Curated Pacific Northwest (PNW) Seismic Dataset for Machine Learning

Yiyu Ni

CRESCENT Technical Short Course

Seattle, WA | May 12, 2025

1. Earthquake Catalog
2. Hands-on: query catalog and waveforms
3. Curated seismic dataset
4. SeisBench ecosystem
5. Hands-on: use PNW dataset

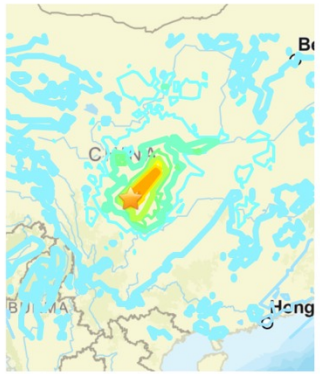
Earthquake Catalog

Earthquake catalog is the collection of earthquakes and attributes

M 7.9 - 58 km W of Tianpeng, China

2008-05-12 06:28:01 (UTC) | 31.002°N 103.322°E | 19.0 km depth

Interactive Map



Contributed by US

Regional Information

Contributed by US

Felt Report - Tell Us!

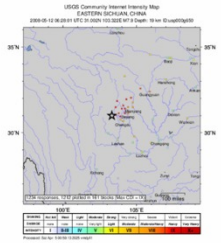
0 0 1 2 3 4

Responses

Contribute to citizen science. Please tell us about your experience.

Citizen Scientist Contributions

Did You Feel It? IX



Community Internet Intensity Map

Contributed by US

ShakeMap IX

Estimated Intensity Map

Contributed by ATLAS

Ground Failure

Landslide Estimate

Extensive area affected

Extensive population exposed

Liquefaction Estimate

Significant area affected

Extensive population exposed

Contributed by US

Origin

Review Status

REVIEWED

Magnitude

7.9 mwc

Depth

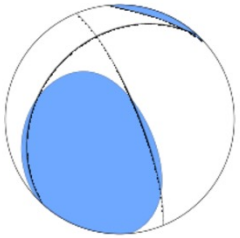
19.0 km

Time

2008-05-12 06:28:01 UTC

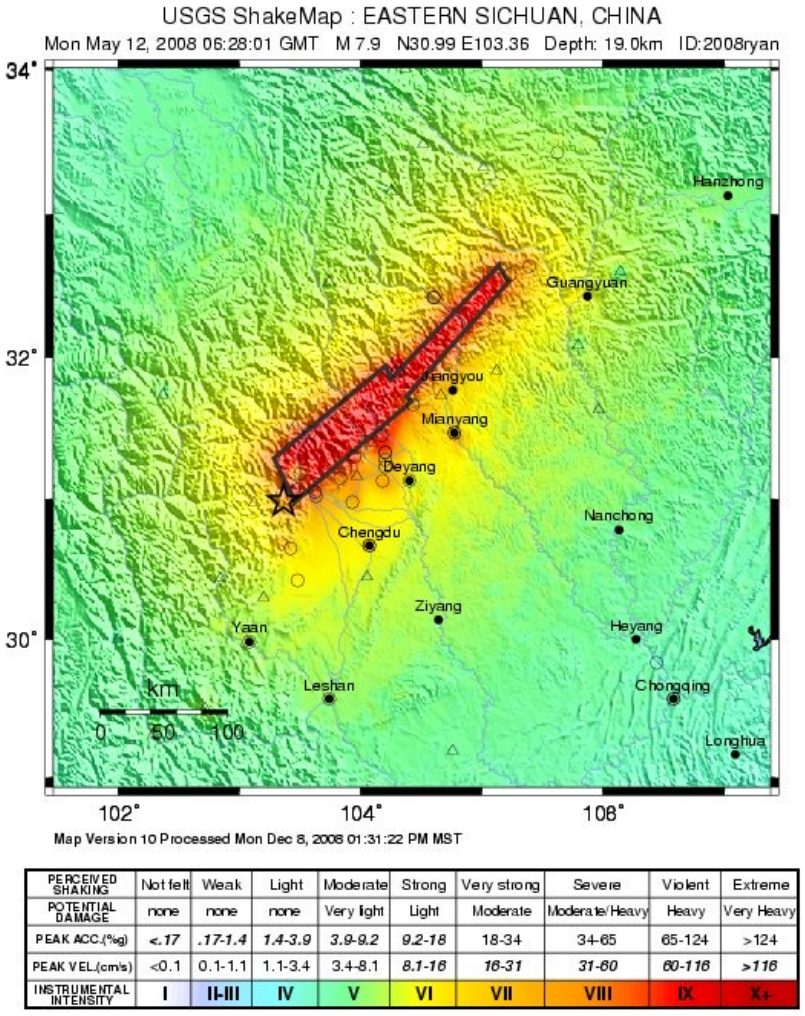
Contributed by US

Moment Tensor



Fault Plane Solution

Contributed by US DUPUTEL

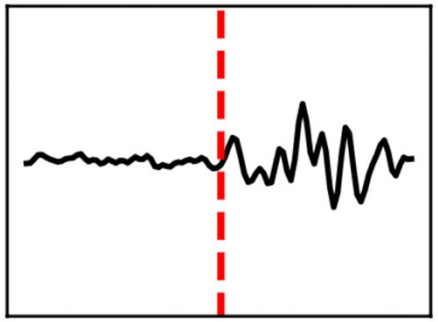


PERCEIVED SHAKING	Not felt	Weak	Light	Moderate	Strong	Very strong	Severe	Violent	Extreme
POTENTIAL DAMAGE	none	none	none	Very light	Light	Moderate	Moderate/Heavy	Heavy	Very Heavy
PEAK ACC.(%g)	<.17	.17-1.4	1.4-3.9	3.9-9.2	9.2-18	18-34	34-65	65-124	>124
PEAK VEL.(cm/s)	<0.1	0.1-1.1	1.1-3.4	3.4-8.1	8.1-16	16-31	31-60	60-116	>116
INSTRUMENTAL INTENSITY	I	II-III	IV	V	VI	VII	VIII	IX	X+

Earthquake catalog is the collection of earthquakes and attributes

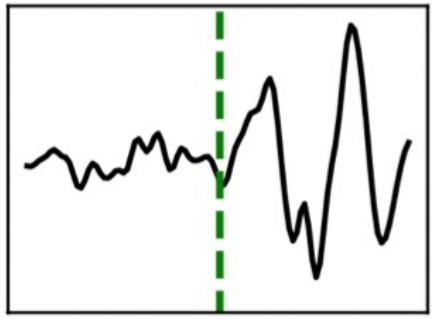
Station	Distance	Azimuth	Phase	Arrival Time
PB B943 EHZ	0.03 °	144.78 °	P	7.4 s
PB B943 EH1	0.03 °	144.78 °	S	13.0 s
UW DOSE HHZ	0.12 °	192.55 °	P	7.6 s
UW DOSE HHE	0.12 °	192.55 °	S	13.4 s
UW COYL HHZ	0.17 °	144.49 °	P	8.1 s
UW COYL HHN	0.17 °	144.49 °	S	14.9 s

P wave pick



09:12:19.352
2021-10-22

S wave pick



09:12:31.252
2021-10-22

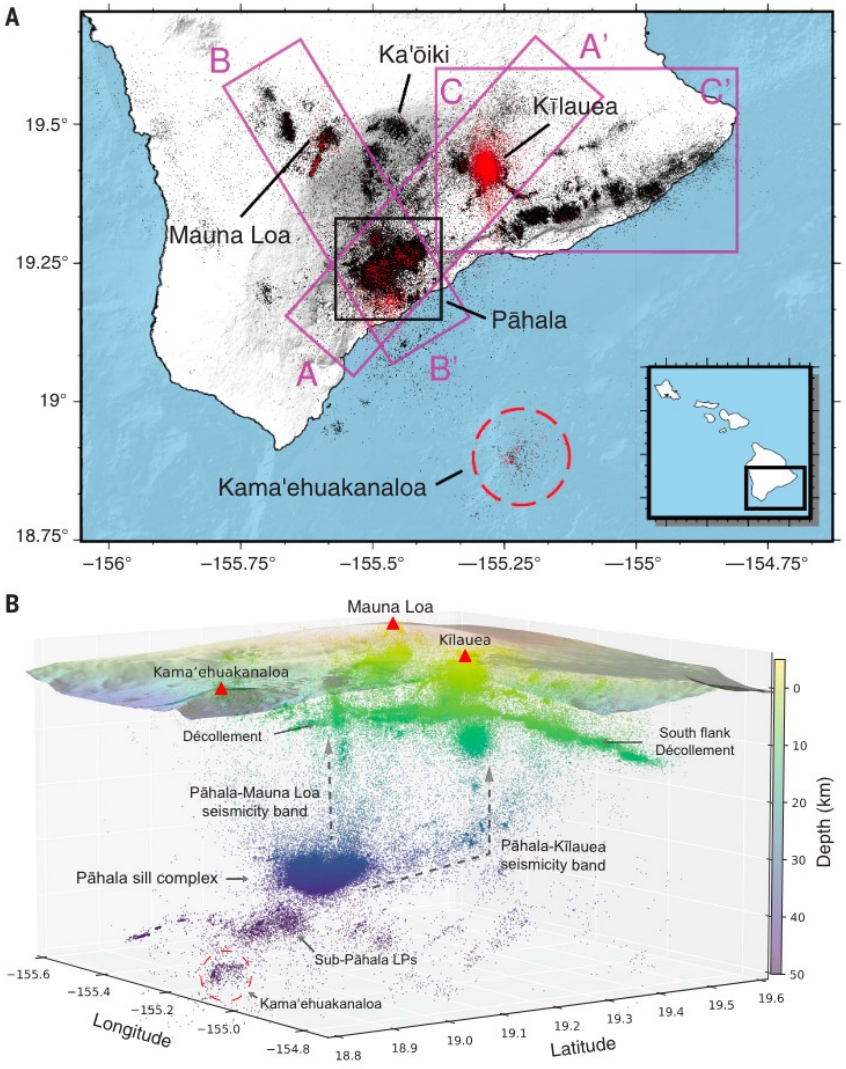
Earthquake catalog is the collection of earthquakes and attributes

Station	Distance	Azimuth	Phase	Arrival Time
PB B943 EHZ	0.03 °	144.78 °	P	7.4 s
PB B943 EH1	0.03 °	144.78 °	S	13.0 s
UW DOSE HHZ	0.12 °	192.55 °	P	7.6 s
UW DOSE HHE	0.12 °	192.55 °	S	13.4 s
UW COYL HHZ	0.17 °	144.49 °	P	8.1 s
UW COYL HHN	0.17 °	144.49 °	S	14.9 s

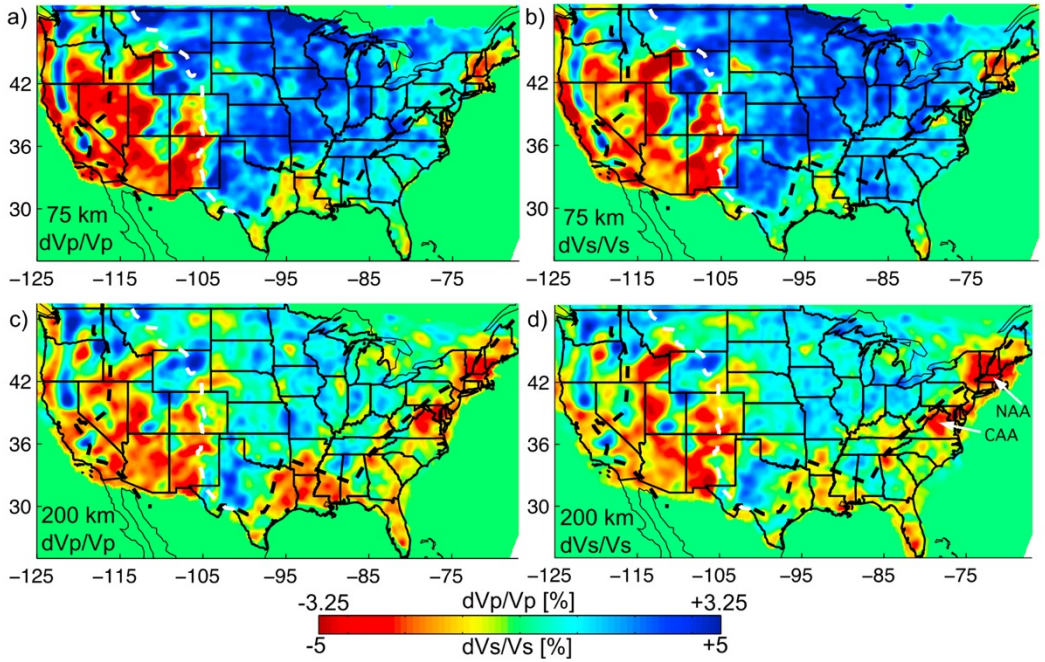
- Event location
- Origin time
- Magnitude
- Focal mechanism
- Event type
- Ground motion

Earthquake catalog

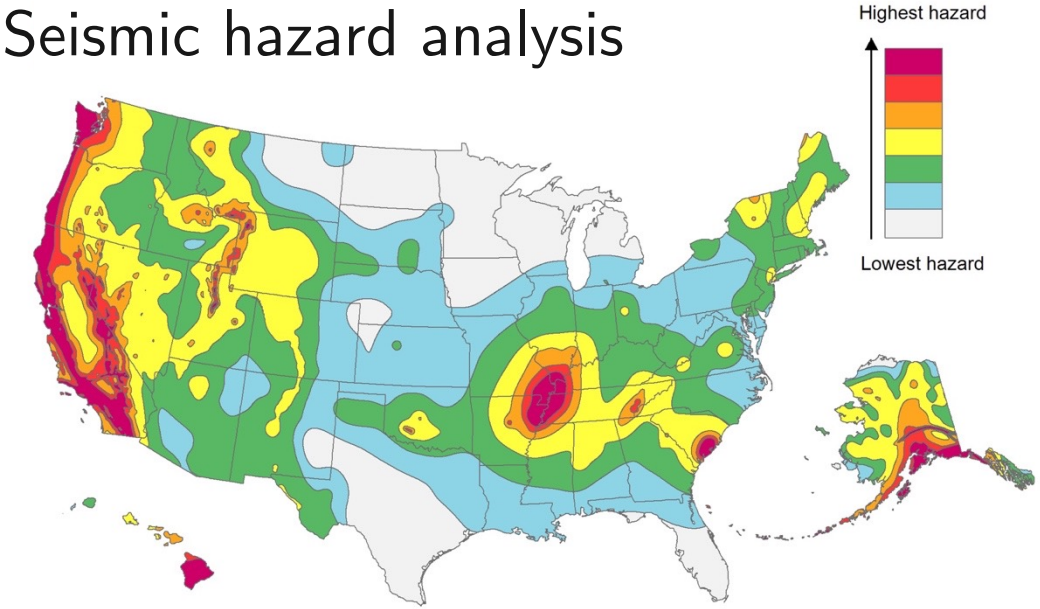
Volcano monitoring



Seismic tomography



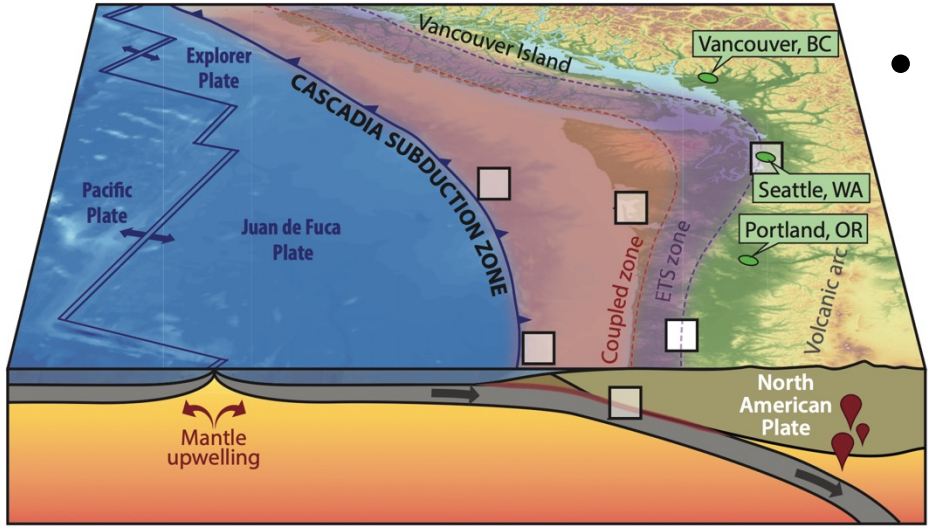
Seismic hazard analysis



Earthquake catalog

- Pacific Northwest (PNW) hosts a variety of earthquake behaviors: megathrust, intra-slab and crustal.

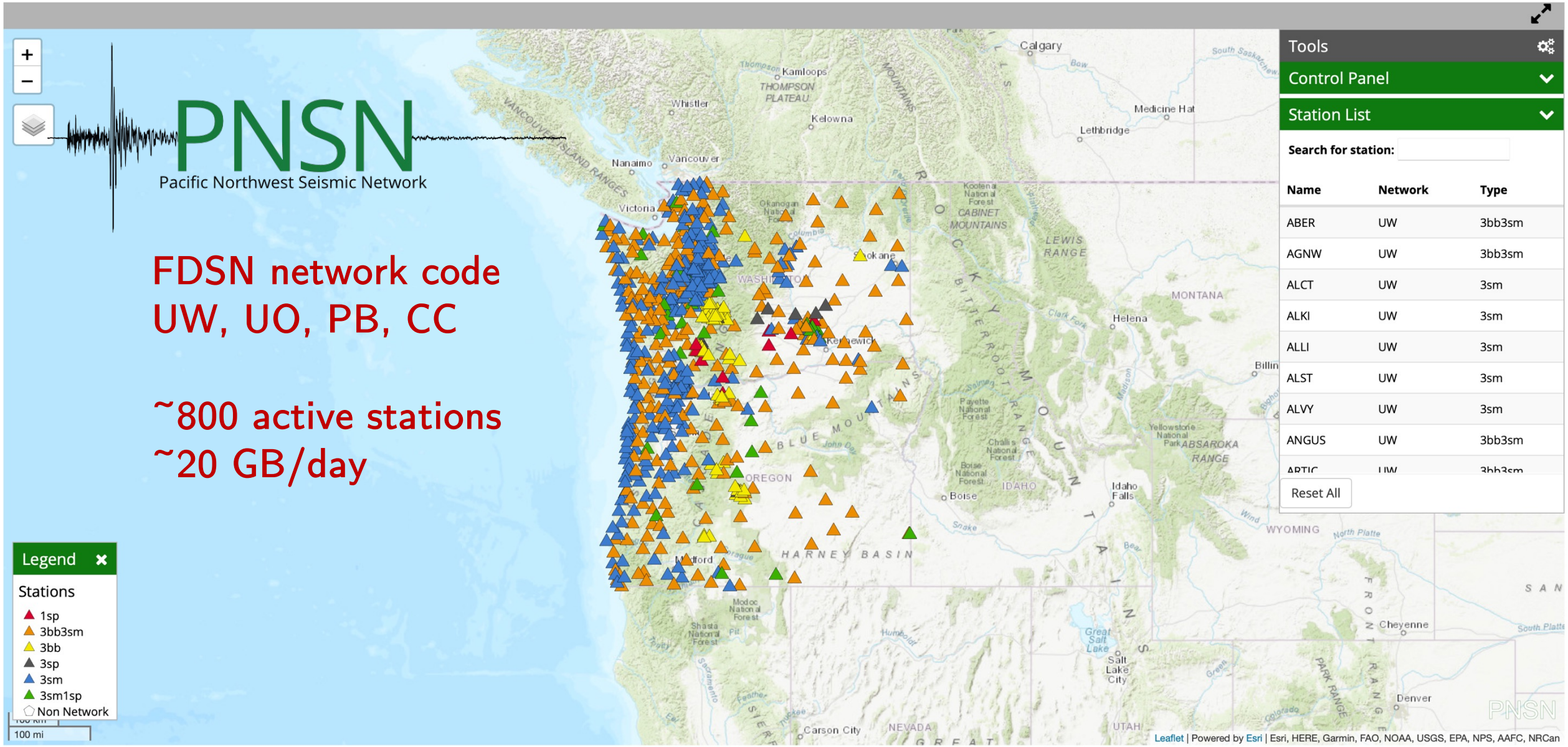
- Regional seismic hazard amplified by the sedimentary basins.



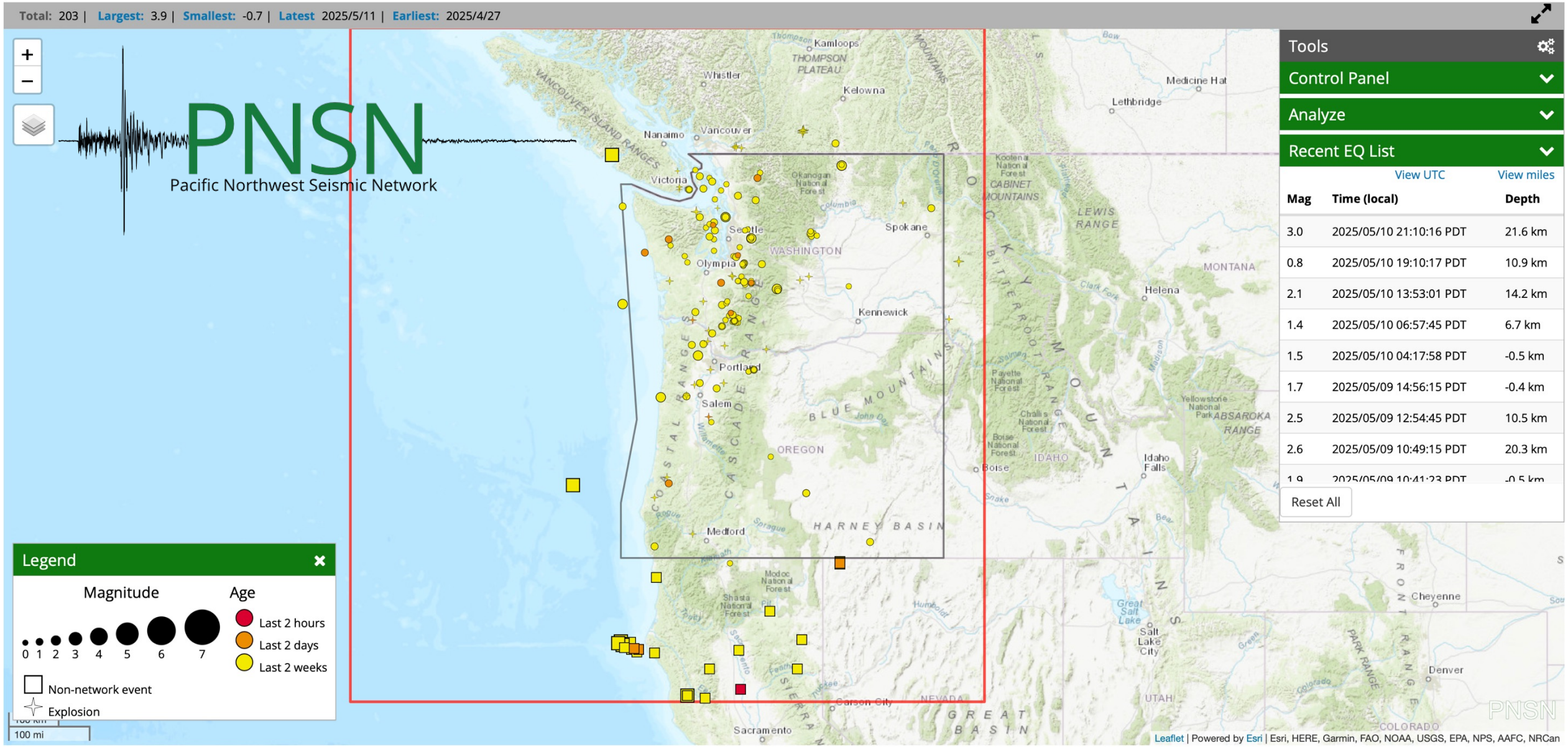
Modified from Walton et al., 2021

- Geohazard associated with landslides and volcanic activities.

Pacific Northwest Seismic Network



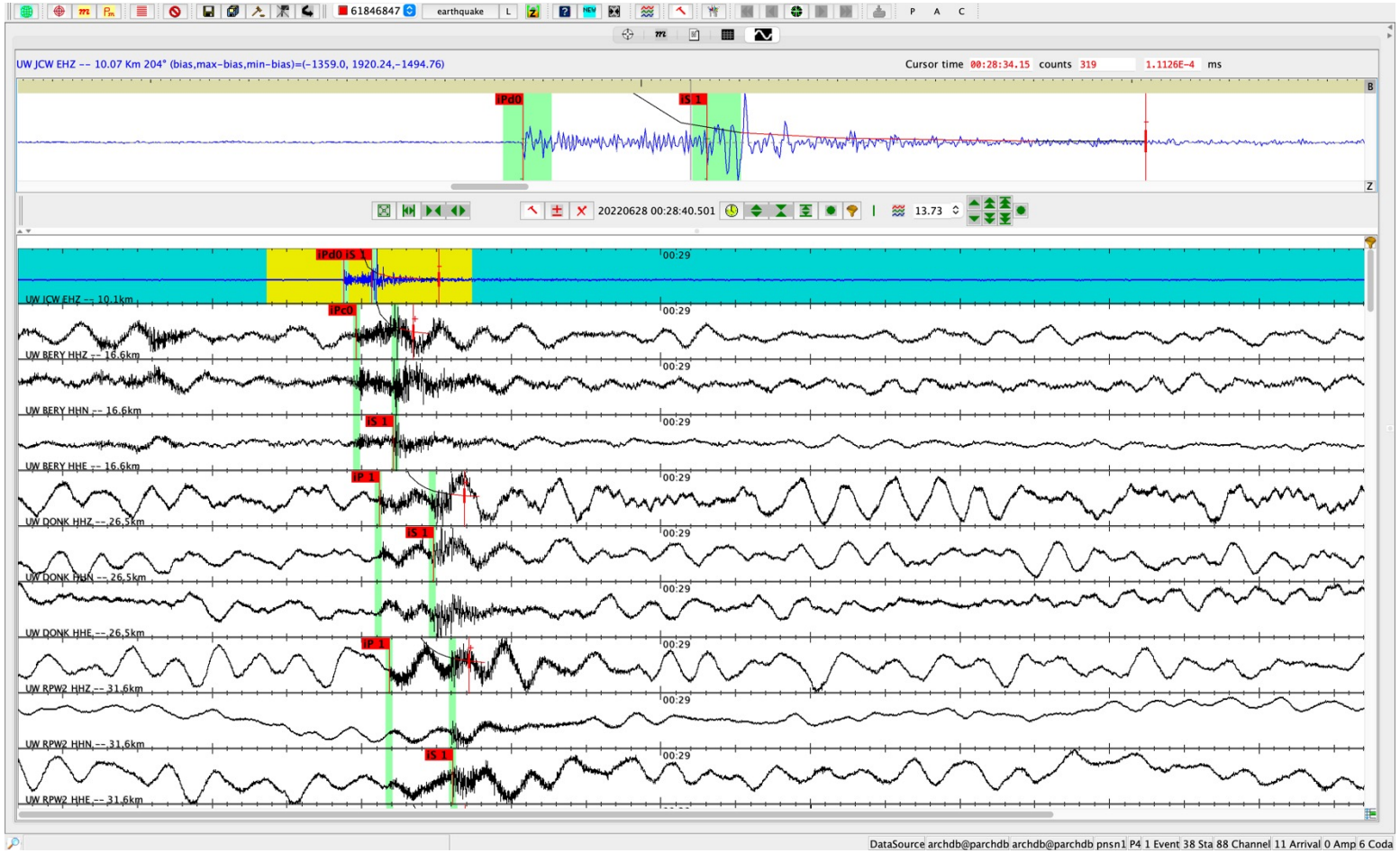
Pacific Northwest Seismic Network (PNSN)



How seismic network process seismic events

PNSN analysts use **Jiggle** to pick phases, locate earthquakes, and calculate magnitude.

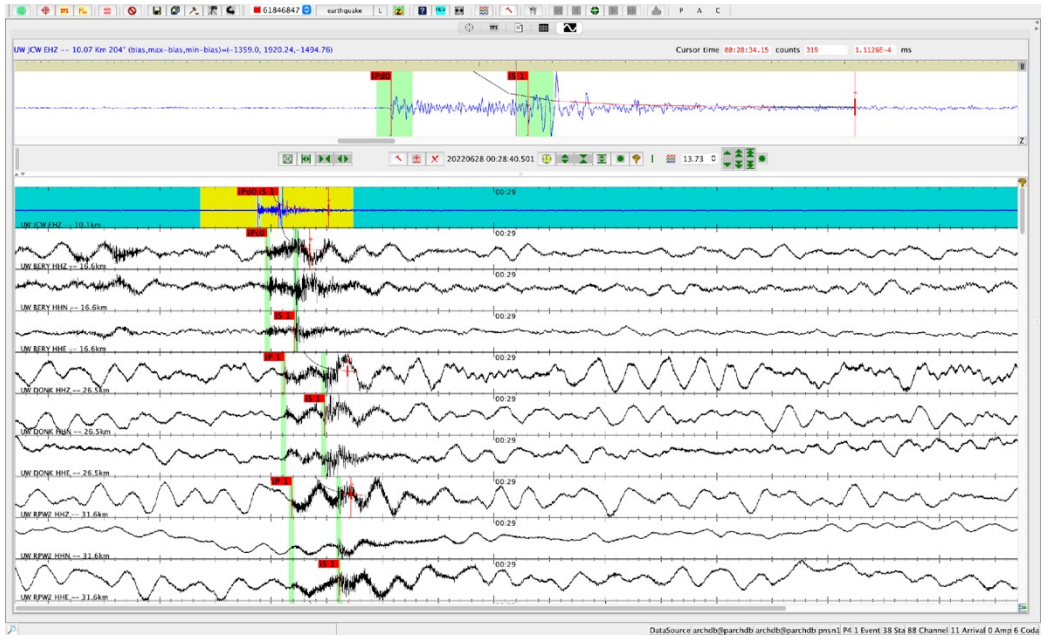
- 0 min — Trigger alarm
- +0.5 min — Duty Seismologist being notified
- +1 min — Jiggle the event
- +3-10 min — Finalize and send to ANSS/ComCat



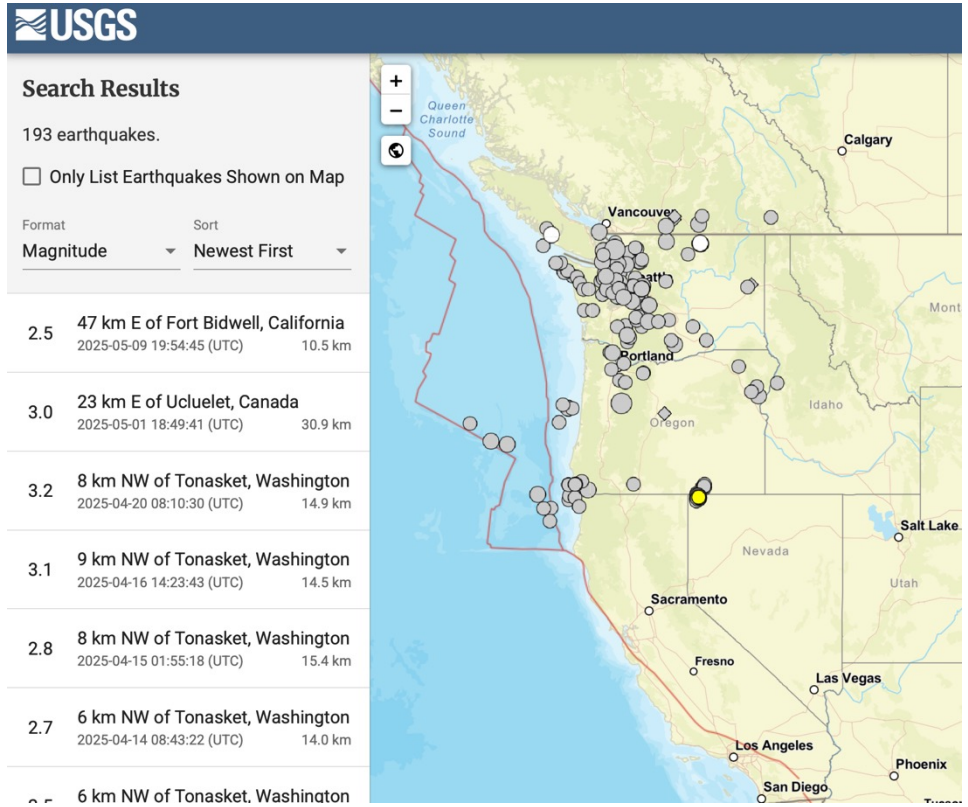
- False triggered events
- Limited stations processed
- Extra training time for new analysts to pick with quality and consistency

How seismic network process seismic events

Data Preparation → Event Detection → Event Classification Phase Picking → Phase Association → Magnitude (Md, Ml, Mw) → Location Focal Mechanism



Analysts picking

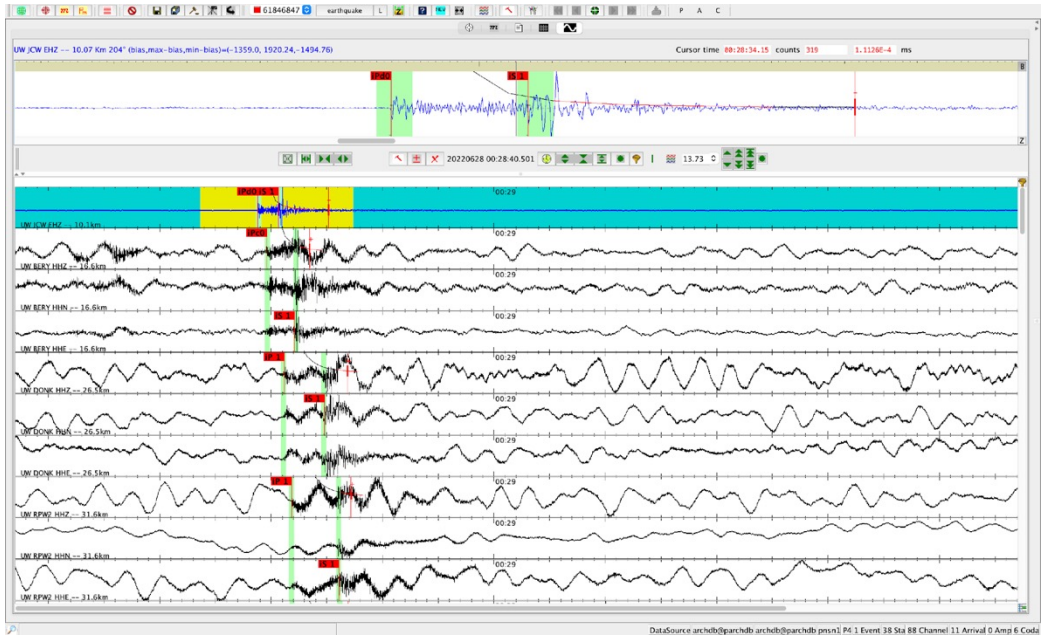


ComCat events

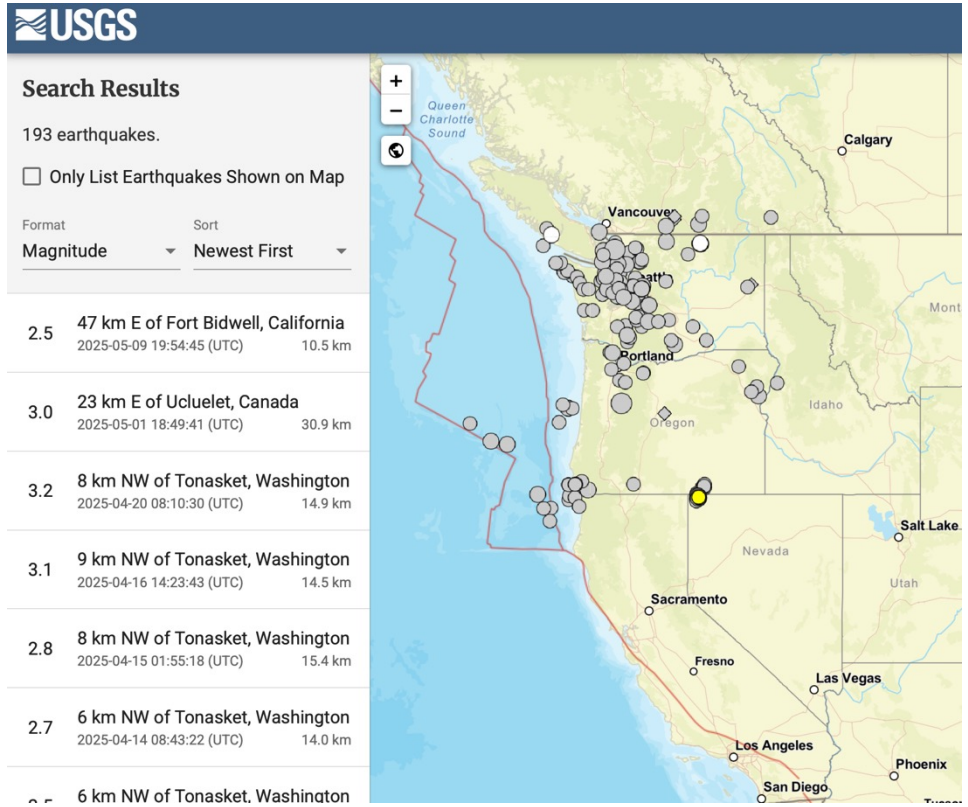
Demo: PNW event picking

How seismic network process seismic events

Data Preparation → Event Detection → Event Classification Phase Picking → Phase Association → Magnitude (Md, Ml, Mw) → Location Focal Mechanism



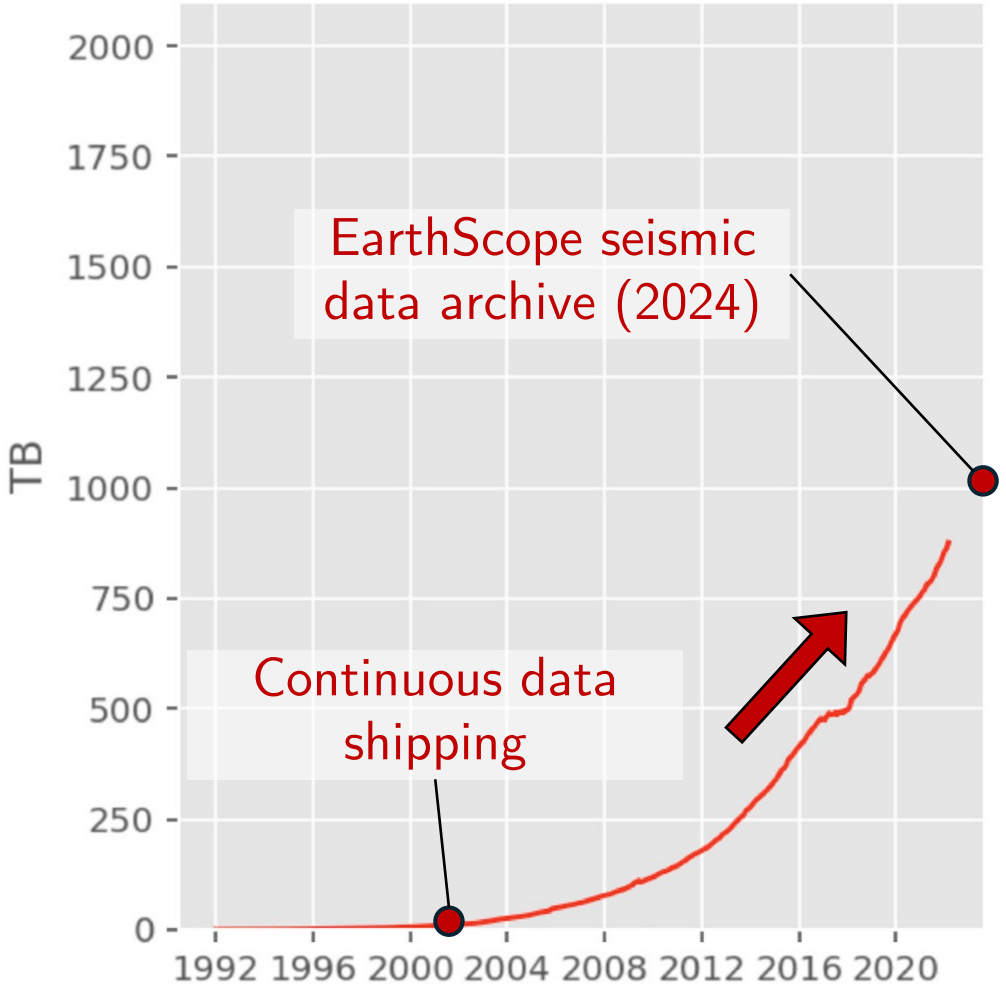
Analysts picking

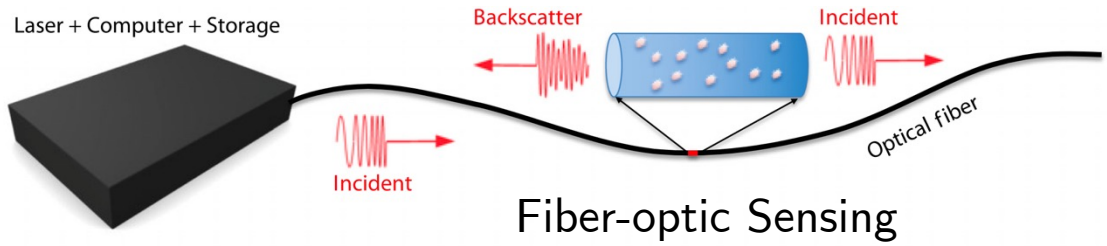


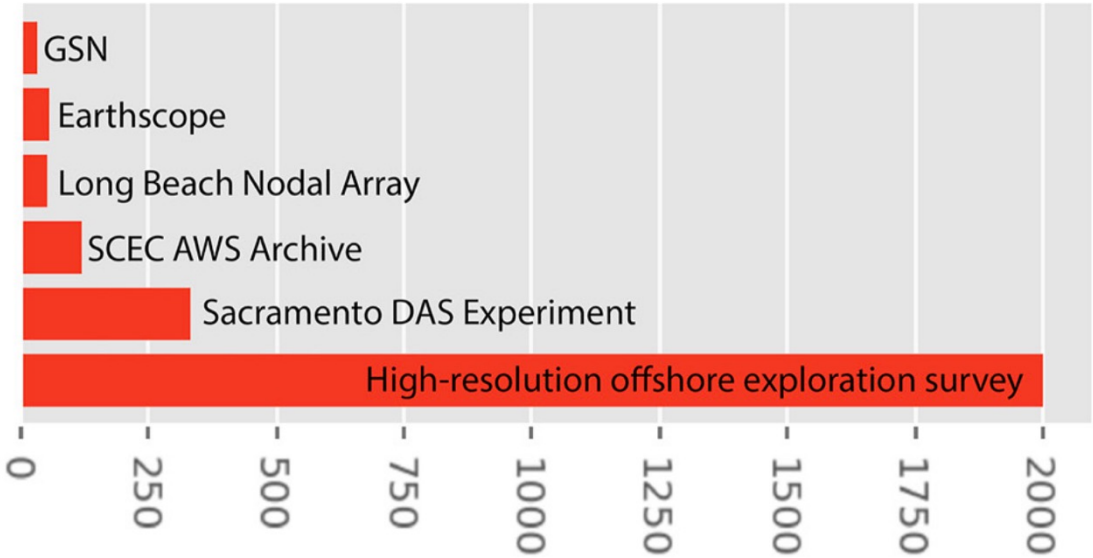
ComCat events

Hands-on: Querying Earthquake Catalog and Waveforms

Seismology is becoming increasingly a big-data discipline...



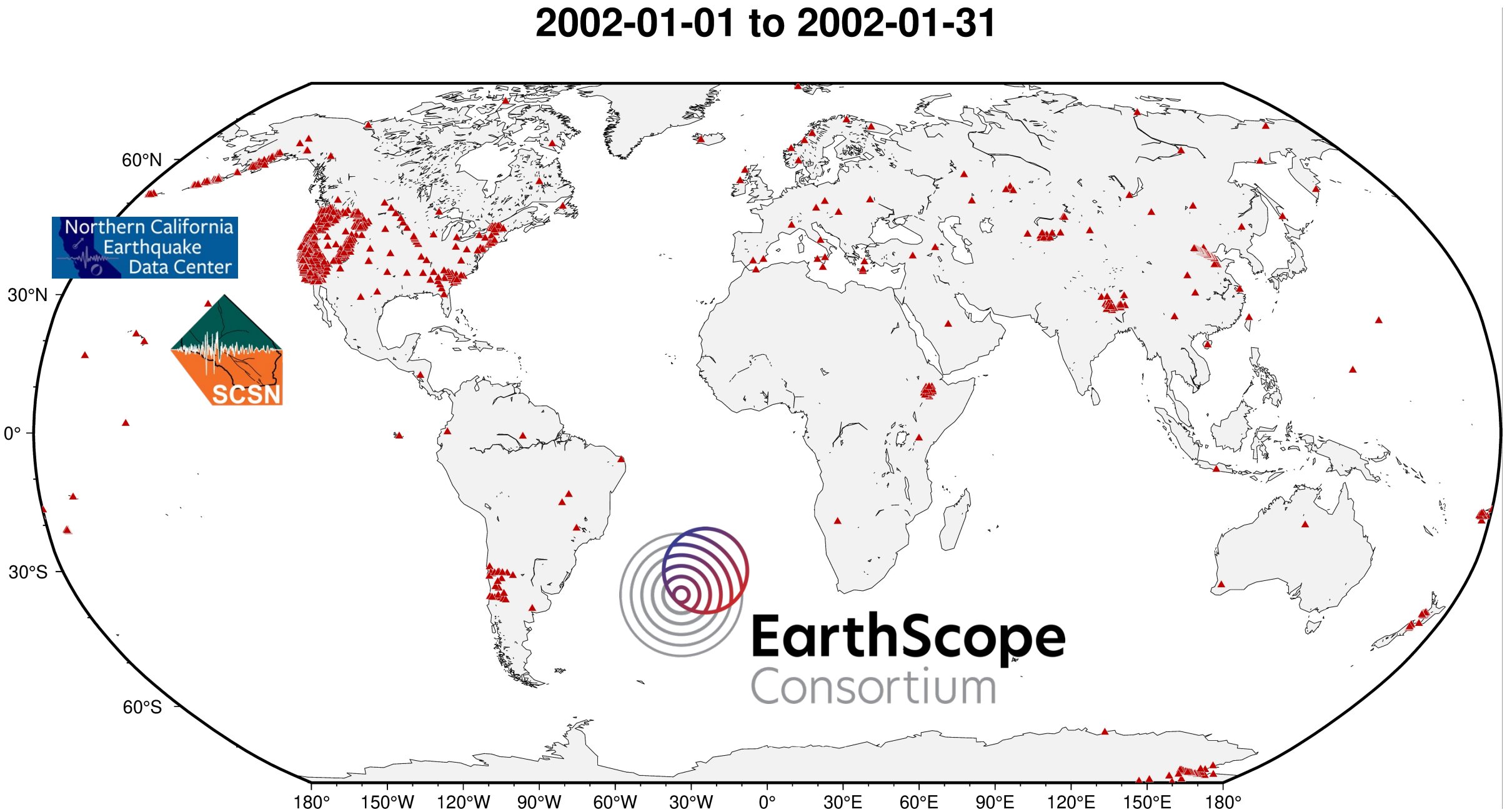




... with the rise of data volume and heterogeneity

Arrowsmith et al., 2021

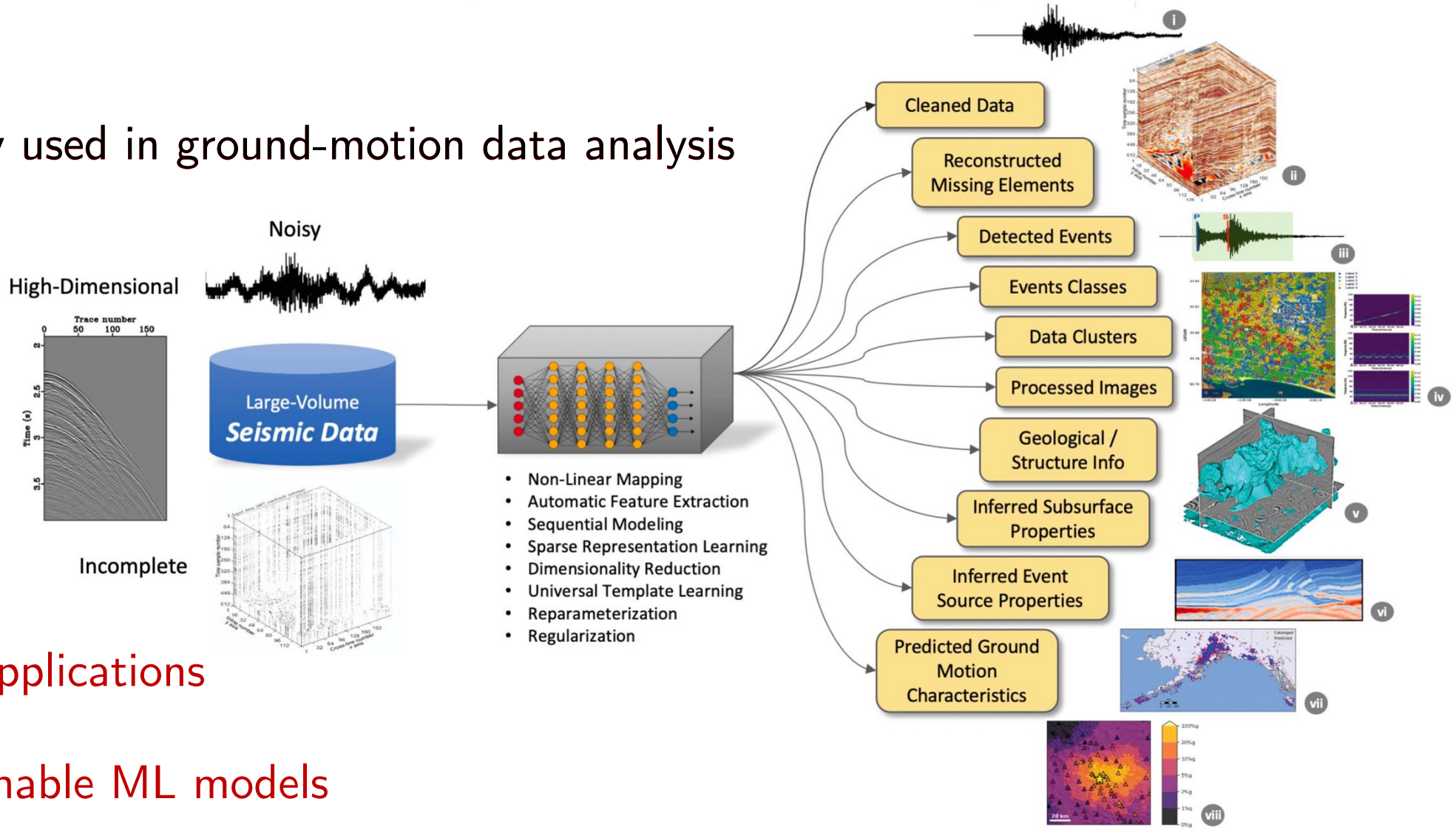
Seismology is becoming increasingly a big-data discipline...



Seismology is becoming increasingly a big-data discipline...

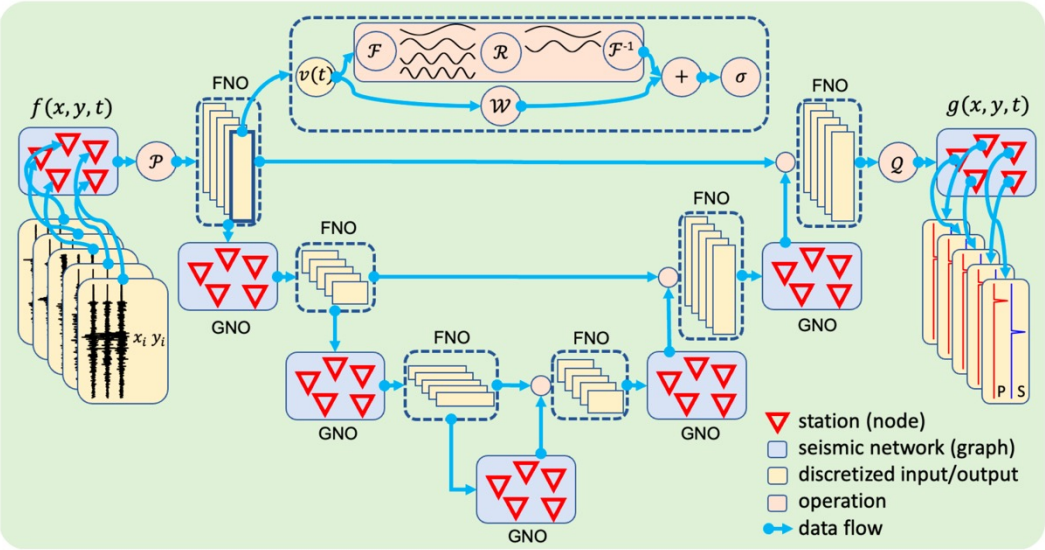
Mousavi and Beroza, 2022

- Widely used in ground-motion data analysis

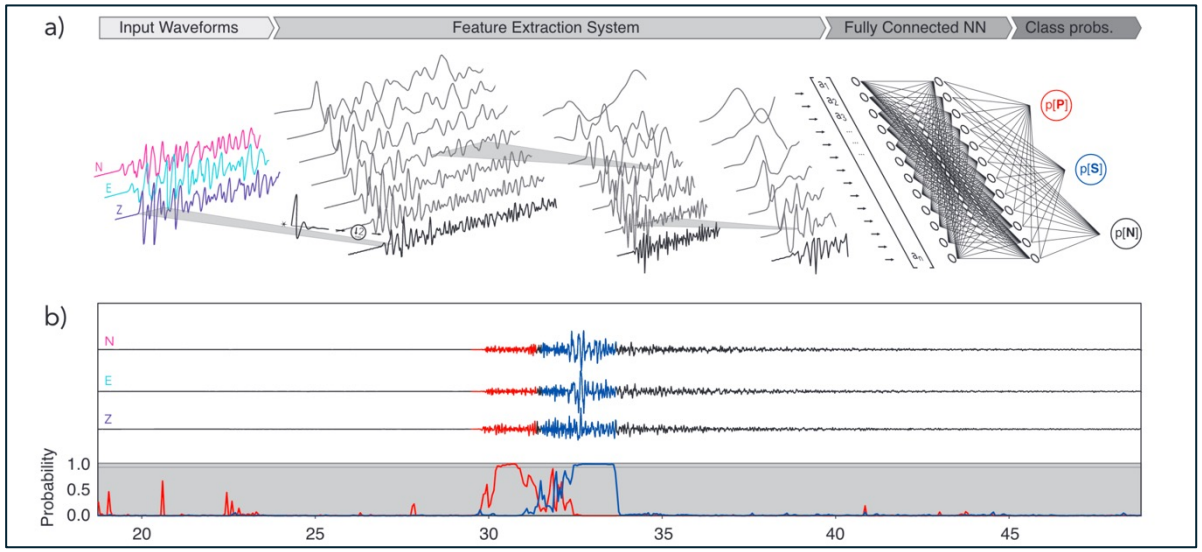


- New applications
- Explainable ML models
- Physics-informed ML models

Machine Learning for earthquake catalog building

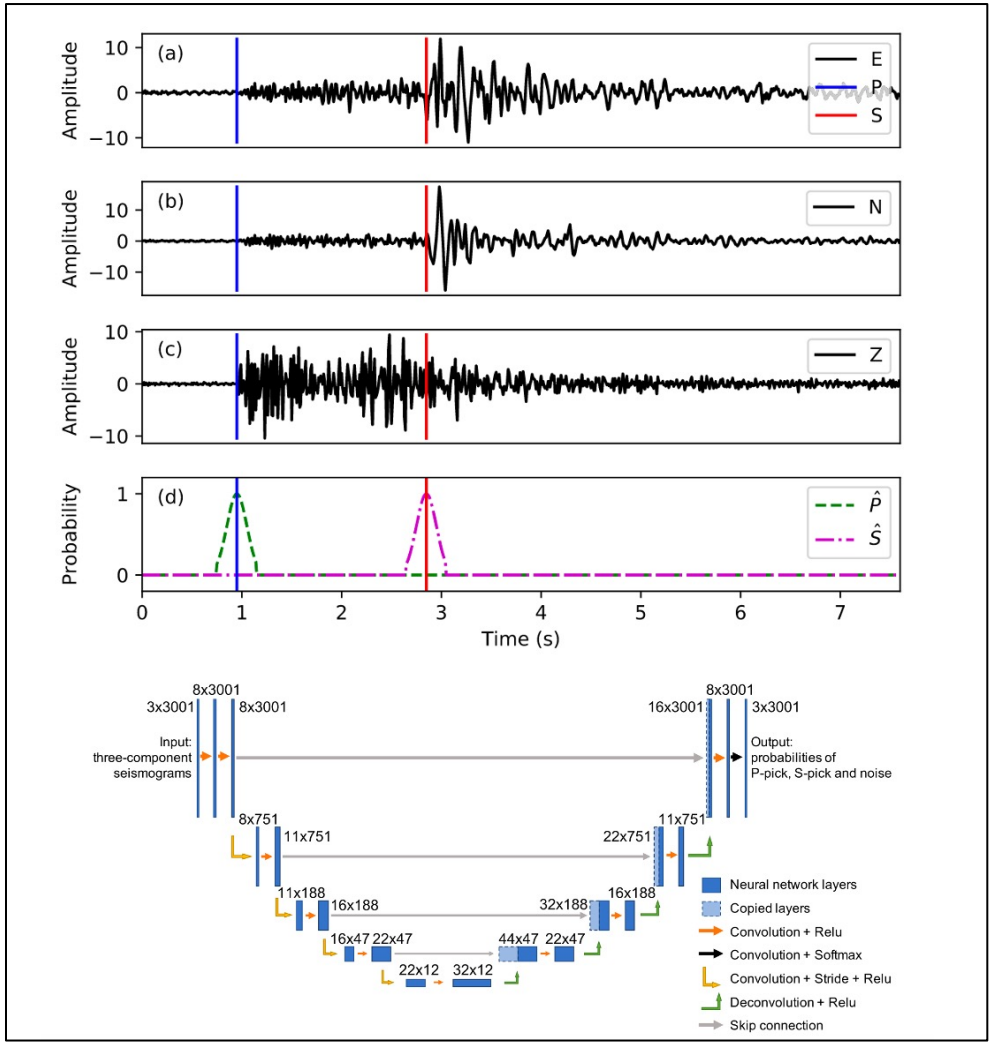


PhaseNO (Sun et al., 2022)



Generalized Phase Detection (GPD, Ross et al., 2018)

PhaseNet (Zhu et al., 2019)



With a laptop, these pickers scan day-long waveform within 10 secs.

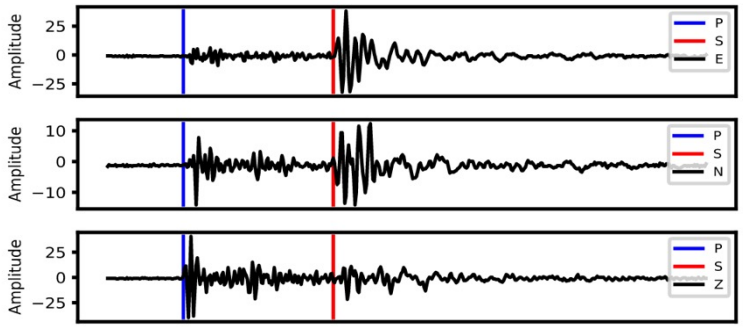
Machine Learning for earthquake catalog building

- ✓ Quick response to new events
- ✓ Less "training" time
- ✓ Stable and consistent results
- ✓ Good scalability and portability

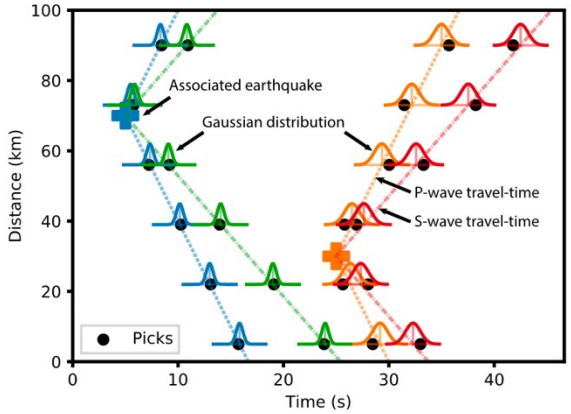
- ❑ Large dataset for model training
- ❑ Computing resource
- ❑ Quality control
- ❑ Operation integration

Data Preparation → Event Detection → Event Classification Phase Picking → Phase Association → Magnitude (Md, Ml, Mw) → Location Focal Mechanism

- GPD (Ross et al., 2018)
- PhaseNet (Zhu et al., 2018)
- EqTransformer (Mousavi et al., 2020)
- PhaseNO (Sun et al., 2022)



- PhaseLink (Ross et al., 2019)
- GaMMA (Zhu et al, 2021)
- Neuma (Ross et al., 2023)
- PyOcto (Münchmeyer, 2023)



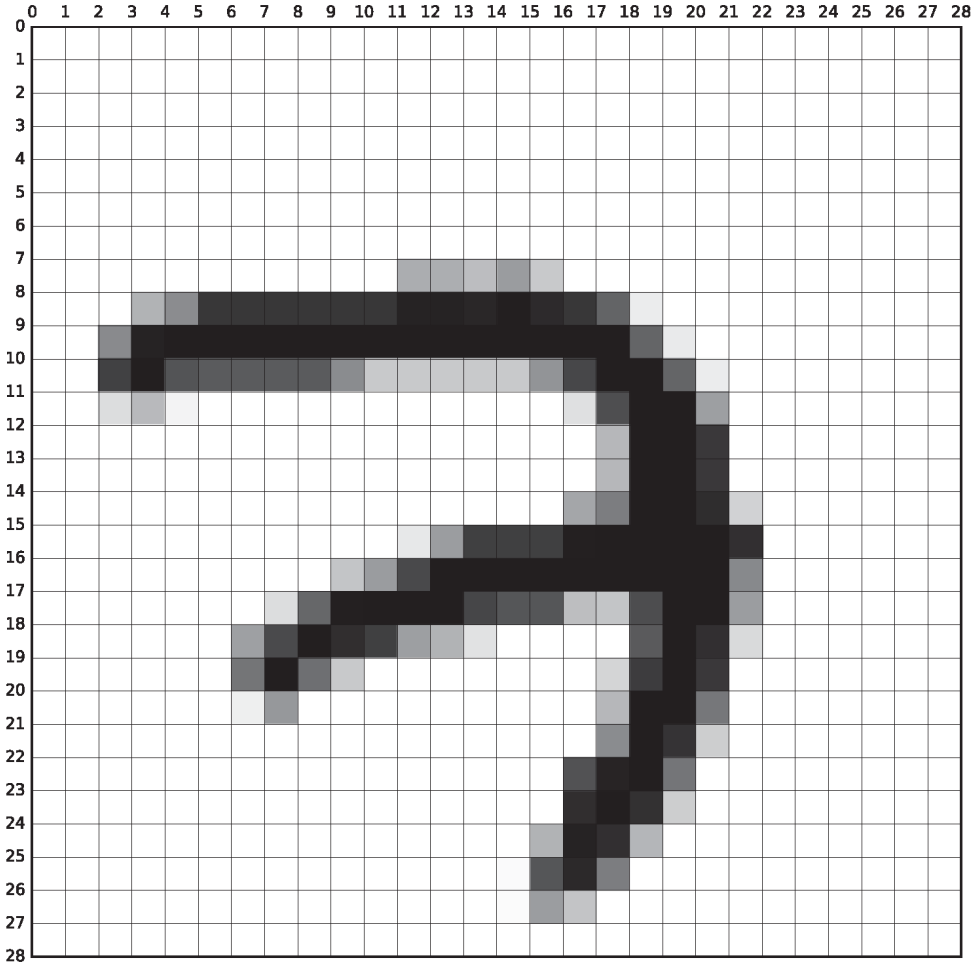
Workflow

- easyQuake (Walter et al., 2021)
- SeisBench (Woollam et al., 2022)
- QuakeFlow (Zhu et al., 2022)
- Loc-Flow (Zhang et al., 2022)

Curated Dataset for Seismology

The MNIST dataset for computer vision

Dataset size: ~11 MB



(**a**) MNIST sample belonging to the digit '7'.

(**b**) 100 samples from the MNIST training set.

Curated seismic datasets

Received August 16, 2019, accepted October 12, 2019, date of publication October 16, 2019, date of current version December 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2947848

STanford EArthquake Dataset (STEAD): A Global Data Set of Seismic Signals for AI

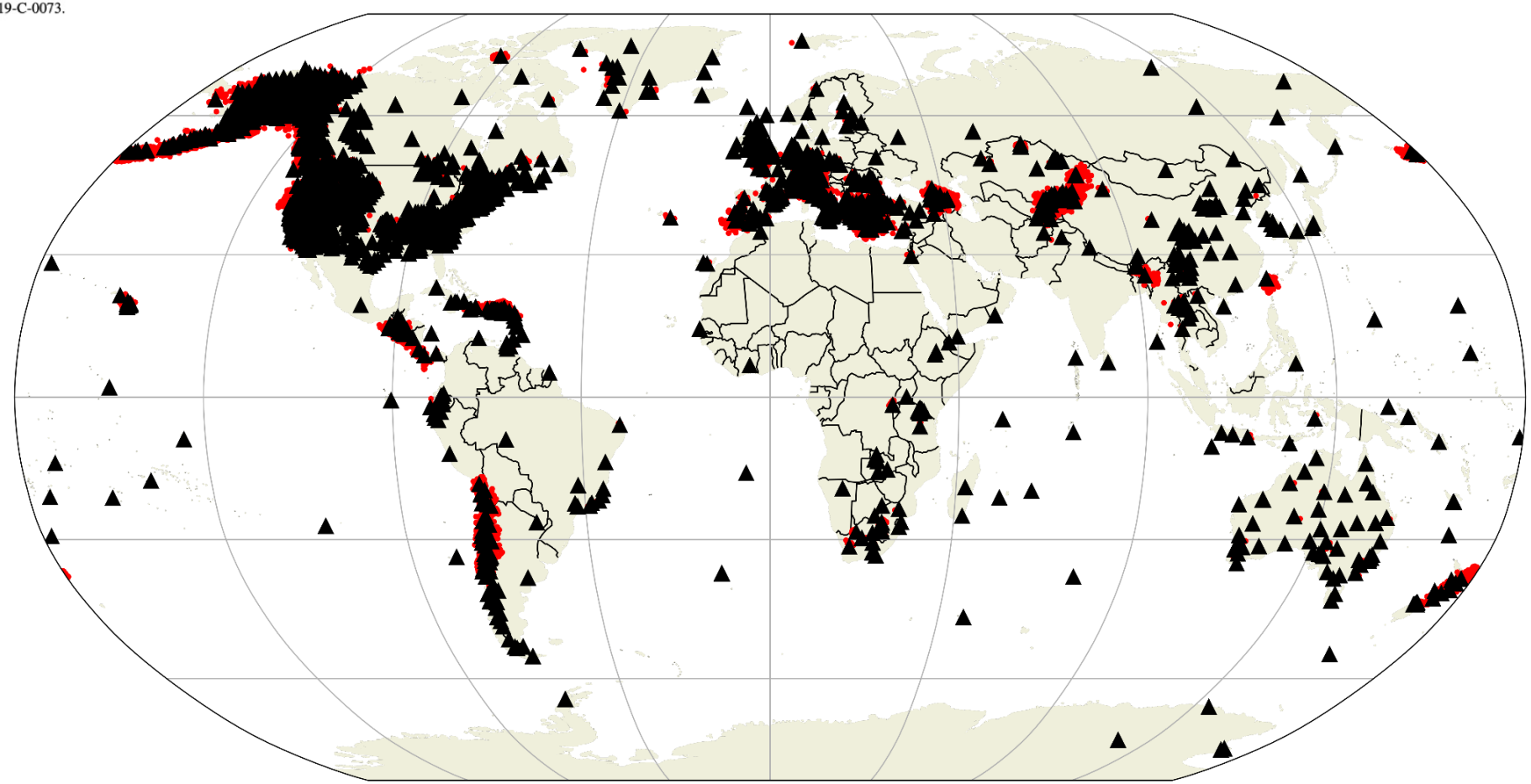
S. MOSTAFA MOUSAVI, YIXIAO SHENG, WEIQIANG ZHU, AND GREGORY C. BEROZA

Geophysics Department, Stanford University, Stanford, CA 94305-2215, USA

Corresponding author: S. Mostafa Mousavi (mmousavi@stanford.edu)

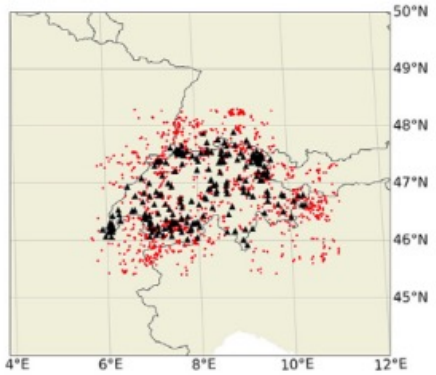
The work of S. M. Mousavi was partially supported by Stanford Center for Induced and Triggered Seismicity (SCITS). The work of G. C. Beroza was supported by AFRL under the contract number FA9453-19-C-0073.

- 1.2 million waveforms and attributes
- Waveforms within 350 km of the origins

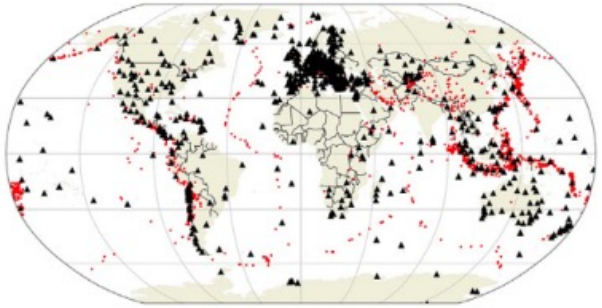


Curated seismic datasets

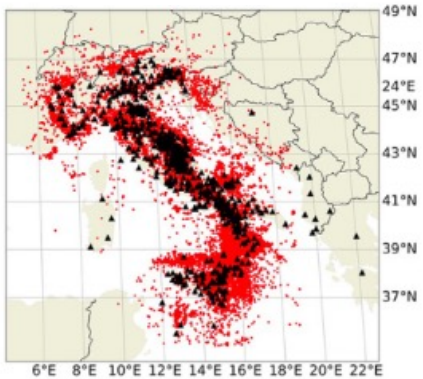
ETHZ



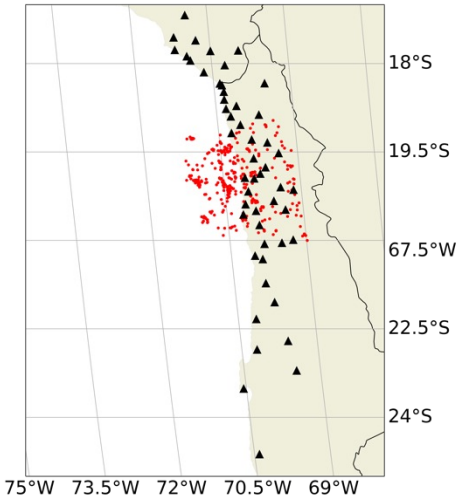
GEOFON



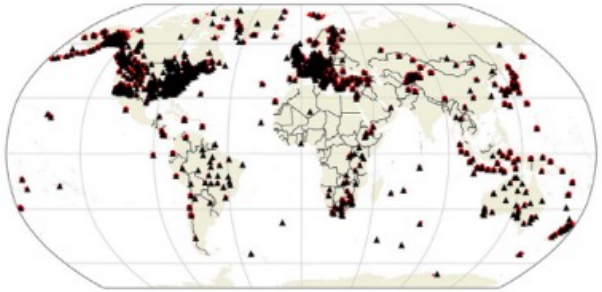
INSTANCE



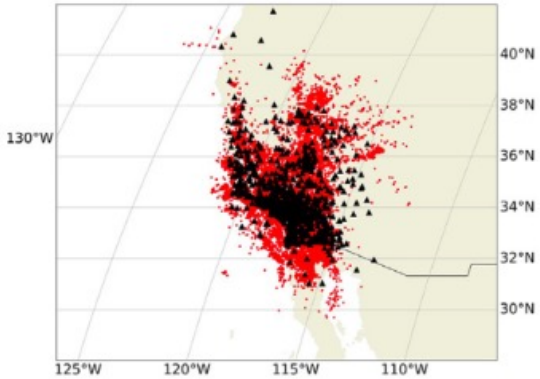
IQUQUE



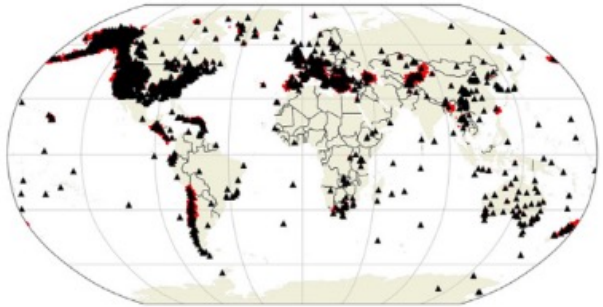
LenDB



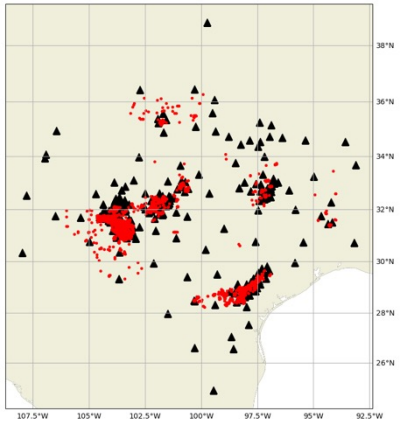
SCEDC



STEAD



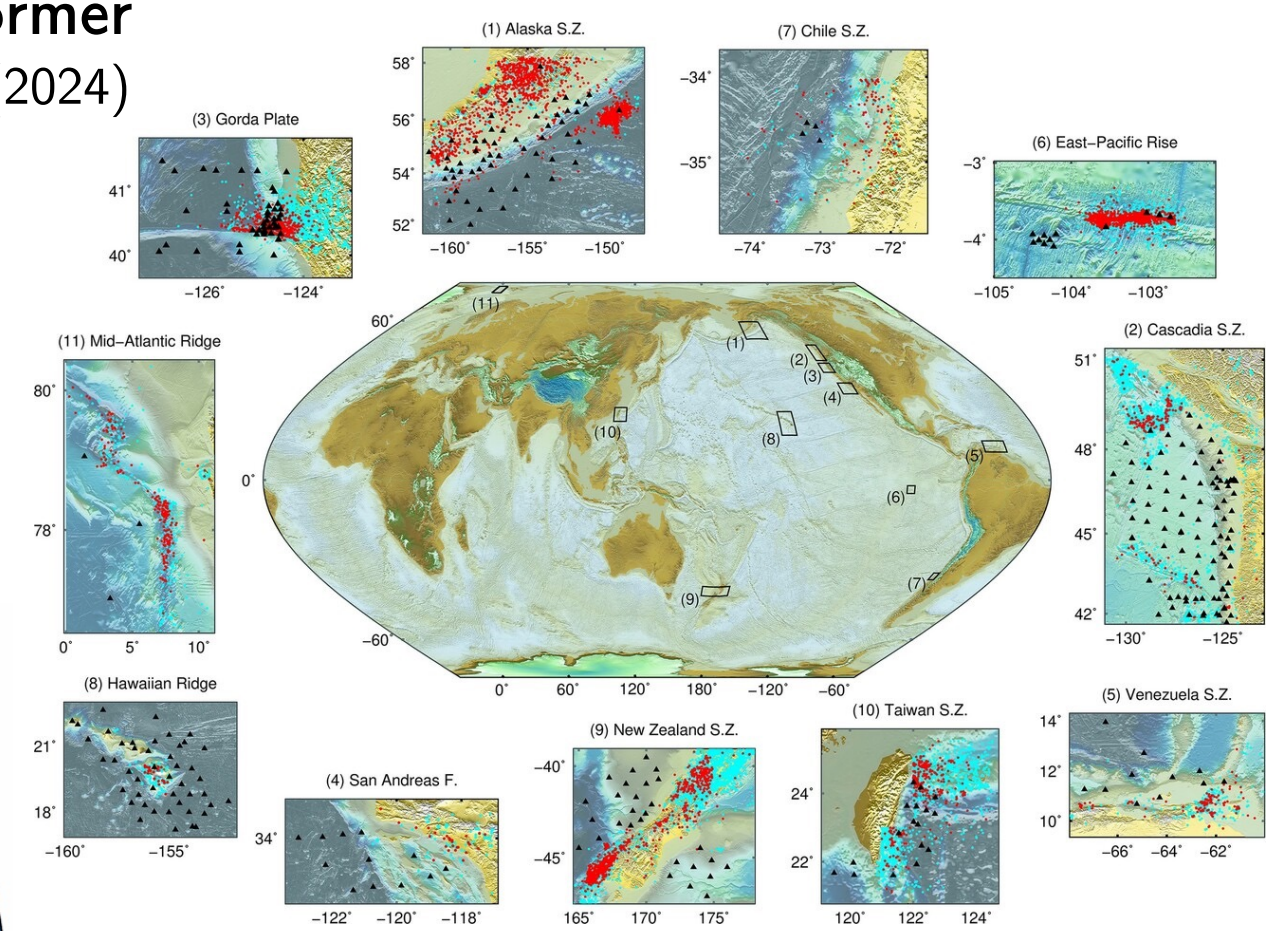
TEXED



Curated seismic datasets

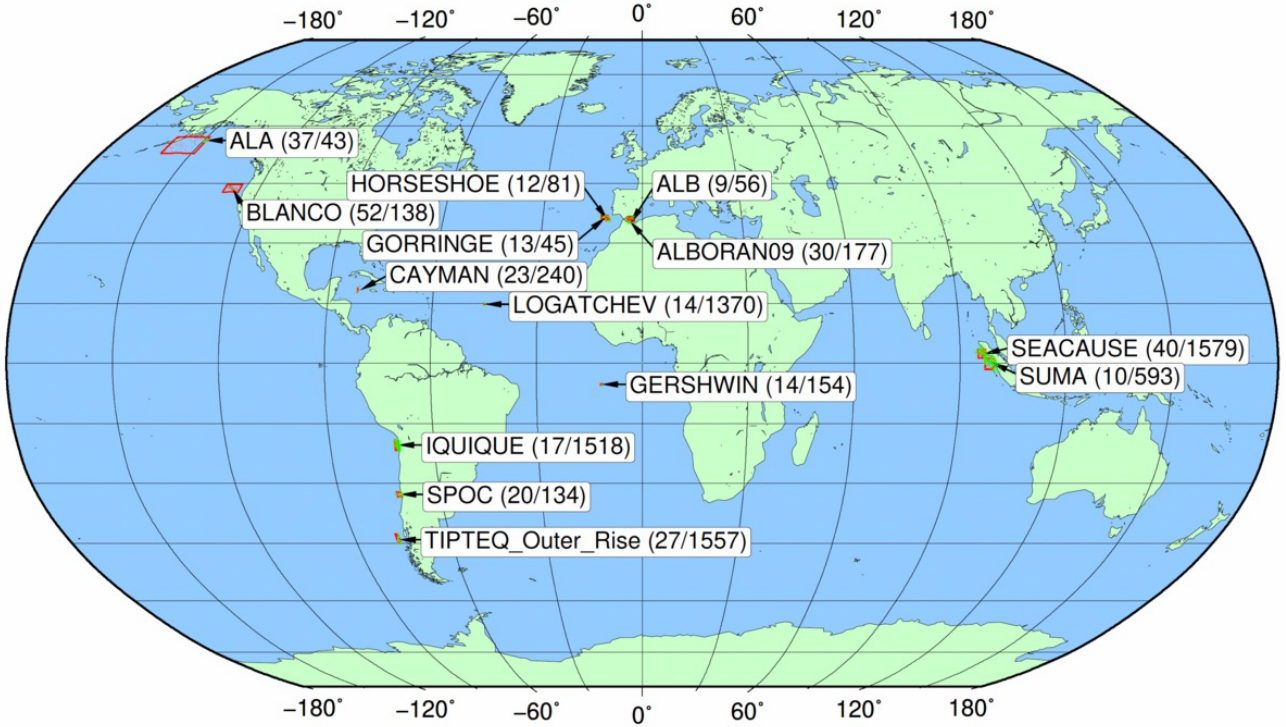
OBSTransformer

Niksejel et al. (2024)



PickBlue

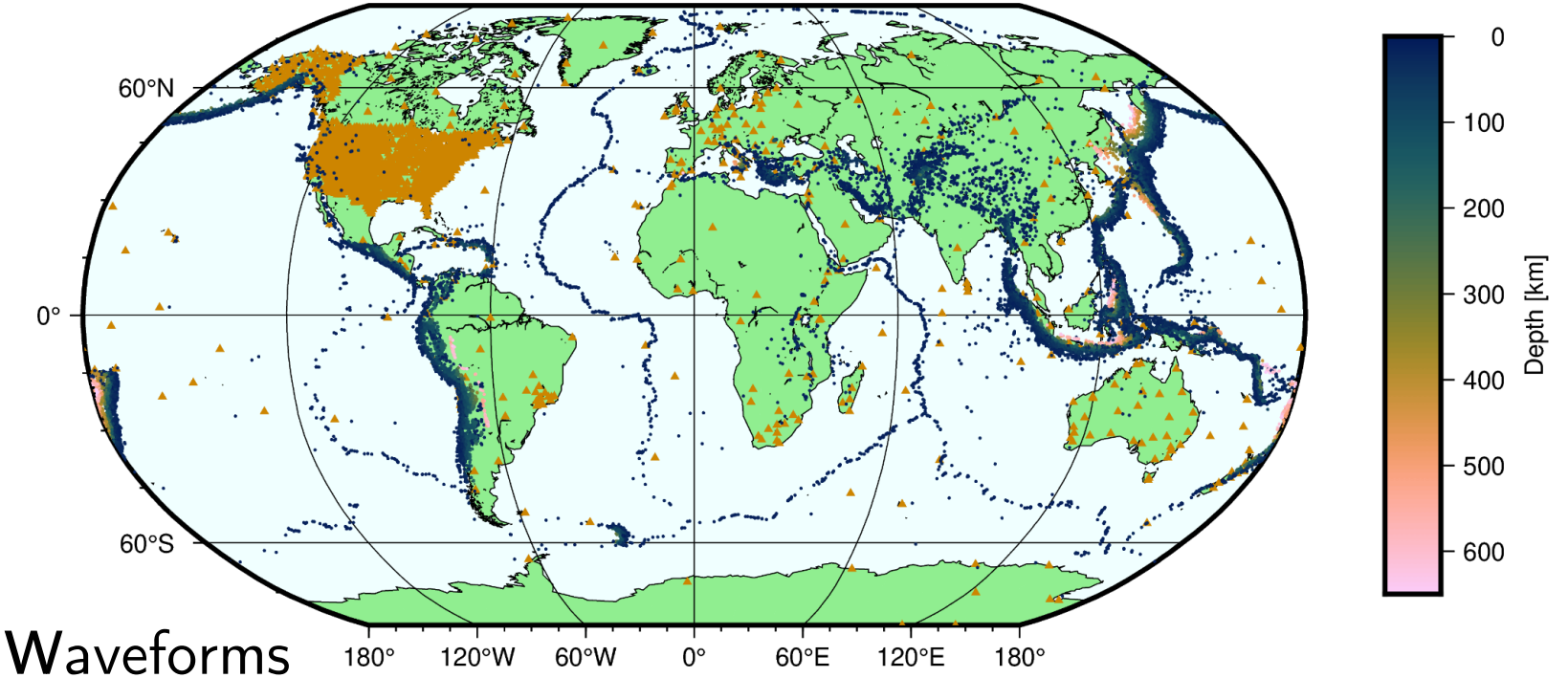
Bornstein et al. (2024)



Curated seismic datasets

ISC-EHB dataset

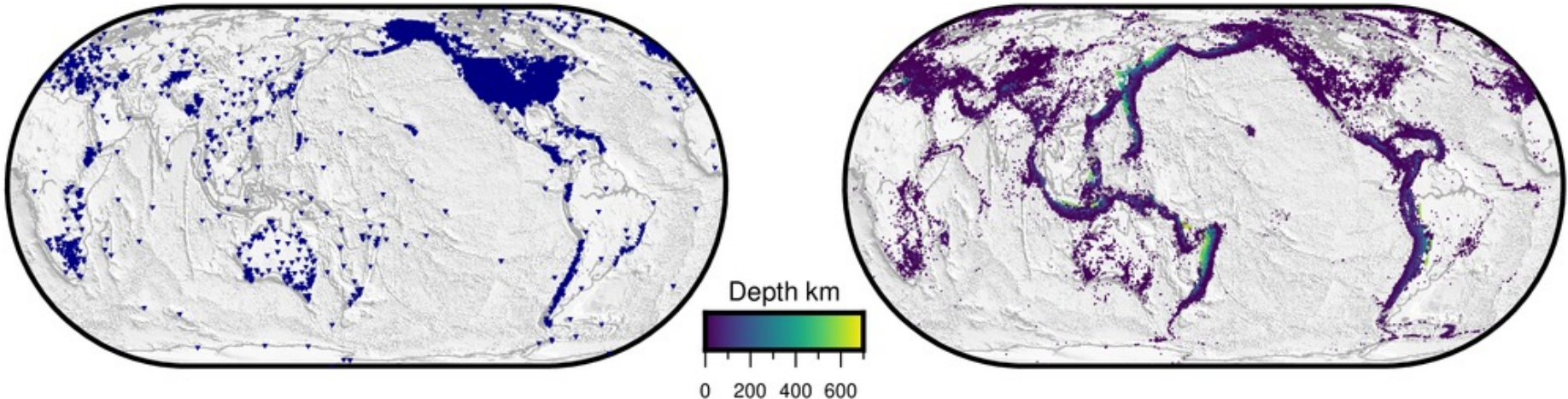
Münchmeyer et al. (2023)



Curated Regional Earthquake Waveforms

CREW: P, Pn, Pg, S, Sn, Sg

Aguilar-Suarez and Beroza (2024)

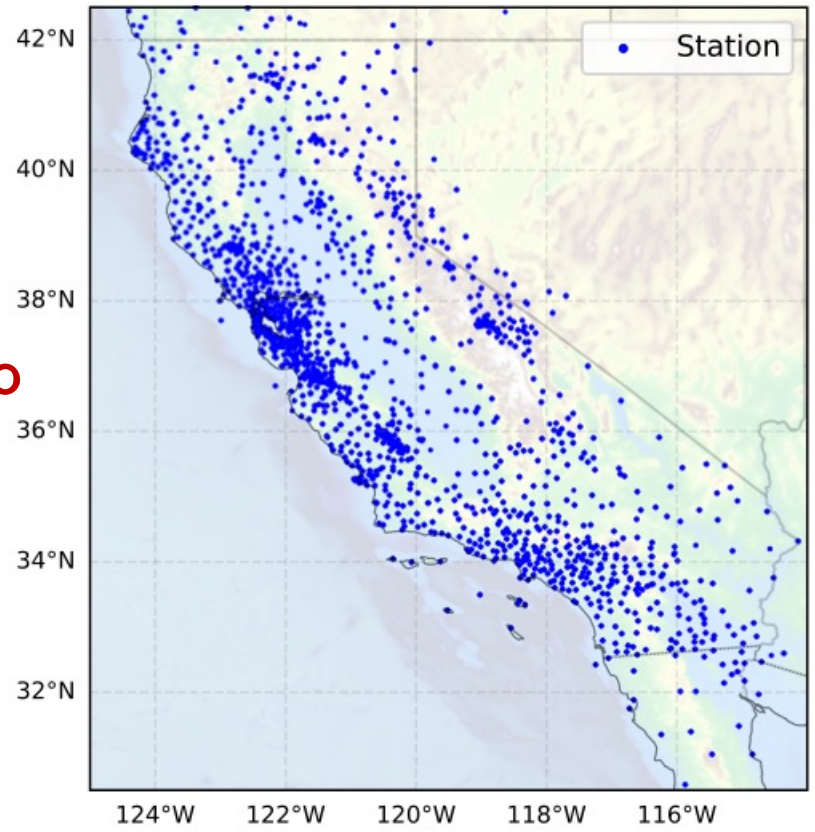


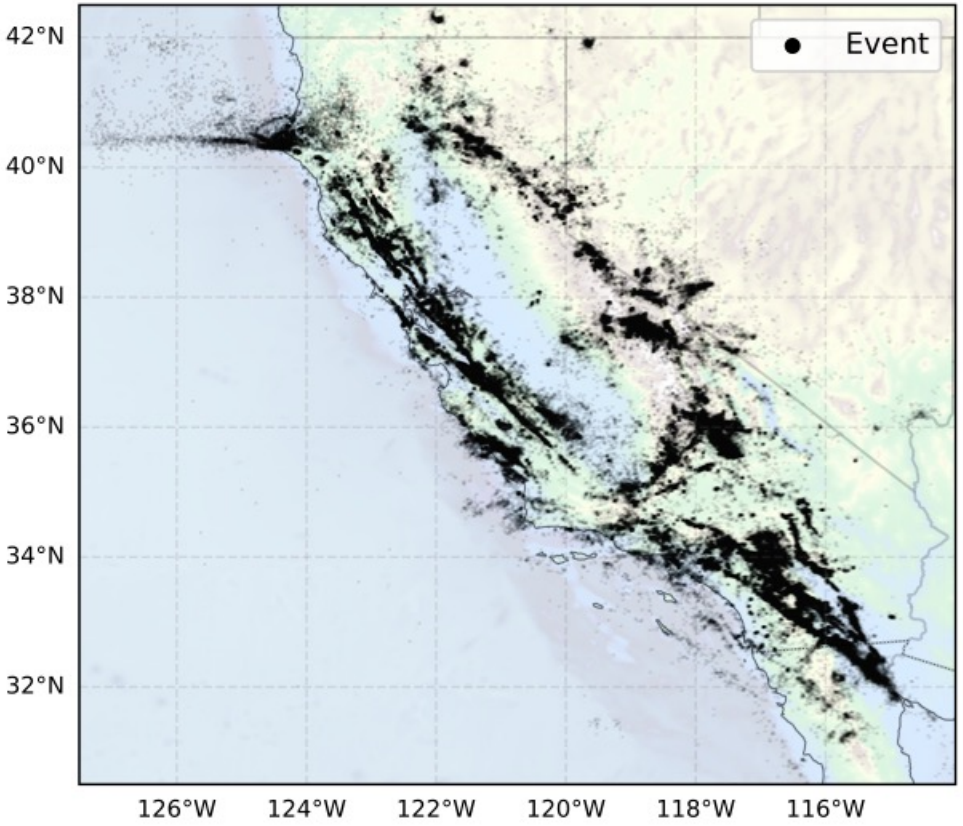
Curated seismic datasets

... and most recently

California Earthquake Event Dataset for ML & Cloud Computing
CEED (Zhu et al., 2025)

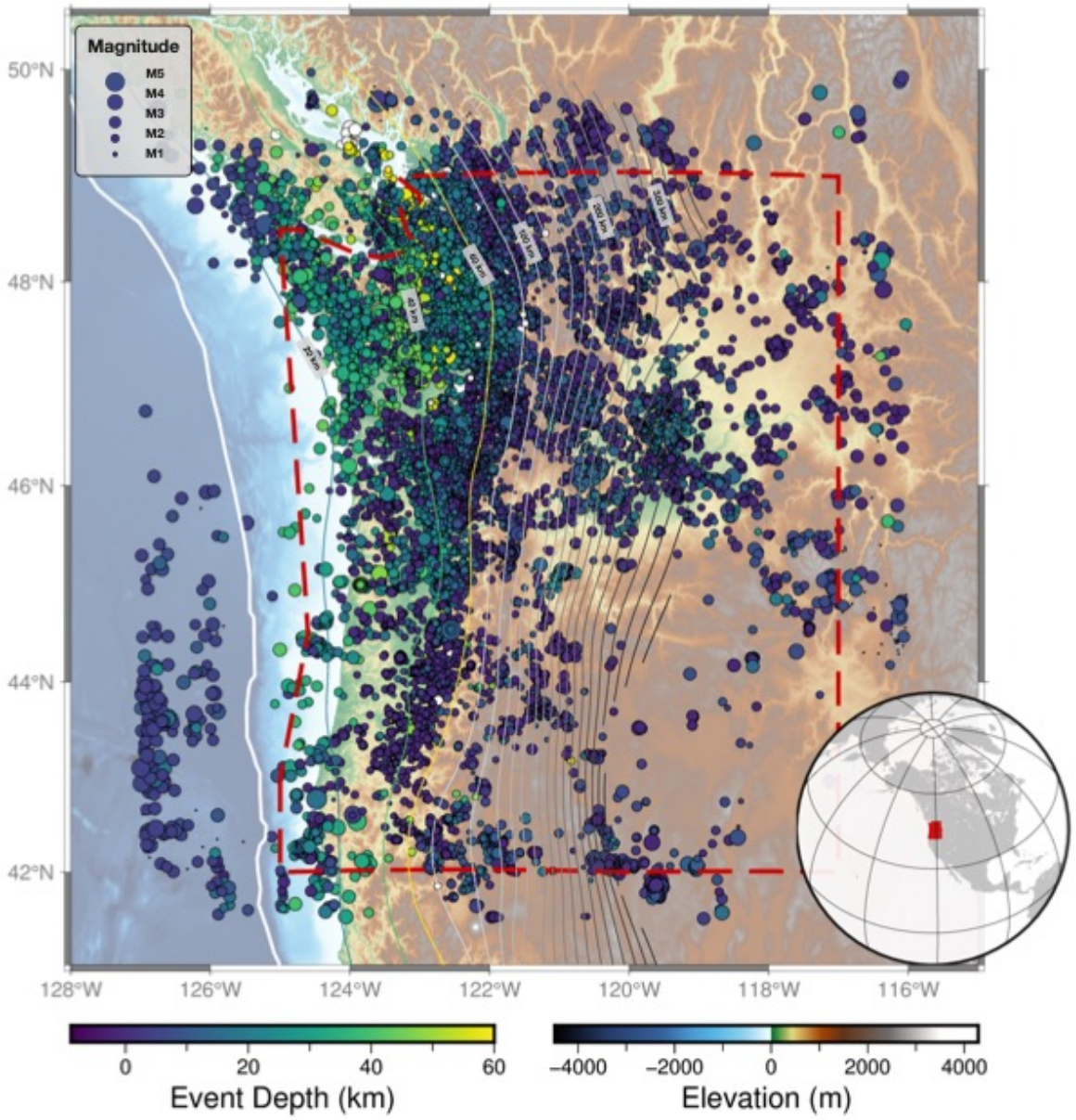
- 4.1 million waveforms
- Elevating datasets into TB scale (~ 1 TB)



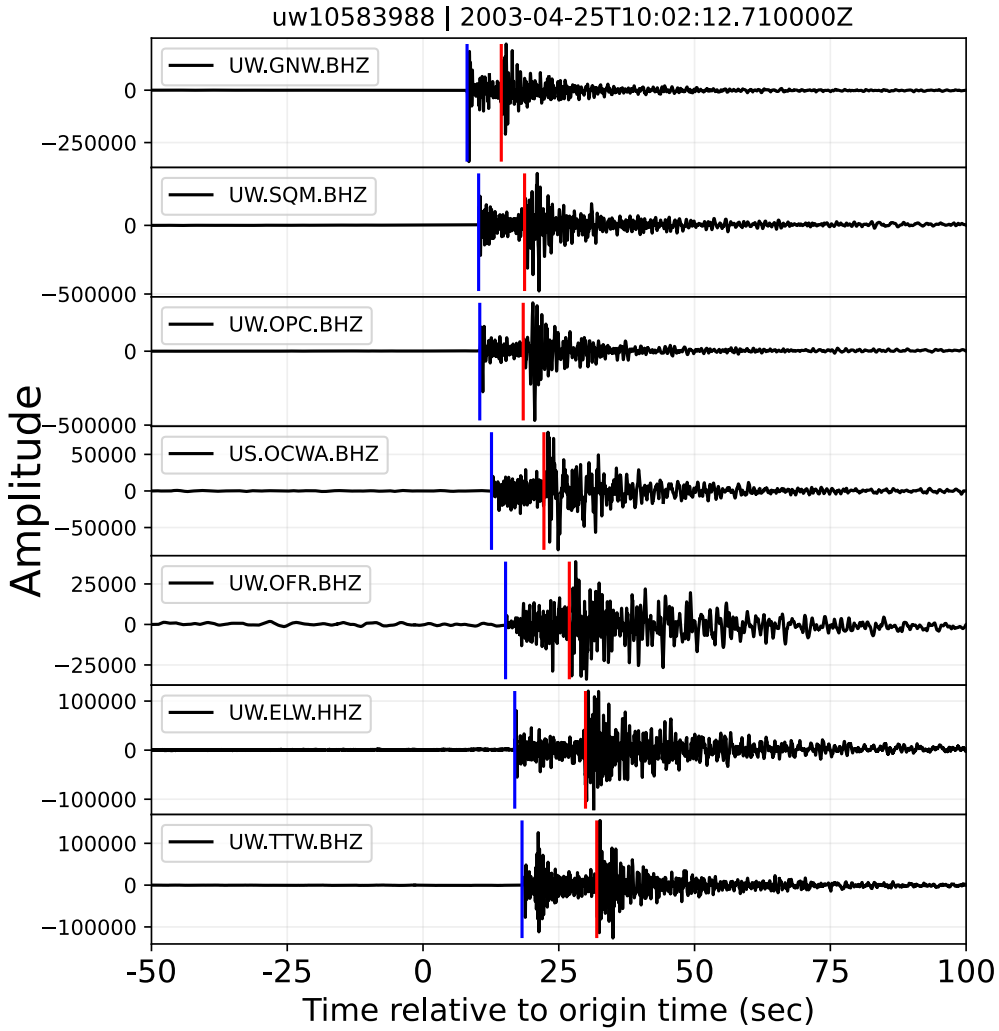


How do researchers easily access these datasets?

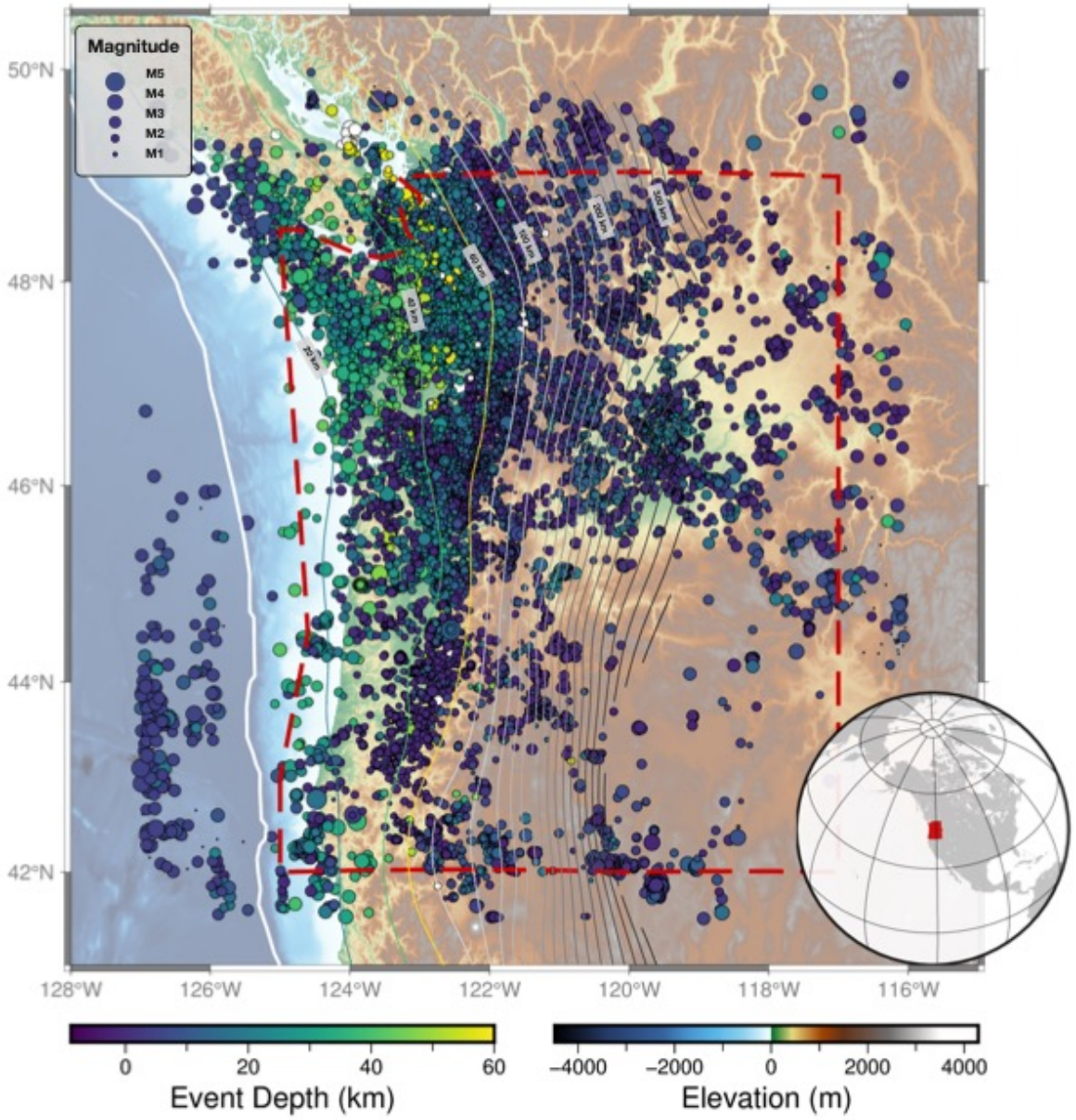
Curated PNW seismic datasets



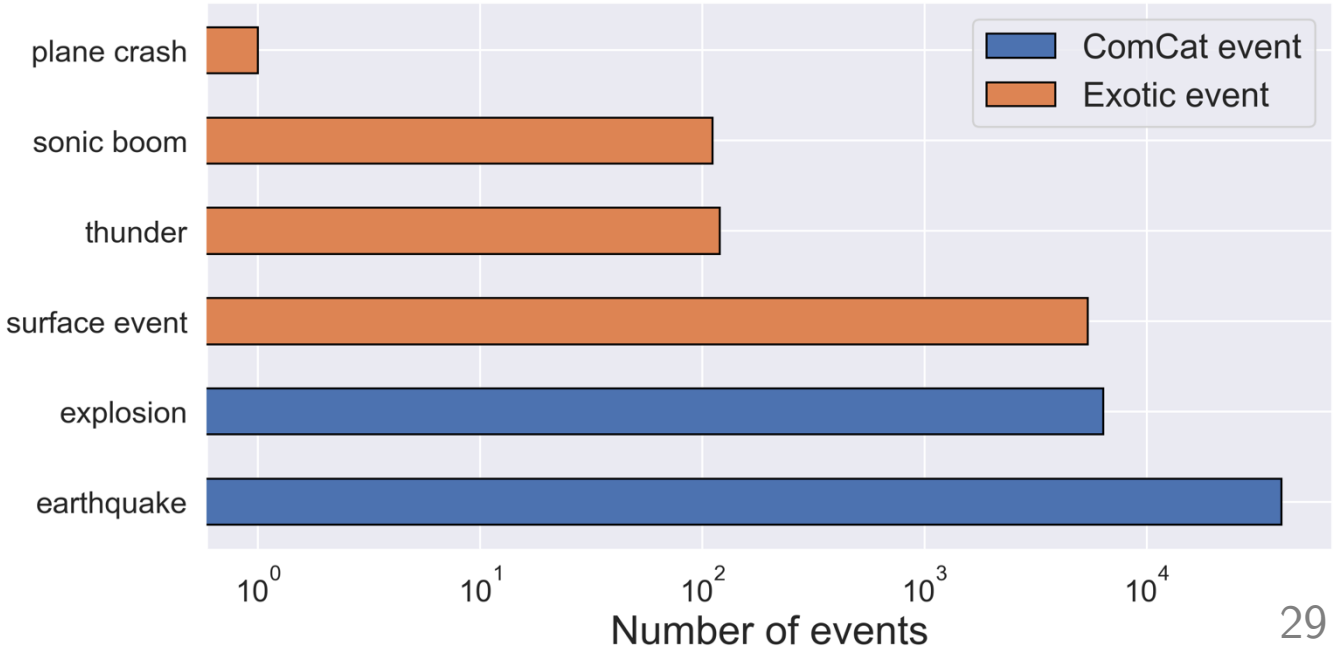
- 44k earthquake and explosion events, 5.6k exotic events, 51k noise waveform
- 150/180-second window length



Curated PNW seismic datasets

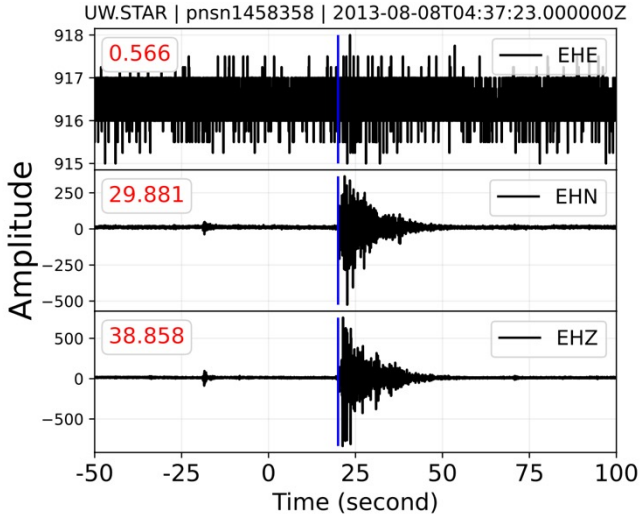


- 44k earthquake and explosion events, 5.6k exotic events, 51k noise waveform
- 150/180-second window length
- Origin and phase information
- Machine readable format (HDF5+CSV)
- **Compatible with SeisBench ecosystem**

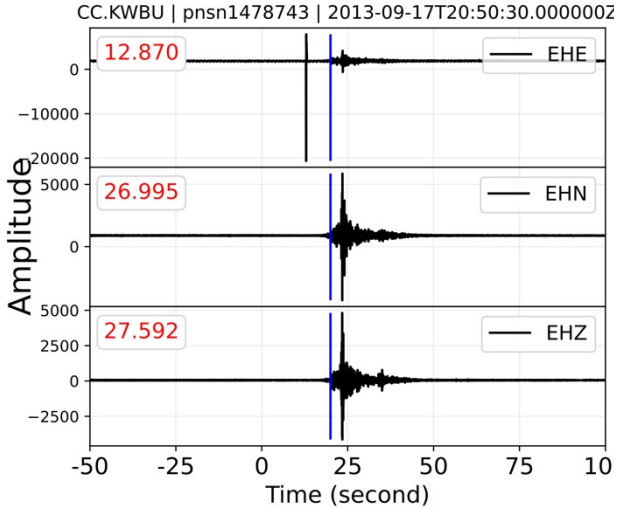


Curated datasets are not perfect

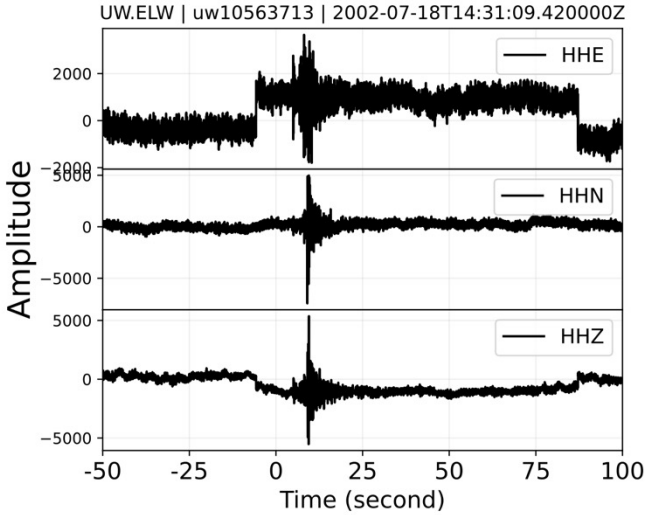
Missing channel

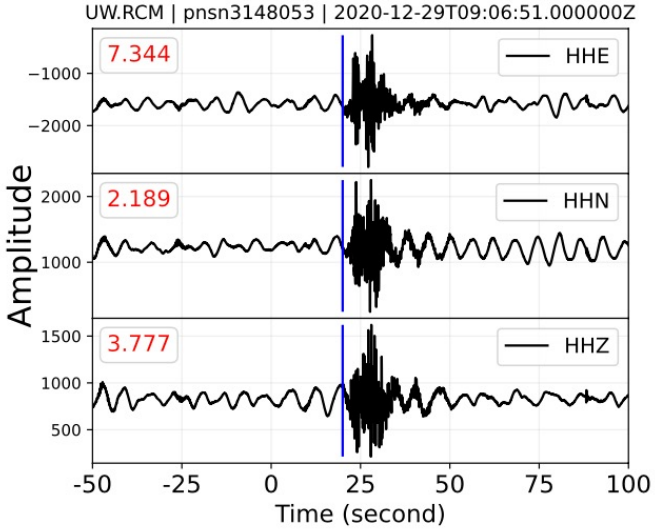


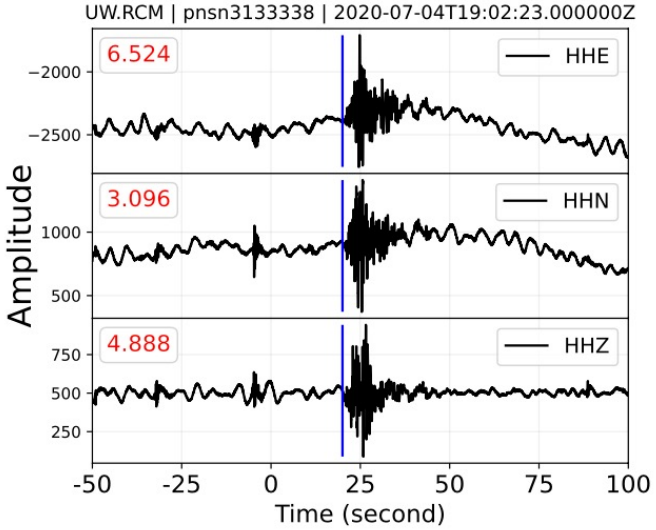
Glitch

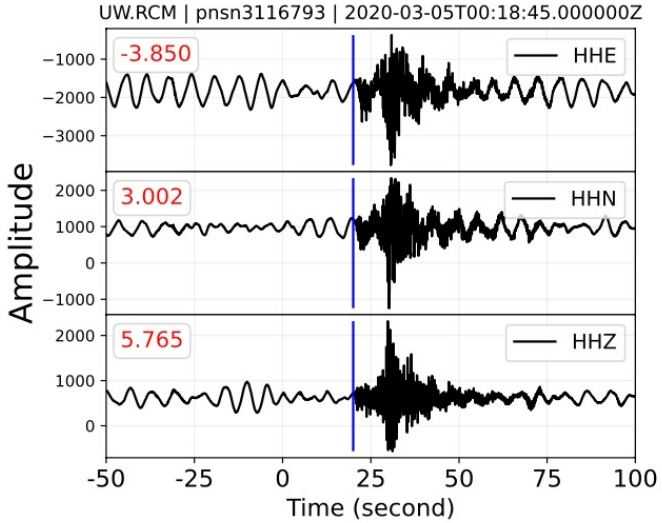


Offsets





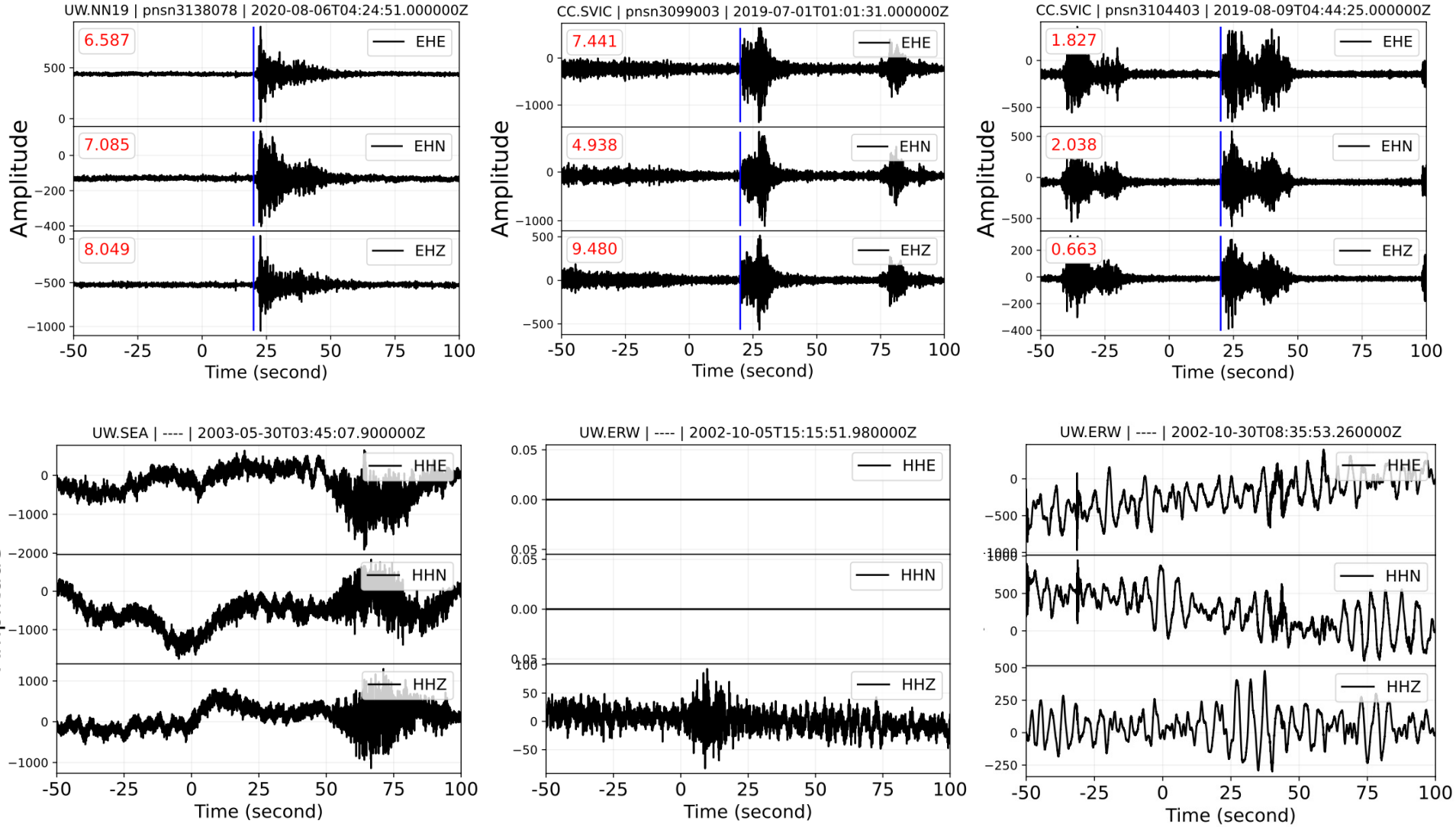




Event type could be misclassified

Curated datasets are not perfect

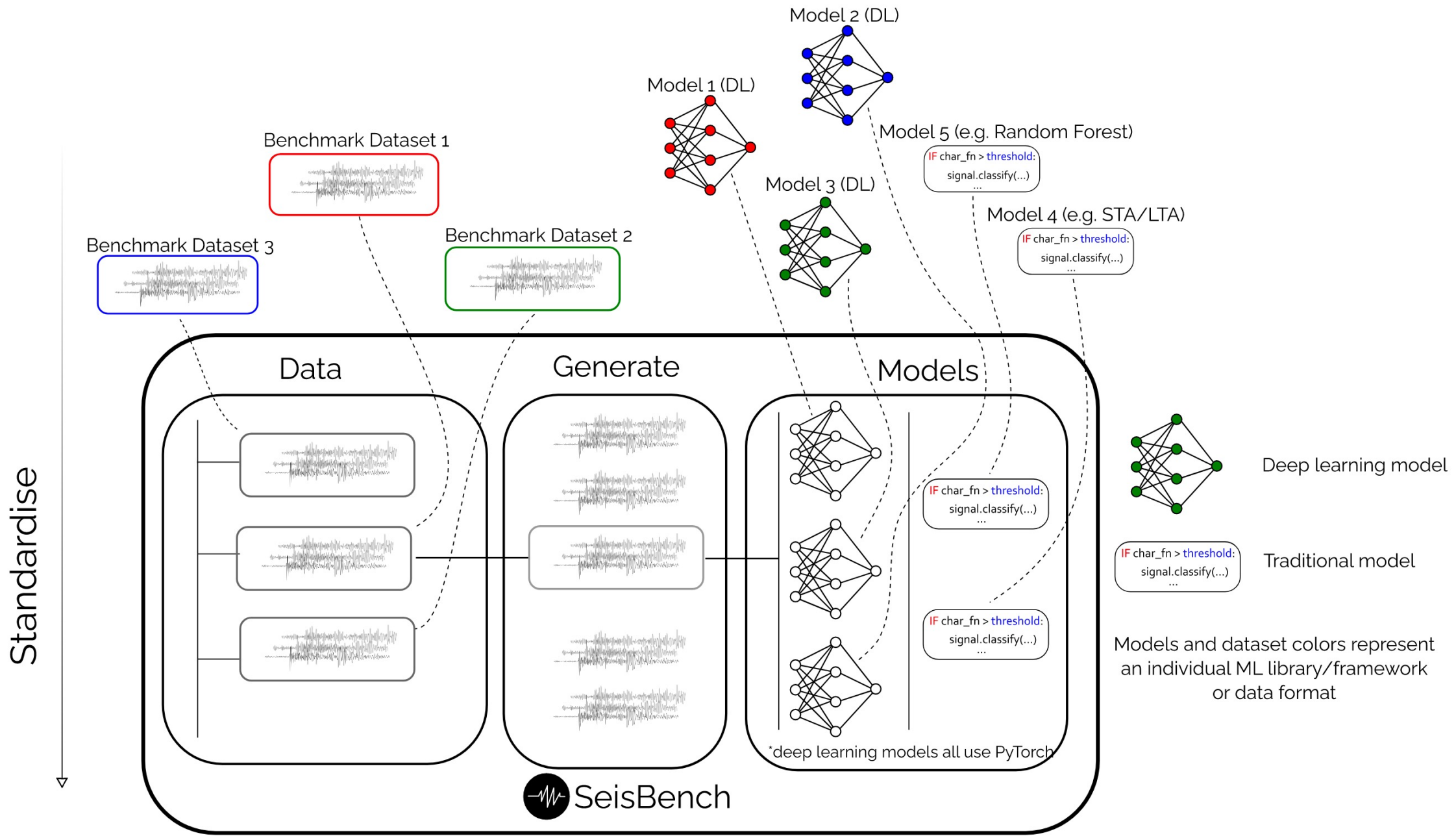
Multiple events: unlabeled phases



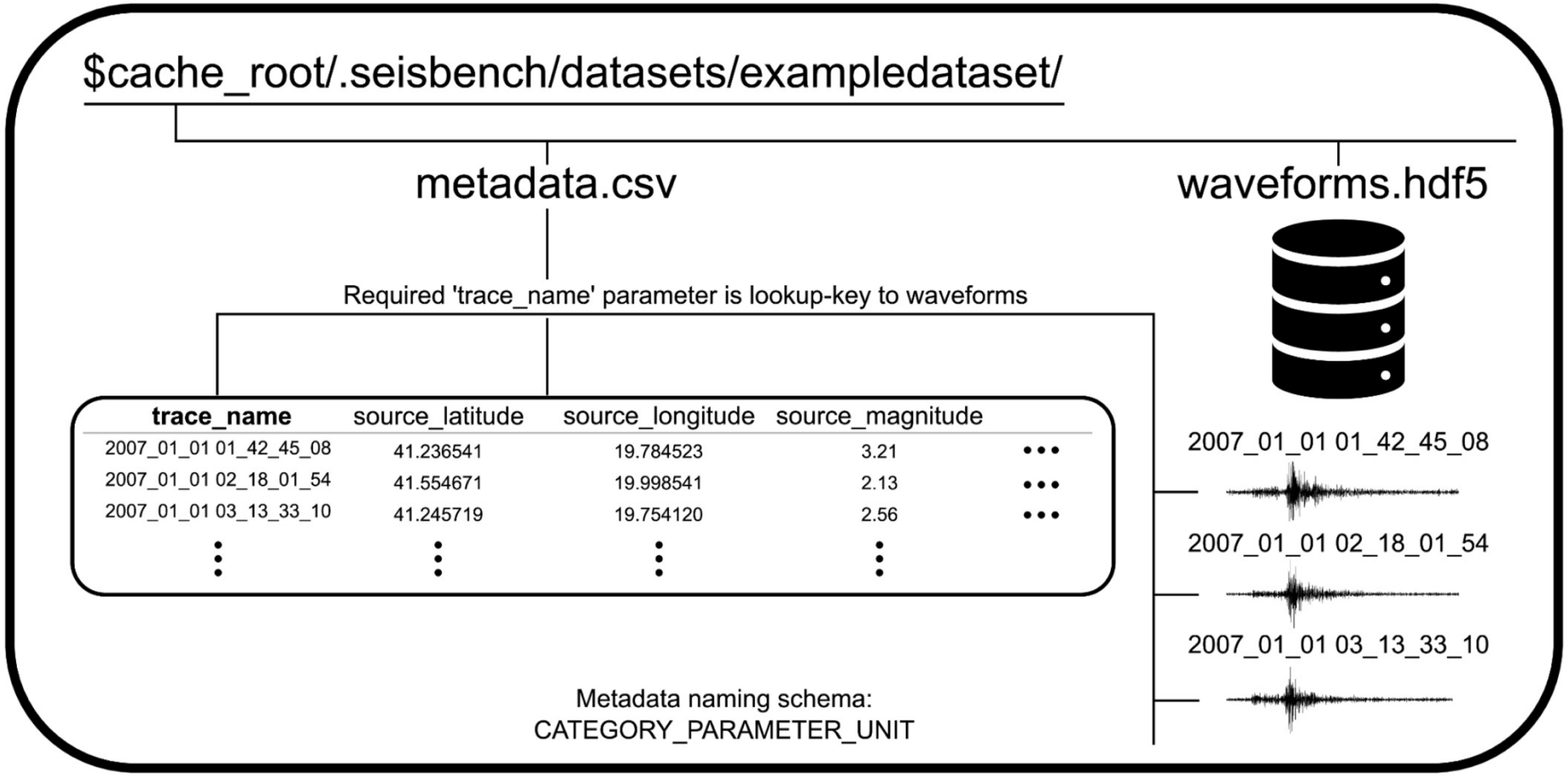
Uncataloged events & padded channels

SeisBench Ecosystem

SeisBench – a toolbox for Machine Learning seismology



SeisBench – a toolbox for Machine Learning seismology



SeisBench data format
HDF5 + CSV

SeisBench: an open-source community

seisbench / seisbench

Type / to search

Code | Issues 18 | Pull requests 2 | Actions | Projects | Security | Insights

seisbench Public

Edit Pins | Watch 21 | Fork 100 | Star 312

main | 14 Branches | 45 Tags | Go to file t | Add file | Code

yetinam Merge pull request #341 from seisbench/python313 — ae21ee2 · 4 days ago — 839 Commits

.github/workflows	Revert previous commit as issue is not fixed	4 days ago
contrib	Minimal benchmark script for annotate	last year
docs	Updated URL of backup repository to new GFZ domain in doc...	2 months ago
examples	Fix typo in dataset doc	6 months ago
seisbench	Fix 404 handling in list_pretrained	4 days ago
tests	Fix 404 handling in list_pretrained	4 days ago
.gitignore	adding pyproject / isort	3 years ago
.pre-commit-config.yaml	Disable pretty-format-json pre-commit hook	2 years ago

About

SeisBench - A toolbox for machine learning in seismology

python | science | machine-learning | deep-learning | seismology

Readme
GPL-3.0 license
Code of conduct
Activity
Custom properties
312 stars
21 watching
100 forks
Report repository

Hands-on: PNW dataset & SeisBench