Improving earthquake detection with data mining & machine learning

Karianne Bergen

kbergen@stanford.edu

Institute for Computational and Mathematical Engineering, Stanford University

[next position: Data Science Initiative Postdoctoral Fellow, Harvard University]

What is Machine Learning?

Machine learning (ML)

A set of tools for *automatically* learning and recognizing complex patterns *from data*

- e.g. linear regression, logistic regression, PCA

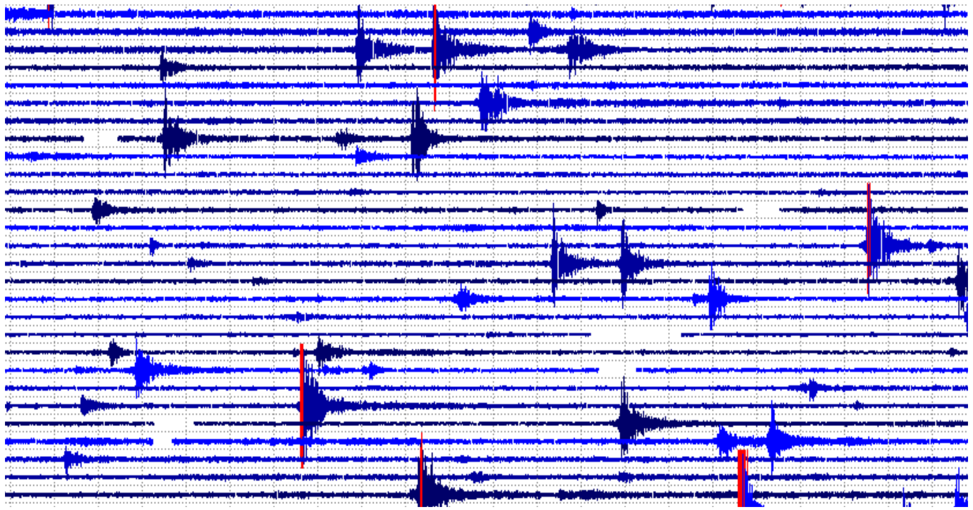
Data mining

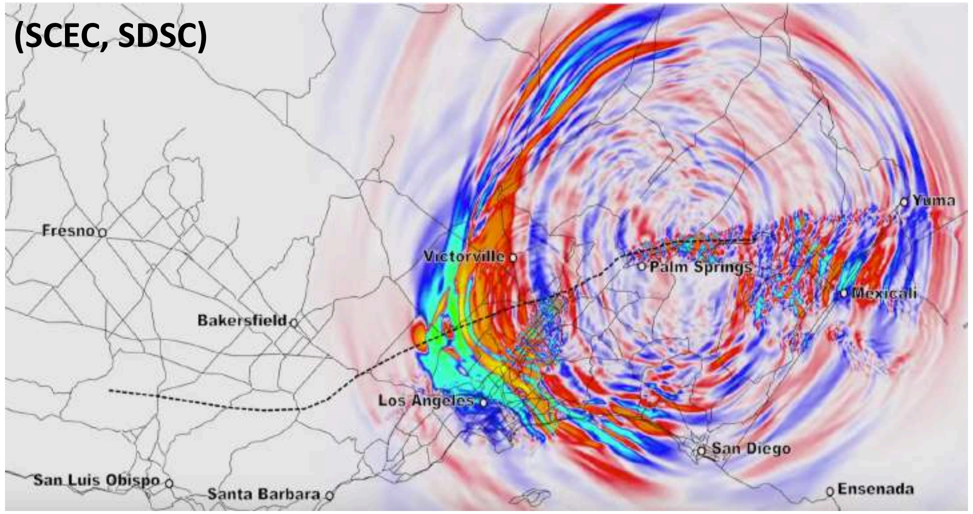
Tools for extracting unknown patterns or information from large data sets

- Closely related to machine learning

Machine learning for data-driven discovery

Scientific discovery depends on ability to ***extract information*** from ***massive data sets.***

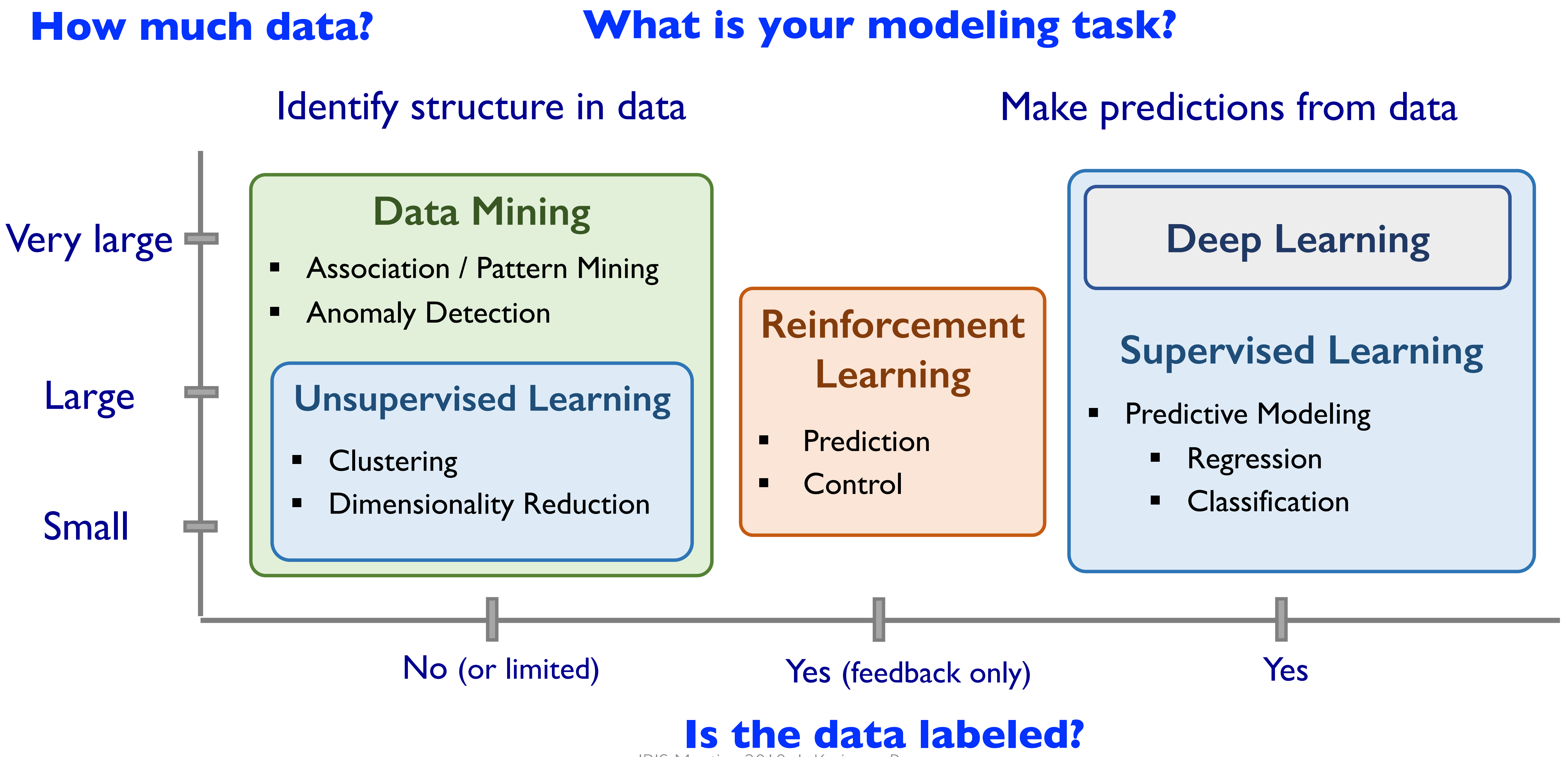




Use machine learning & data mining to:

- *Automate* large-scale data processing or specialized, repetitive tasks
- *Model* complex relationships
- *Discover* interesting or unexpected patterns

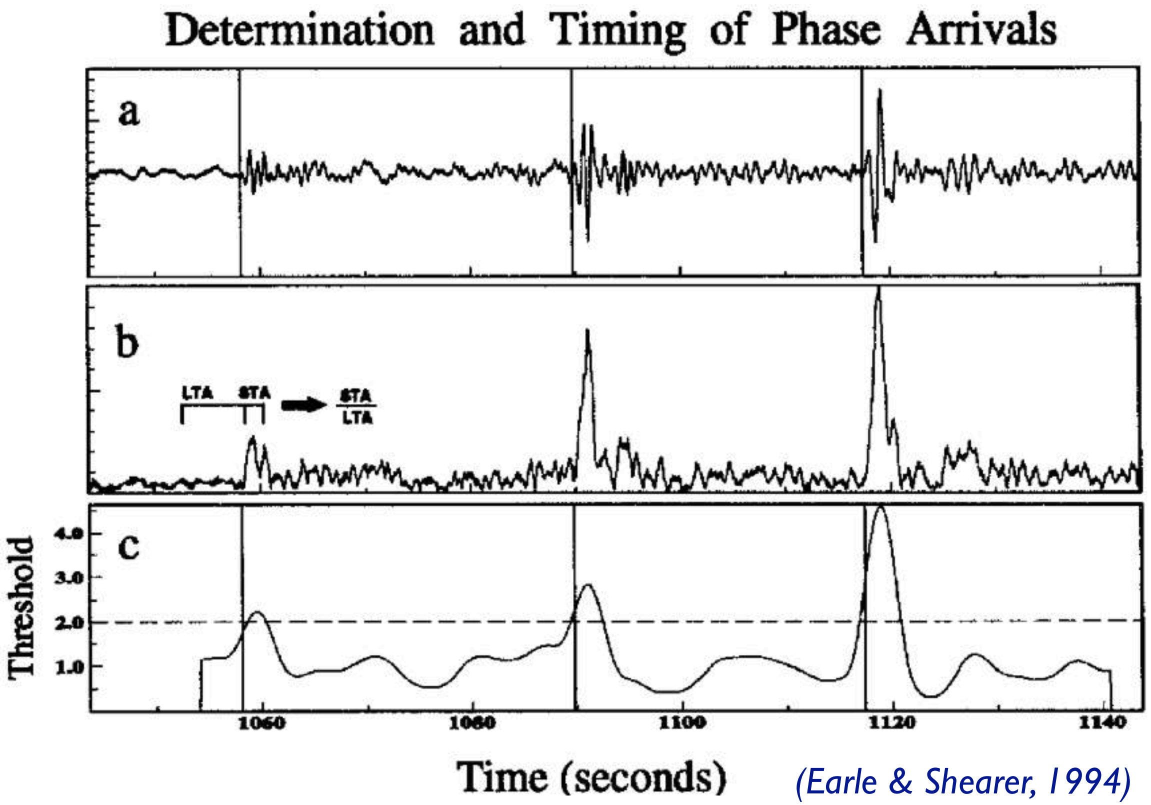
Selecting a machine learning approach



Why data mining & ML for earthquake detection?

- Two key properties:
  - Data-driven outcomes

Energy detectors (STA/LTA)



(Earle & Shearer, 1994)

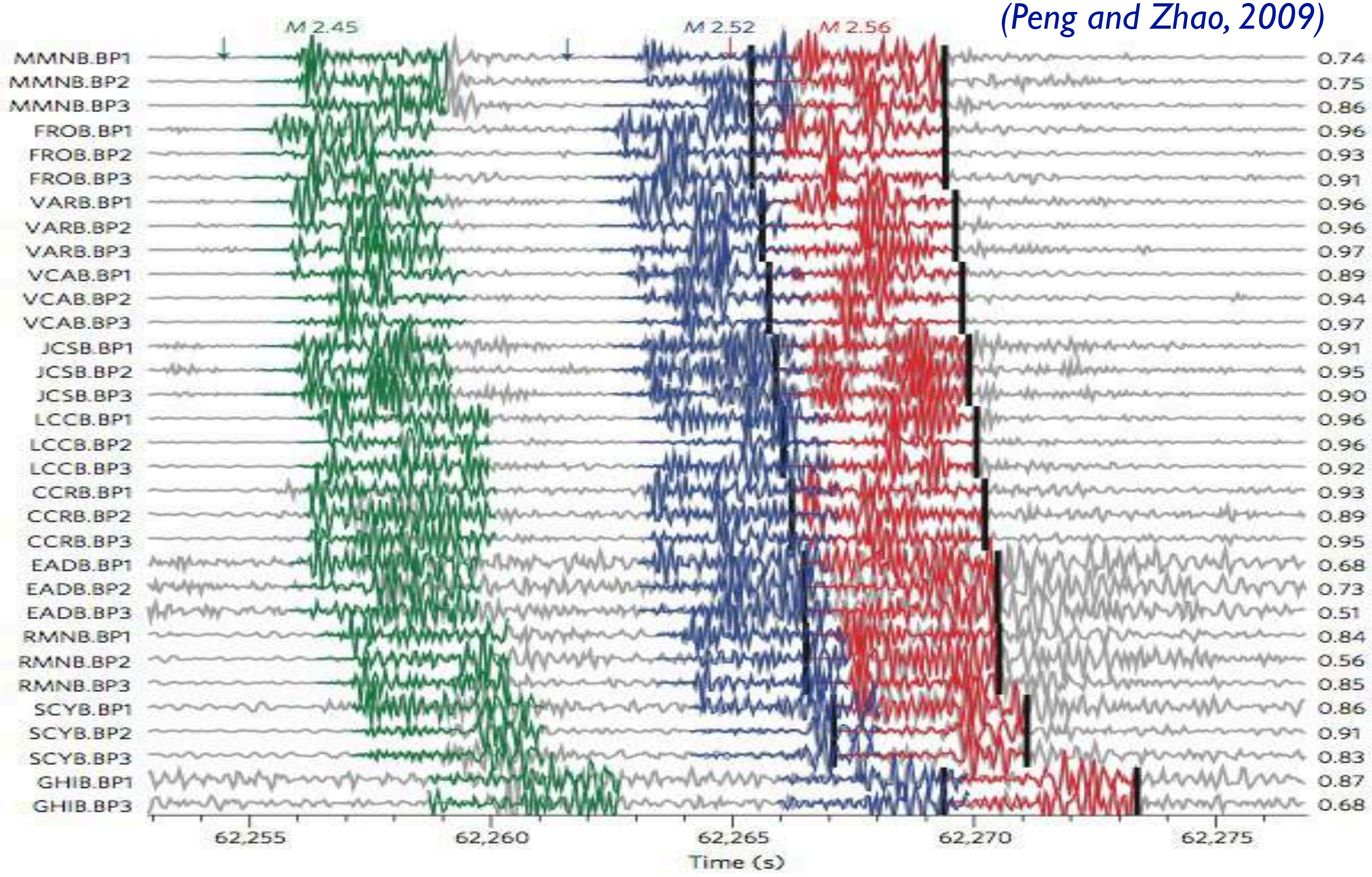
Does not adapt / improve based on past observations

Why data mining & ML for earthquake detection?

- Two key properties:
  - Data-driven outcomes
  - Ability to generalize

Template Matching

(Peng and Zhao, 2009)

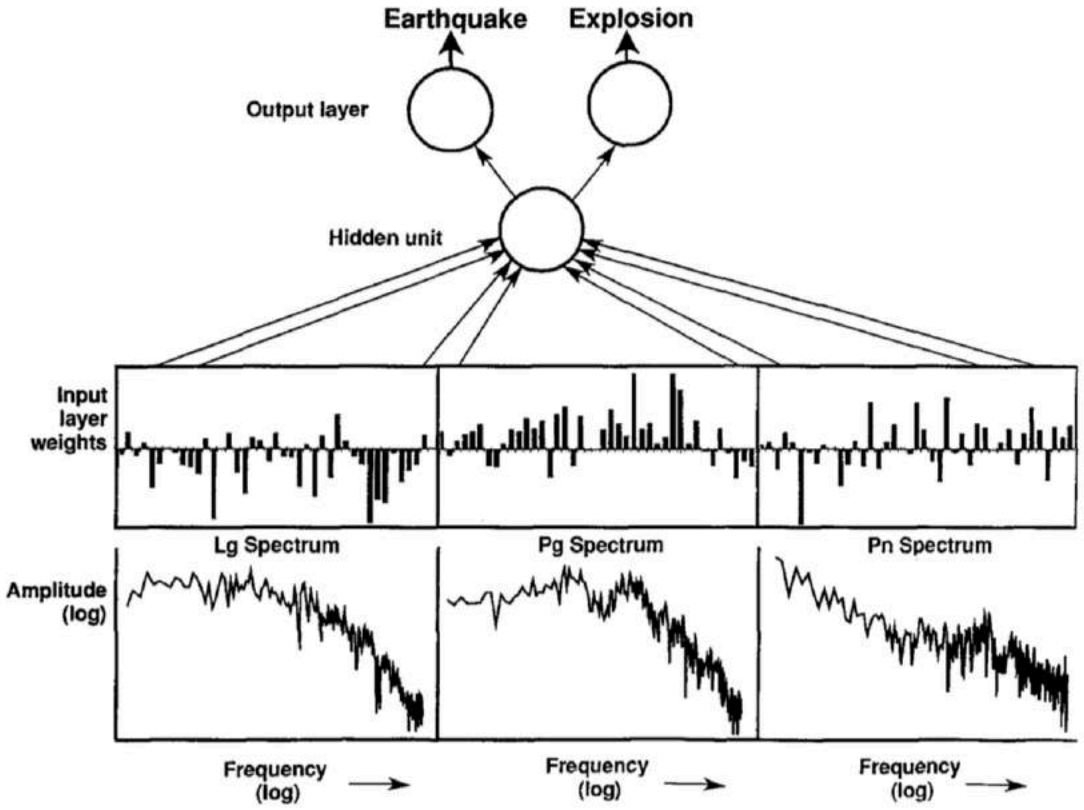


Memorizes template waveforms – no new sources

Seismologists have been using ML for > 20 years

- Artificial Neural Networks

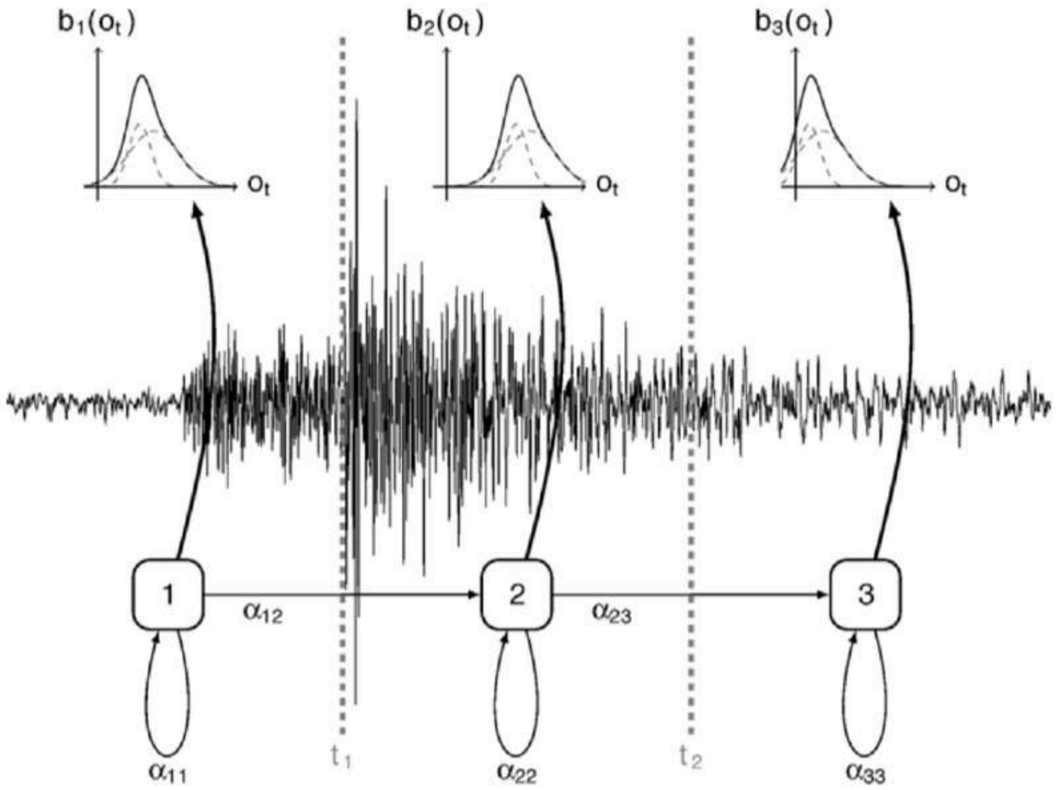
(e.g. Dowla et al., 1990; Dysart & Pulli, 1990)



Dowla et al., (1990)

- Hidden Markov Models

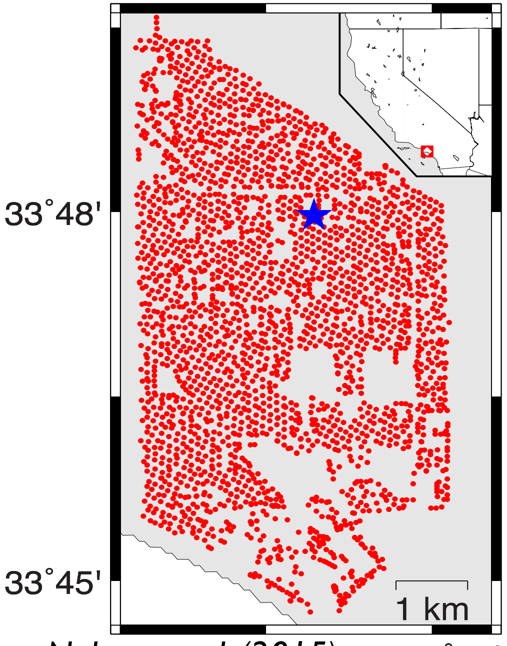
(e.g. Ohrnberger, 2001; Beyreuther et al., 2008)



Beyreuther et al., (2008)

Recent developments → New opportunities in seismology

- Massive seismic data sets
- New ML algorithms and models
- Improvements in computing technology



Nakata et al. (2015)

Long Duration (Large-T)

>10 years continuous waveform data

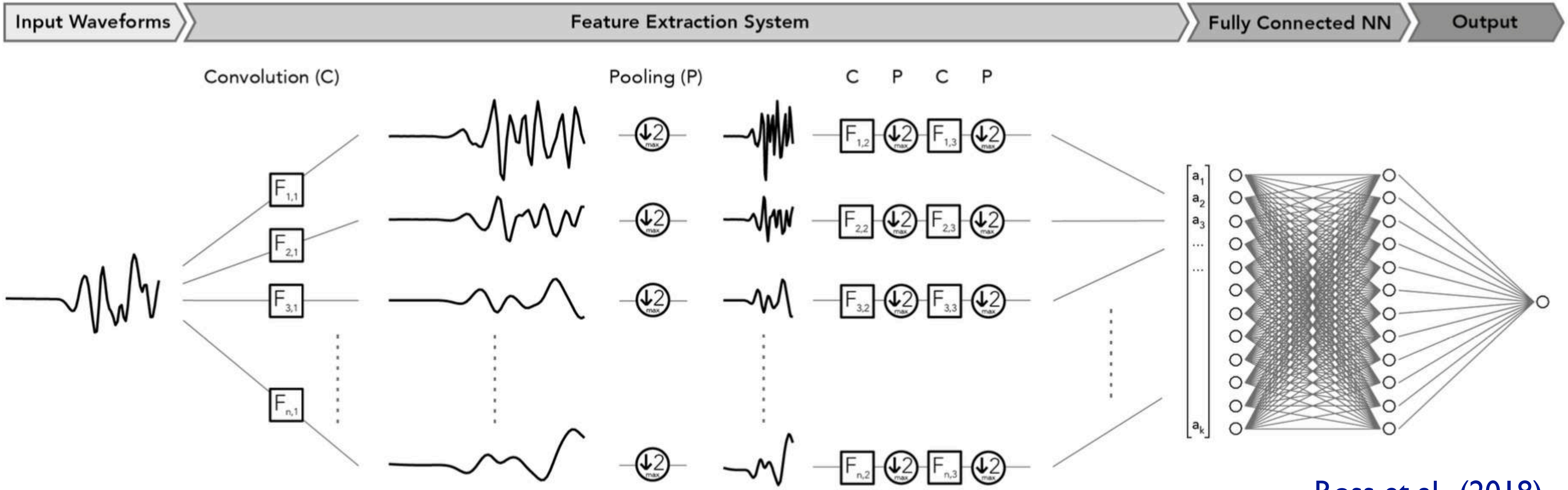
Big Networks (Large-N)

1000's of sensors

New Data Sources

Recent developments → New opportunities in seismology

- Massive seismic data sets
- New ML algorithms and models
- Improvements in computing technology



Ross et al., (2018)

Recent developments → New opportunities in seismology

- Massive seismic data sets
- New ML algorithms and models
- Improvements in computing technology

GPU Computing

Open Source Tools

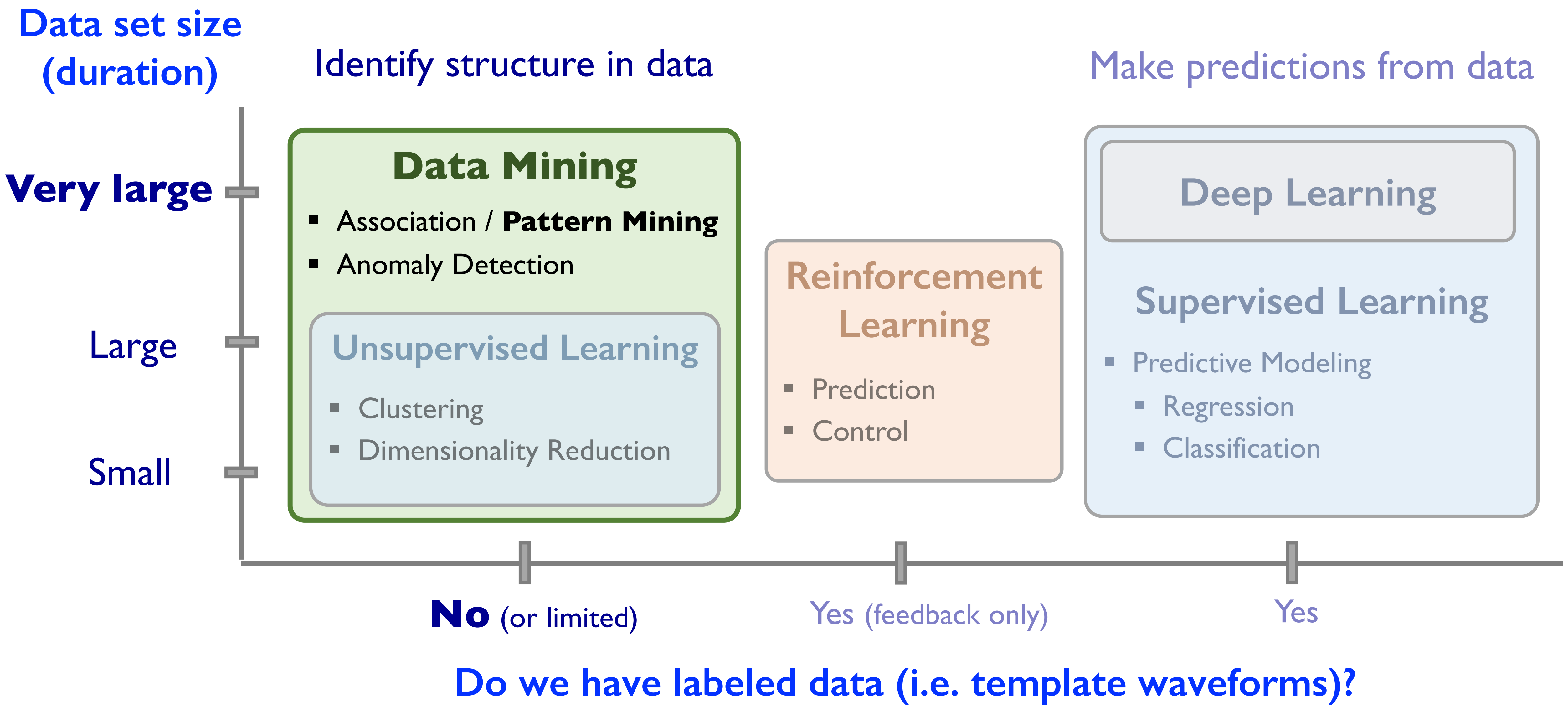
TensorFlow

Caffe2

Keras

PYTORCH

Data mining for earthquake detection



FAST: a data mining approach to earthquake detection

Stanford EARTH

Stanford Computer Science

ICME


C.E. Yoon


O. O'Reilly


G.C. Beroza


H. Elezabi


K. Rong


P. Bailis


P. Levis

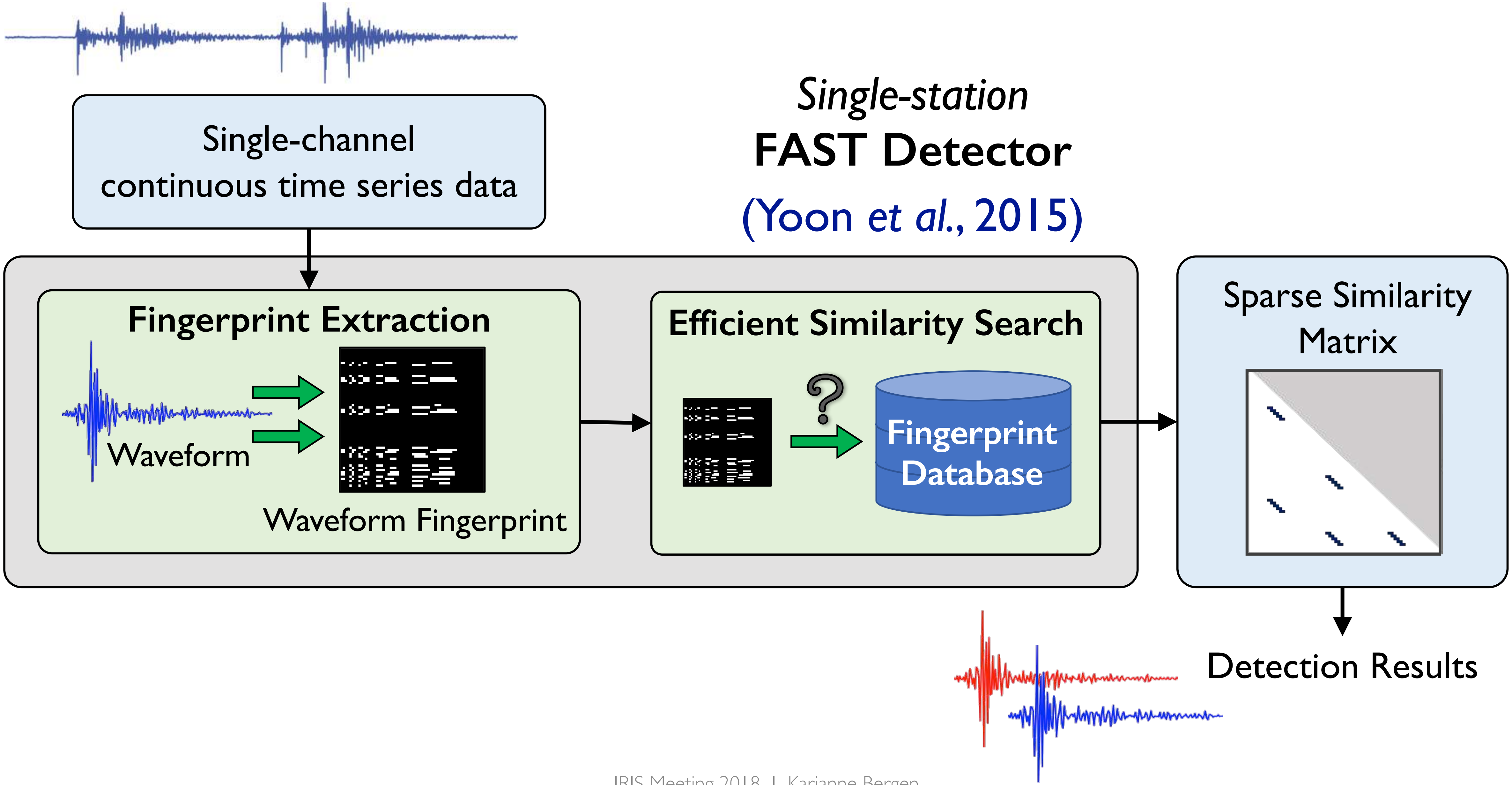
Fingerprint and Similarity Thresholding (FAST)

- Uses waveform-similarity as basis for detection
- Unsupervised technique – does not require templates

- **Detection task:** find all pairs of similar waveforms in continuous data
  - Data mining – similarity search / near neighbor search
  - Computational efficiency – locality-sensitive hashing, not exhaustive search
- Similar to technology for audio clip identification

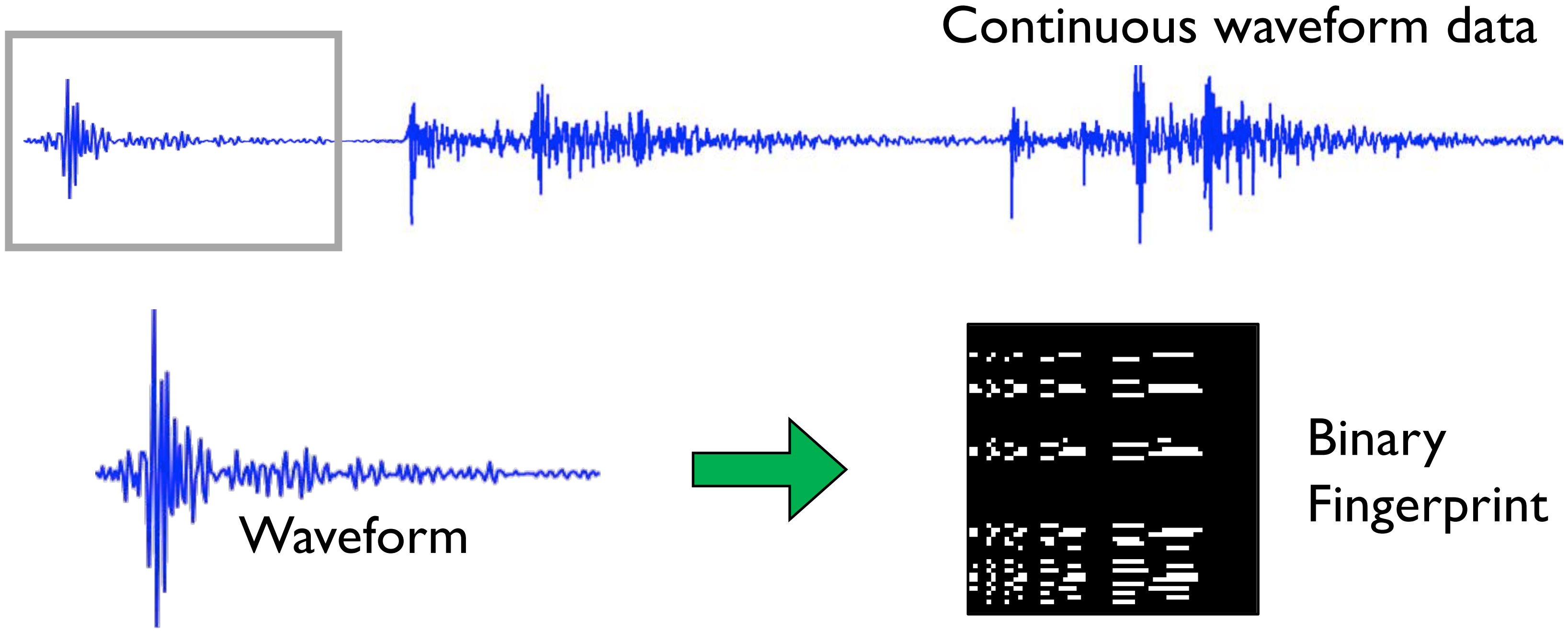
Yoon et al. (2015), *Sci. Adv.*

Fingerprint and Similarity Thresholding (FAST)

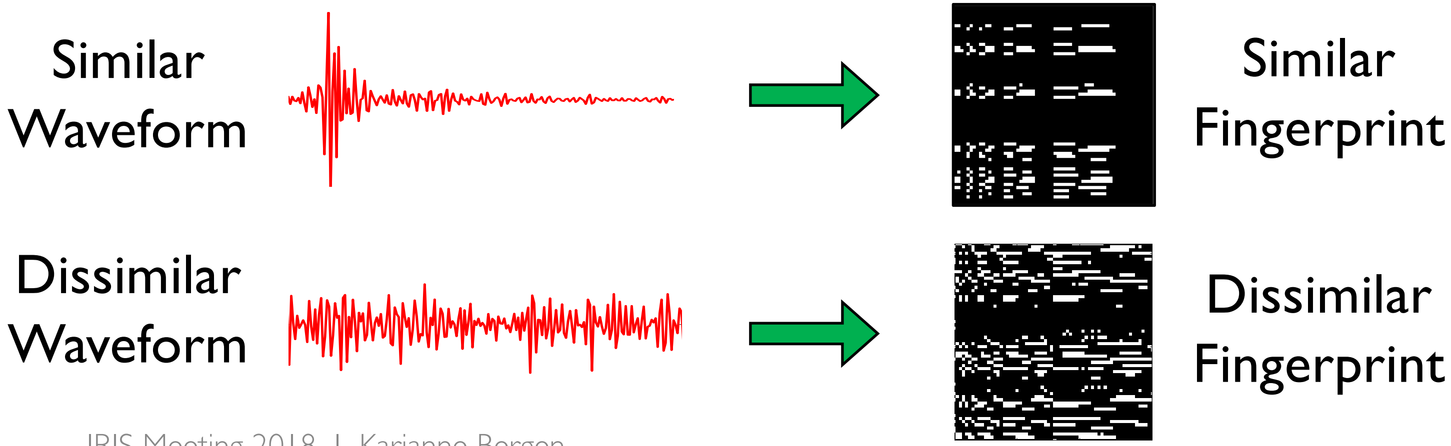


FAST Detection Pipeline

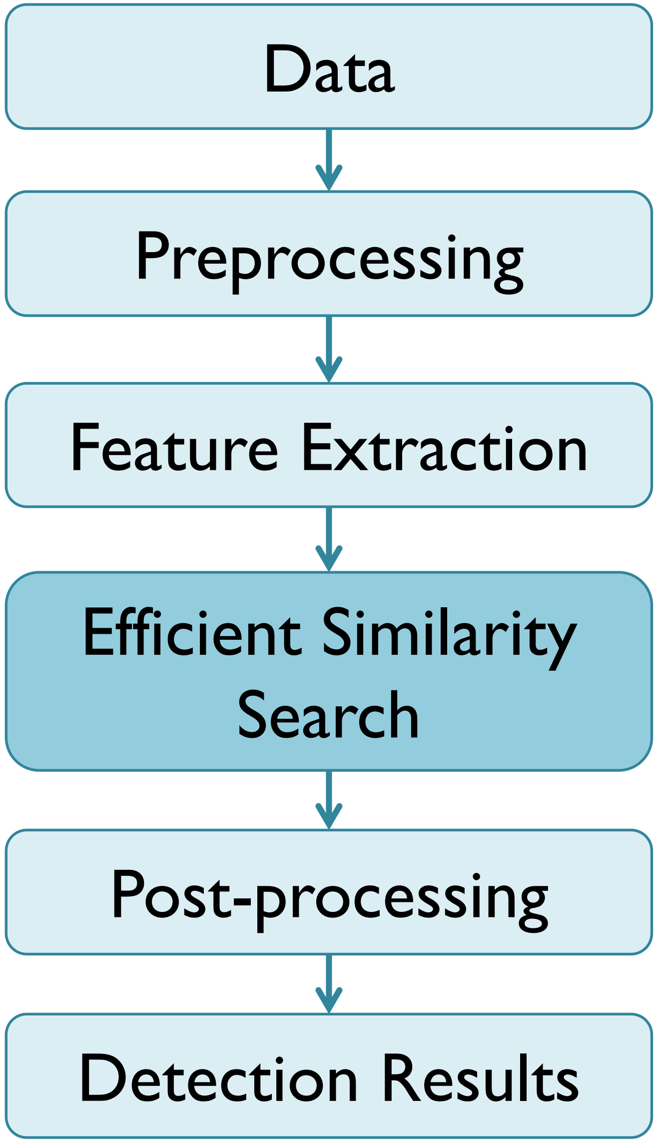
- Data
- Preprocessing
- **Feature Extraction**
- Efficient Similarity Search
- Post-processing
- Detection Results



Fingerprints should be discriminative

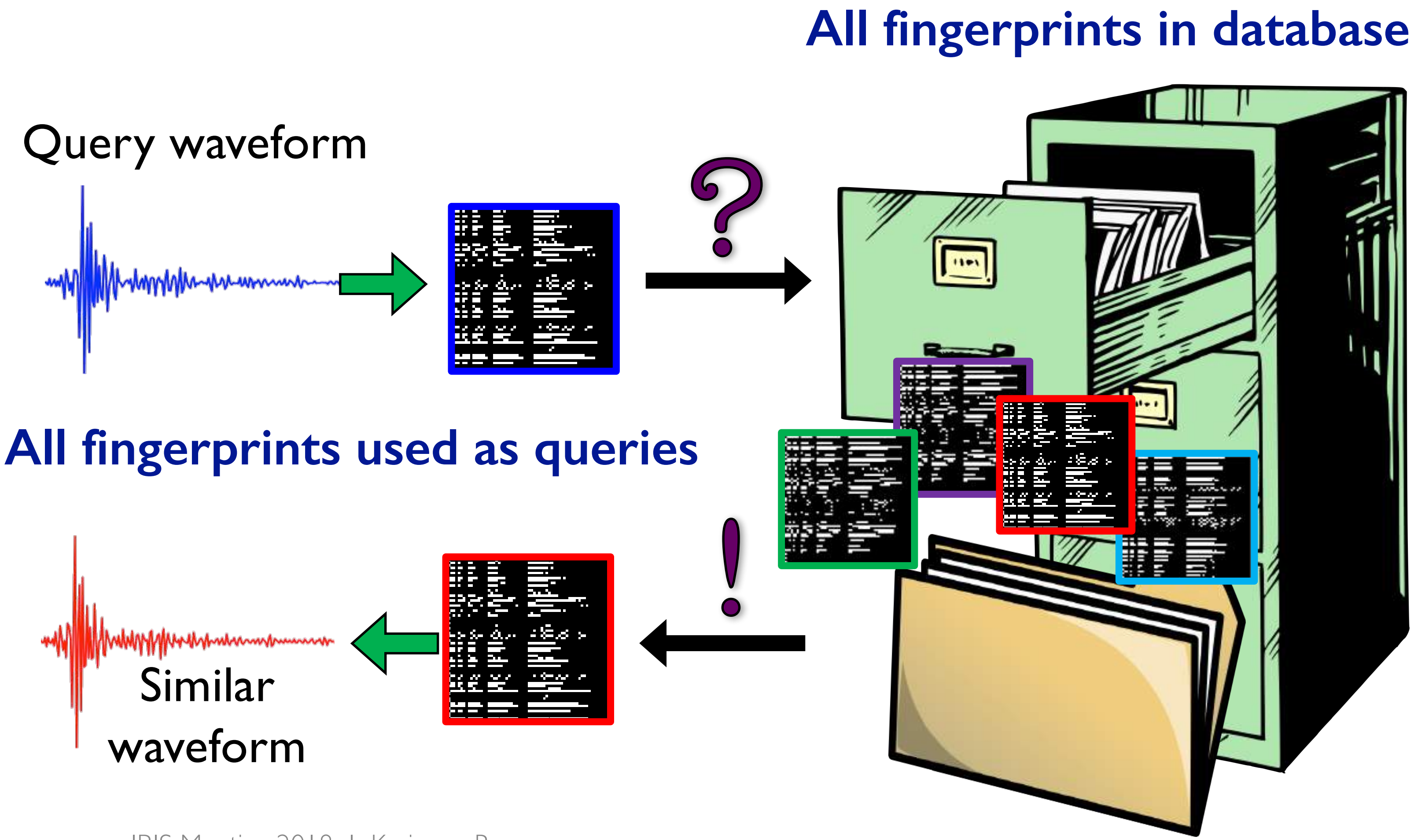


FAST Detection Pipeline

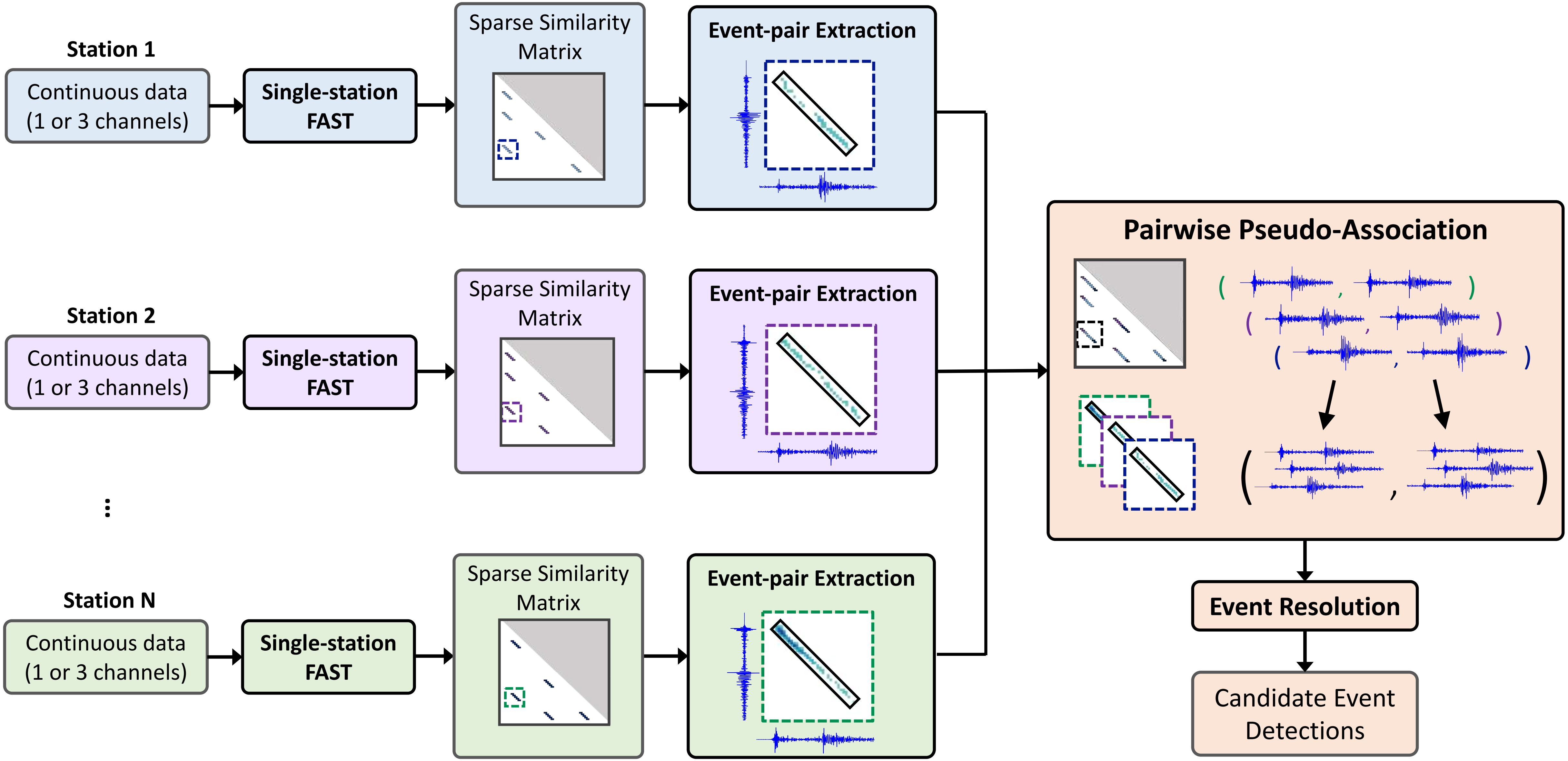


Fast approximate similarity search

- **MinHash** and **Locality Sensitive Hashing (LSH)**

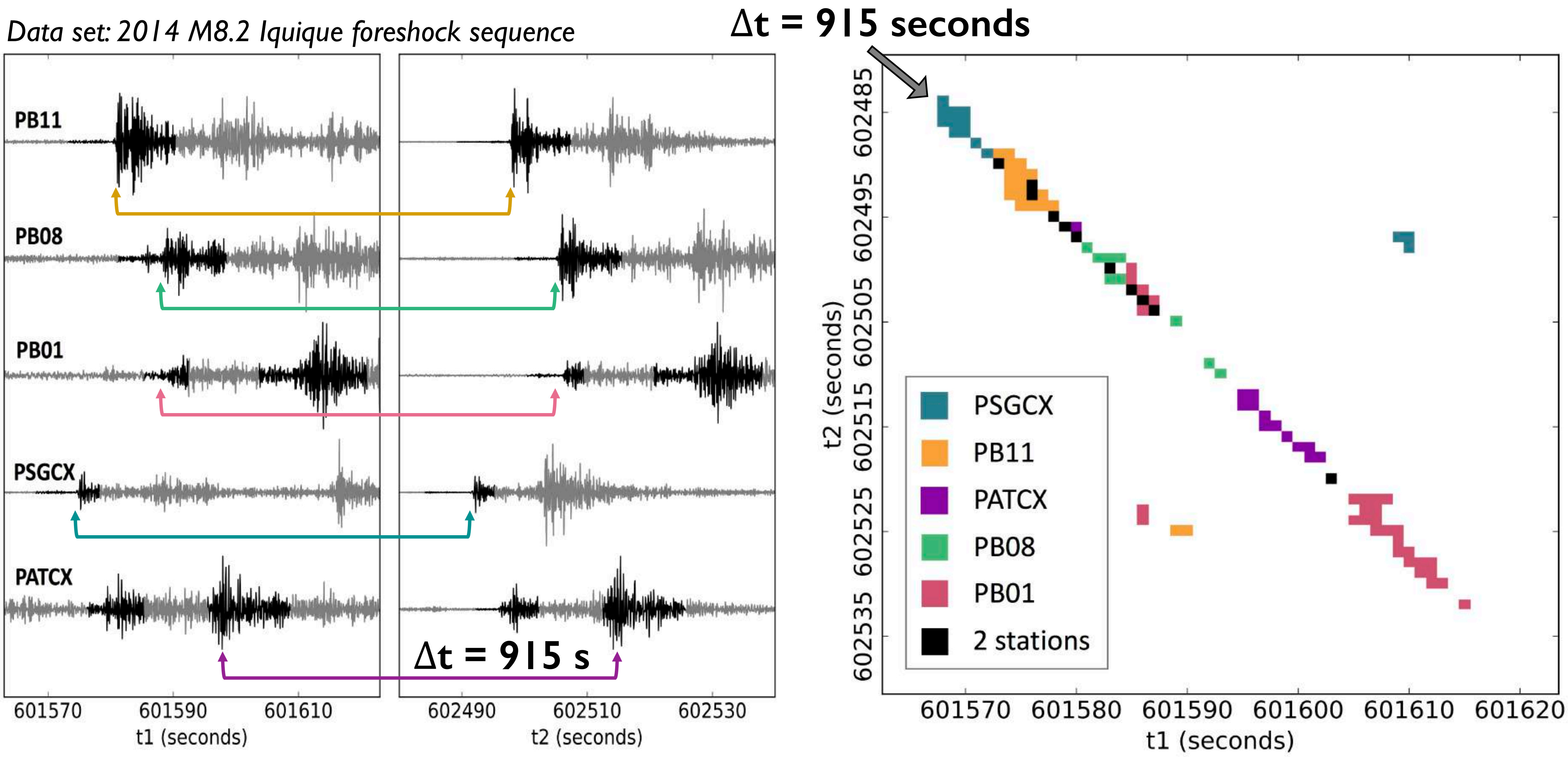


Network (Multi-station) Detection with FAST



Association of pairwise detections

Data set: 2014 M8.2 Iquique foreshock sequence

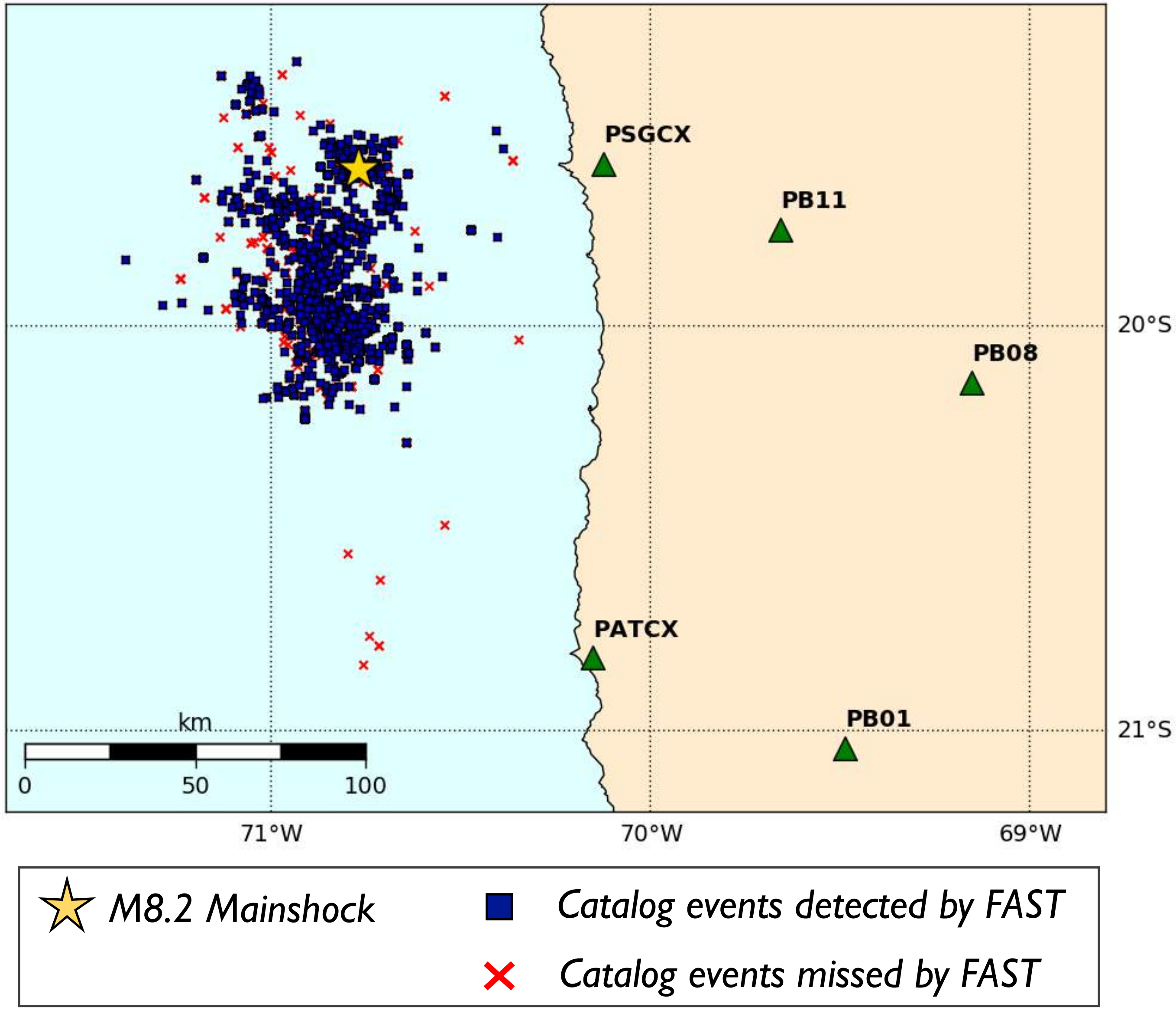


Results: 2014 M8.2 Iquique foreshock sequence

2788
candidate events identified by FAST
(at 4+ of 5 stations)

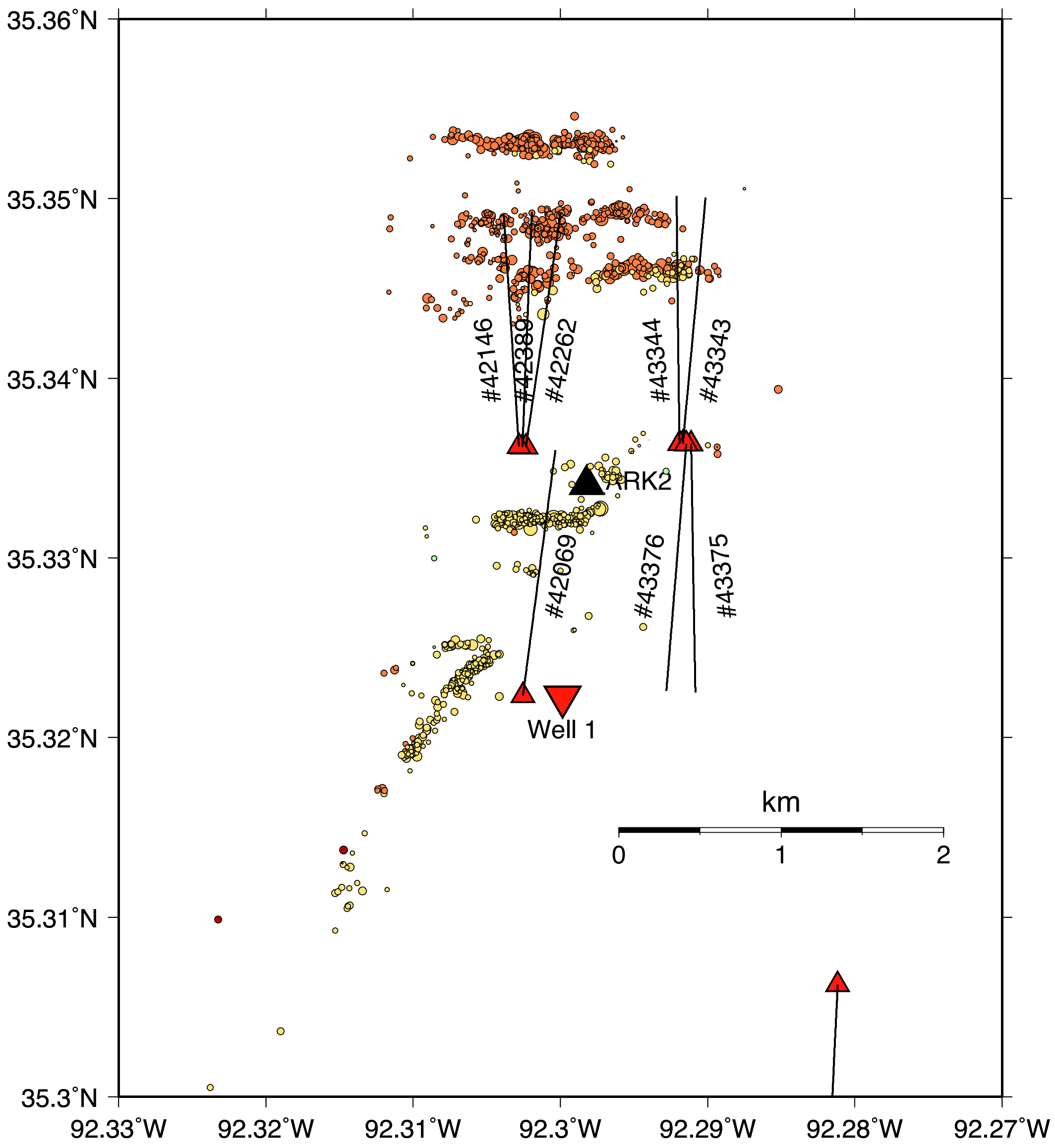
571
events in local (CSN) catalog

<1%
false discovery rate



Bergen & Beroza (2018), *Geophys. J. Int.*

Results: Induced seismicity in Guy-Greenbrier, AK



75

Events in catalog,
$1.2 < M_L < 2.9$

13,026

Events detected by FAST,
$-1.5 < M_L < 2.9$

FAST reveals spatial and temporal correlations between events and individual stages of hydraulic fracturing stimulation

Yoon et al. (2017), *J. Geophys. Res.*

FAST in long-duration (large-T) data

10 years
waveform data

300 million
fingerprints

16 hours
similarity search runtime

Better memory management
Parallel queries in similarity search

FAST software: https://github.com/stanford-futuredata/FAST

Rong et al. (2018), *arXiv.*

Recent work: data mining & ML in seismology

Automation

- Earthquake detection and phase-picking with deep neural networks
  [e.g. Perol et al (2018), Wu et al, (2018), Ross et al. (2018), Zhu & Beroza (2018)]

Modeling

- Synthetic seismograms with deep generative models [Krischer & Fichtner (2017)]
- Ground motion prediction with random forests [Trugman & Shearer (2018)]

Discovery

- Identifying temporal patterns in seismic source spectra with unsupervised learning [Holtzman et al. (2018)]

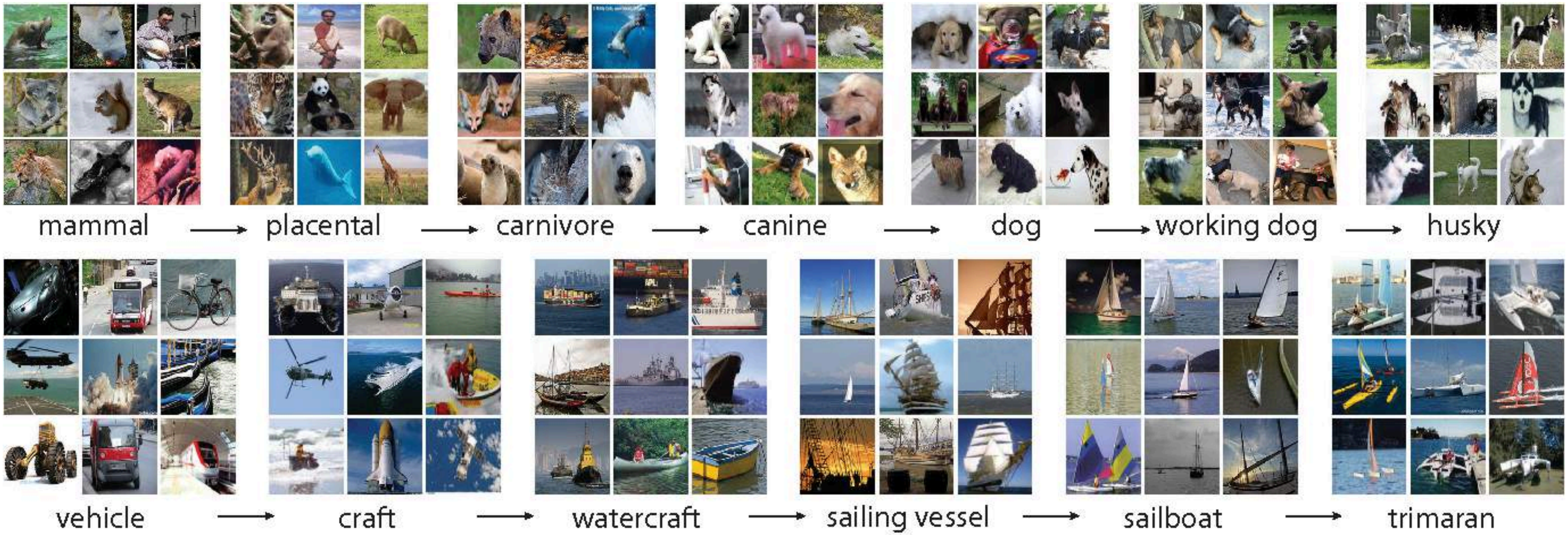
The future of data mining & ML in seismology

- Benchmark data sets
- Open source code & data
- Data science education

Benchmark data sets

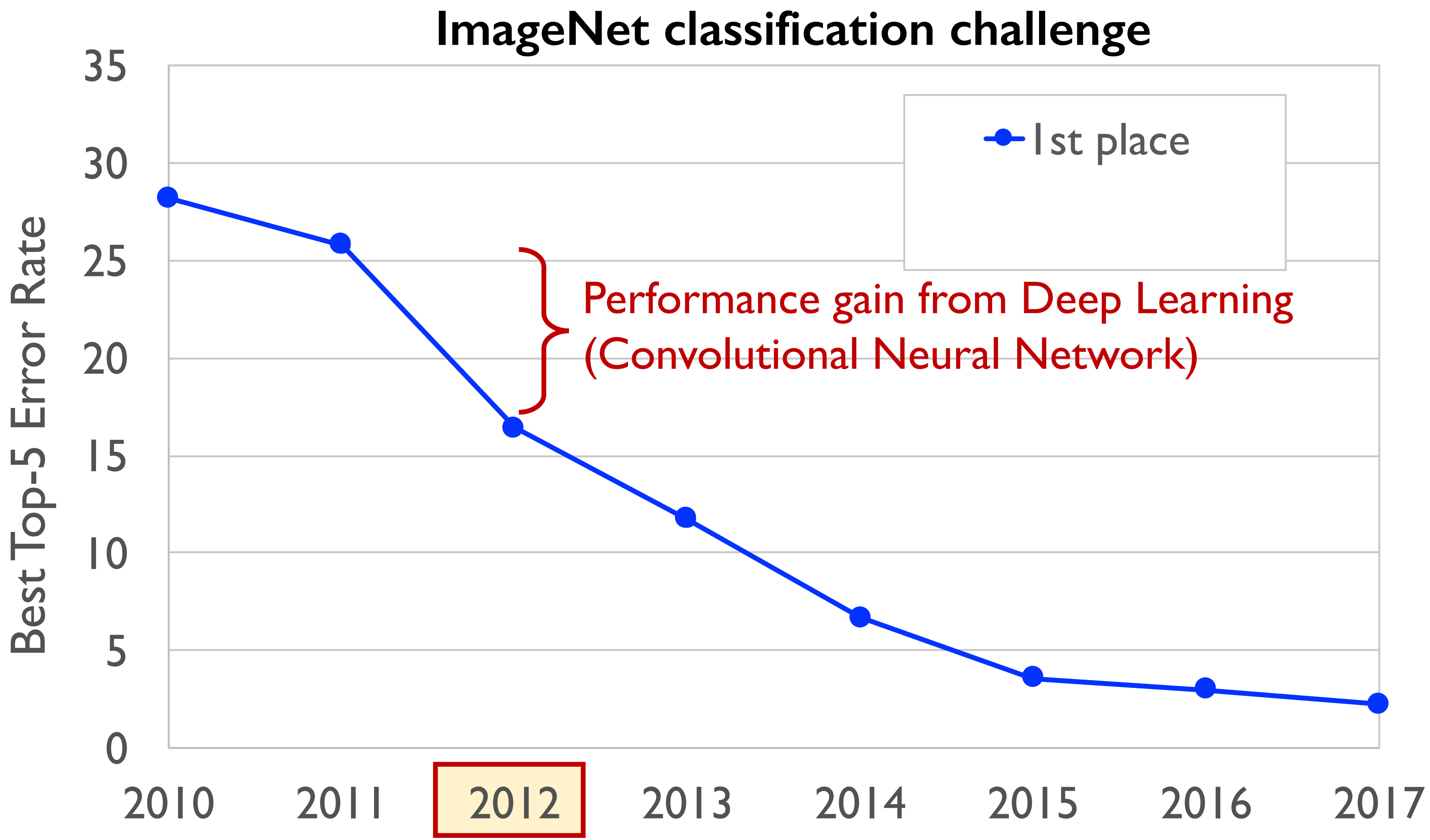
- Benchmark data sets & competitions drive progress in ML/AI communities
  - High quality data set available to community
  - Compare algorithms & identify best methods

IMAGENET

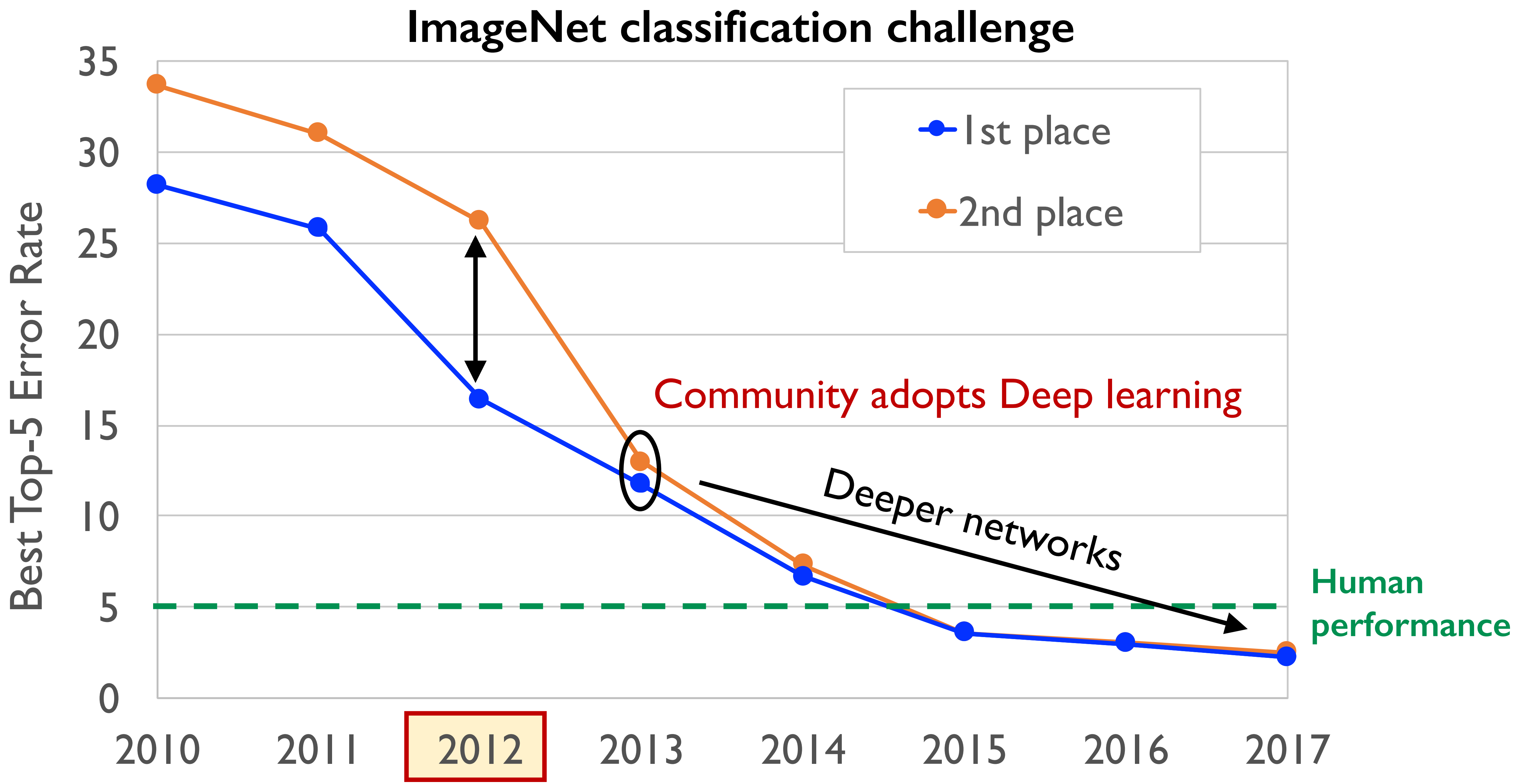


Deng et al. (2009), *CVPR.*

Benchmarks: Moving toward better algorithms

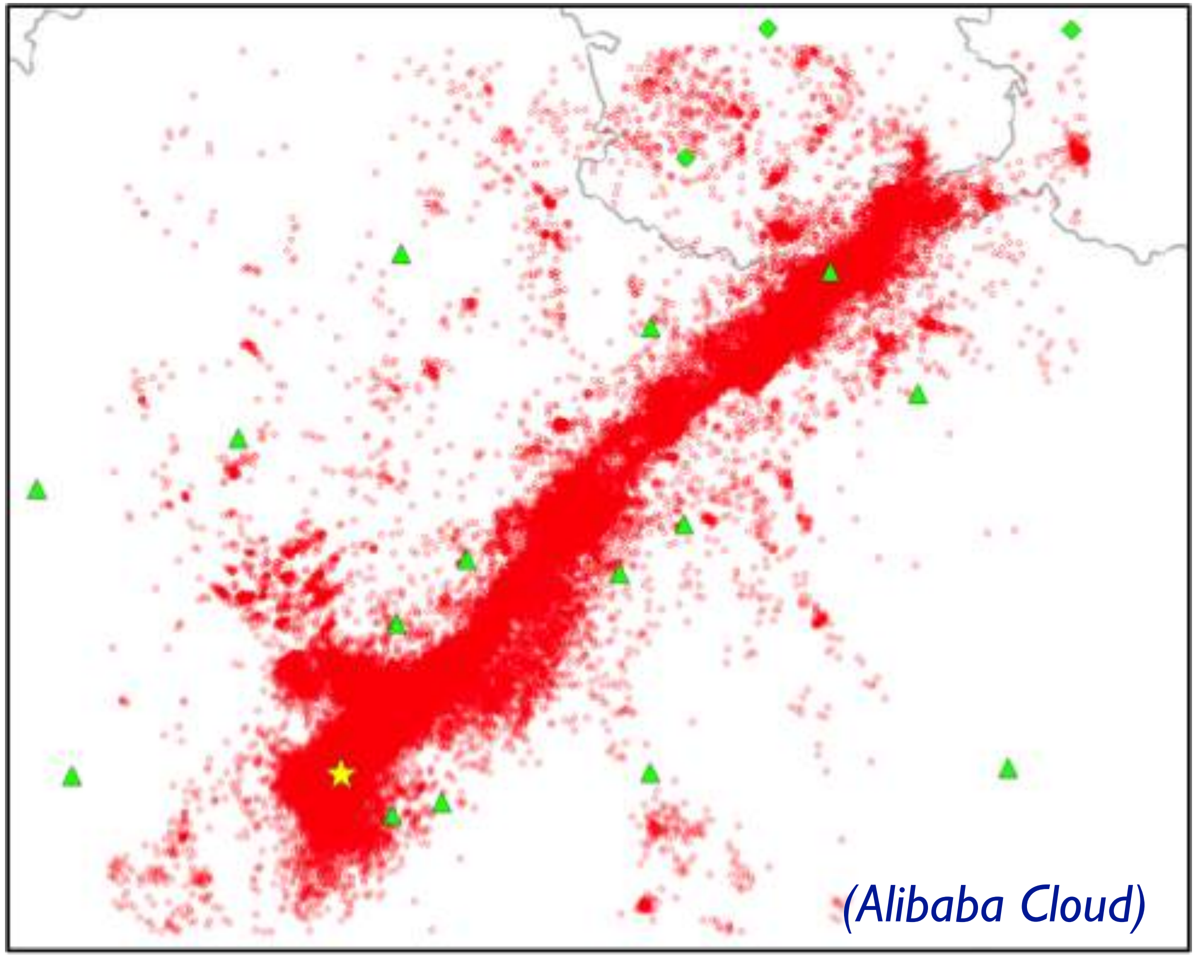


Benchmarks: Moving toward better algorithms



Benchmark data sets – earthquake detection

- SeismOlympics: Aftershock detection contest – 2008 Wenchuan Earthquake
- Task: detection and phase-picking
- Data: 16 stations, 5 months
- Ground truth: from CEA analysts
- 1000+ teams competed
- Opportunity for researchers to test their algorithms
- Need more benchmarks/contests
  - Challenge: ground truth, bias
  - Diversity of tasks & data sets



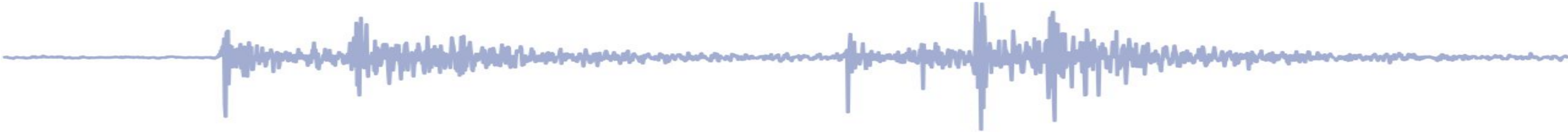
Fong et al. (2017), *SRL.*

The future of data mining & ML in seismology

- Benchmark data sets
- Open source code & data
- Data science education

Questions?

kbergen@stanford.edu



FAST software available at: https://github.com/stanford-futuredata/FAST

References:

- Yoon et al. (2015). Earthquake detection through computationally efficient similarity search. Science Advances.
- Bergen & Beroza (2018). Detecting Earthquakes over a Seismic Network using Single-Station Similarity Measures. Geophys. J. Int.
- Rong et al. (2018). Locality-sensitive hashing for earthquake detection: A case study scaling data-driven science. arXiv:1803.09835.