

GEOLOGY

Machine learning reveals cyclic changes in seismic source spectra in Geysers geothermal field

Benjamin K. Holtzman,^{1*} Arthur Paté,¹ John Paisley,² Felix Waldhauser,¹ Douglas Repetto¹

The earthquake rupture process comprises complex interactions of stress, fracture, and frictional properties. New machine learning methods demonstrate great potential to reveal patterns in time-dependent spectral properties of seismic signals and enable identification of changes in faulting processes. Clustering of 46,000 earthquakes of $0.3 < M_L < 1.5$ from the Geysers geothermal field (CA) yields groupings that have no reservoir-scale spatial patterns but clear temporal patterns. Events with similar spectral properties repeat on annual cycles within each cluster and track changes in the water injection rates into the Geysers reservoir, indicating that changes in acoustic properties and faulting processes accompany changes in thermomechanical state. The methods open new means to identify and characterize subtle changes in seismic source properties, with applications to tectonic and geothermal seismicity.

INTRODUCTION

Understanding earthquake physics through in situ observations is difficult because the recorded seismic signals are convolutions of the earthquake source and the wave propagation process. Seismicity induced by human activity provides semicontrolled experiments in which to study the tightly bound processes of fracture, faulting, and frictional failure. Geothermal reservoirs provide important cases because extraction of thermal energy from Earth's crust ("heat mining") to generate electricity is a practically inexhaustible supply of CO₂-free power that could be critical for a post-oil economy [for example, the study of Tester *et al.* (1)]. Numerous practical limitations arise from the difficulty of drilling into hot rock, controlling the flow of water through fracture networks as illustrated in Fig. 1A, maximizing the recovery of heated water or steam, and minimizing the number of resulting felt earthquakes. Figure 1B is a mechanism map, showing a space defined by three primary mechanisms that produce different patterns in seismicity, namely, shear fracture, hydraulic fracture, and thermal cracking. Shear faulting can be driven by tectonic, gravitational, and tidal forces as well as hydraulic pressure and thermal stresses. Changes in the friction and failure mechanism during shear faulting will change the frequency content of the seismic signal [for example, the study of McLaskey *et al.* (2)].

Machine learning (ML) methods show enormous potential for rapidly improving our understanding of earthquake physics by identifying much more subtle patterns in seismic signals than standard methods of seismic waveform analysis. Most applications of ML to seismology thus far have focused on detection of increasingly small events, with great success [for example, the studies of Yoon *et al.* (3), Li *et al.* (4), Provost *et al.* (5), Perol *et al.* (6), Aguiar and Beroza (7), and Reynen and Audet (8)]. With the exception of the study of Aguiar and Beroza (7), these methods are mostly "supervised" approaches in which the goal is to learn a model that predicts an output corresponding to a given input. Recently, a random forest method has been used to forecast stick-slip events in laboratory friction experiments (9), raising the important possibility that precursory signals exist that could be used to forecast natural earthquakes.

Here, our aim is characterization, not detection, that is only possible after a catalog has been built. We take an "unsupervised" learning approach, in which the machine is given no predefined prediction task for pattern identification. The patterns of interest are subtle changes in the spectral content of a large number of recorded earthquakes that are likely due to some combination of wave propagation and changes in the faulting process. On the basis of previous successes of ML methods applied to audio in the field of music information retrieval, for example, the studies of Downie (10) and Li *et al.* (11), the analysis input in our study is the event spectrogram, computed as short-time Fourier transforms. Previous work on human perception of sonified seismic data demonstrated that humans are surprisingly sensitive to subtle spectral aspects of the signals that reflect physical aspects of the earthquake source and wave propagation (12–18). The use of the spectrogram (instead of the waveform) as input enables certain pattern matching techniques that lead to unprecedented levels of pattern identification in and among signals [for example, the study of Cotton and Ellis (19)].

We apply our approach to induced earthquakes in the Geysers geothermal field, in Sonoma and Lake Counties, CA, one of the largest (~800 MW) and longest-running geothermal fields in the world. Water is continuously injected (poured, not pumped) into the crust via 71 injection wells, where it is heated and vaporized at depths of up to 5 km. The steam is tapped by 335 production wells and drives multiple turbine generators [for example, the studies of Majer and Peterson (20) and Hartline *et al.* (21)]. The water injection rates increase significantly in winter months as runoff and greywater from nearby cities are reclaimed. The Geysers has among the highest rates of seismicity in a geothermal field, with ~15,000 events/year in recent years, digitally recorded at 30 seismometers by the U.S. Geological Survey, with measurable magnitudes in the range of $-0.7 < M_L < 4.7$. The seismicity tracks the fluid injection volume, but is likely induced predominantly by thermal, not hydraulic or tectonic, stress [for example, see Hartline *et al.* (21), Martínez-Garzón *et al.* (22), and Jeanne *et al.* (23)]. We gathered 3 years of earthquake waveforms (January 2012 to December 2014, 20 s each, including the *P* and *S* wave trains) at two borehole seismic stations (SQK and STY) in the central part of the Geysers field, with over 46,000 precisely located events with magnitudes $-0.7 \leq M_L \leq 4.5$ and a *b* value of 1.13, with catalog completeness (M_c) falling quickly below 0.6 (fig. S3B) (24).

The three-stage method, illustrated in Fig. 2 and explained in detail in Materials and Methods, is as follows: First, nonnegative matrix

¹Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY 10964, USA.

²Department of Electrical Engineering, Data Science Institute, Columbia University, New York, NY 10025, USA.

*Corresponding author. Email: benh@ldeo.columbia.edu

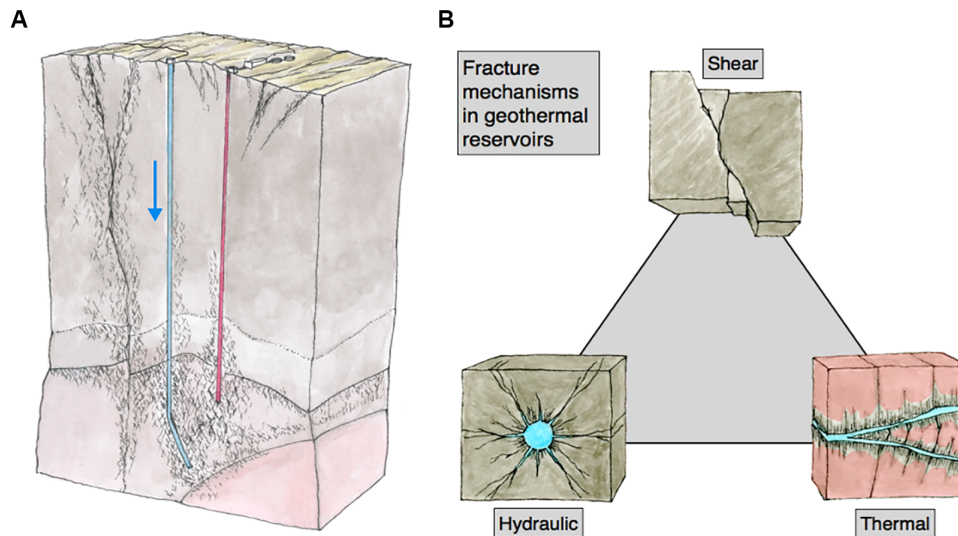


Fig. 1. Geothermal mechanics. (A) General geothermal energy production: Cold water percolates through fracture networks in hot rock, extracts heat, and returns to the surface, where that heat (in vapor) is used to drive a turbine to generate electricity. (B) Qualitative mechanism map of fracture processes that cause earthquakes. Note that the stresses associated with the bottom two processes can cause shear fracture/faulting as well.

factorization (NMF; Fig. 2B) [for example, the study of Lee and Seung (25)] is performed on spectrograms to reduce dimensionality and to isolate spectral patterns in the signals that are common to all. Each spectrogram is reconstructed via the product of a dictionary of frequency components (a left matrix common to all spectrograms) and a right matrix of activation coefficients (one for each spectrogram). We developed a Bayesian nonparametric Poisson matrix factorization model for this research, which builds off of previous NMF modeling techniques (25–27). This algorithm extracts the main frequency components shared by all signals, the number of which was inferred by the model, and reduces each spectrogram to a lower-dimensional sequence of mixtures over these components (section S1).

Second, hidden Markov models (HMMs; Fig. 2C) (28) are then learned on the NMF activation coefficient matrices (leaving the NMF dictionary behind) to further isolate and remove commonalities between signals and reduce the dimensionality, focusing on temporal patterns. The HMM represents each NMF activation matrix as a sequence of hidden states in time, where states are defined by patterns of frequency components that tend to occur together (emissions matrix B in Fig. 2C). From the state sequence matrix (Fig. 2C), a fingerprint is calculated by counting transitions between states in time, capturing subtle differences in signals in small matrices, which are directly comparable across signals. The two Bayesian models (NMF and HMM) are learned using stochastic variational inference (SVI) [for example, the study of Hoffman *et al.* (29)], which makes this study possible by drastically reducing the computation time (see Materials and Methods).

Finally, the K -means clustering method (28) is run on the fingerprints to define clusters of similar signals (Fig. 2D). No previous information is given to guide clustering, other than the number of clusters, J . Results from runs with J between 2 and 20 show similar temporal patterns (figs. S5 to S7). Looking at marginal improvement with increasing J , we found that $J > 10$ shows minimal improvement, whereas $J > 3$ was necessary to capture the basic clustering diversity. In the following, we discuss results for $J = 4$.

RESULTS

The resulting clusters for signals recorded at station SQK differ from each other in the time–frequency content of their signals, distinguishable by comparative listening (section S2.6). After clustering, we look for correlations between clusters and known physical properties of each earthquake, such as spatial location, magnitude, and temporal occurrence. A map of epicenters (Fig. 3B) and a cross section of hypocenters (fig. S1A) colored by cluster show no apparent reservoir-scale spatial patterns, although histograms of depth by cluster (fig. S1B) show a slight increase in depth in C3 and C4 relative to C1 (discussed further in section S2.1). Histograms of magnitudes in each cluster (fig. S3A) show that for $J = 5$, one cluster gathers the largest events, whereas all other events are grouped into C1 to C4. These four clusters have events in the range of $0.3 < M_L < 1.5$, with similar distributions and b values (fig. S3B). When plotted by time (Fig. 3C), the distributions of earthquakes in each cluster reveal temporal pulses of high concentrations of events. These pulses have durations of about 1 to 4 months, separated by about 1 year, and the temporal location of these annual pulses is phase-shifted for each cluster. In Fig. 3C, we order the clusters vertically by the timing of the first pulse, such that the cluster with the earliest occurrence is placed at the top (section S2.3). The same pattern is found when analyzing the signals recorded at station STY (section S2.5).

DISCUSSION

To better understand the temporal pattern of variation of the spectral properties of the earthquakes, we overlay in Fig. 3C the average monthly water mass injected over the whole Geysers area. A strong correlation appears between the injection rate and the occurrence of earthquake concentrations in each cluster; peak injection rate is correlated with one cluster (C1, winter) and the minima with another (C4, summer/fall). In Fig. 3D, we show monthly averages of injection rate and earthquake occurrence in each cluster that indicate association with injection rate (C1, maximum; C4, minimum). C3 suggests an additional sensitivity in clustering to the sign of the slope of the injection rate; C2 is

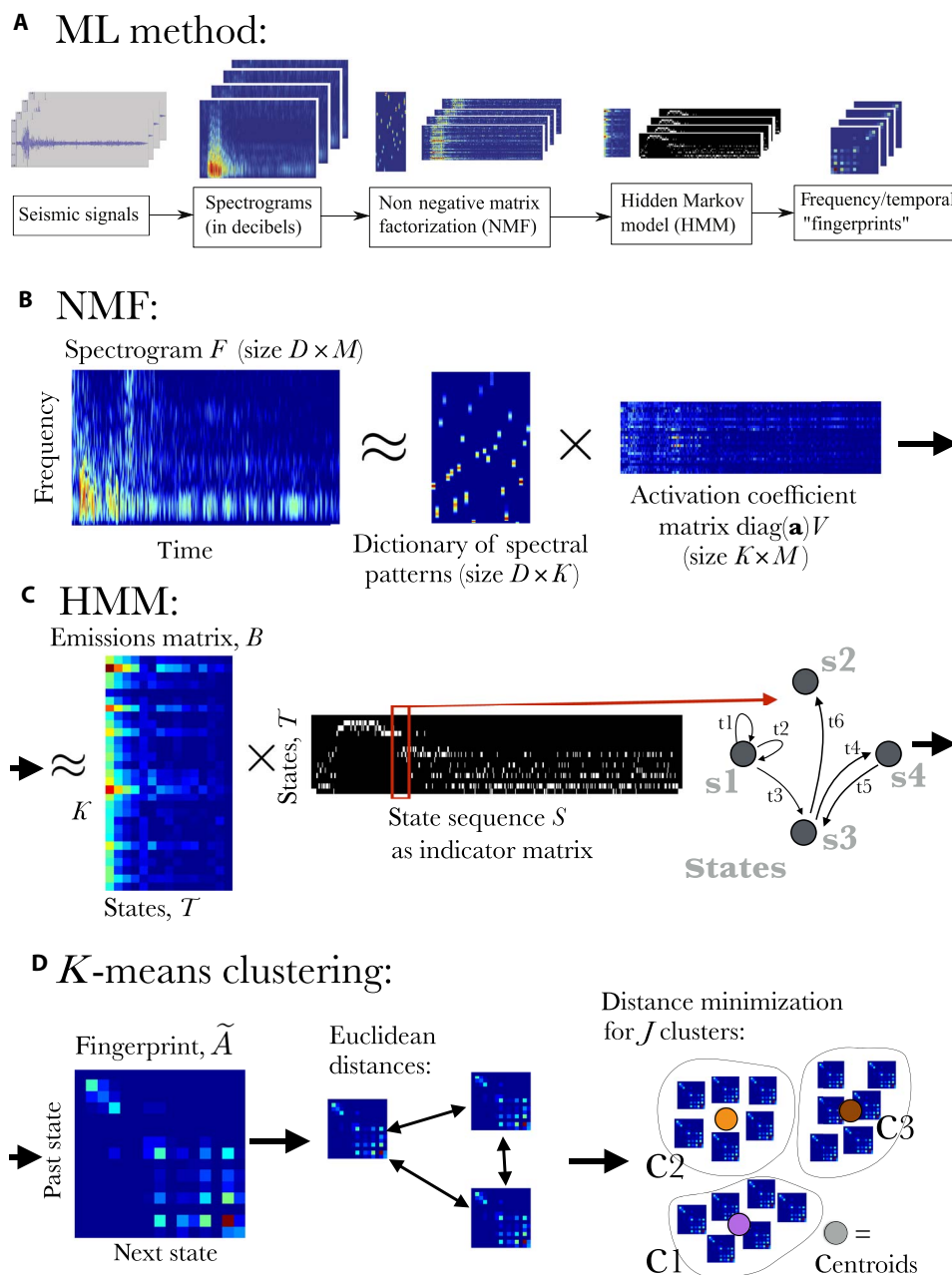


Fig. 2. Machine learning methods. (A) Flowchart of the ML approach, from waveform to fingerprint. NMF and HMM methods both reduce dimensionality and remove features common to all signals. (B) Example of the NMF decomposition of a spectrogram F into the product of matrices, dictionary U , and the activation coefficient matrix $\text{diag}(\mathbf{a})V$; (notation used in Materials and Methods). (C) In the HMM, each NMF activation matrix can be described as the product of a learned emissions matrix B and a state sequence S . To get to the fingerprint, the algorithm counts every time (t_1, t_2, \dots) a certain state follows a previous state; brighter spots mean that one state follows another more frequently. (D) Euclidean (-like) distances are calculated between each fingerprint pair, and then K -means produces J clusters.

contained entirely in 2014 and may be different because that year saw intensifying drought, with sustained low injection rate. Thus, it is inferred that changes in fluid content in the reservoir fracture network are the cause for the observed changes in the spectral content of the earthquakes over time, but by what mechanisms?

The null explanation for the correlation of clusters and injection rate is a wave propagation effect: A change in attenuation (by scattering or intrinsic dissipation) occurs with changes in fluid content of the

reservoir between the earthquake source and the seismometer. Vapor in fractures attenuates more effectively than fluid in fractures, and injection of cool water causes condensation, so lower attenuation (higher Q) is expected during high injection rates (C1 in Fig. 3C). However, if attenuation were the dominant cause of variations in spectra, then spatial clustering should appear, as events with longer source-station distances would accrue more high frequency loss [discussed further in section S2.1 and fig. S2; for example, the study of Zucca *et al.* (30)].

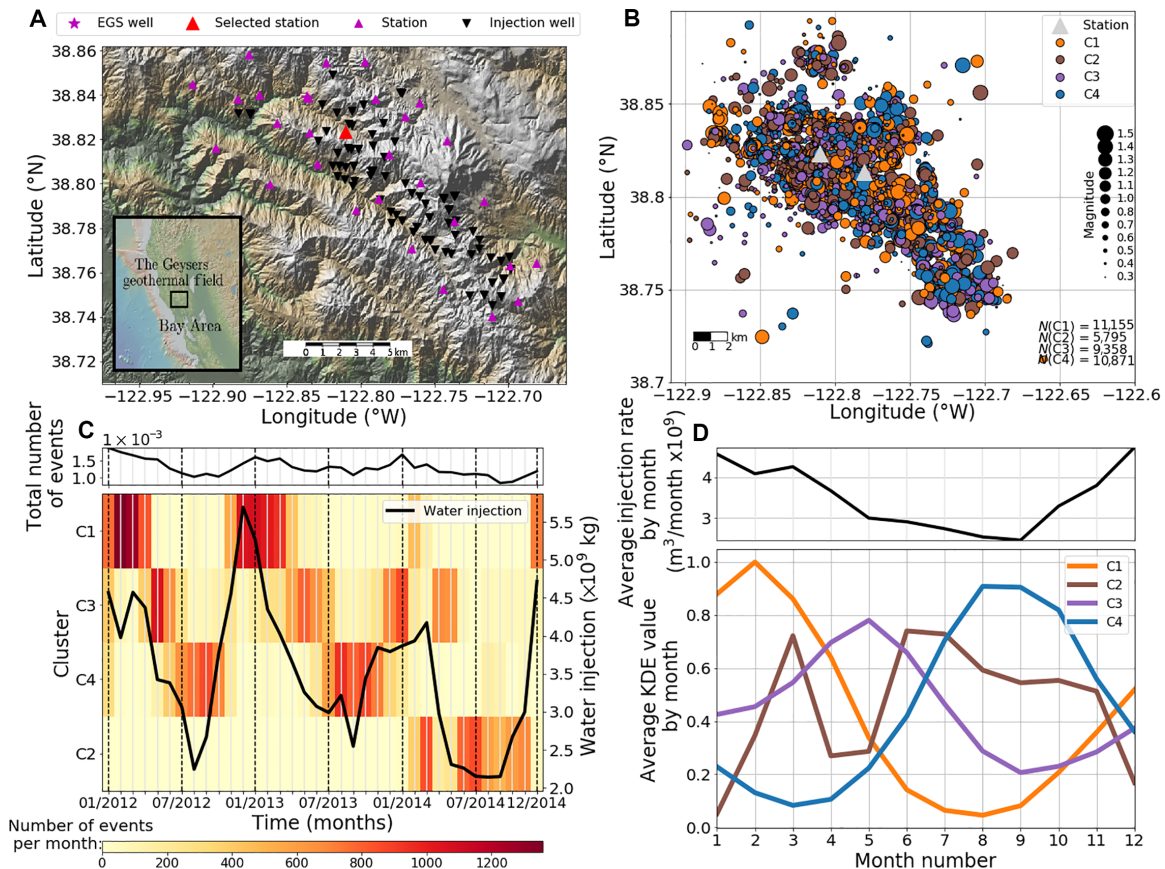


Fig. 3. Results. (A) Map of the Geysers geothermal field region. EGS, Enhanced Geothermal Systems. (B) Map view of earthquake epicenters scaled by magnitude and colored by cluster ID (C1 to C4). (C) Cluster gram shows clear clustering of increased earthquake occurrence in time. Black curve indicates average monthly mass of injected fluid over 71 injection wells at the Geysers geothermal field. Top panel shows total number of earthquakes, which correlates with injection rate (21). (D) Monthly stacked averages over the 3 years in (C) of injection rate (top) and earthquake occurrences in each cluster (bottom), using the kernel density estimator (KDE) of each cluster and averaging the monthly values. The peaks indicate associations of clusters with injection rate (C1, maximum; C4, minimum). C2 and C3 are discussed in the text.

Because reservoir-scale spatial clustering is not observed, we infer that dispersion by attenuation is not the dominant cause of clustering. Temporal changes in seismic velocity are documented [for example, Gritto and Jarpe (31) and Li and Wu (32)], but any accompanying changes in Q associated with injection rate would have to be reservoir-scale and dominate effects of dispersion by distance to explain the temporal clustering alone. Alternatively, local changes in the mechanical properties around the station (station effects) could be caused by seasonal changes in shallow groundwater saturation. However, station effects should affect amplitude fairly uniformly across the bandwidth we are investigating. Because the signals are amplitude-normalized before calculating the spectrograms, the observed seasonal variations are likely not primarily reflecting station effects.

Another compelling possibility is that ML methods are identifying changes in the faulting mechanism (Fig. 1B), due to changes in the driving force for faulting and/or fault properties. Local changes in properties of individual earthquakes have already been identified in the northwestern section of the Geysers field using standard seismic analysis methods that operate on the waveform or on spectra of the whole signal (for example, determinations of moment tensor, amplitude attenuation, corner frequency/stress drop). These changes are associated with fluid injection rates, during both passive injection (22, 33, 34) and pumping

at elevated fluid pressure [for example, Dreger *et al.* (35) and Boyd *et al.* (36)], to stimulate new fracture formation. Martínez-Garzón *et al.* (22) demonstrated that increased injection rate results in an increased number of normal faults to strike slip faults, increased total moment release, and decreased b values (more large earthquakes relative to small ones), for seismicity catalogs of $M > 1.0$. In more recent work, Martínez-Garzón *et al.* (34) also demonstrated an increase in the volumetric component of moment tensors with increased injection rate. Johnson *et al.* (37) found that larger earthquakes ($M \geq 1.5$) migrate to increasing depths (from 2 to 6 km) in 6-month periods after peak injection in both the northwestern and southeastern sections of the Geysers, consistent with our findings for smaller events (shown in fig. S1B). These focused studies suggest that the spectra-temporal patterns our ML methods are identifying reflect changes in earthquake source physics in a large catalog of earthquakes, most of which are too small for standard seismic analysis. It is hypothesized that injection of cool fluid causes transient stress changes in the reservoir (23, 38): First, thermal contraction could activate one fracture mechanism (illustrated in Fig. 1B), as observed in thermal fracturing experiments by Wang *et al.* (39); second, as the water draws more heat out of the rocks, vaporizes, and changes the pore pressure on the fault/fracture interfaces [for example, the study of Chen *et al.* (40)], effective stress and/or frictional

properties could change, possibly forming new or activating existing faults and fractures. Together, these studies suggest that the temporal patterns in microseismicity observed here correlate with changes in the thermomechanical state of the reservoir and resulting changes in failure process and/or source mechanism.

Here, we identify subtle changes in 46,000 earthquake spectra from two seismometers using a novel ML approach, analyzing 3 years of seismicity in the Geysers. Evidence suggests that these spectral variations are due in part to changes in the faulting mechanism, as identified in the detailed studies of the northwest Geysers. Further work is needed to relate the ML-identified patterns to changes in faulting mechanisms identified by standard seismic analysis methods. The convergence of ML methods with standard seismic methods on natural and laboratory data will lead to supervised learning methods by associating template fingerprints with different fault properties or processes. When used in a real-time environment for the purpose of operational reservoir monitoring, these new tools could enable vast improvements in informed reservoir engineering decision-making, to mitigate hazardous seismicity and to optimize small-scale seismicity that enhances fracture networks and improves the efficiency of energy production. An ability to rapidly identify these mechanisms and transitions among them would greatly improve our understanding of the thermal-mechanical state of a geothermal reservoir. Furthermore, these methods, when applied to natural tectonic faults, provide an entirely new way to identify and characterize spectral changes in microseismicity and have the potential to improve our understanding of the earthquake process.

MATERIALS AND METHODS

Here, we provide a mathematical description of the methodologies in Fig. 2. Our motivation comes from previous successes of ML methods applied to audio in the emergent field of “machine listening” and the widespread field of music information retrieval [for example, the studies of Downie (10) and Ogiwara and Tzatenakis (11)], as well as new advances in scalable probabilistic modeling [for example, the study of Hoffman *et al.* (29)]. Hence, the catalog of 20-s-long seismic signals (that include the *P* and *S* wave trains) is represented by the spectrogram (series of overlapping short-time Fourier transforms; Fig. 2A), not the waveform. This representation of an audio signal is closer than a waveform is to the pattern analyzed by the ear-brain system due to frequency separation performed by the cochlea and the auditory neurons [for example, the study of Carlile (41)].

About our notation

To increase transparency in the following derivations, we avoided indexing and products/summations when possible while still remaining mathematically unambiguous. For example, we let all functions of matrices be element-wise unless they have a long and standard interpretation otherwise, for example, if X is a matrix, then $\ln X$ is an element-wise logarithm. We let $X \sim p(\Lambda)$ be the element-wise generation of the values in X given a matrix Λ of equal size, that is, $X_{ij} \sim p(\Lambda_{ij})$ for all values of i and j . We similarly wrote the product likelihood using matrices, for example, $\text{Pois}(X|\Lambda)$ stands for $\prod_{i,j} \text{Pois}(X_{ij}|\Lambda_{ij})$, and likewise for vectors. When undefined, parameters are the same size as the relevant random variable or piece of data. For example, Λ is implicitly the same size as X above. In this case, setting $\Lambda = 1$ means creating a matrix entirely of 1s. For two matrices A and B of the same size, A/B indicates element-wise division, and $A \times B$ indicates element-wise multiplication, but AB indicates standard matrix multiplication.

Nonnegative matrix factorization

The first stage of modeling entails performing NMF on functions of all $N = 46,000$ spectrograms in the data set for each of the two stations (SQK and STY). NMF is a fundamental technique in ML for performing “parts-based learning” of data (25). For example, initial experiments on images of faces showed how the factors learned by NMF corresponded to parts of faces, whereas its application to text uncovers the themes or “topics” contained in a corpus. For spectrograms, we anticipated that these parts would be frequency regions that often co-occur, which was borne out by our experiments. As a result, NMF performs dimensionality reduction, mapping the columns of the original spectrogram into a much smaller dimension ($K < D$). There are many ways to learn NMF in a probabilistic setting, the most basic corresponding to maximizing the likelihood of a Poisson model on the data (42). Using variational inference, a fundamental ML technique for approximate posterior inference in Bayesian models (43), one can scale this learning up to large data sets using stochastic optimization. However, a prior model must first be defined, the most popular being that of the study of Cemgil (26). We slightly modified this prior to allow for a Bayesian nonparametric approach that determines a suitable rank of the factorization automatically from data. Our stochastic algorithm is closely related to that of the study of Paisley *et al.* (27).

Let F_i be the $D \times M$ spectrogram for the i th signal, illustrated in Fig. 2B. In our experiments, the index i is of a seismic event recorded by a single seismometer. For our data, D is the number of frequencies of the spectrogram, and M is the number of discrete time points ordered sequentially along the columns. (M could change per signal, but for our experiments, we extracted a time window of the same length around each event.) We first performed a function of this matrix as follows, for the i th signal

$$X_i = \max \left[\mathbf{0}, 20 \log_{10} \frac{F_i}{\text{median}(F_i)} \right] \quad (1)$$

We modeled this matrix using a K -rank factorization

$$X_i \sim \text{Pois}(U \text{diag}(\mathbf{a}) V_i), \quad i = 1, 2, \dots \quad (2)$$

$$U \sim \text{Gamma}(\mathbf{1}, \mathbf{1}), \quad \mathbf{a} \sim \text{Gamma}(\mathbf{1}/K, \mathbf{1}), \quad V_i \sim \text{Gamma}(\boldsymbol{\gamma}, \mathbf{1})$$

Here, U is $D \times K$, V_i is $K \times M$, and \mathbf{a} is a K -dimensional vector. The basis matrix U is shared by all signals and picks out the common frequencies, referred to in Fig. 2B as a “dictionary”. The diagonal matrix $\text{diag}(\mathbf{a})$ has a sparse gamma prior that will shrink out unnecessary columns of U by setting values to zero, or equivalently, it will shrink the corresponding rows of each V_i . This Bayesian nonparametric prior automatically infers these dimensions from the data. V_i contains the dimension-reduced representation for the i th signal. We could view the matrix product $\text{diag}(\mathbf{a}) V_i$ as the matrix of activations for the i th signal, where many of the activations will be learned to always be zero, indicating that a particular basis function was unnecessary.

We used SVI to learn these parameters. We defined approximating distributions $q(U)$, $q(\mathbf{a})$, and $q(V_i)$ and stochastically optimized the variational objective function

$$\mathcal{L} = \mathbb{E}_q \left[\ln \frac{p(U)}{q(U)} \right] + \mathbb{E}_q \left[\frac{p(\mathbf{a})}{q(\mathbf{a})} \right] + \sum_{i=1}^N \mathbb{E}_q \left[\frac{p(X_i|V_i, U, \mathbf{a}) p(V_i)}{q(V_i)} \right] \quad (3)$$

We defined these q distributions as follows

$$\begin{aligned} q(U) &= \text{Gamma}(U_1, U_2), \quad q(\mathbf{a}) = \text{Gamma}(\mathbf{a}_1, \mathbf{a}_2), \\ q(V_i) &= \text{Gamma}(V_{i1}, V_{i2}) \end{aligned} \quad (4)$$

The dimensions of $U_1, U_2, V_{i1}, V_{i2}, \mathbf{a}_1,$ and \mathbf{a}_2 are all equal to the variables they correspond to in $q(\cdot)$, which are given above. We sketched the algorithm for learning the parameters of these q distributions in Algorithm 1 (see the Supplementary Materials). As a step size, we set $\rho_t = (10^4 + t)^{-0.6}$, which ensures the Robbins-Monro conditions that $\sum_{t=1}^{\infty} \rho_t = \infty$ and $\sum_{t=1}^{\infty} \rho_t^2 < \infty$. Scalability is a result of only performing the inner loop on one signal at a time rather than all signals at once. After the algorithm converges, we used the learned distributions $q(U)$ and $q(\mathbf{a})$ to find the final values for each $q(V_i)$ by running the inner loop of Algorithm 1 (shown in pseudocode in section S1) for each signal.

Hidden Markov modeling

HMMs constitute a fundamental approach to modeling data sequences, including speech and other audio (44). They treat each output in a sequence as being generated from a distribution that is dependent on the hidden, imaginary state to which the observation belongs. The sequence of these states is generated according to a first-order Markov assumption, meaning that the probability of the next state is only dependent on the current state. SVI algorithms have been defined to learn HMMs (45), again greatly reducing computational time (as with NMF). We extended this algorithm to the stochastic setting using the standard framework (29). In this step in our process, we took from the NMF step the signal-specific $q(V_i)$ and learned a sequential model to uncover how the signals evolve in time. We made no further use of the dictionary U , which represents frequency components common to all signals.

To this end, for each signal, we defined

$$\hat{X}_i = \mathbb{E}_q[\text{diag}(\mathbf{a})] \mathbb{E}_q[V_i] = \text{diag}(\mathbf{a}_1/\mathbf{a}_2) V_{i1}/V_{i2}, \quad i = 1, 2, \dots$$

which takes each signal and maps it to its lower dimensional representation in U . This can be seen because, from NMF, the original signal $X_i \approx (U_1/U_2)\text{diag}(\mathbf{a}_1/\mathbf{a}_2)(V_{i1}/V_{i2})$ and $\mathbb{E}_q[U] = U_1/U_2$ according to its approximate variational posterior distribution. Because \hat{X}_i is a $K \times M$ matrix and $K \ll D$, the dimensionality of the columns of X_i has been reduced.

We then learned a T -state HMM on these \hat{X}_i of the form

$$\hat{X}_i(:, j) \sim \text{Pois}(B_{S_i(j)}), \quad S_i \sim \text{MC}(\pi, A_i) \quad (5)$$

$$\pi_i \sim \text{Dirichlet}(\pi_0/K), \quad A_i \sim \text{Dirichlet}(\alpha/T), \quad B \sim \text{Gamma}(\beta/T, 1)$$

This condensed notation for the generative process can be understood as follows: First, create a $T \times K$ matrix B with i.i.d. (independent and identically distributed) samples from the indicated gamma distribution. Then, for each signal, generate a T dimensional initial-state distribution π_i and a $T \times T$ Markov transition matrix A_i , where each row of A_i is independently Dirichlet-distributed. For each signal \hat{X}_i , generate a corresponding Markov chain S_i of the same length. Then, model each column of \hat{X}_i as a vector of Poisson random variables using the row of B indicated by the Markov chain. We observed that each signal has its own transition parameters but shares the same parameters for the matrix \hat{X} .

Using SVI, we defined the following q distributions

$$\begin{aligned} q(\pi_i) &= \text{Dirichlet}(\pi'_i), \quad q(A_i) = \text{Dirichlet}(A'_i), \\ q(B) &= \text{Gamma}(B_1, B_2) \end{aligned} \quad (6)$$

In this notation, the second Dirichlet distribution is a factorized distribution on each row of A_i using the corresponding row of A'_i as parameters. Because we learned T states, π_i and π'_i are T dimensional vectors, and A_i and A'_i are $T \times T$ matrices. The gamma distribution on B is factorized element-wise over this matrix, with the parameters for entry $B_{a,b}$ equal to $(B_1(a, b), B_2(a, b))$. The dimensionality of B_1 and B_2 are $T \times K$, the same as B . We sketched the SVI algorithm for learning these parameters in Algorithm 2 (shown in the Supplementary Materials). As with the NMF algorithm, after the HMM algorithm converges, we used the learned distributions $q(B)$ to find the final values for each $q(A)$ and $q(\pi)$ by running the inner loop of Algorithm 2 for each signal.

K-means clustering

The final step is to cluster signals in an unsupervised manner. We used the statistics from each signal's transition matrix distribution $q(A'_i)$ by defining the fingerprint for each signal, $\tilde{A}_i = \sqrt{A'_i/\sum_{a,b} A'_i(a, b)}$ (Fig. 2, C and D). In the Supplementary Materials (figs. S10 and S11), we show the differences among the most representative fingerprints for two clusters. We then performed K -means on these fingerprints using the Euclidean distance. The K -means objective function is $\mathcal{L}(\mathbf{c}, \boldsymbol{\mu}) = \sum_{n=1}^N \sum_{j=1}^J 1(c_n = j) \|\tilde{A}_i - \boldsymbol{\mu}_j\|^2$. The indicator c_i defines the cluster to which X_i is assigned. The number of clusters is J , which must be chosen. In the Supplementary Materials, we show the objective function as a function of J (fig. S6) and results for different J (figs. S7 and S8). These clusters are then analyzed by looking for correlations with physical parameters, as discussed in the main text.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at <http://advances.sciencemag.org/cgi/content/full/4/5/eaao2929/DC1>

section S1. Supplemental text: Algorithms
 section S2. Supplemental text: Further discussion of results
 fig. S1. Distribution of earthquake depths by cluster.
 fig. S2. Energy loss due to attenuation, as a function of distance across the area of the Geysers reservoir, parameterized by Q_p and ω or f .
 fig. S3. Distribution of earthquake magnitudes by cluster.
 fig. S4. Kernel density estimators of the histogram values for each cluster.
 fig. S5. HMM with 49 states, station SQK.
 fig. S6. K -means objective function versus J , the number of clusters chosen.
 fig. S7. HMM with 15 states, station SQK.
 fig. S8. HMM with 49 states, station STY.
 fig. S9. Clustering with signals from both stations SQK and STY combined (49 HMM states, 5 clusters).
 fig. S10. Three most characteristic waveforms and associated spectrograms and fingerprints (from top to bottom) from the two clusters associated with maximum (C1) and minimum (C4) fluid injection rates for station SQK.
 fig. S11. Three most characteristic waveforms and associated spectrograms and fingerprints (from top to bottom) from the two clusters associated with maximum (C1) and minimum (C4) fluid injection rates for station STY.
 table S1. List of attached audified sounds and their characteristics.
 audio S1 to S24. Audified seismic data from the three most characteristic earthquakes in each of clusters C1 to C4 (table S1).
 movie S1. Animation of the Geysers earthquake catalog, without and with clustering information.

REFERENCES AND NOTES

- J. W. Tester, B. Anderson, A. Batchelor, D. Blackwell, R. DiPippo, E. Drake, J. Garnish, B. Livesay, M. C. Moore, K. Nichols, in *The Future of Geothermal Energy: Impact of Enhanced Geothermal Systems (EGS) on the United States in the 21st Century* (Massachusetts Institute of Technology, 2006), pp. 372.
- G. C. McClaskey, A. M. Thomas, S. D. Glaser, R. M. Nadeau, Fault healing promotes high-frequency earthquakes in laboratory experiments and on natural faults. *Nature* **491**, 101–104 (2012).
- C. E. Yoon, O. O'Reilly, K. J. Bergen, G. C. Beroza, Earthquake detection through computationally efficient similarity search. *Sci. Adv.* **1**, e1501057 (2015).
- Z. Li, Z. Peng, X. Meng, A. Inbal, Y. Xie, D. Hollis, J.-P. Ampuero, Matched filter detection of microseismicity in long beach with a 5200-station dense array, in *SEG Technical Program Expanded Abstracts 2015*, R. V. Schneider, Ed. (Society of Exploration Geophysicists, 2015), pp. 2615–2619.
- F. Provost, C. Hibert, J.-P. Malet, Automatic classification of endogenous landslide seismicity using the Random Forest supervised classifier. *Geophys. Res. Lett.* **44**, 113–120 (2017).
- T. Perol, M. Gharbi, M. Denolle, Convolutional neural network for earthquake detection and location. *Sci. Adv.* **4**, e1700578 (2018).
- A. C. Aguiar, G. C. Beroza, PageRank for earthquakes. *Seismol. Res. Lett.* **85**, 344–350 (2014).
- A. Reynen, P. Audet, Supervised machine learning on a network scale: Application to seismic event classification and detection. *Geophys. J. Int.* **210**, 1394–1409 (2017).
- B. Rouet-Leduc, C. Hulbert, N. Lubbers, K. Barros, C. J. Humphreys, P. A. Johnson, Machine learning predicts laboratory earthquakes. *Geophys. Res. Lett.* **44**, 9276–9282 (2017).
- J. S. Downie, Music information retrieval. *Annu. Rev. Inf. Sci. Technol.* **37**, 295–340 (2003).
- T. Li, M. Oghihara, G. Tzatanakis, *Music Data Mining* (CRC Press, 2011).
- S. D. Speeth, Seismometer sounds. *J. Acoust. Soc. Am.* **33**, 909–916 (1961).
- G. E. Frantti, L. A. Levereault, Auditory discrimination of seismic signals from earthquakes and explosions. *Bull. Seismol. Soc. Am.* **55**, 1–25 (1965).
- F. Dombois, *Proceedings of the International Conference on Auditory Display (ICAD)*, Kyoto, Japan, 2 to 5 July 2002.
- Z. Peng, C. Aiken, D. Kilb, D. R. Shelly, B. Enescu, Listening to the 2011 magnitude 9.0 Tohoku-Oki, Japan, earthquake. *Seismol. Res. Lett.* **83**, 287–293 (2012).
- A. Paté, L. Boschi, J.-L. Le Carrou, B. Holtzman, Categorization of seismic sources by auditory display: A blind test. *Int. J. Hum. Comput. Stud.* **85**, 57–67 (2016).
- A. Paté, L. Boschi, D. Dubois, J.-L. Le Carrou, B. Holtzman, Auditory display of seismic data: On the use of experts' categorizations and verbal descriptions as heuristics for geoscience. *J. Acoust. Soc. Am.* **141**, 2143–2162 (2017).
- L. Boschi, L. Delcor, J.-L. Le Carrou, C. Fritz, A. Paté, B. Holtzman, On the perception of audified seismograms. *Seismol. Res. Lett.* **88**, 1279–1289 (2017).
- C. V. Cotton, D. P. W. Ellis, Spectral vs. spectro-temporal features for acoustic event detection, in *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, (IEEE, 2011), pp. 69–72.
- E. L. Majer, J. E. Peterson, The impact of injection on seismicity at the Geysers, California Geothermal Field. *Int. J. Rock Mech. Min. Sci.* **44**, 1079–1090 (2007).
- C. S. Hartline, M. A. Walters, M. C. Wright, C. K. Forson, A. J. Sadowski, Three-dimensional structural model building and induced seismicity analysis at the Geysers Geothermal Field, Northern California, in *Proceedings of the 41st Workshop on Geothermal Reservoir Engineering*, 22 to 24 February 2016 (Stanford Univ., 2016).
- P. Martínez-Garzón, G. Kwiątek, H. Sone, M. Bohnhoff, G. Dresen, C. Hartline, Spatiotemporal changes, faulting regimes, and source parameters of induced seismicity: A case study from the Geysers geothermal field. *J. Geophys. Res. Solid Earth* **119**, 8378–8396 (2014).
- P. Jeanne, J. Rutqvist, P. F. Dobson, M. Walters, C. Hartline, J. Garcia, The impacts of mechanical stress transfers caused by hydromechanical and thermal processes on fault stability during hydraulic stimulation in a deep geothermal reservoir. *Int. J. Rock Mech. Min. Sci.* **72**, 149–163 (2014).
- F. Waldhauser, D. P. Schaff, Large-scale relocation of two decades of Northern California seismicity using cross-correlation and double-difference methods. *J. Geophys. Res. Solid Earth* **113**, B08311 (2008).
- D. D. Lee, H. S. Seung, Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).
- A. T. Cemgil, Bayesian inference for nonnegative matrix factorisation models. *Comput. Intell. Neurosci.* **2009**, 785152 (2009).
- J. Paisley, D. Blei, M. Jordan, Bayesian nonnegative matrix factorization with stochastic variational inference, in *Handbook of Mixed Membership Models and Their Applications*, E. M. Airoldi, D. Blei, E. A. Eroshva, S. E. Fienberg, Eds. (Chapman and Hall/CRC, 2015).
- C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)* (Springer Verlag, 2006).
- M. D. Hoffman, D. M. Blei, C. Wang, J. Paisley, Stochastic variational inference. *J. Mach. Learn. Res.* **14**, 1303–1347 (2013).
- J. J. Zucca, L. J. Hutchings, P. W. Kasameyer, Seismic velocity and attenuation structure of the Geysers geothermal field, California. *Geothermics* **23**, 111–126 (1994).
- R. Gritto, S. P. Jarpe, Temporal variations of V_p/V_s -ratio at The Geysers geothermal field, USA. *Geothermics* **52**, 112–119 (2014).
- G. Lin, B. Wu, Seismic velocity structure and characteristics of induced seismicity at the Geysers geothermal field, eastern California. *Geothermics* **71**, 225–233 (2018).
- P. Martínez-Garzón, G. Kwiątek, M. Bohnhoff, G. Dresen, Impact of fluid injection on fracture reactivation at The Geysers geothermal field. *J. Geophys. Res. Solid Earth* **121**, 7432–7449 (2016).
- P. Martínez-Garzón, G. Kwiątek, M. Bohnhoff, G. Dresen, Volumetric components in the earthquake source related to fluid injection and stress state. *Geophys. Res. Lett.* **44**, 800–809 (2017).
- D. S. Dreger, O. S. Boyd, R. Gritto, Automatic moment tensor analyses, in-situ stress estimation and temporal stress changes at The Geysers EGS Demonstration Project, in *Proceedings of the 42nd Workshop on Geothermal Reservoir Engineering*, 13 to 15 February 2017 (Stanford Univ., 2017).
- O. S. Boyd, D. S. Dreger, V. H. Lai, R. Gritto, A systematic analysis of seismic moment tensor at the Geysers geothermal field, California. *Bull. Seismol. Soc. Am.* **105**, 2969–2986 (2015).
- C. W. Johnson, E. J. Totten, R. Bürgmann, Depth migration of seasonally induced seismicity at the Geysers geothermal field. *Geophys. Res. Lett.* **43**, 6196–6204 (2016).
- J. Rutqvist, P. F. Dobson, J. Garcia, C. Hartline, P. Jeanne, C. M. Oldenburg, D. W. Vasco, M. Walters, The Northwest Geysers EGS Demonstration Project, California: Pre-stimulation modeling and interpretation of the stimulation. *Math. Geosci.* **47**, 3–29 (2015).
- H. F. Wang, B. P. Bonner, S. R. Carlson, B. J. Kowallis, H. C. Heard, Thermal stress cracking in granite. *J. Geophys. Res. Solid Earth* **94**, 1745–1758 (1989).
- J. Chen, A. Niemeijer, L. Yao, S. Ma, Water vaporization promotes coseismic fluid pressurization and buffers temperature rise. *Geophys. Res. Lett.* **44**, 2177–2185 (2017).
- S. Carlile, *Psychoacoustics* (Logos Verlag, 2011), chap. 3, pp. 41–61.
- D. D. Lee, H. S. Seung, Algorithms for non-negative matrix factorization, in *Proceedings of the 13th International Conference on Neural Information Processing Systems* (MIT Press, 2000), pp. 535–541.
- M. Wainwright, M. Jordan, Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* **1**, 1 (2008).
- L. R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77**, 257–286 (1989).
- A. Zhang, S. Gultekin, J. Paisley, Stochastic Variational Inference for the HDP-HMM, in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, *PMLR* **51**, 800–808 (2016).

Acknowledgments: This article is LDEO contribution #8212. Our paper benefited from discussions with D. Ellis, B. Bonner, K. Nihei, P. Dobson, L. Hutchings, M. Walters, C. Hartline, H. Savage, Y. J. Tan, L. Boschi, B. Rouet-Leduc, C. Hulbert, and P. Johnson and reviews from R. Bürgmann and two anonymous reviewers. **Funding:** The project was funded by a seed grant from Columbia University (RISE: Research Initiatives in Science and Engineering). **Author contributions:** B.K.H., J.P., and F.W. designed the project. F.W. developed the earthquake catalog. J.P. developed the ML algorithms. A.P. analyzed and visualized the results and developed the auditory perception aspects. B.K.H., J.P., and A.P. wrote the manuscript and supplement. B.K.H. drew Fig. 1. D.R. and B.K.H. produced movie S1. All authors discussed interpretations and contributed to the final manuscript. **Competing interests:** All authors declare that they have no competing interests. **Data and materials availability:** Waveform data, metadata, and data products for this study are publicly available, accessed through the Northern California Earthquake Data Center and the Induced Seismicity Data Website at the Lawrence Berkeley National Laboratory, supported by the U.S. Department of Energy Office of Geothermal Technology. Fluid injection data are publicly available at the California Department of Conservation, Department of Oil and Gas and Geothermal Resources. Additional data and information can be requested from the authors, addressed to the corresponding author.

Submitted 4 July 2017

Accepted 11 April 2018

Published 23 May 2018

10.1126/sciadv.aao2929

Citation: B. K. Holtzman, A. Paté, J. Paisley, F. Waldhauser, D. Repetto, Machine learning reveals cyclic changes in seismic source spectra in Geysers geothermal field. *Sci. Adv.* **4**, eaa02929 (2018).

Machine learning reveals cyclic changes in seismic source spectra in Geysers geothermal field

Benjamin K. Holtzman, Arthur Paté, John Paisley, Felix Waldhauser and Douglas Repetto

Sci Adv 4 (5), eaao2929.
DOI: 10.1126/sciadv.aao2929

ARTICLE TOOLS	http://advances.sciencemag.org/content/4/5/eaao2929
SUPPLEMENTARY MATERIALS	http://advances.sciencemag.org/content/suppl/2018/05/21/4.5.eaao2929.DC1
REFERENCES	This article cites 33 articles, 4 of which you can access for free http://advances.sciencemag.org/content/4/5/eaao2929#BIBL
PERMISSIONS	http://www.sciencemag.org/help/reprints-and-permissions

Use of this article is subject to the [Terms of Service](#)

Science Advances (ISSN 2375-2548) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. 2017 © The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. The title *Science Advances* is a registered trademark of AAAS.