Center for Analysis Systems and Applications

Developing next-generation analysis capabilities

SAND2017-7293C

New Performance Metric for Approximate Nearest Neighbor Search

Cara Monical, University of Illinois at Urbana-Champaign

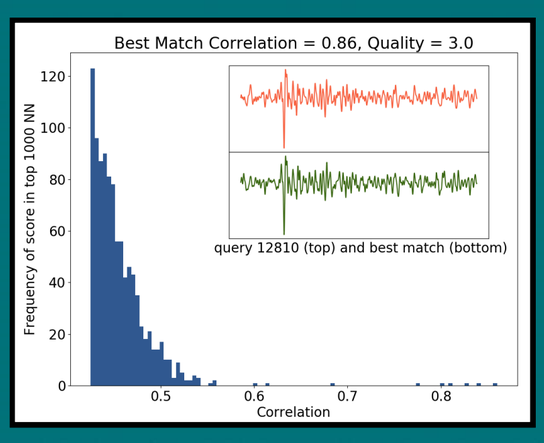
Project Mentors: Antonio Gonzales and Nicholas Blazier, Org. 06362

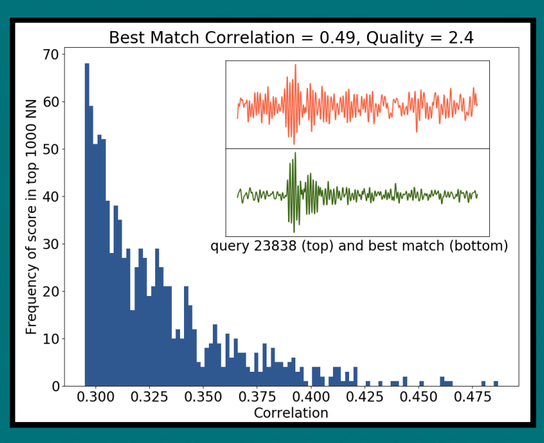
Problem Statement:

Goal: find nearest neighbors (NN) of a seismic waveform in a historical database

Motivation: potentially allows for quick identification of the source of the event

One obstacle: only close matches are useful – but what makes a match close?





Objective:

Develop a way to evaluate matches on a per-waveform basis and use that to judge performance of a system, both with and without ground-truth.

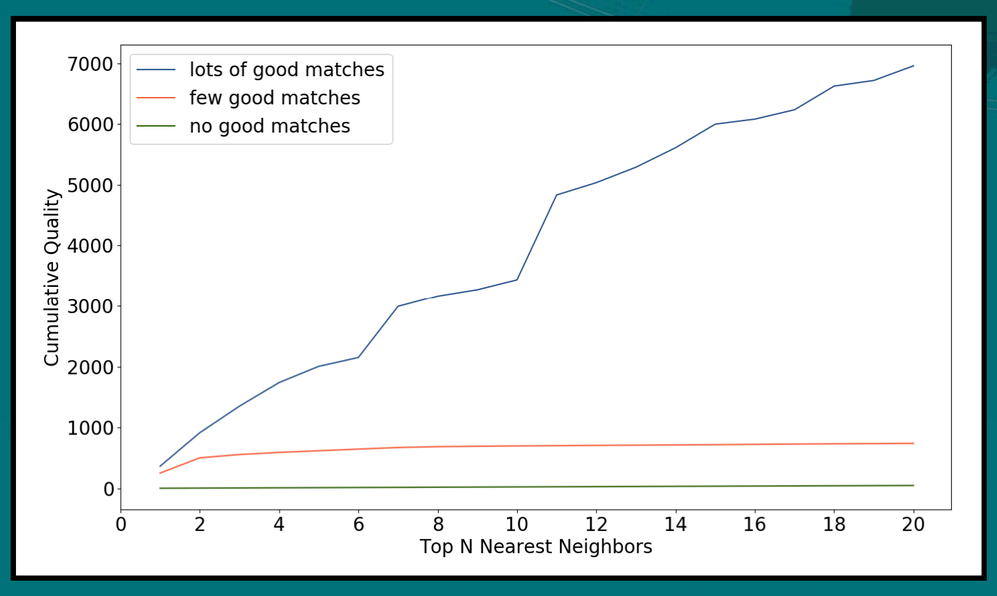
Approach:

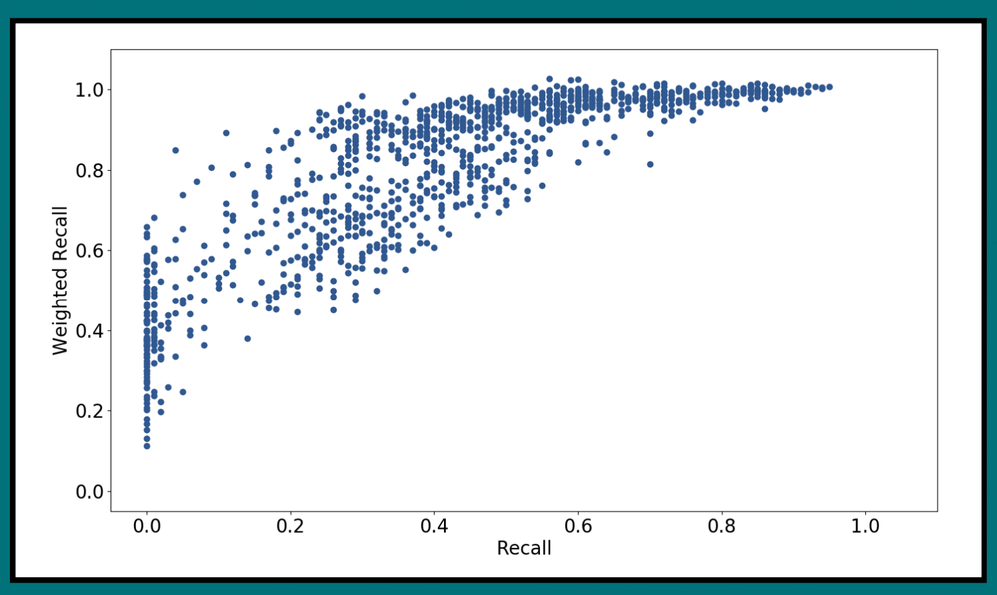
Three new notions that capture this:

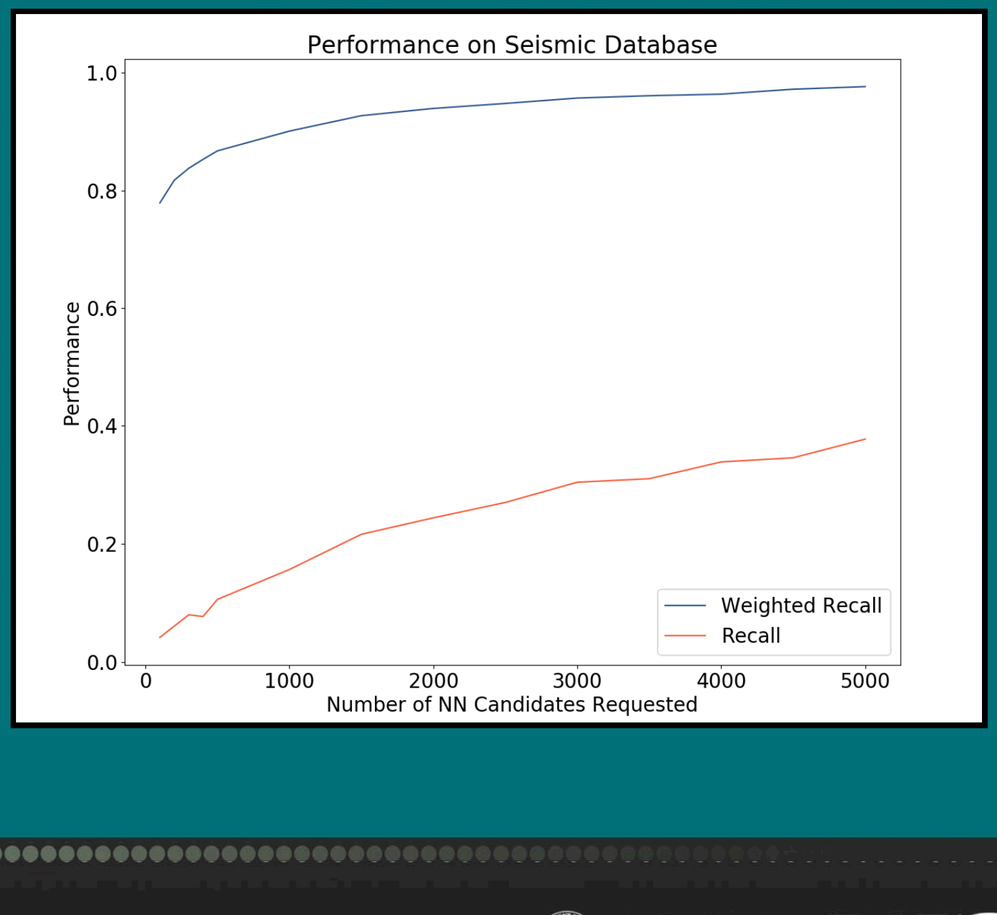
- *quality of a result* measures how "good" a match is on a per-waveform basis based on the statistics of the comparison of the waveform to a "typical" waveform
- *quality of a set of results* measures the average quality of each results, weighted so that "good" results contribute more
- *weighted recall* measures the percentage of "good" true matches returned

Metric	High	Low
Recall	Found high % of true NN	Found low % of true NN
Set Quality	Has good matches	Has no/few good matches
Weighted Recall	Found similar quality results to true NN	Missed good matches

Results:







Impact and Benefits:

- Significantly reduces the false positive rate compared to using correlation as a threshold
- Quality can be used in a real system to determine how likely the matches are to be true matches
- Weighted recall allows more nuanced performance analysis – can distinguish between missing the best results vs missing mediocre results and getting equivalent ones

Sandia National Laboratories

U.S. DEPARTMENT OF ENERGY

National Nuclear Security Administration