FastMapSVM

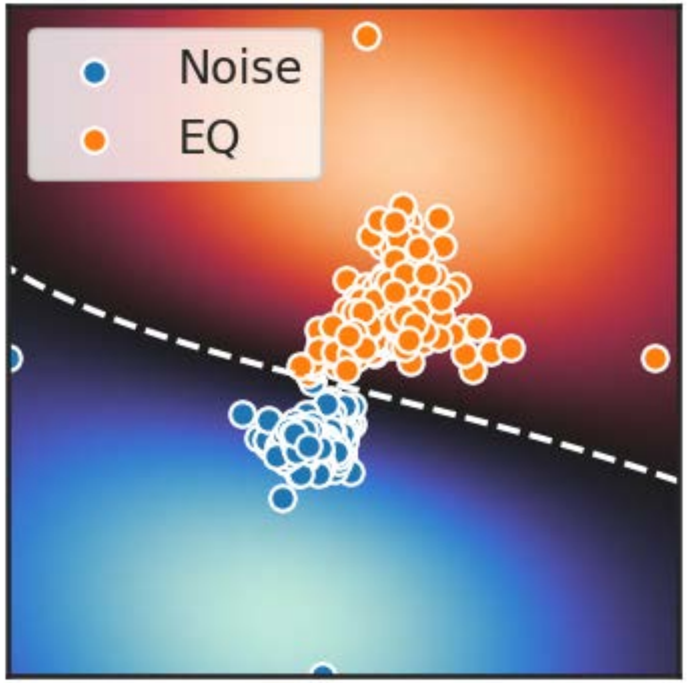
Classifying Complex Objects Using the FastMap Algorithm and Support-Vector Machines

Malcolm C. A. White

Postdoctoral Associate, Department of Earth, Atmospheric and Planetary Sciences

In collaboration with Nori Nakata, Kushal Sharma, Ang Li, and T. K. Satish Kumar

May 26, 2022



FastMapSVM performs comparably to state-of-the-art NNs, but uses two orders of magnitude less data and time for training.

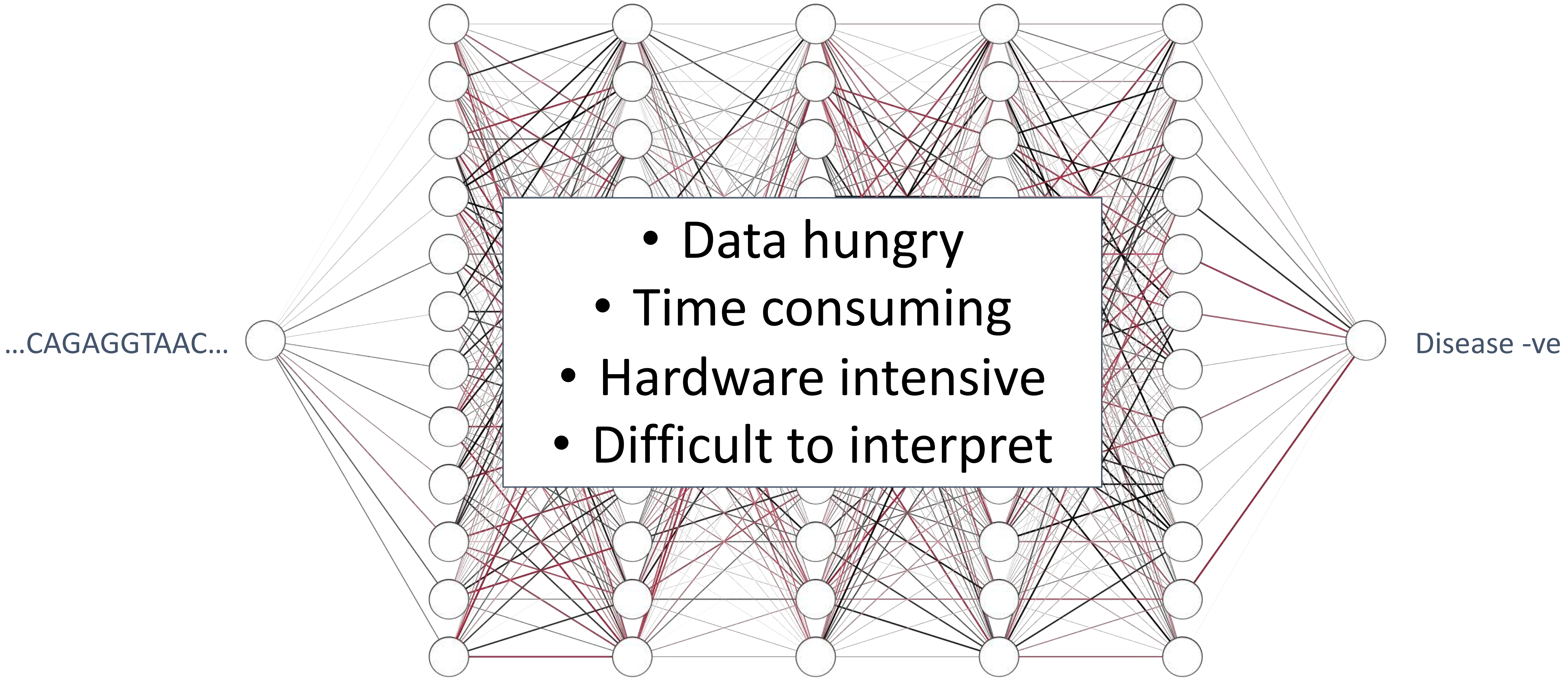
Model	Precision	Recall	F1	Training Size	Train Time (s)
EQTransformer	1.0	1.0	1.0	10^6	10^5
CRED	1.0	0.96	0.98	10^6	NA
FastMapSVM	1.0	0.97	0.98	10^4	10^3

Consider a general problem of classifying complex objects such as DNA sequences.

Disease -ve	Disease +ve
CATCAGATTCTACGTCGAAGATTCAAACGCACAGAGGTAAC	TGGCTGGATCCGTATTTGCCCGCTTCGTAACCCTACTTGCT
TGCAATCGGCATCCTCCCGTGGAACTCCCGTTAATTTTAAC	GCTCTACCACGGAGAAGGTCAGTACGATAATCCCGTTAAAA
CCGTTTTATCCAAGCGCCCAGAGCTCATTGGCAGGCATATG	CCATGTTGCCGTCCGTCCCGCAGGCGCACCTCACGCCCTCA
…	…
GGAGGGCTGGCGTATCCGTACAGGTGAGGCATTGGCCTTAG	TGACGGATAAACGTCCTCAGATCGGCCCGCTAGAGTTGCGG
TGTACAGCGCATGTGTAGTGCTCAATAAACCGGCTAACTCC	TGGGAATCTTATCTTCCAAGCTTCTTCAACTAGCAGCACTG
TCTGGGTACAGCTTAACAATTATAGCGGAGGATAGTGATCT	TTAAAGTCTATTCGCCGTGTGATCCGGTTCGGCCTTGCGCG

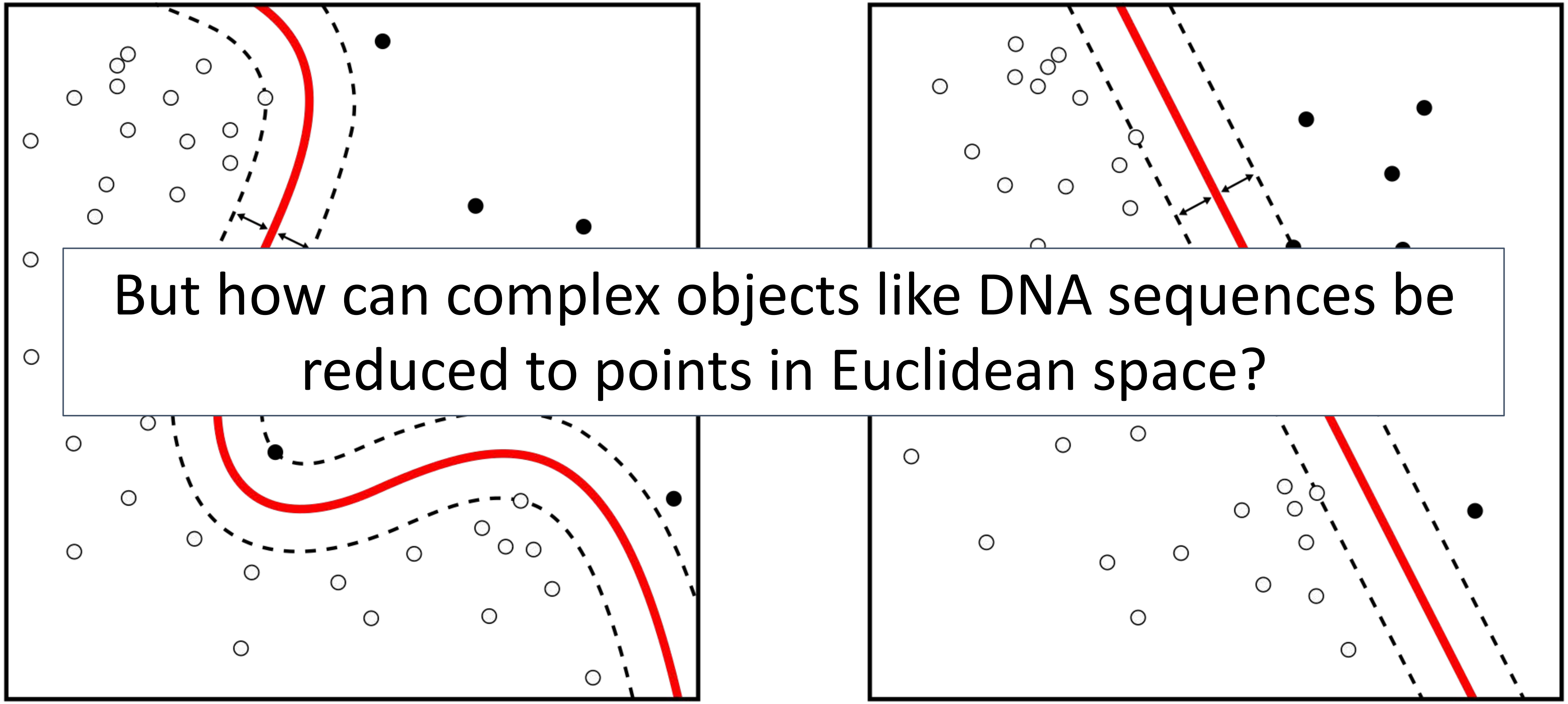
Unknown
CACTATACGATGCTTGGCCCGAGATAACTCCGTCGGAGTTC
…
AATGCAGTAGACCCTCAGGTATACAGTGAACAAACGTTAAA

NNs have become a go-to solution for such tasks.



- Data hungry
- Time consuming
- Hardware intensive
- Difficult to interpret

Support-vector machines and kernel methods yield optimal solutions for classifying geometric points.

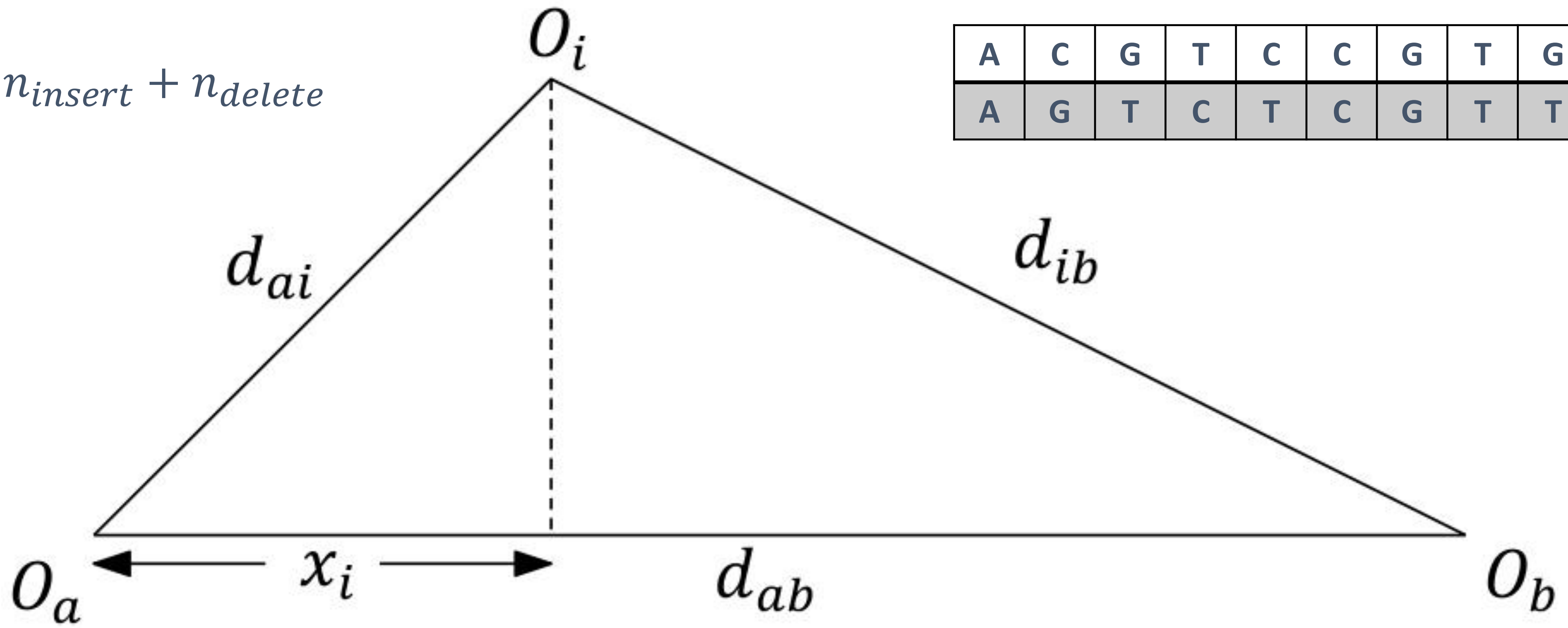


But how can complex objects like DNA sequences be reduced to points in Euclidean space?

Source: https://en.wikipedia.org/wiki/Support-vector_machine

The FastMap algorithm (Faloutsos and Lin, 1995) embeds complex objects in Euclidean space using a distance function defined for pairs of objects.

$$d_{ij} = n_{insert} + n_{delete}$$

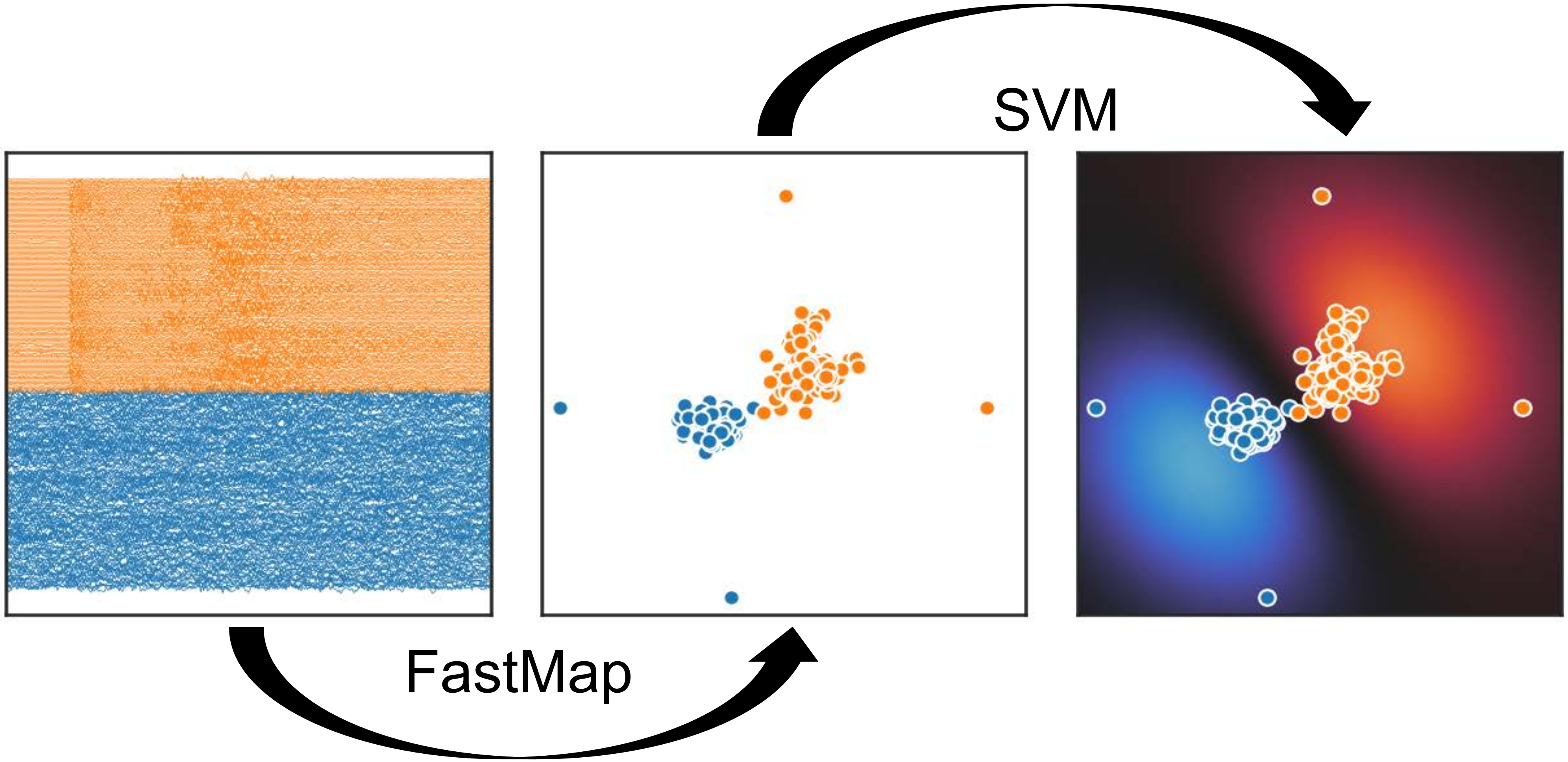


A	C	G	T	C	C	G	T	G	C
A	G	T	C	T	C	G	T	T	G

Two perennial problems in Earthquake Science that need robust, adaptable solutions and can be formulated as classification problems:

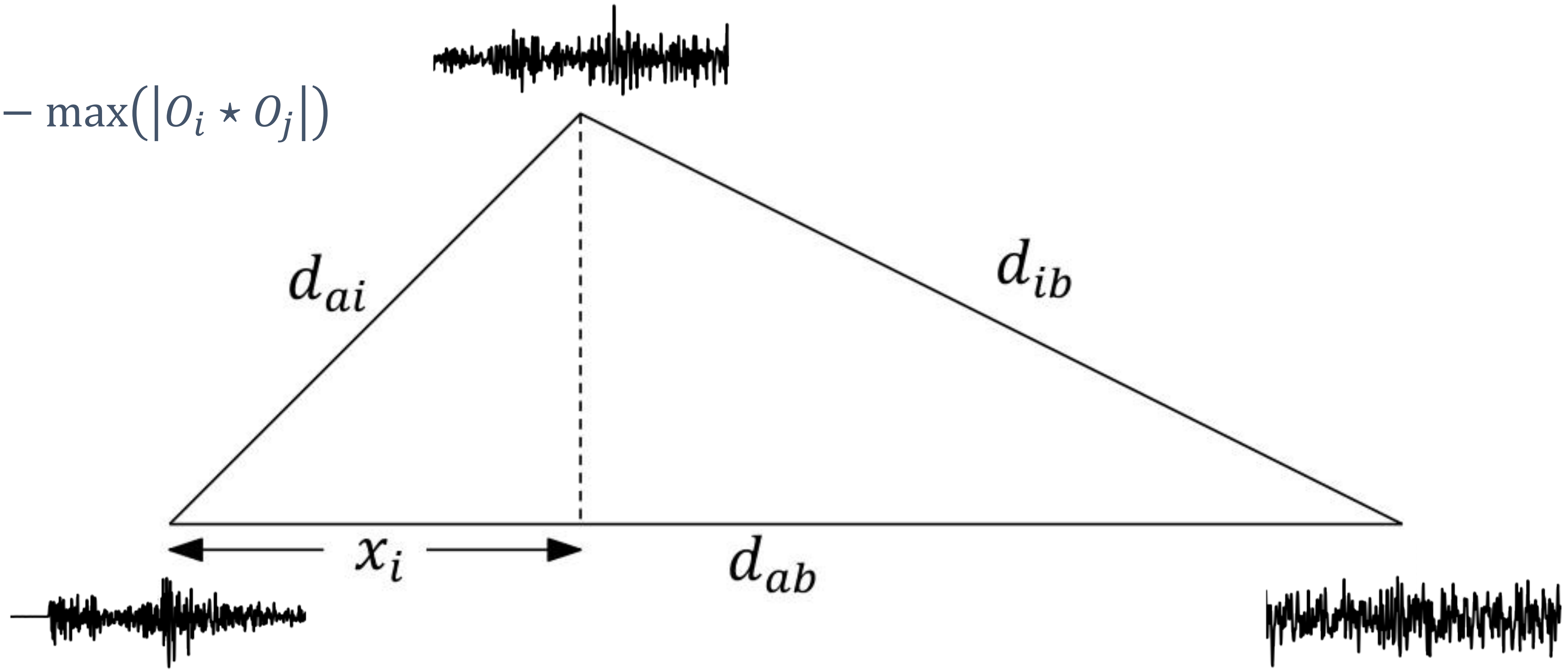
1. Detecting earthquakes (i.e., *noise* versus *EQ*)
2. Identifying phases (i.e., *P* versus *S*)

FastMap embeds seismograms in Euclidean space and SVMs with kernel methods classify embedded seismograms.

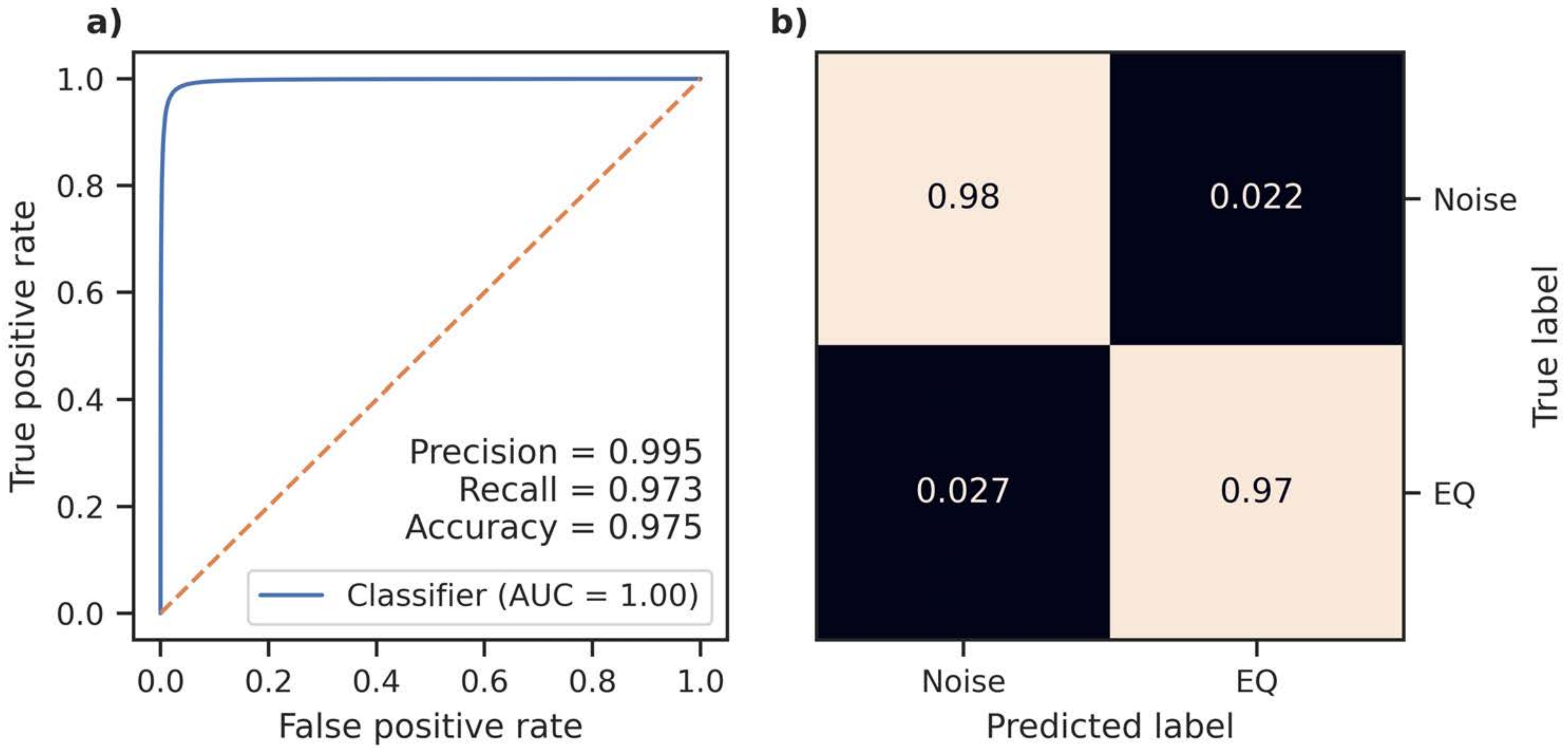


The normalized cross-correlation operator effectively quantifies seismogram (dis)similarity (i.e., “distance”).

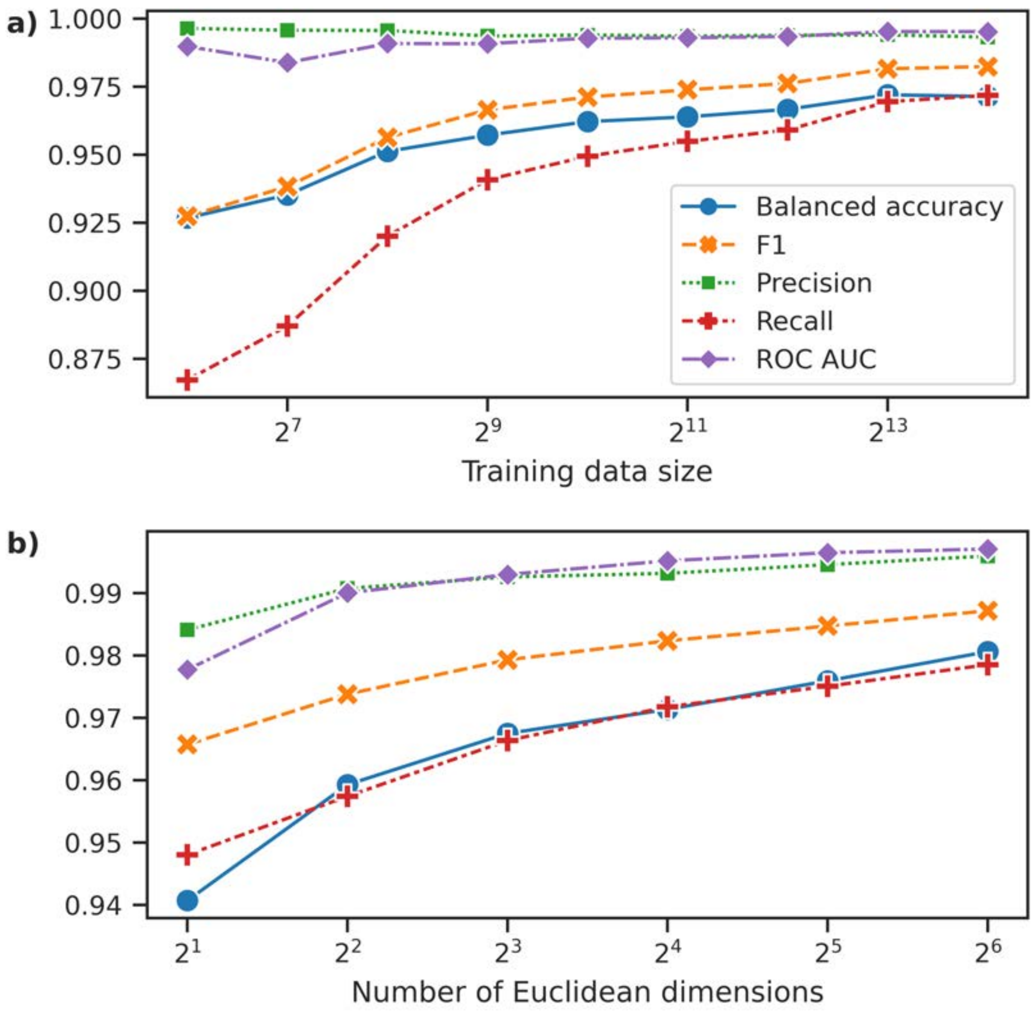
$$d_{ij} = 1 - \max(|o_i \star o_j|)$$



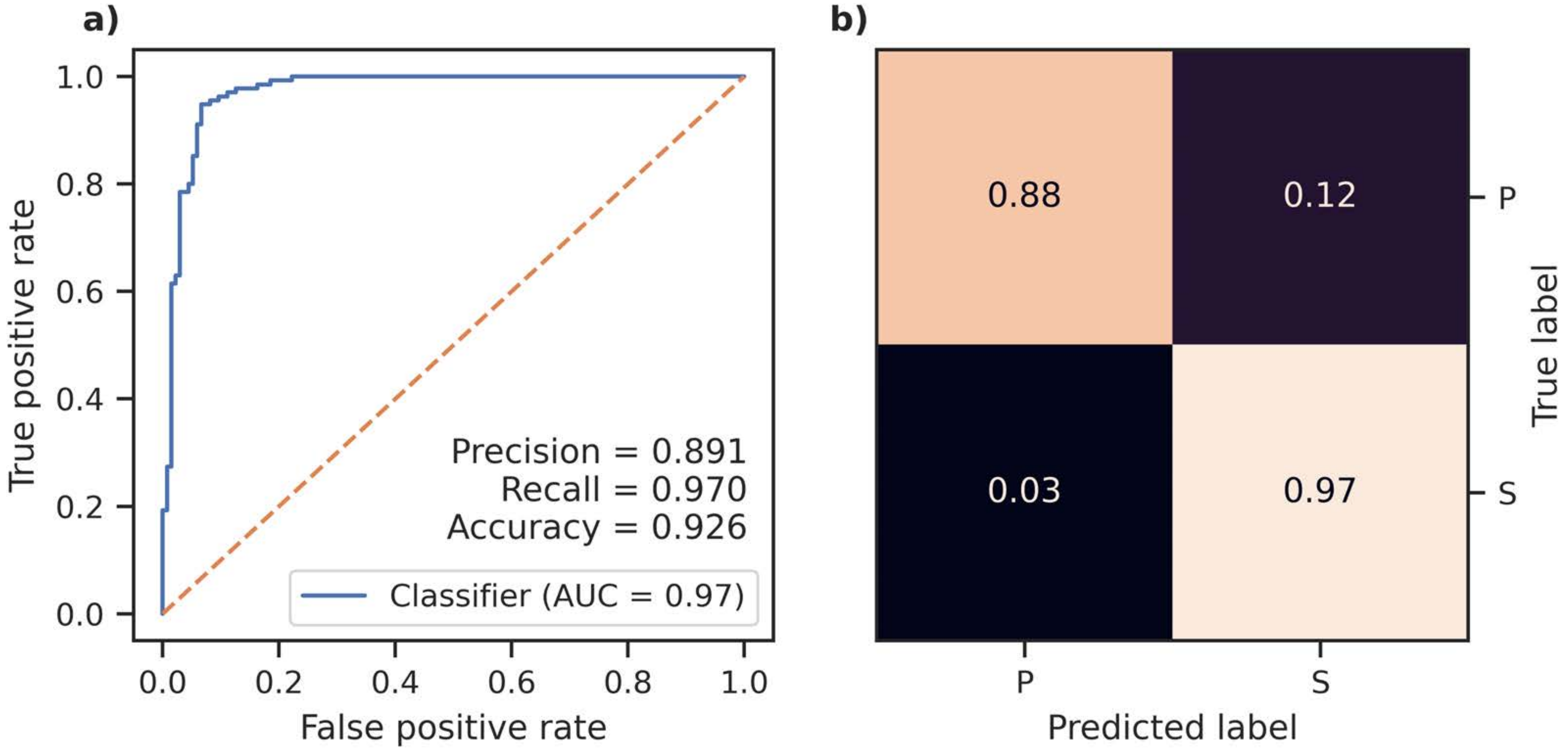
FastMapSVM robustly detects earthquakes in STEAD using ~1% of the data set for training (compared to 85% for EQTransformer and CRED).



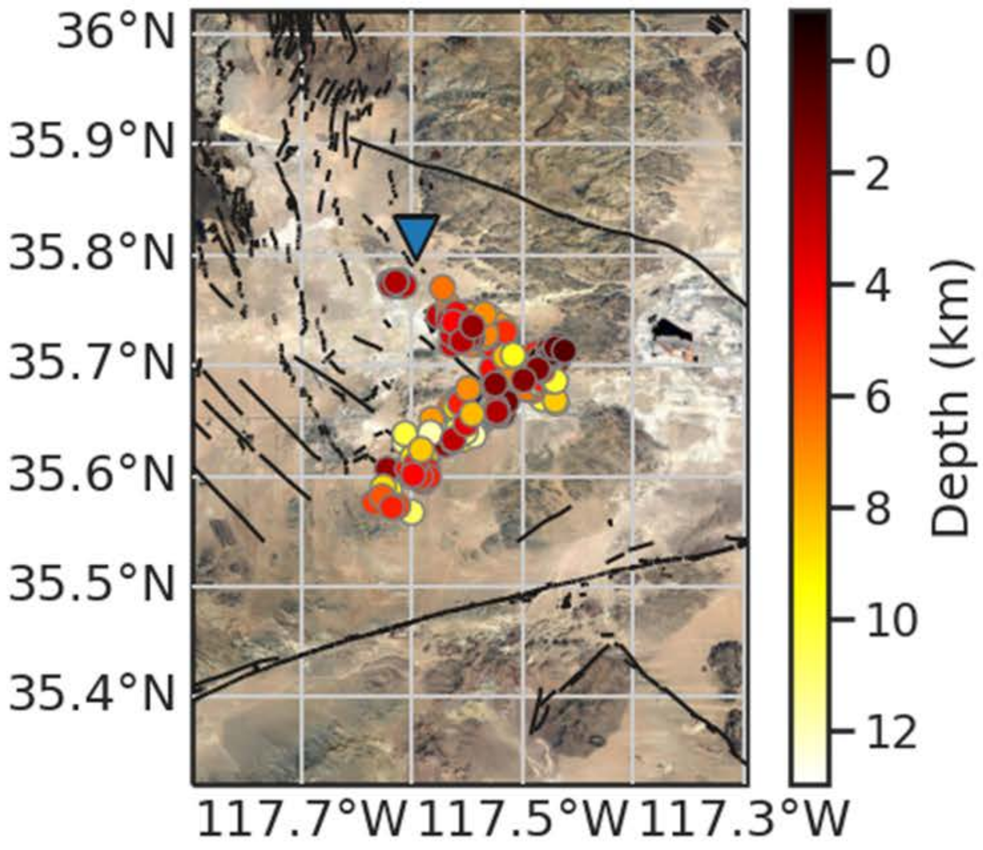
Returns diminish for increasing size of training data and Euclidean embedding

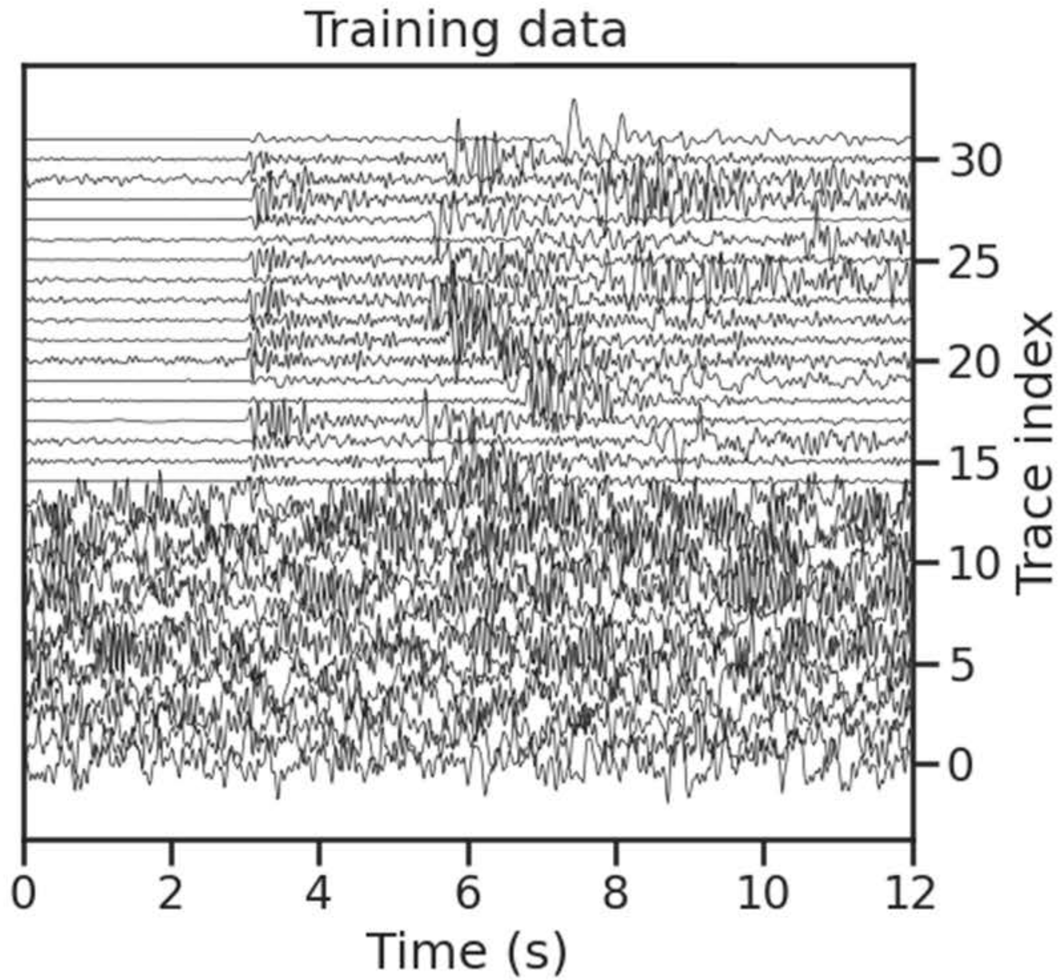


FastMapSVM can be easily tailored to different classification tasks, such as identifying phases.

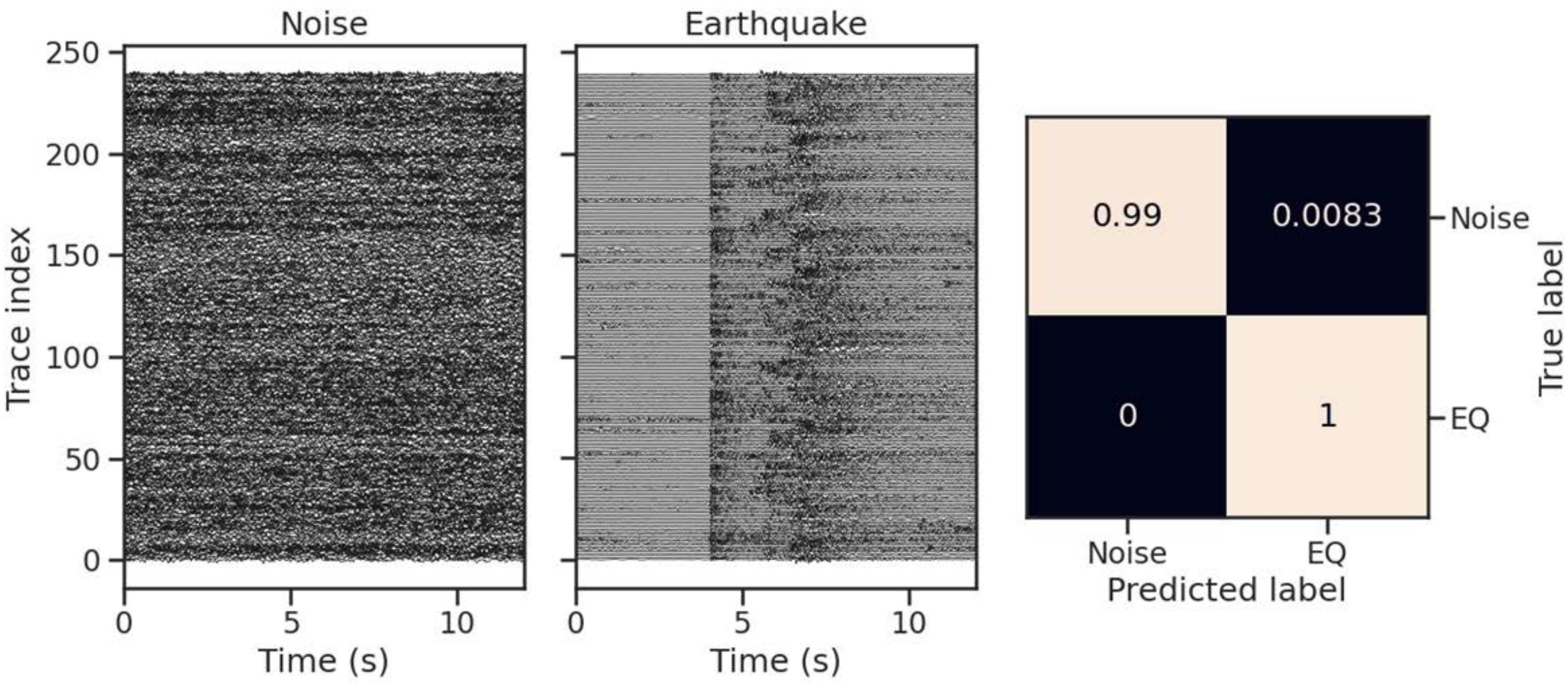


FastMapSVM trained to detect Ridgecrest aftershocks using only 32 training instances.

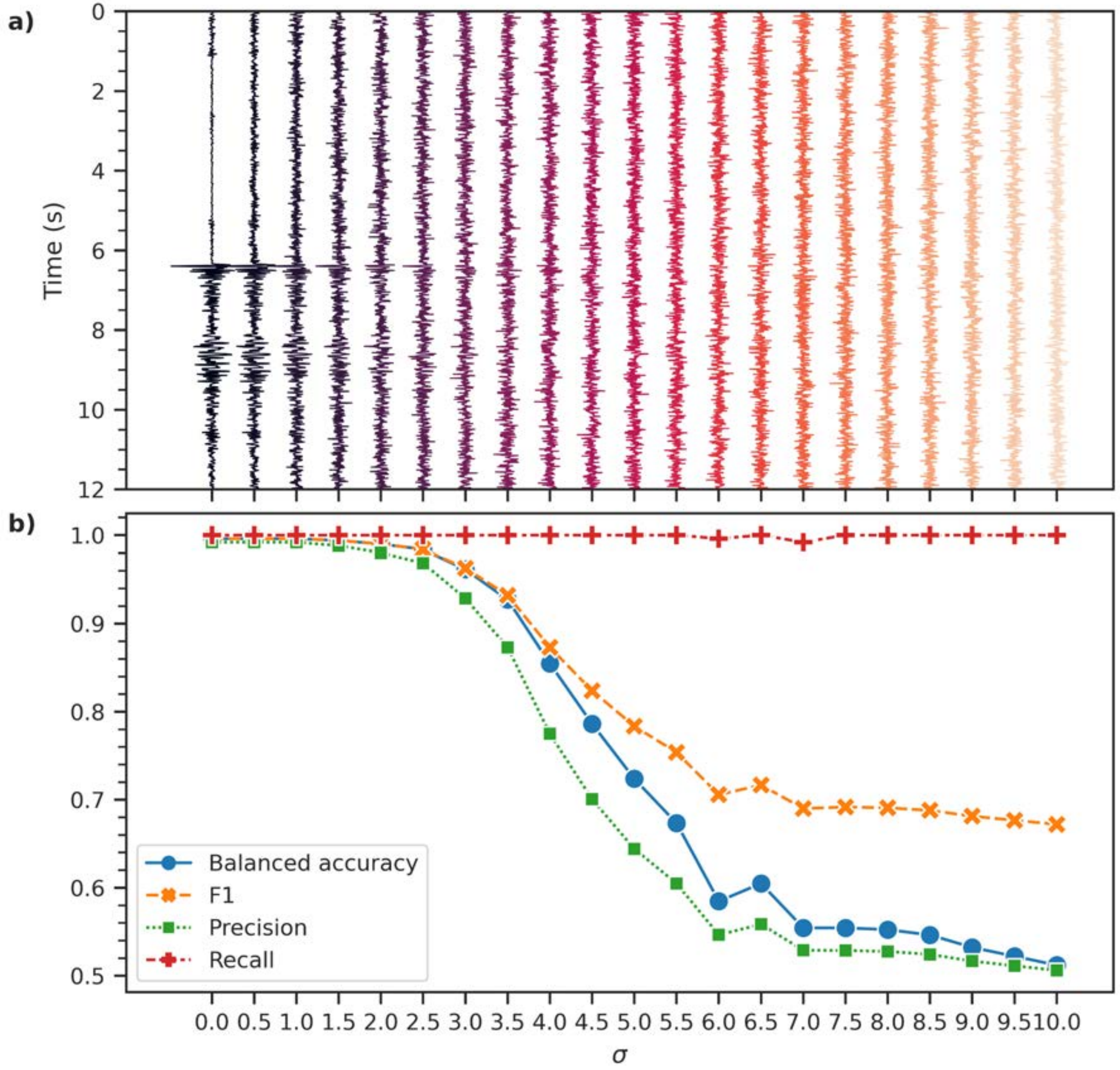




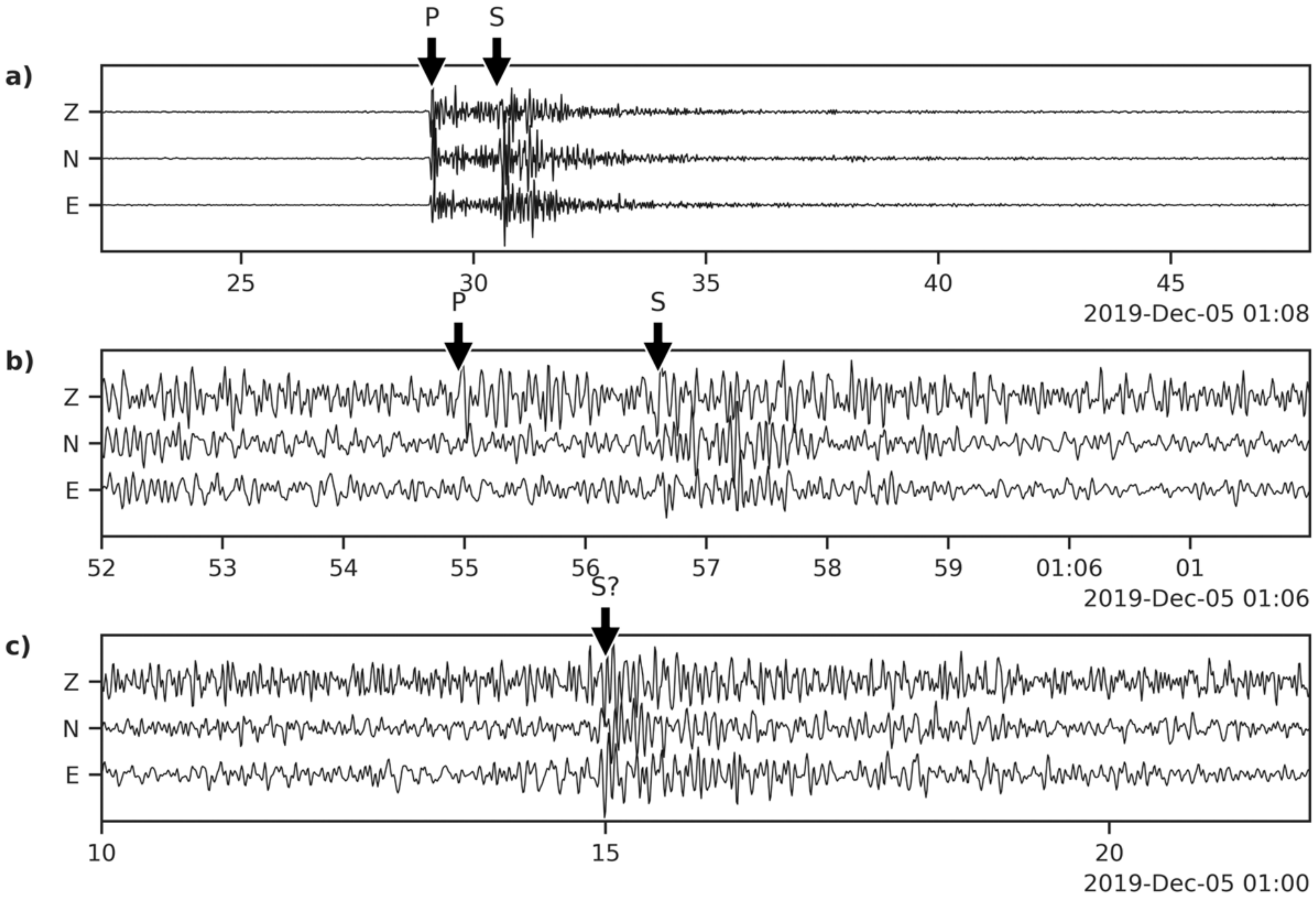
Classifies ~500 seismograms with nearly perfect accuracy.



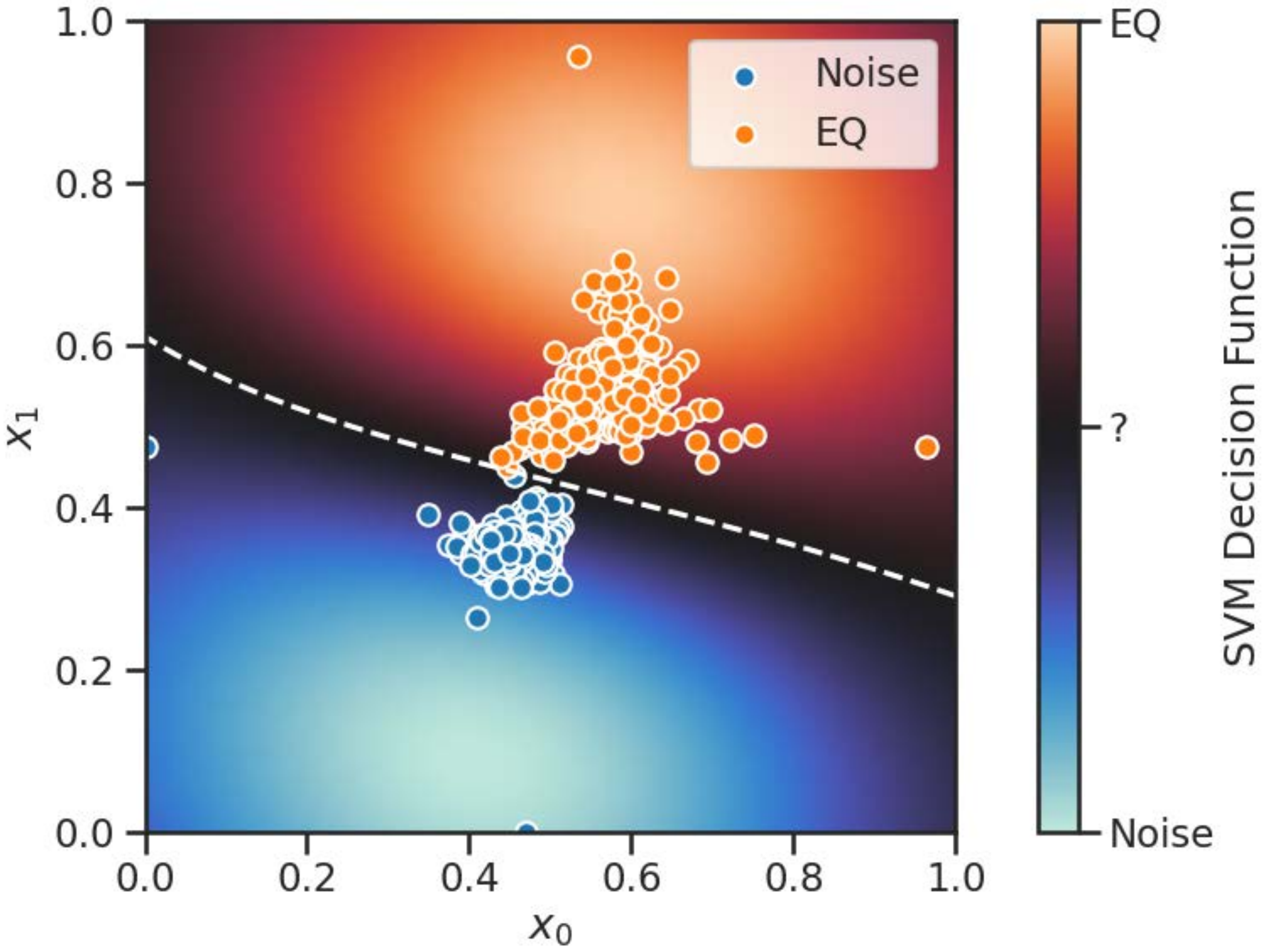
FastMapSVM is resilient against noisy perturbations.



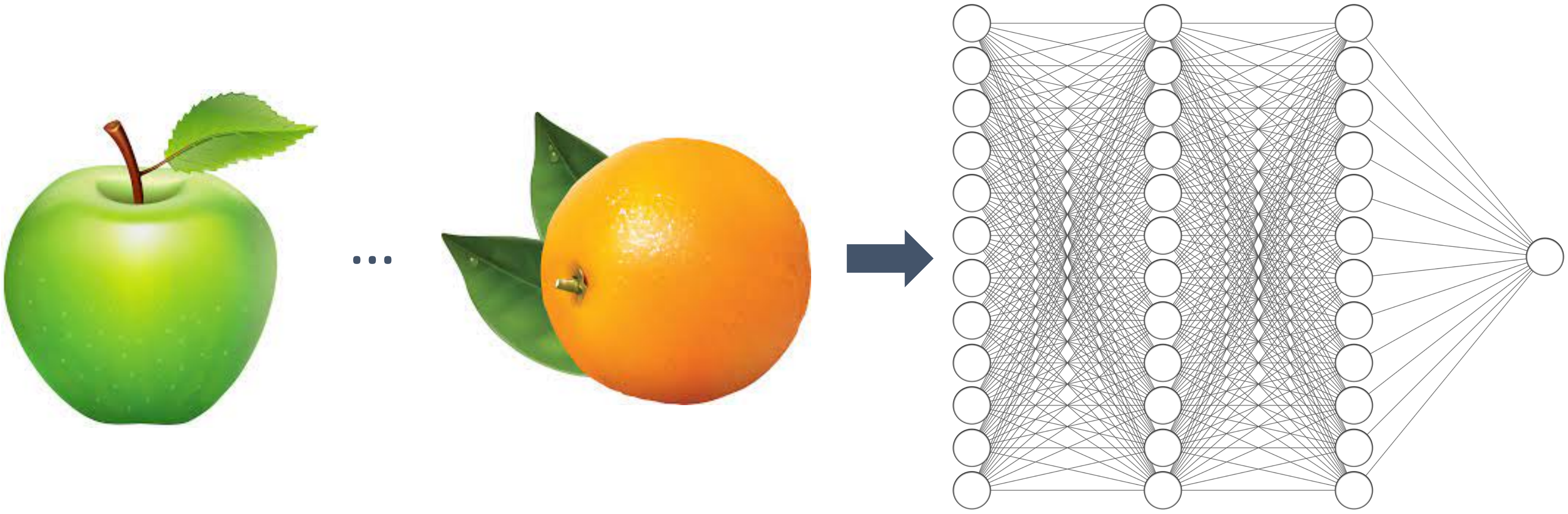
We identify 19 new events in 10 minutes of continuous data recorded by CI.CLC 5 months after Ridgecrest mainshocks.



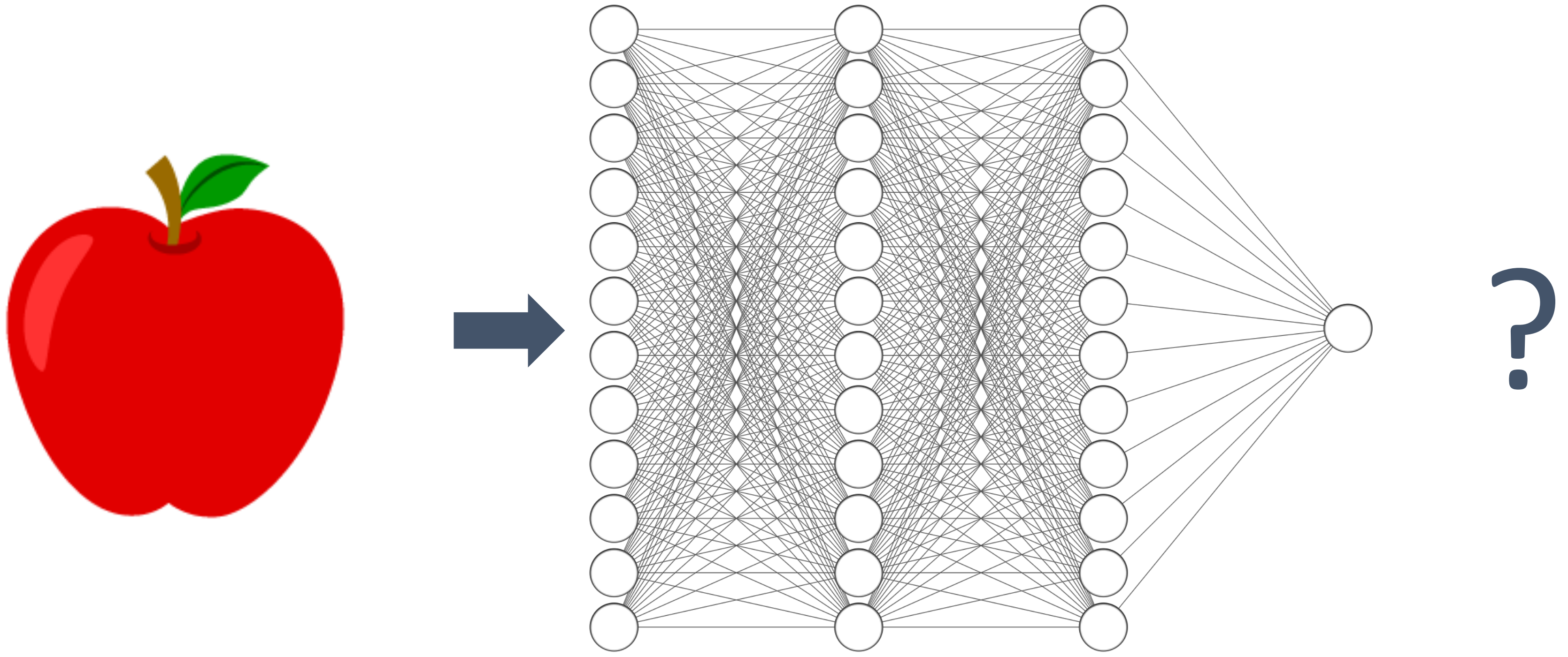
FastMapSVM yields a perspicuous visualization of objects and decision boundaries.



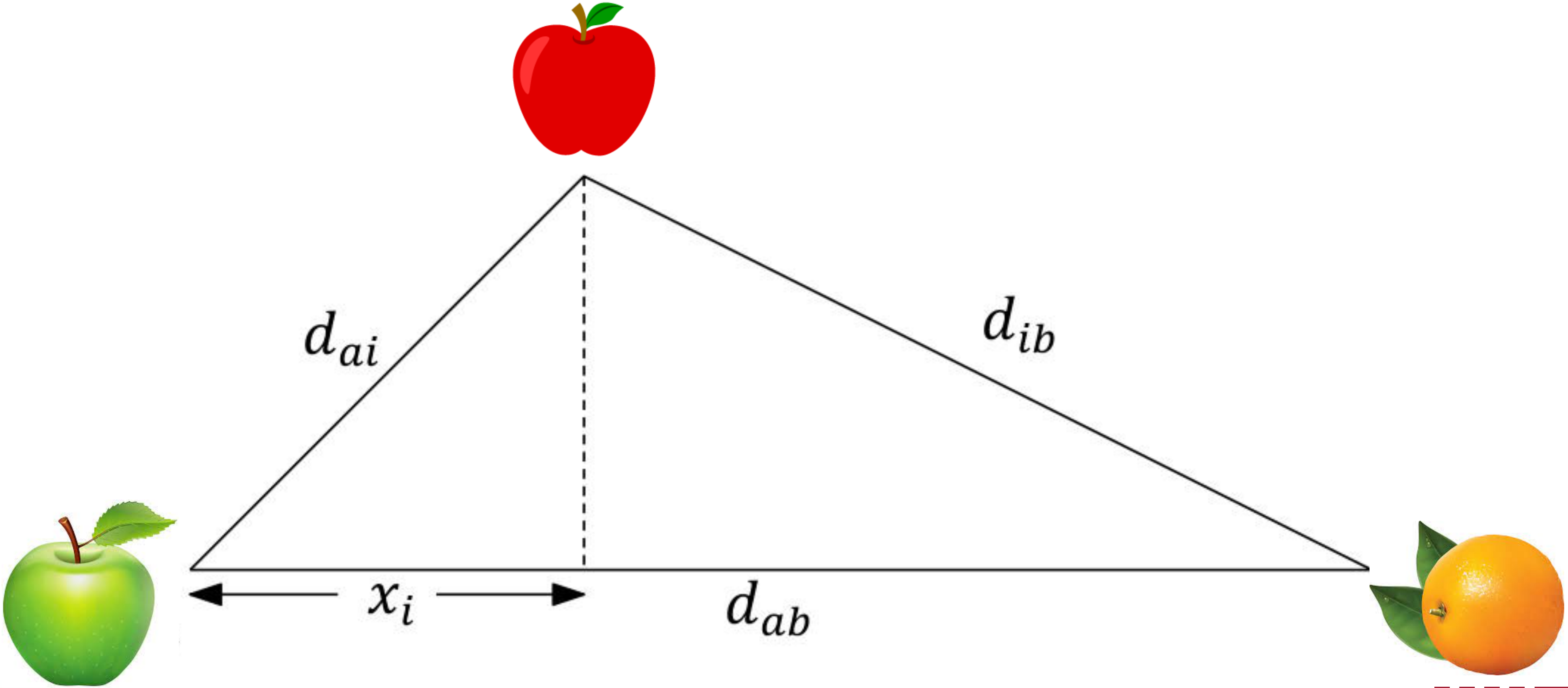
NNs learn from *individual* training instances; O(N) pieces of information.



NNs store an abstract representation of training instances.



FastMapSVM considers *pairs* of objects ; O(N^2) pieces of information.



In summary,

- FastMapSVM
  a) can be trained quickly using little training data;
  b) leverages $O(N^2)$ pieces of information in $O(N)$ time;
  c) integrates domain knowledge through the distance function;
  d) explicitly compares test objects against reference objects in original data domain; and
  e) yields a perspicuous visualization of objects and decision boundaries.
- https://github.com/malcolmw/FastMapSVM
- Preprint available on arXiv.