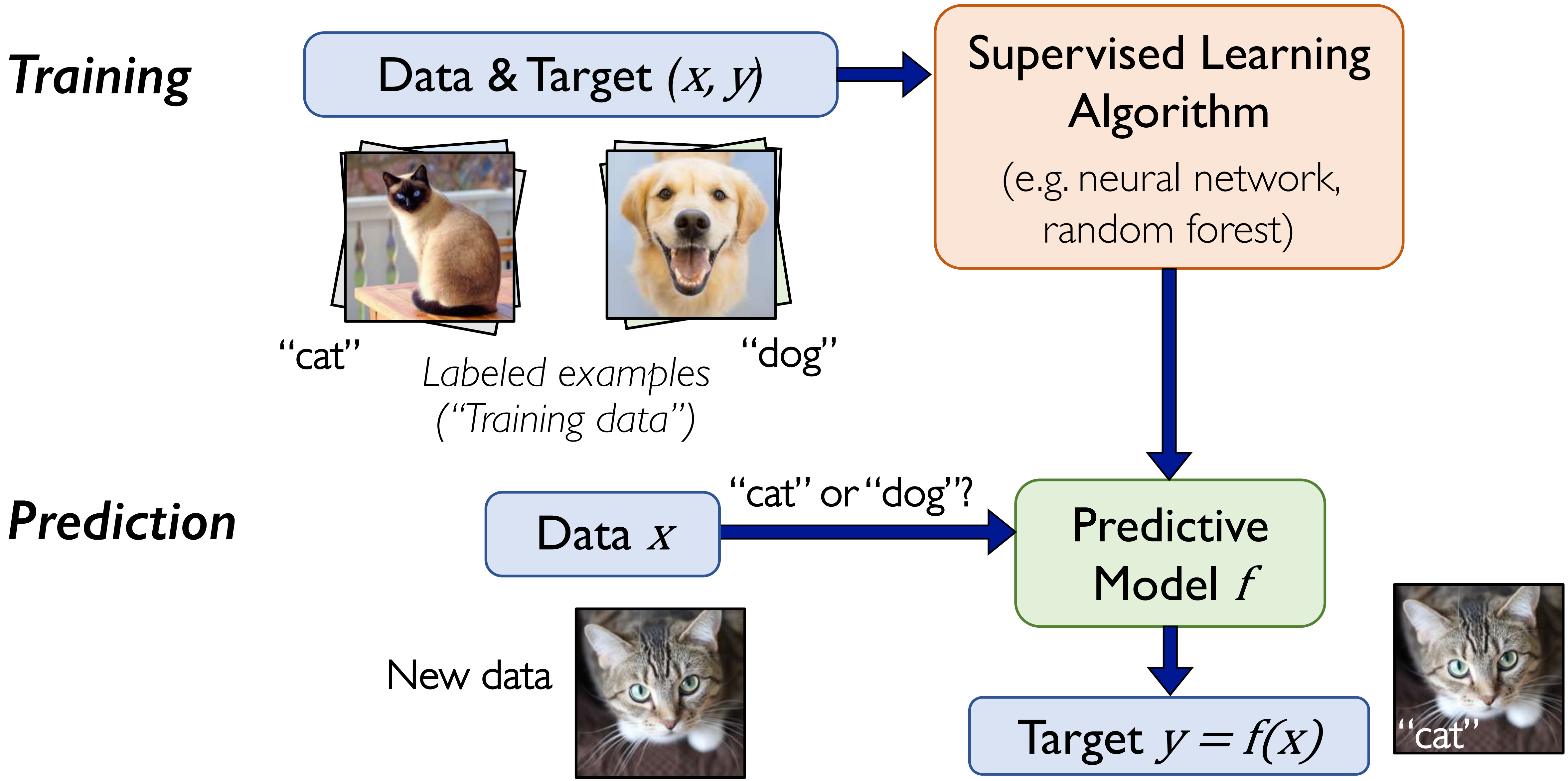
a set of tools for recognizing complex patterns and building predictive models *automatically from data*

- linear regression, logistic regression, PCA





Building models from examples *(Supervised Learning)*

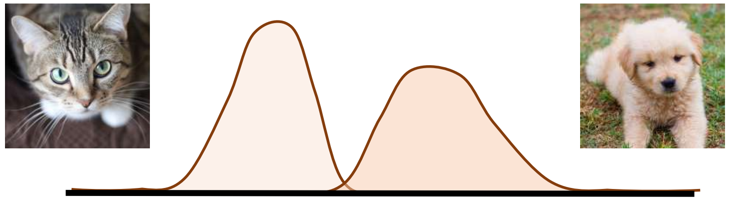


Finding patterns in data *(Unsupervised Learning)*

Data X → **Unsupervised Learning Algorithm** (e.g. PCA, k-means) → Structure in X

Data only
(no labels)

Common patterns or features in data



Model of data distribution

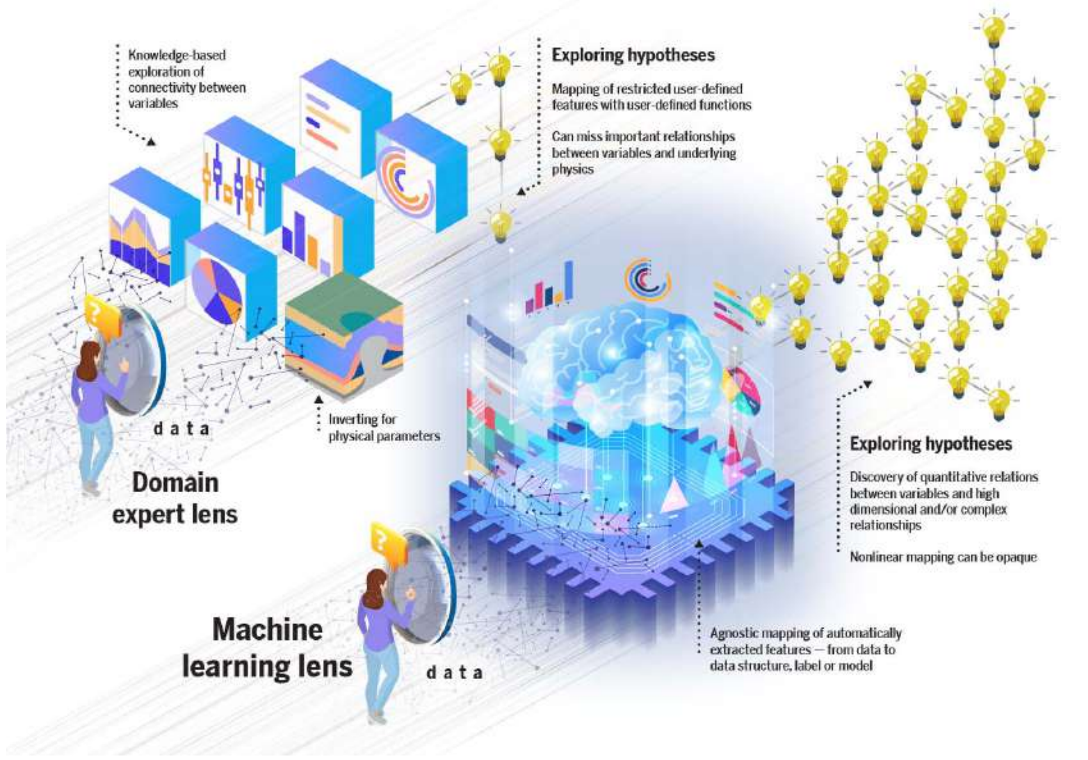
Groups of similar objects

For specific examples of ML applied to solid Earth geoscience:

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]*



Algorithms for big scientific data

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

FAST: scalable "Large-T" earthquake detection

Leverages technology for efficient audio recognition

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

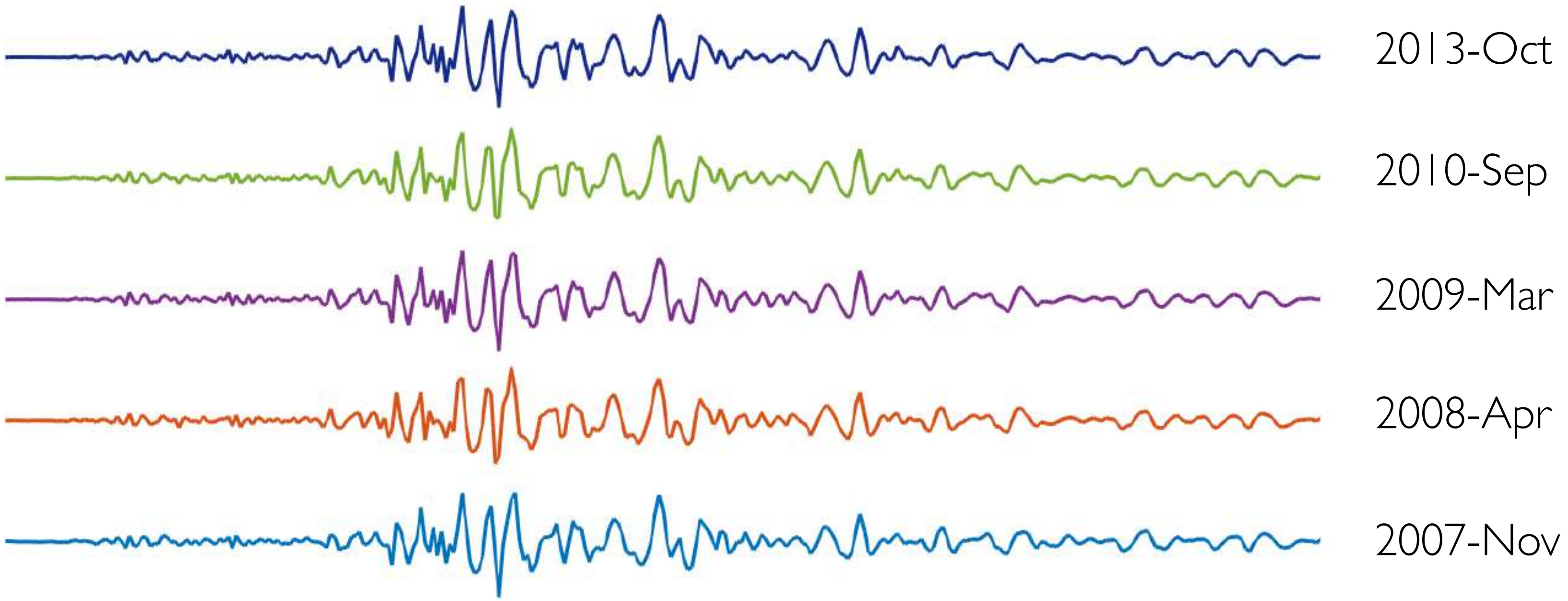
Computationally efficient:
500× more data with reduced runtime

[Yoon et al., Sci, Adv, 2015]

FAST: scalable "Large-T" earthquake detection

Data mining approach: extracting patterns from large data sets

- ***What pattern?*** Repeating or similar waveforms



Exact search for similar waveforms: quadratic scaling – small data only

Source: NCSN station CCOB.EHN

FAST: scalable "Large-T" earthquake detection

Data mining approach: extracting patterns from large data sets

- ***Scaling to big data?*** Locality-sensitive hashing (LSH) for similarity search

Find duplicate web pages
Search for copyright content
Identify songs

FAST: efficient similarity search with LSH

Key ideas:

1. Searching a well-organized data collection is faster.
   LSH clusters together similar waveforms for quick retrieval.

vs.

2. Sacrificing (a little) accuracy can substantially reduce runtime.
   FAST uses a *fast, approximate* rather than *slow, exact* similarity search.

Exact search:	**7** days of data	in	**215** hours runtime
FAST:	**10** years of data	in	**3–16** hours

[Rong et al., PVLDB 2018; Yoon et al., BSSA 2019]

DAS and Big Scientific Data analysis

- DAS produces larger data sets than traditional seismic arrays
- Solutions will leverage modern data science & computing:
  - Machine learning
  - Efficient algorithms
  - Parallel and distributed computing, data reduction, etc.

Questions?

Available on this afternoon (4pm ET)
Zoom Meeting ID: **991 7999 4192**

karianne_bergen@brown.edu