SAND2016-7122C

Using Approximate Nearest Neighbor Methods to Rapidly Search Large Seismic Signal Archives

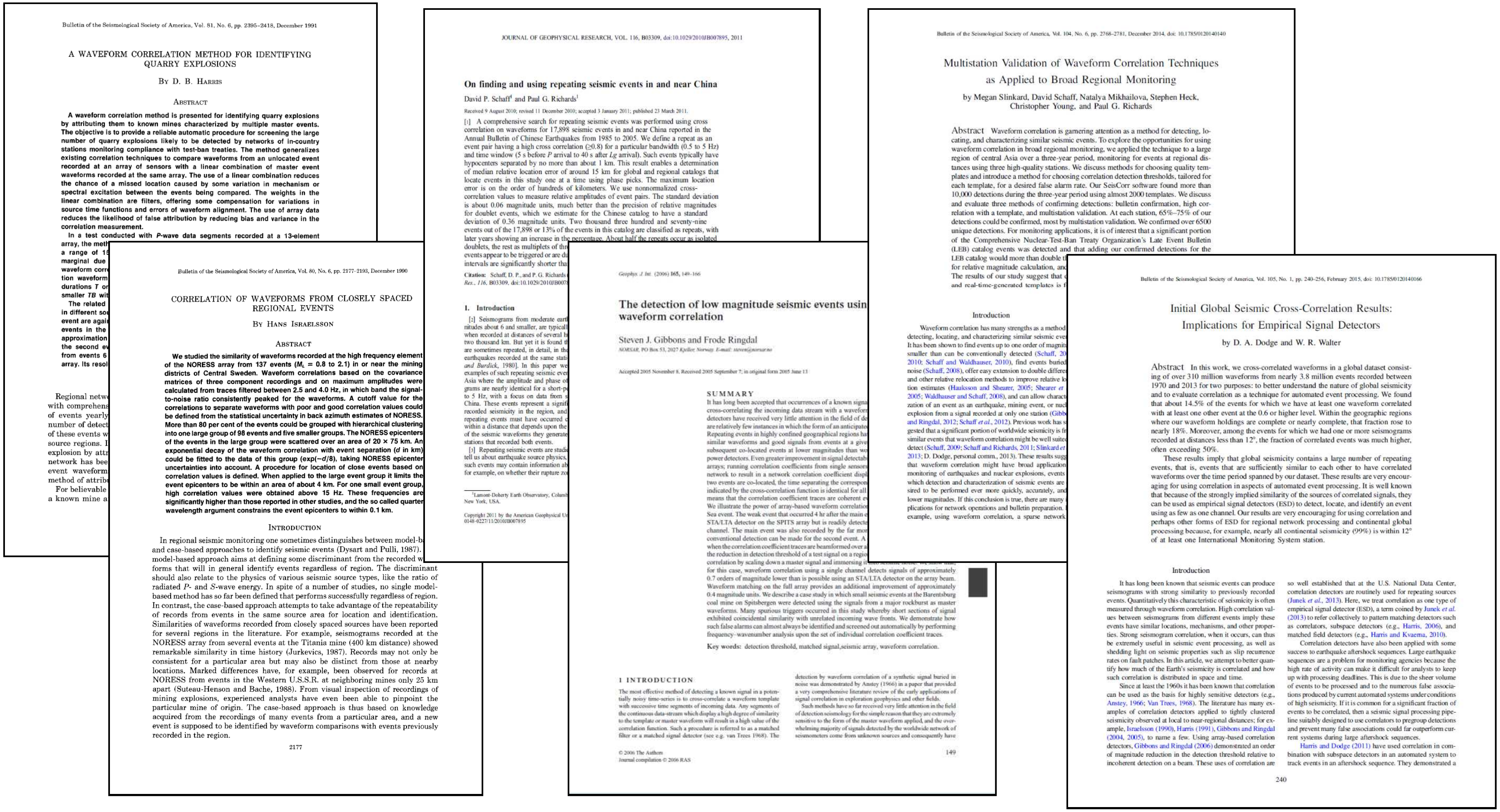
Chris Young & Antonio Gonzales

Sandia National Laboratories

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Motivation (1)

- **Waveform correlation has proven to be a very effective means to detect and identify signals from various types of events…**



Motivation (2)

- **However, the basic calculation (waveform correlation) is computationally expensive, so either the size of waveform archive used has been modest…**
  - ~3 templates, NK Test Site Monitoring (Schaff et al., 2012; Zhang & Wen, 2014)
  - ~200 templates, small aftershock sequence (Dodge & Harris, 2011)

- **…or complex distributed computing systems have been used**
  - ~5000 templates, broad regional monitoring (Slinkard et al., 2014 – used SNL developed distributed)
  - ~500M templates**, global scale research (Dodge & Walter, 2015)

***this study did NOT perform scanning of continuous data but rather performed cross-correlations of windowed signal detection waveforms in the LLNL GNEMRD database*

Objective

- **Utilize Approximate Nearest Neighbor (ANN) methods developed in other domains (e.g. searching for images) to enable searching of very large waveform archives without requiring a complex distributed computing system.**

Test Case

- **Our test case: 14 Years of IDC LEB MKAR Signal Detections (2002-2014)**
  - Archive: 2002-2013
    - 308K analyst reviewed arrivals (240K events)
  - Test Set: 2014
    - 26K analyst reviewed arrivals
  - 60% teleseismic P, 14% regional
  - Used only MK01/SHZ channel (no array processing)
  - 0.5-5.0 Hz bandpass filter (generic multi-purpose)

Results

– **20x speedup** finding highly correlated waveforms

– How? In exchange for finding 80% of the events in the archive

– Quick to set up: ANN Index for 308K waveforms builds in 180 seconds

– Quick to query: ~30msec for each search, ~1sec to calculate correlation scores

– All on a standard desktop computer

Approximate Nearest Neighbor Search Methodology

Approximate Nearest Neighbor (ANN) Search

Return an approximation of the true nearest neighbor set from an archive of data …

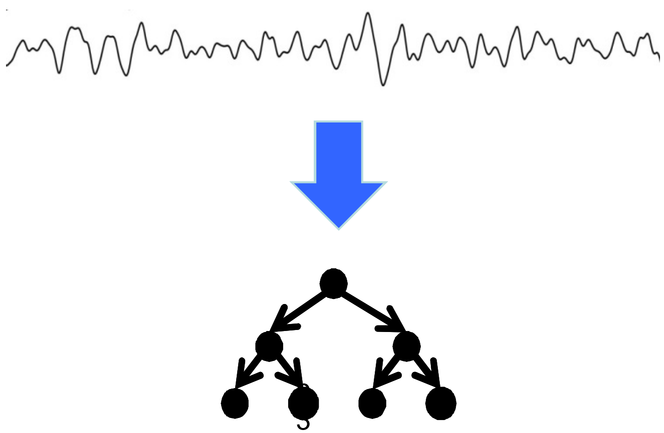
In exchange for large increases in query speed over standard linear search

ANN Application

- Best suited for tasks where the database is expected to have large number of useful neighbors and the specific subset of those returned is not critical
  - Quickly eliminating events from common seismic zones or known mining operations
- Not suited for “Needle in a Haystack” Problems
  - Identifying NK Test Events

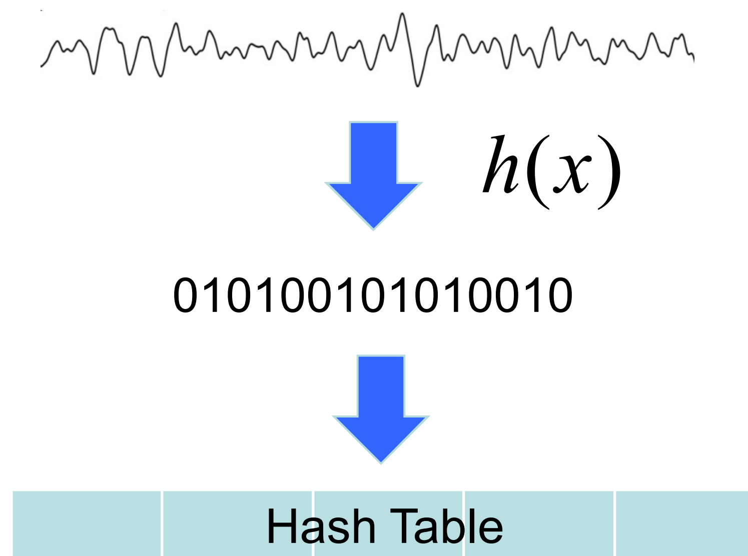
ANN Index Types

Tree Based



Euclidean Similarity Only

Hashing Based



ANN Index Types

Challenge for Seismic Waveforms

Tree Based

Find projection function $p(x)$

$$euclidean(p(x), p(y)) \propto correlation(x, y)$$

Hashing Based

Develop hash function

$$h(x)$$

KPCA Data Projection

Raw Data (19s at 40hz) = 760 dimensions

Kernalized Principal Component Analysis

$$euclidean(kpca(x), kpca(y)) \propto correlation(x, y)$$

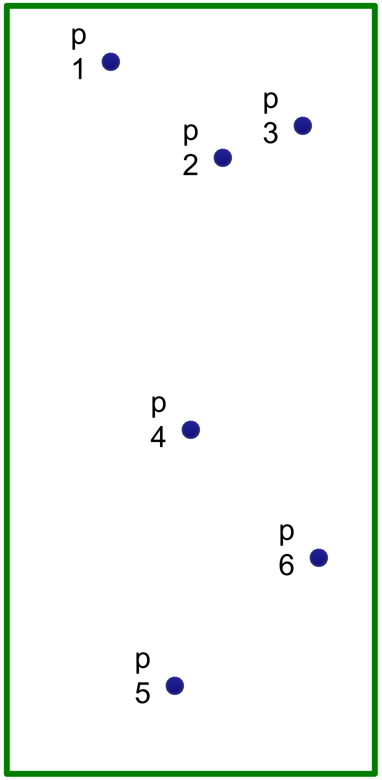
dimensionality reduction

Projected Data = 200 dimensions

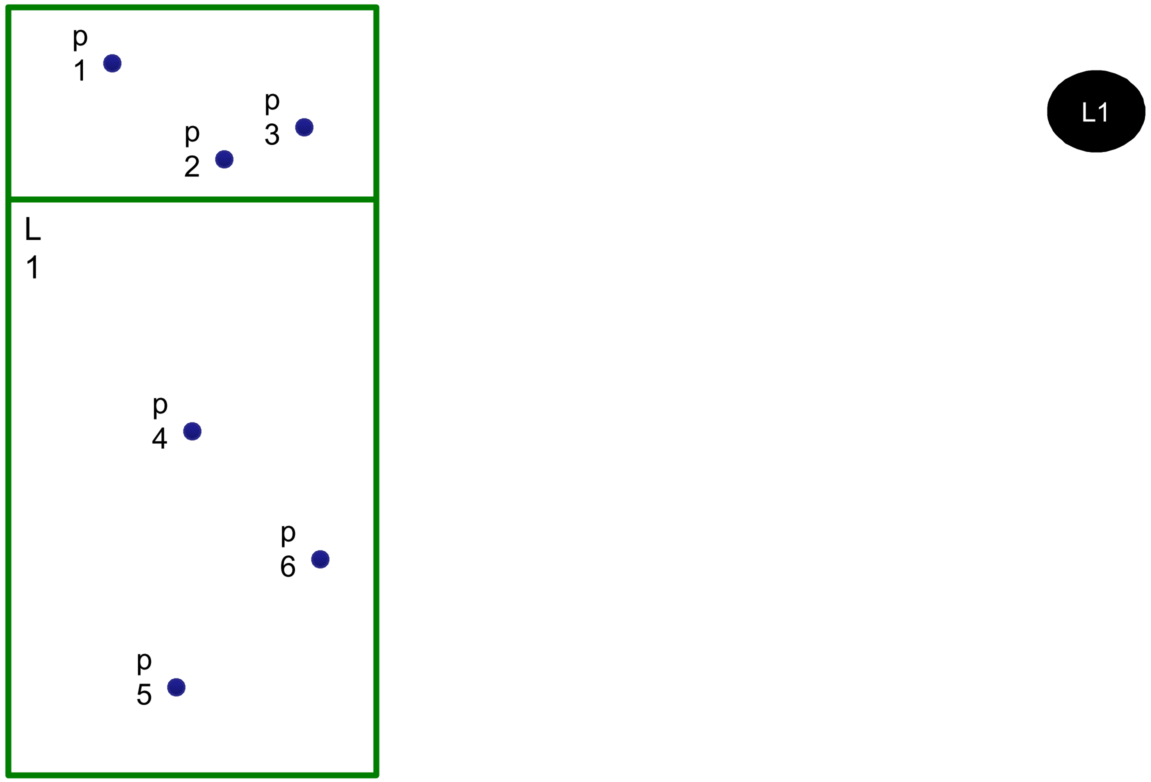
KD-Trees

- K-Dimensional Tree (KD-Tree)
  – Partition a large database of N k-dimensional objects using a tree structure
  – log(N) search time

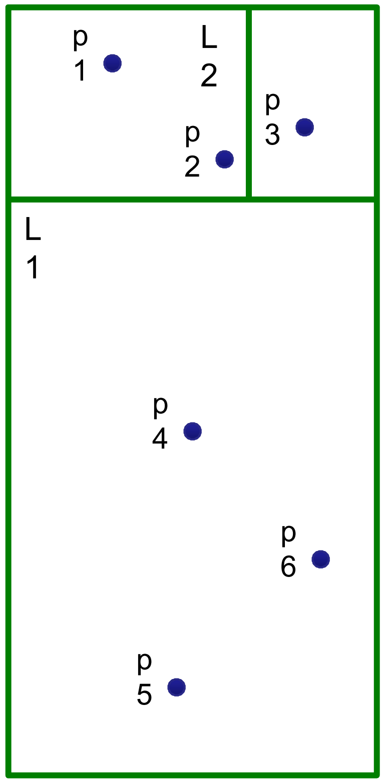
Building a KD-Tree

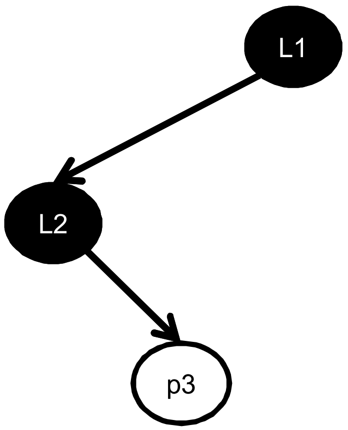


Building a KD-Tree

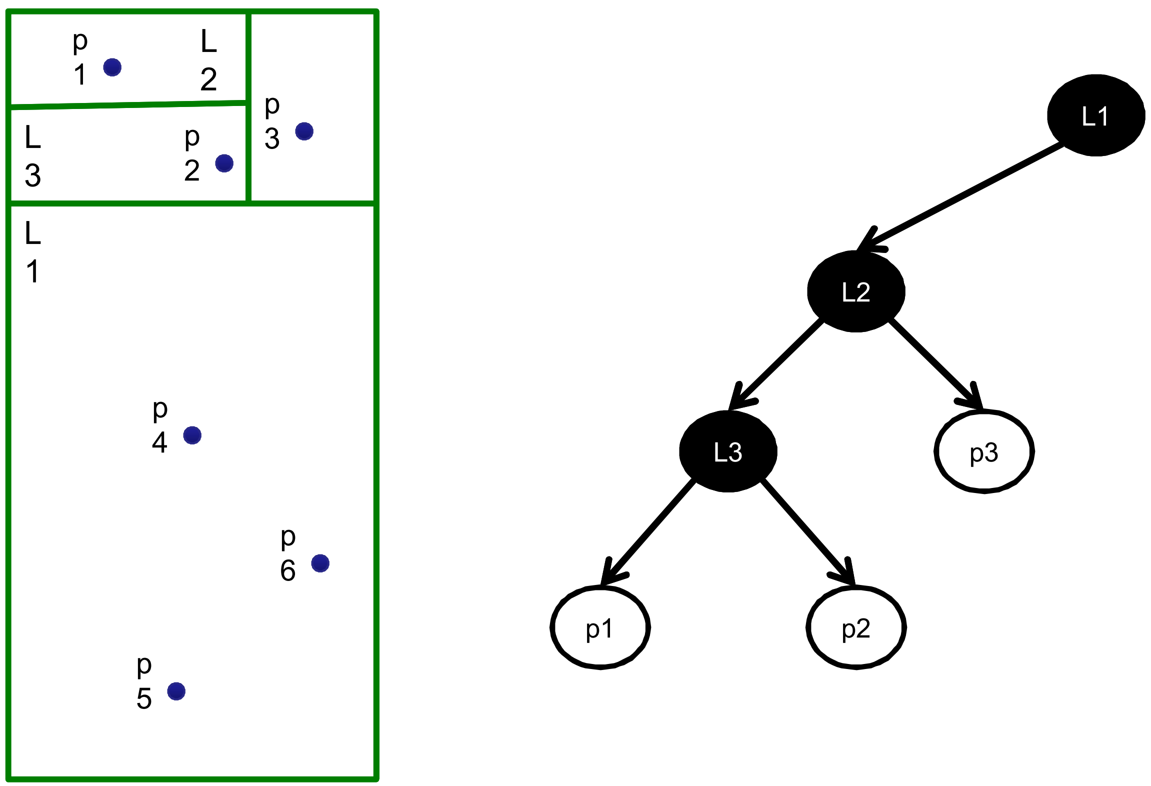


Building a KD-Tree

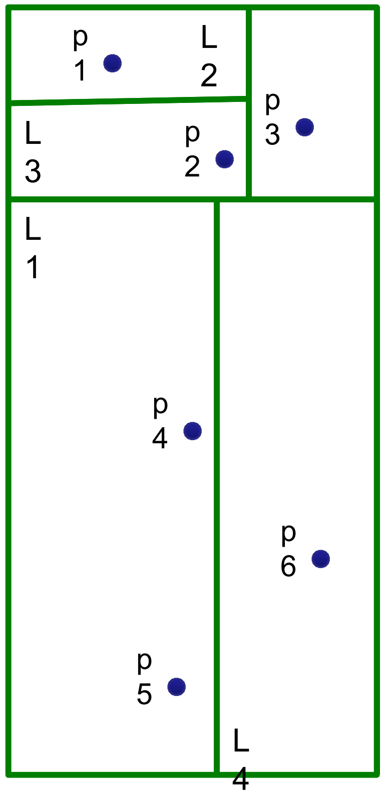


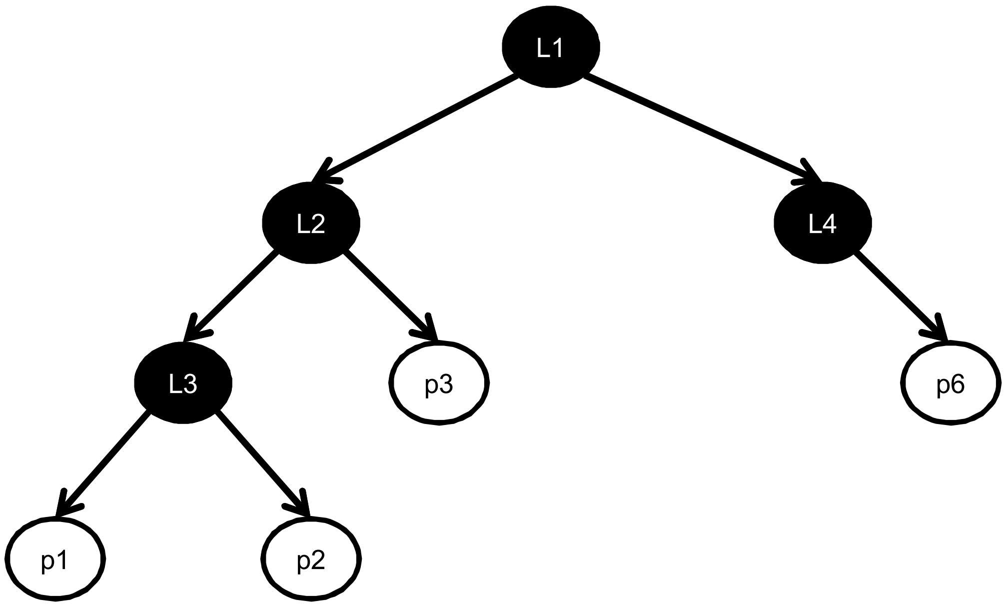


Building a KD-Tree

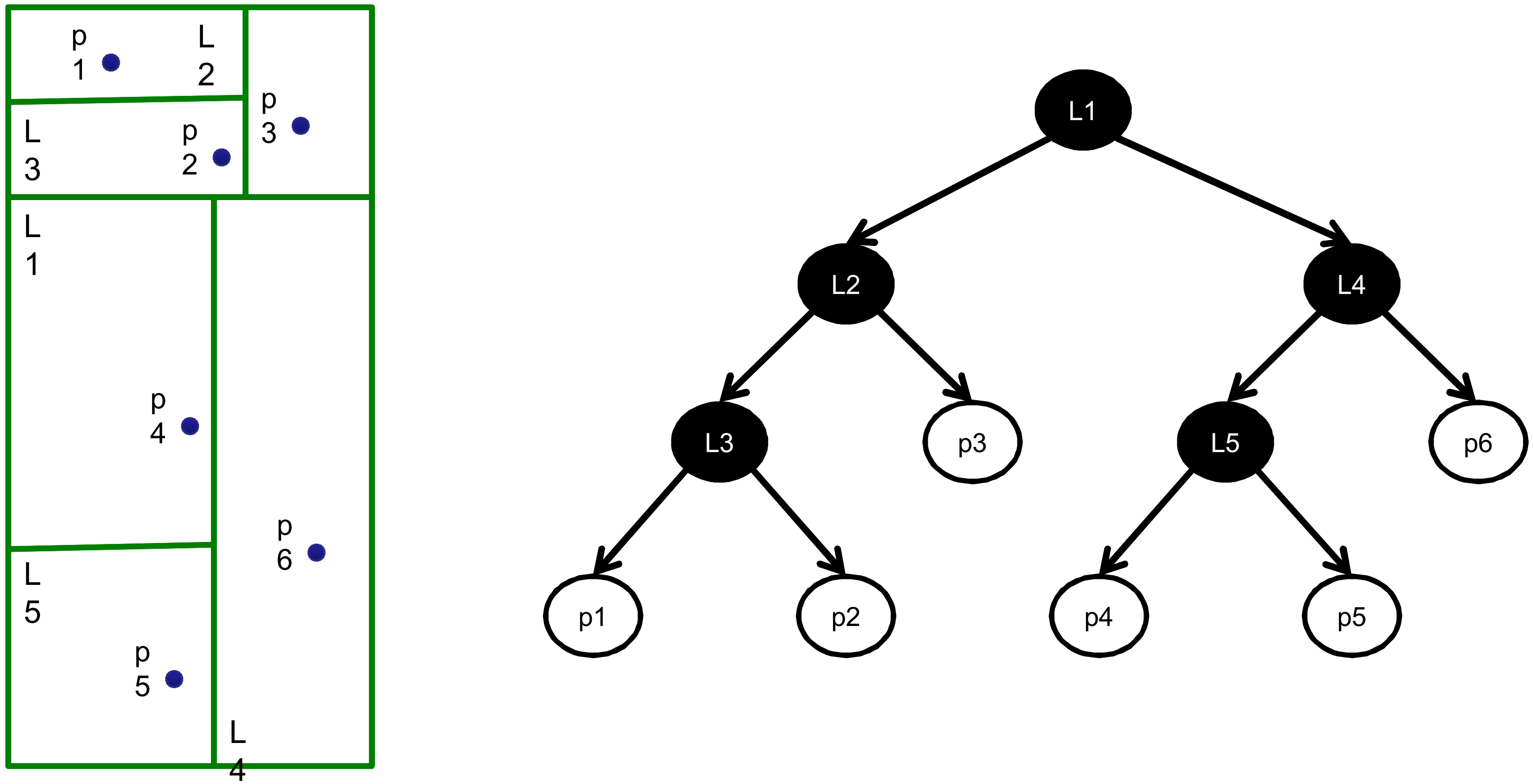


Building a KD-Tree

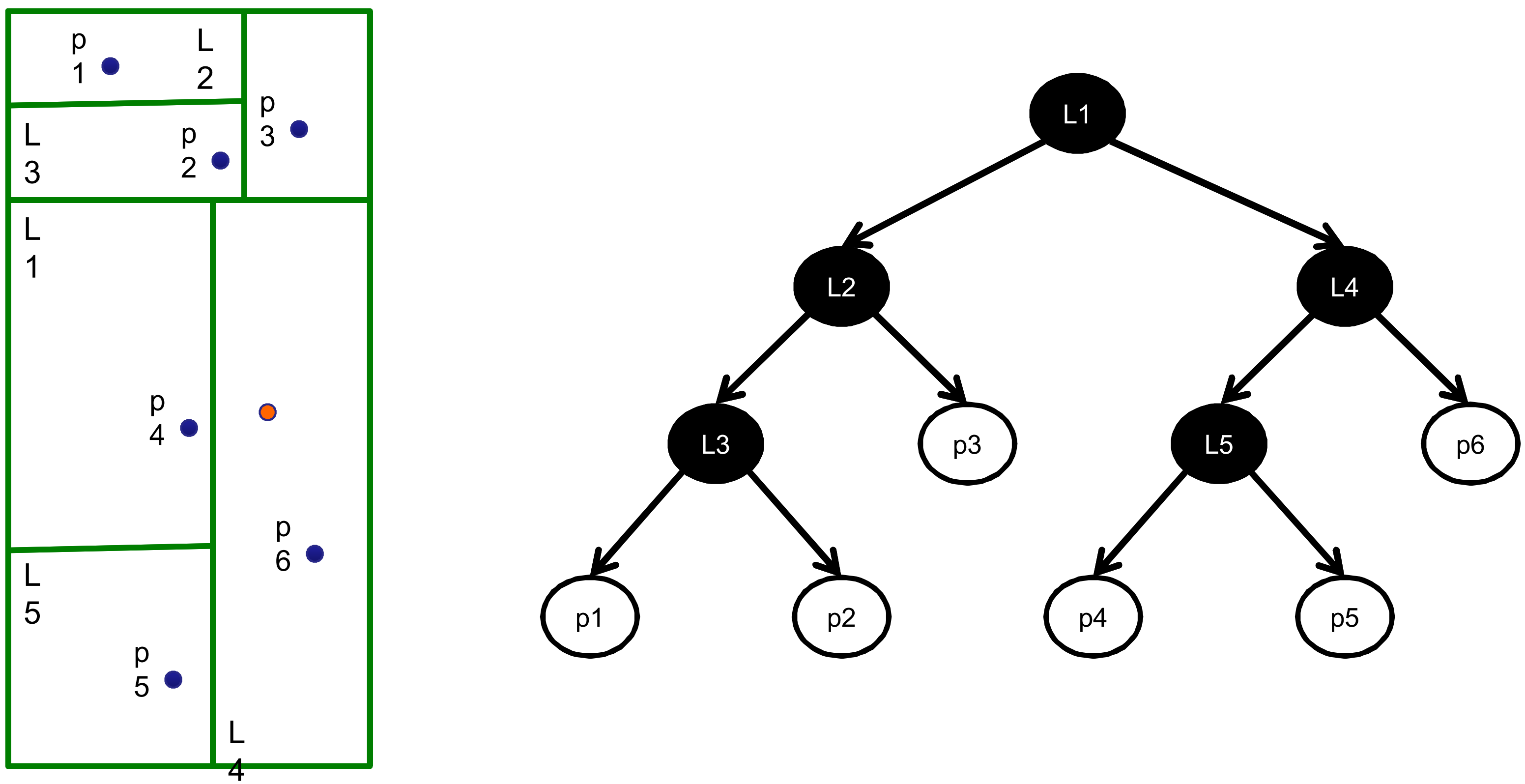




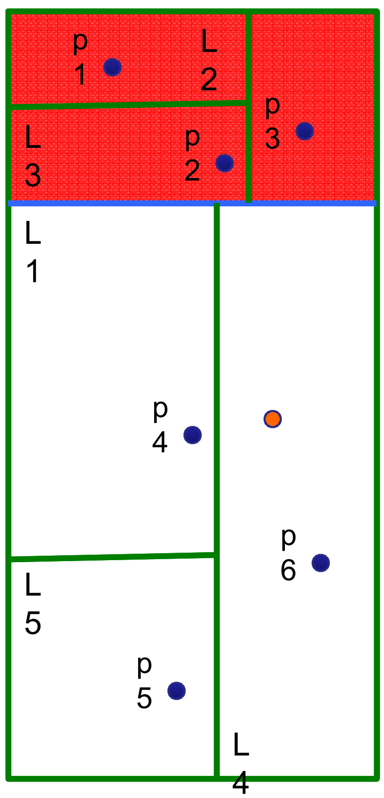
Building a KD-Tree

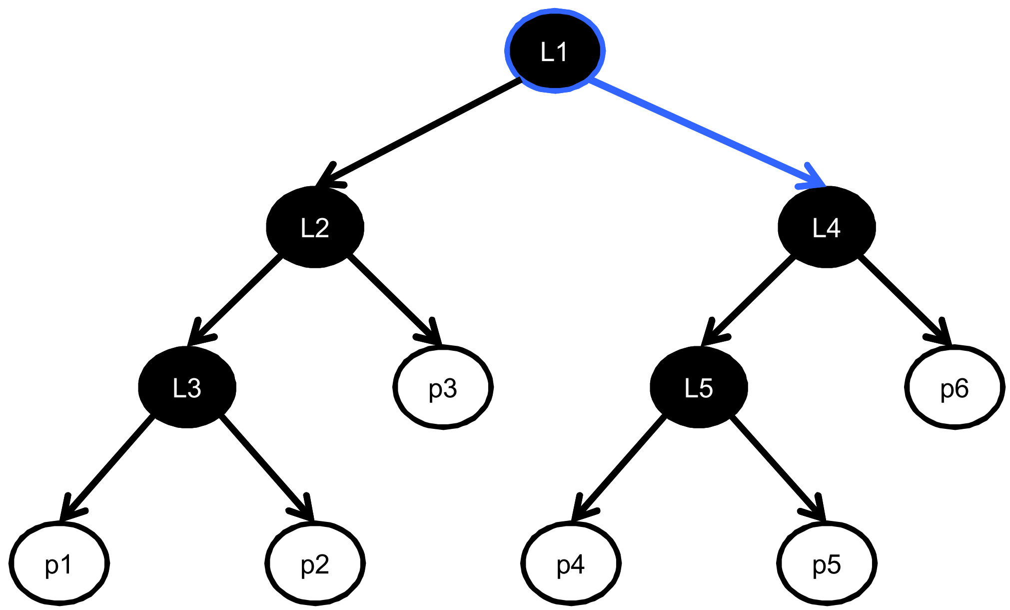


Querying a KD-Tree

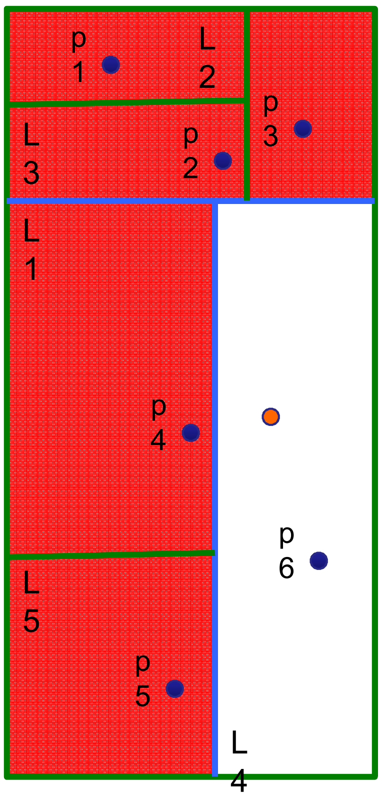


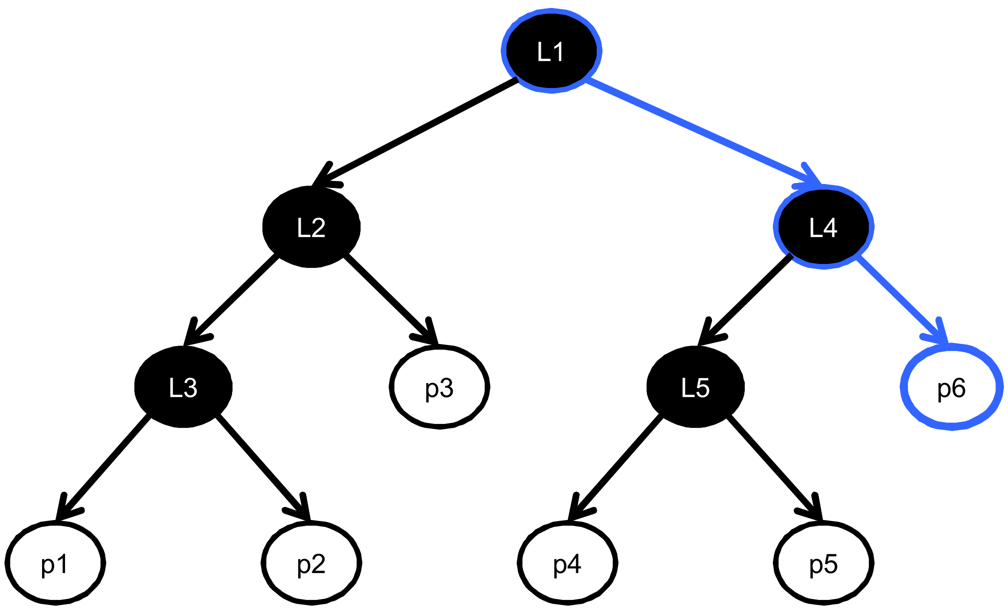
Querying a KD-Tree





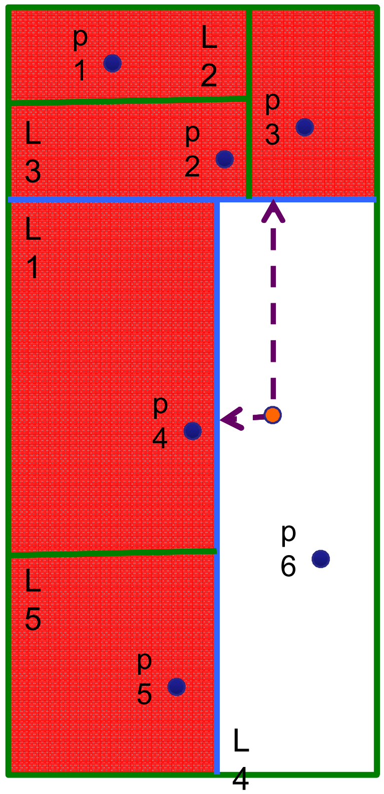
Querying a KD-Tree

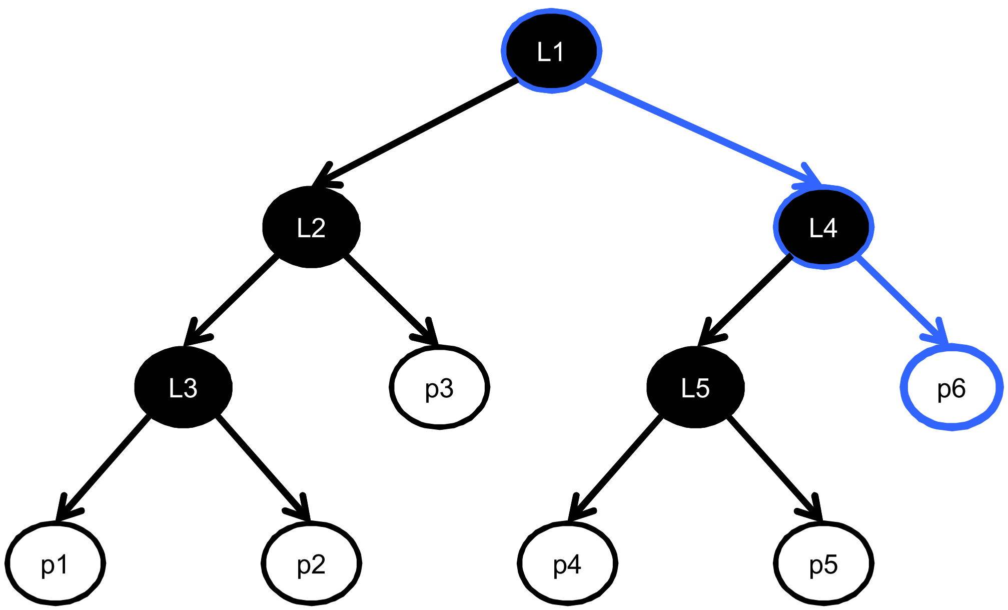




Found Neighbors: p6

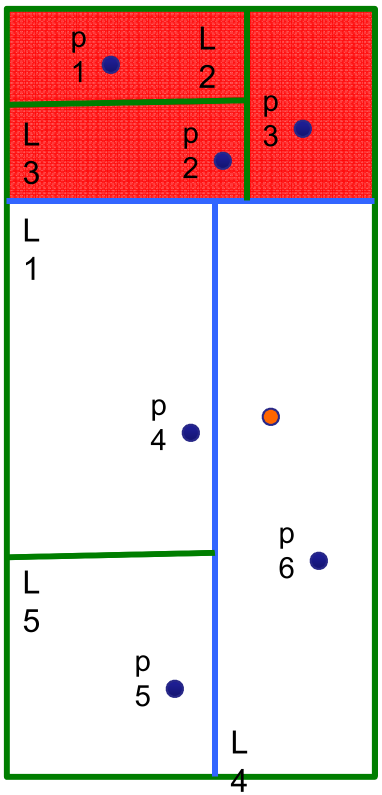
Querying a KD-Tree: Best Bin First

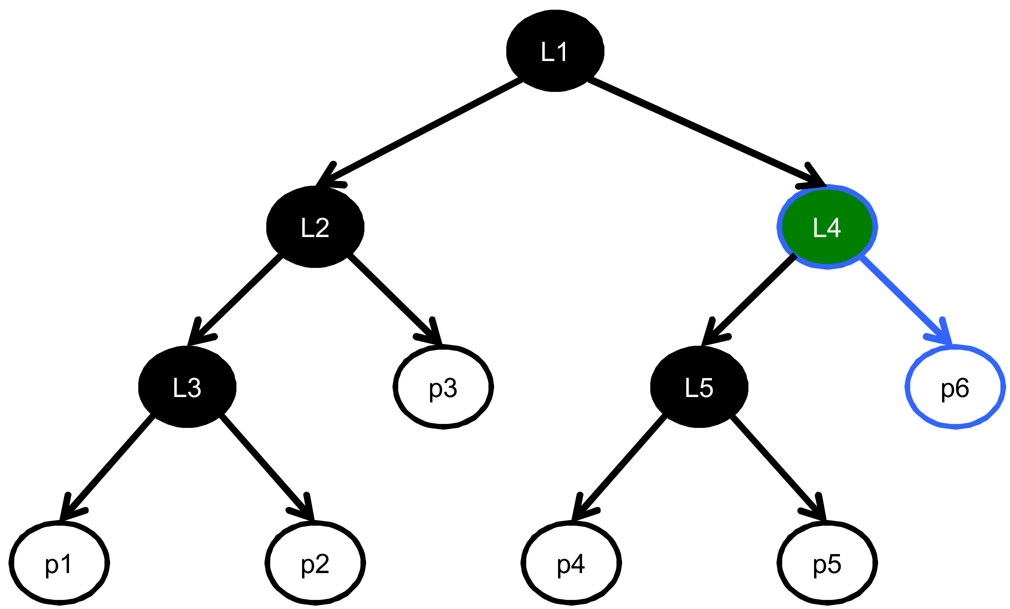




Found Neighbors: p6

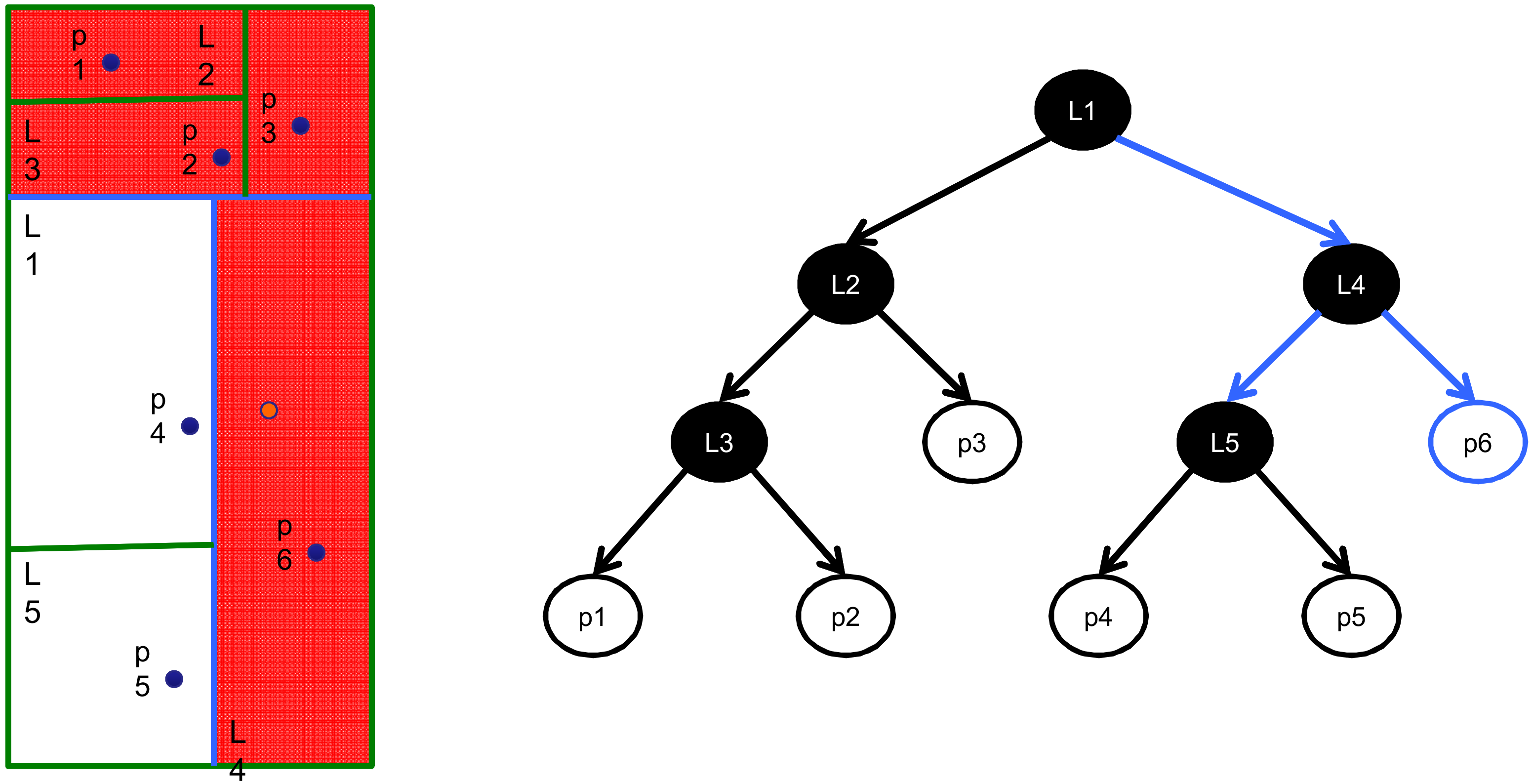
Querying a KD-Tree: Best Bin First





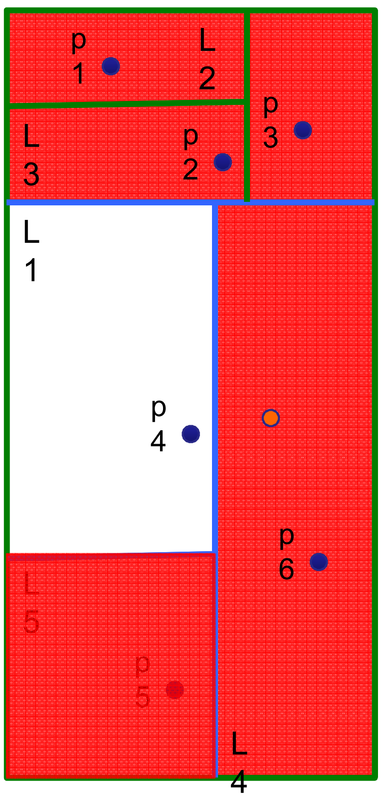
Found Neighbors: p6

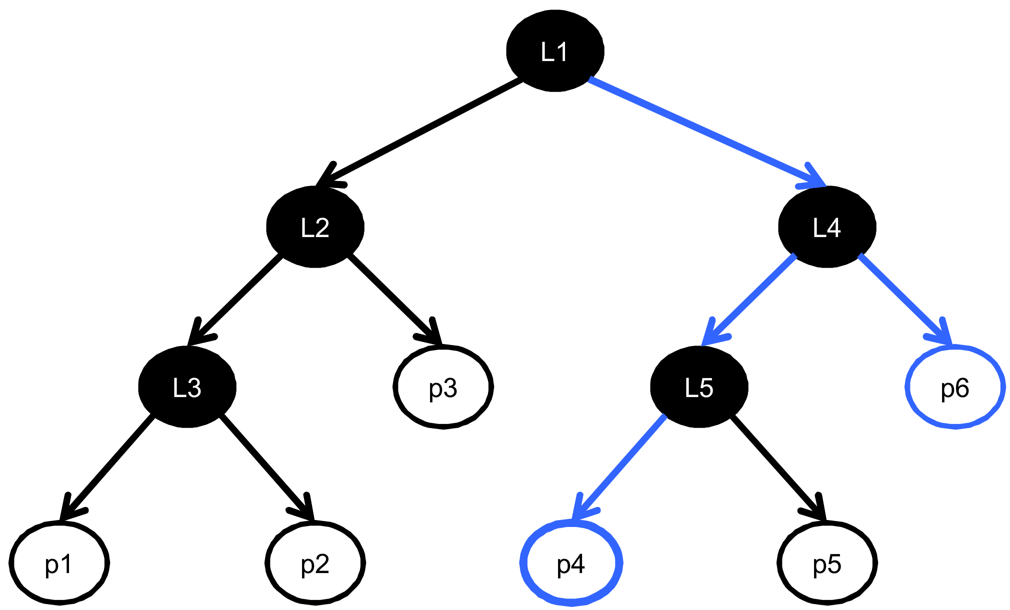
Querying a KD-Tree



Found Neighbors: p6

Querying a KD-Tree





Found Neighbors: p6, p4

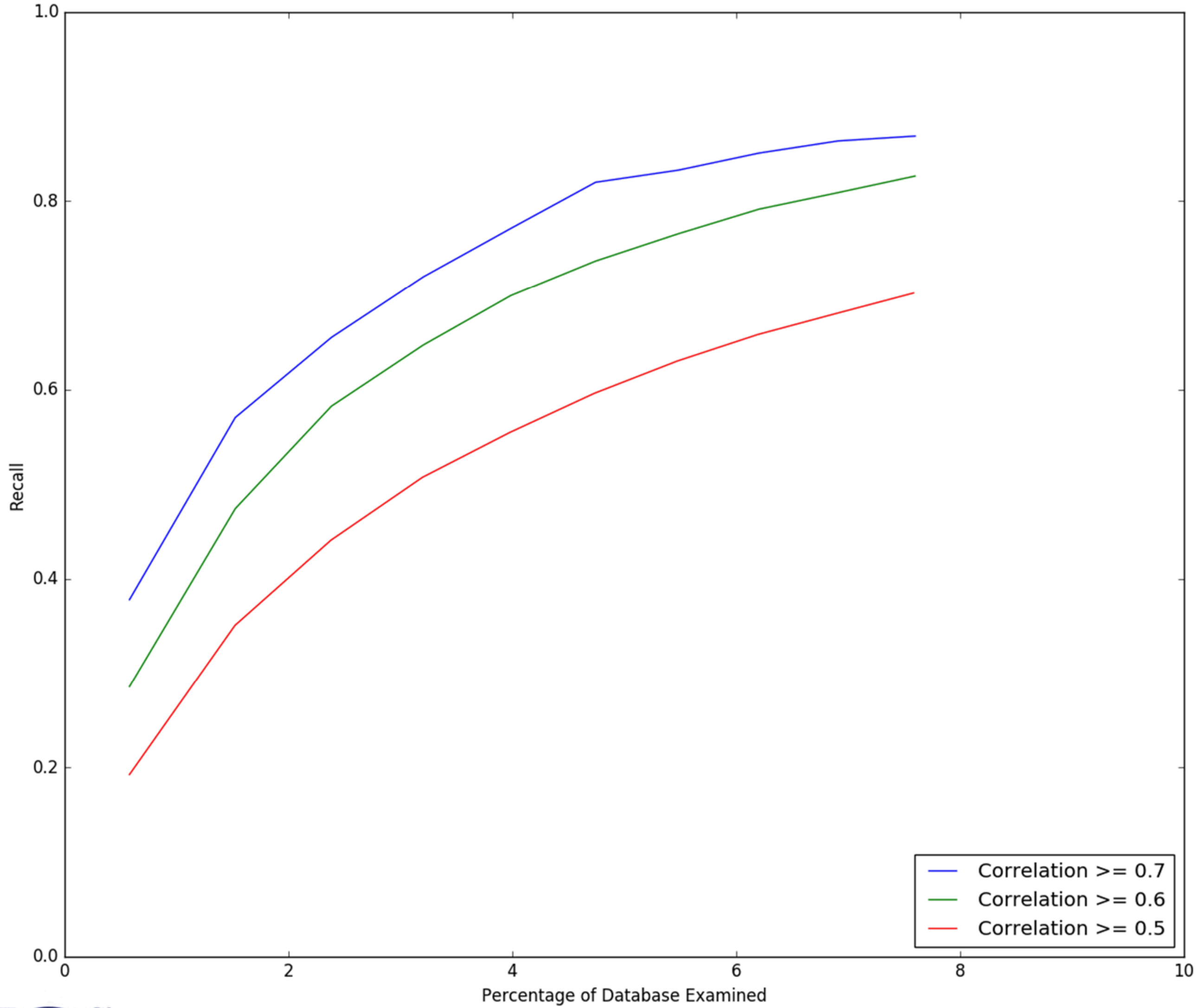
KD-Tree Forests

- KD-Tree Forest: build an index from a group of randomly varied KD-Trees and query them collectively
  - During each split decision, randomly select from the top *n* highest variance dimensions
  - During query, perform Best-Bin-First collectively across all trees
- For seismic case, we use forests with 100 trees

ANN Queries in Practice

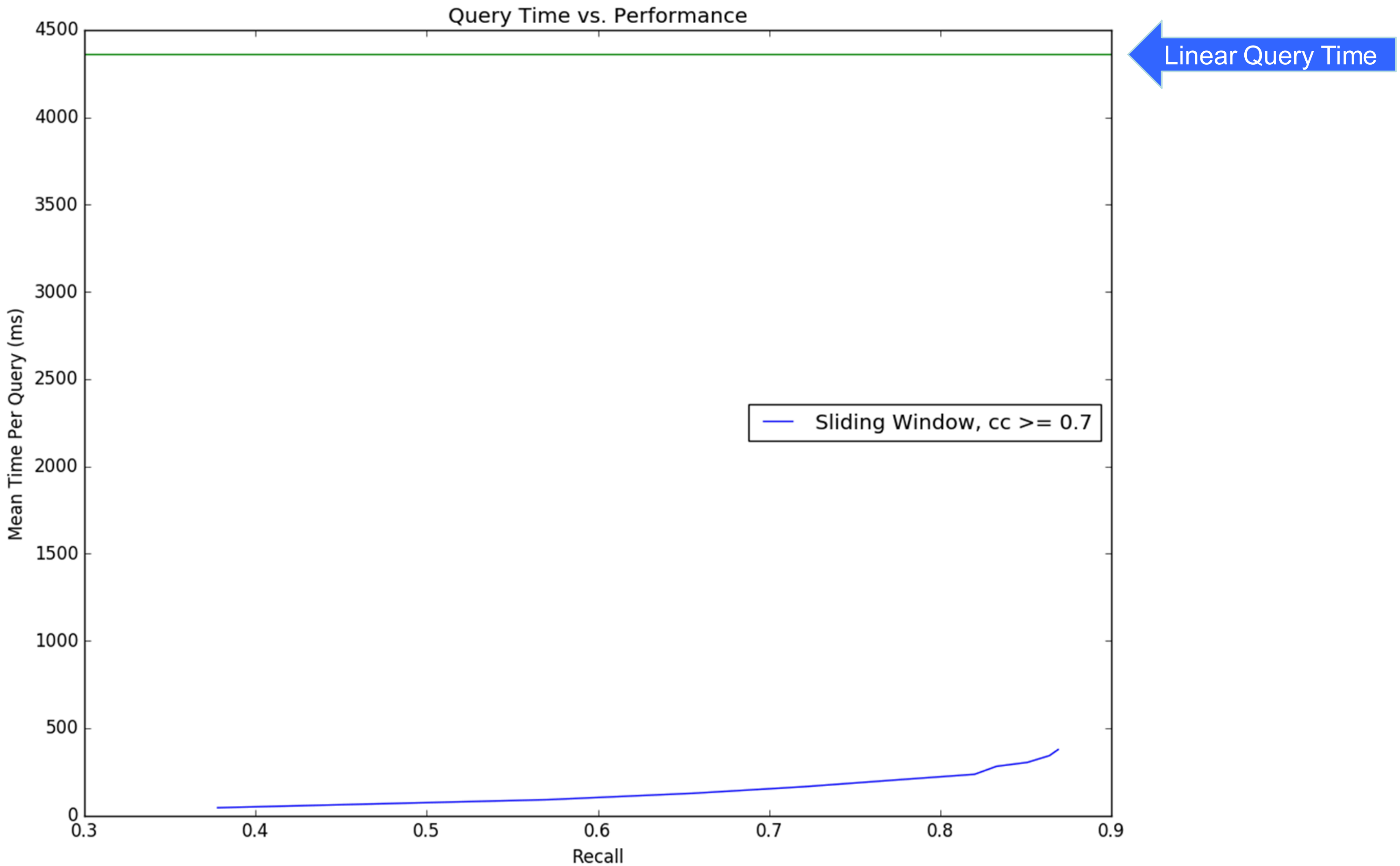
- Due to the approximate nature of ANN, a query into the index must:
  - **Over-sample:** Return a set of candidate nearest neighbors larger than what is need to perform the given task
  - **Prune:** Compute the true correlation between the query and all members of the candidate set, sort, and discard those that do not meet some quality threshold

ANN Search: Performance Evaluation

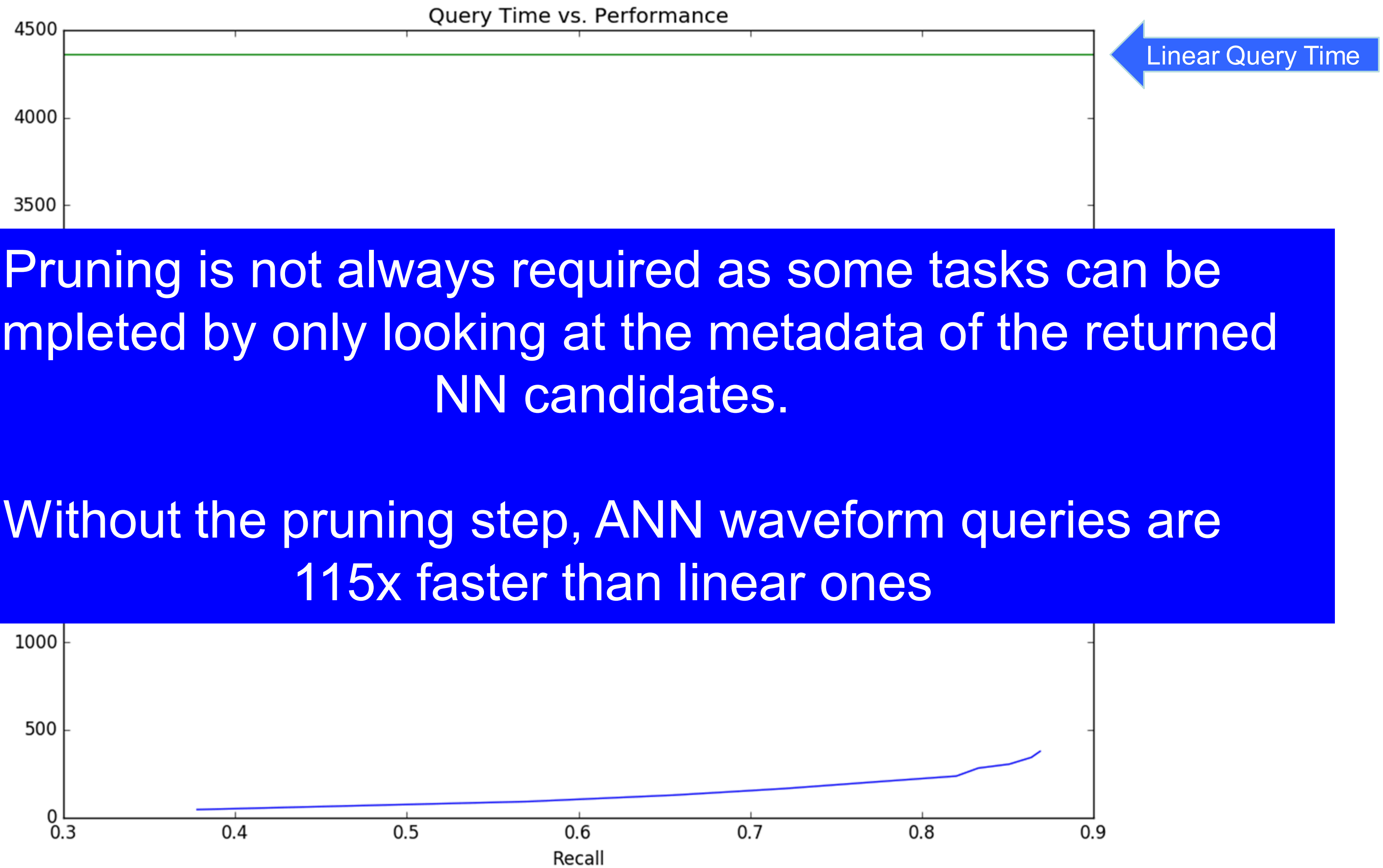


Performance metric is *recall*: the percentage of high quality neighbors returned by ANN query

ANN Search: Performance Evaluation



ANN Search: Performance Evaluation

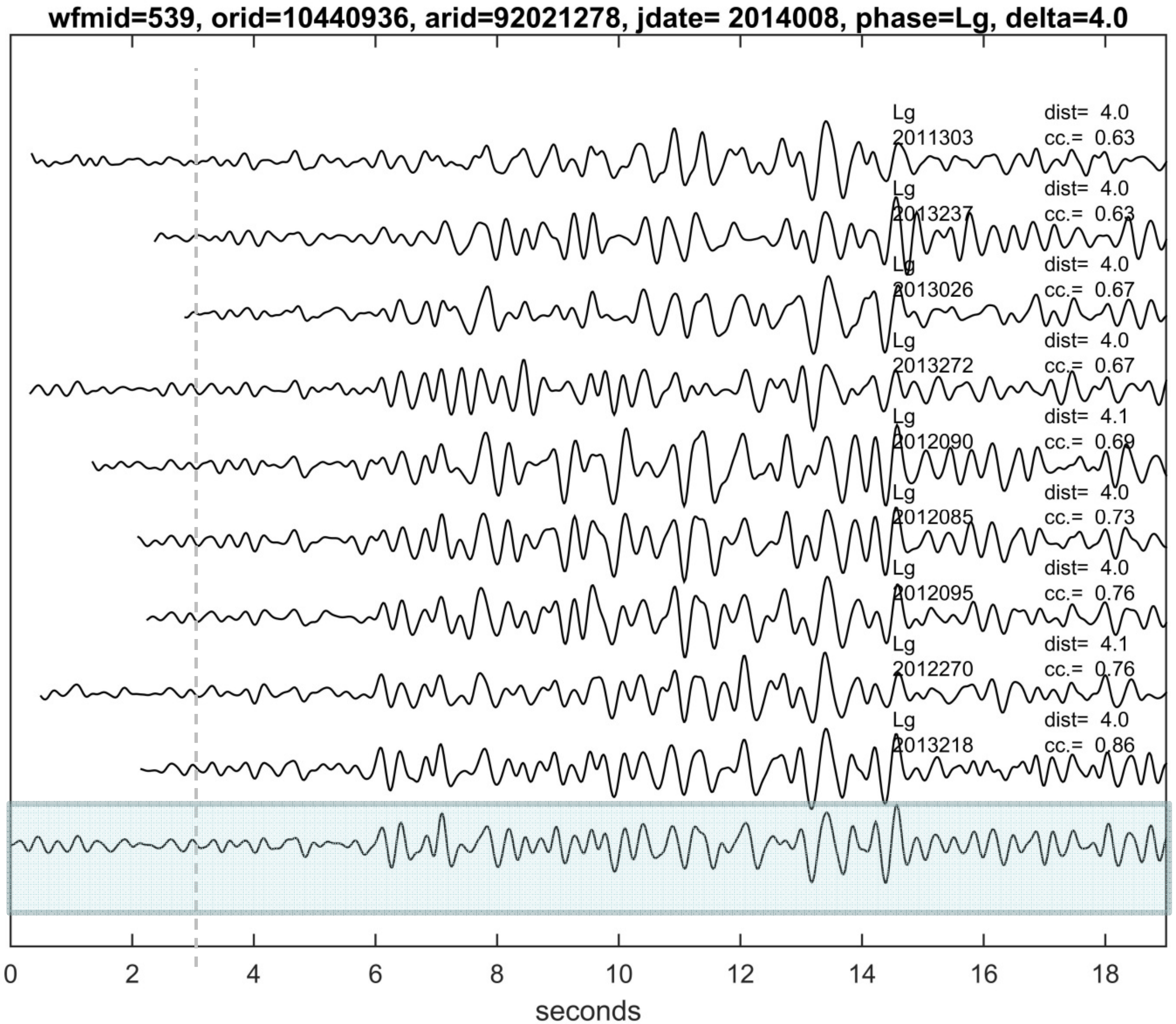


Pruning is not always required as some tasks can be completed by only looking at the metadata of the returned NN candidates.

Without the pruning step, ANN waveform queries are 115x faster than linear ones

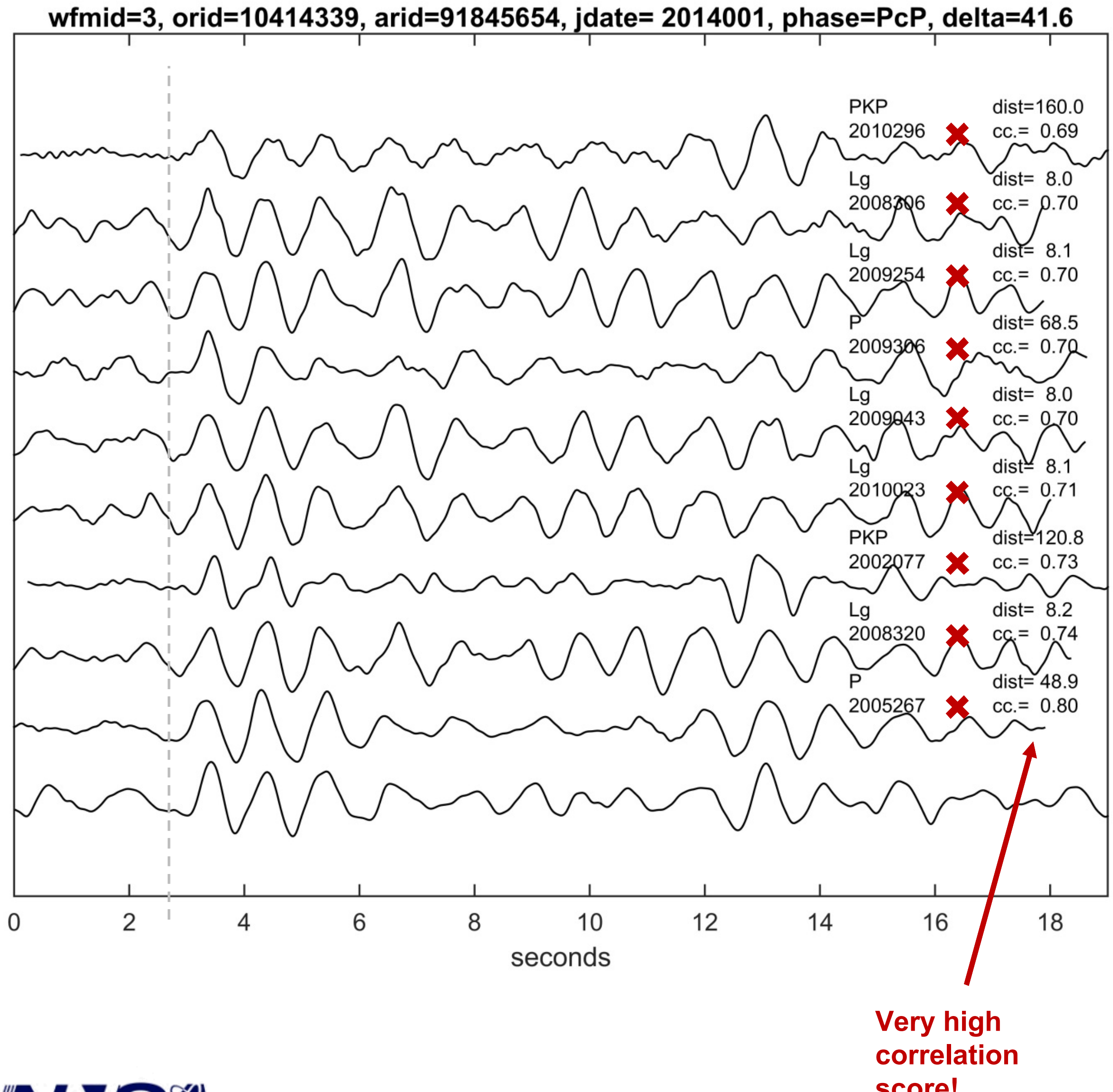
Test Results

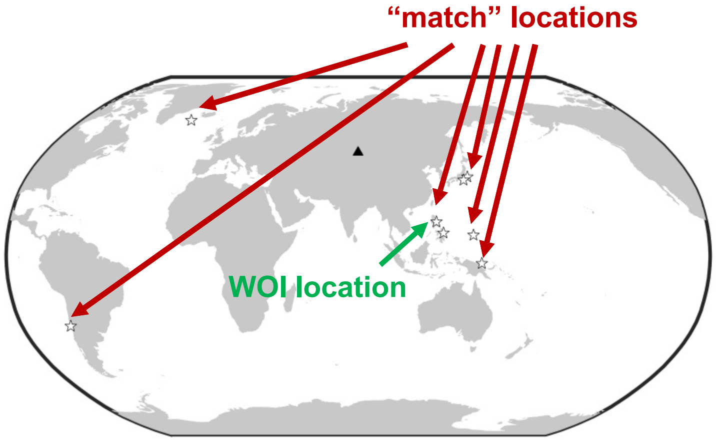
Regional Example #1 (Lg)



- Waveform to be identified is shown at bottom
- Matches shown above, in decreasing order of correlation
- Match pool span 2011-2013
- Mining explosions?

Screening Example #1: Rejected "Match"





- Only showing top 9 (there are136 with cc > 0.60 in this case)
- None of these is in the right location, but we wouldn't know that
- No consensus in location among match pool, so we reject the match

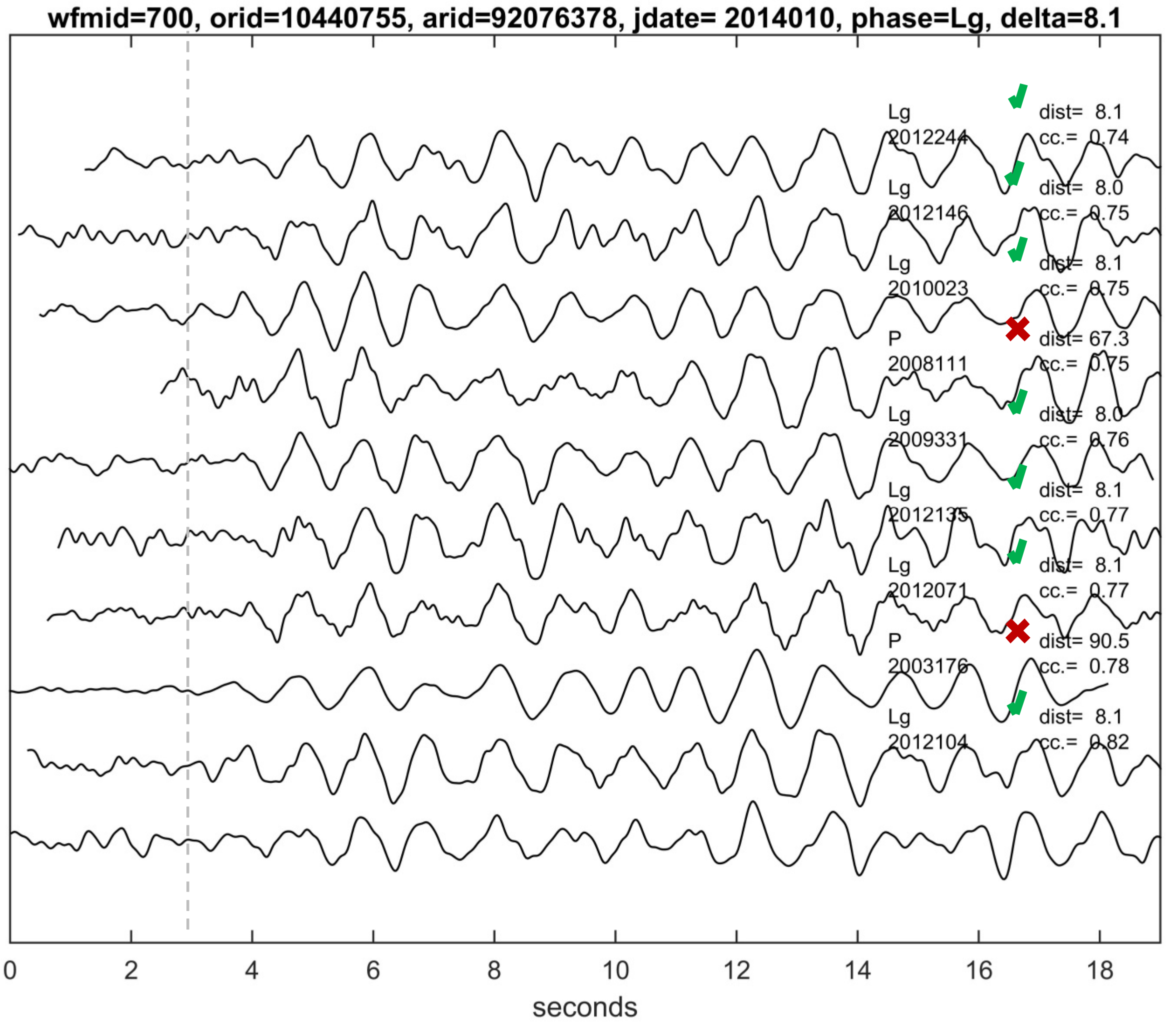
National Nuclear Security Administration

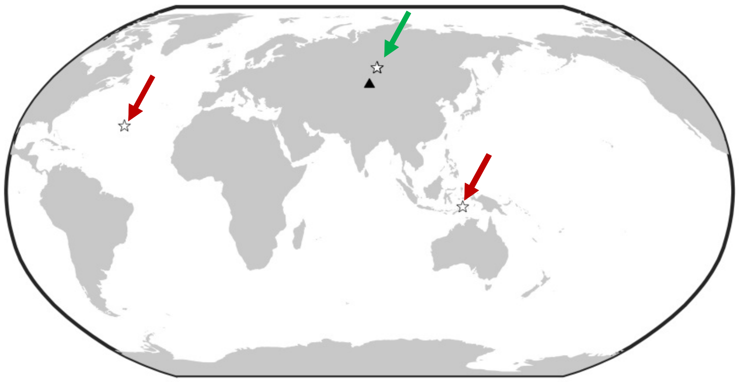
Sandia National Laboratories

Match Results Metadata Screening: An Essential Addition

- High correlation score isn't enough
- Can use *consistency* in event location of match pool to decide whether to accept match
- Developed a single algorithm based on event location consistency of 3 of the top 4 results
- More sophisticated approaches could be tried, but this worked quite well for this first simple study.

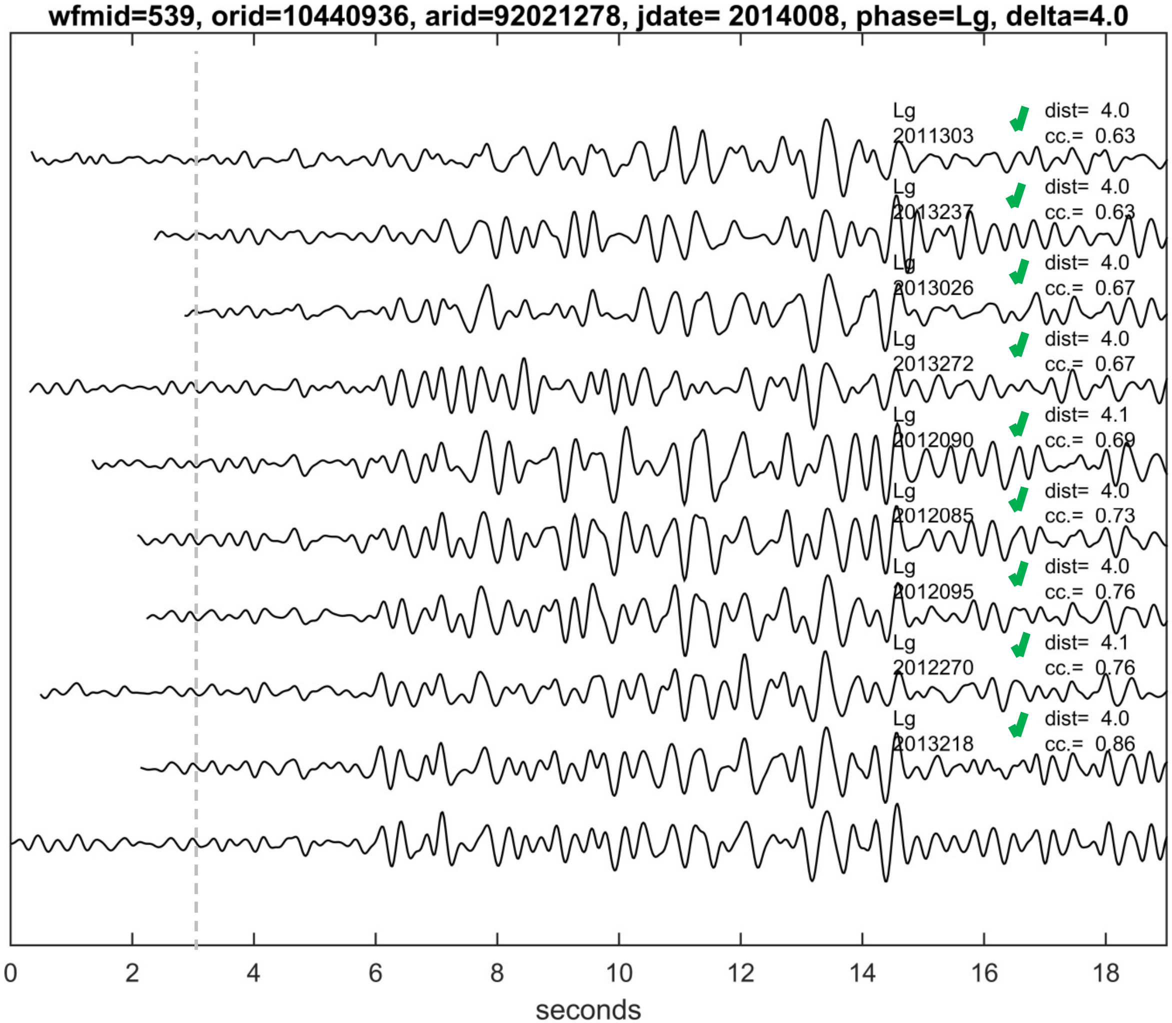
Screening Example #2: Accepted Match"





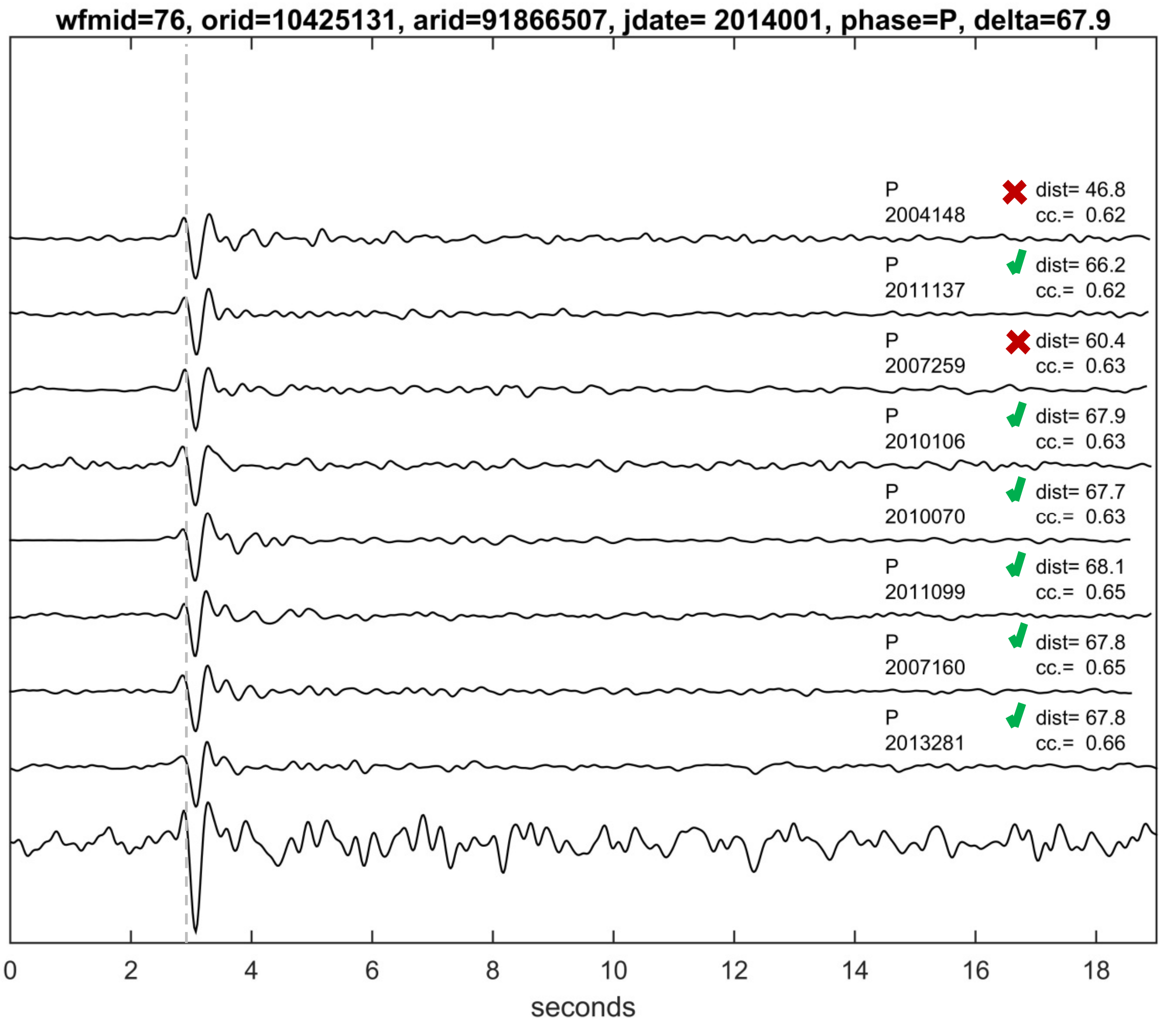
- Results include correct and incorrect matches
- This is a confirmed match: algorithm tolerates some dissent (3 of top 4 have excellent agreement)

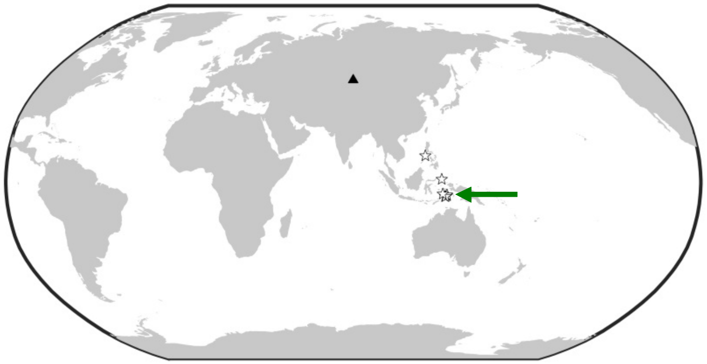
Regional Accepted Match Example #1 (Lg)



- Waveform to be identified is shown at bottom
- Matches shown above, in decreasing order of correlation
- Match pool span 2011-2013
- Mining explosions?

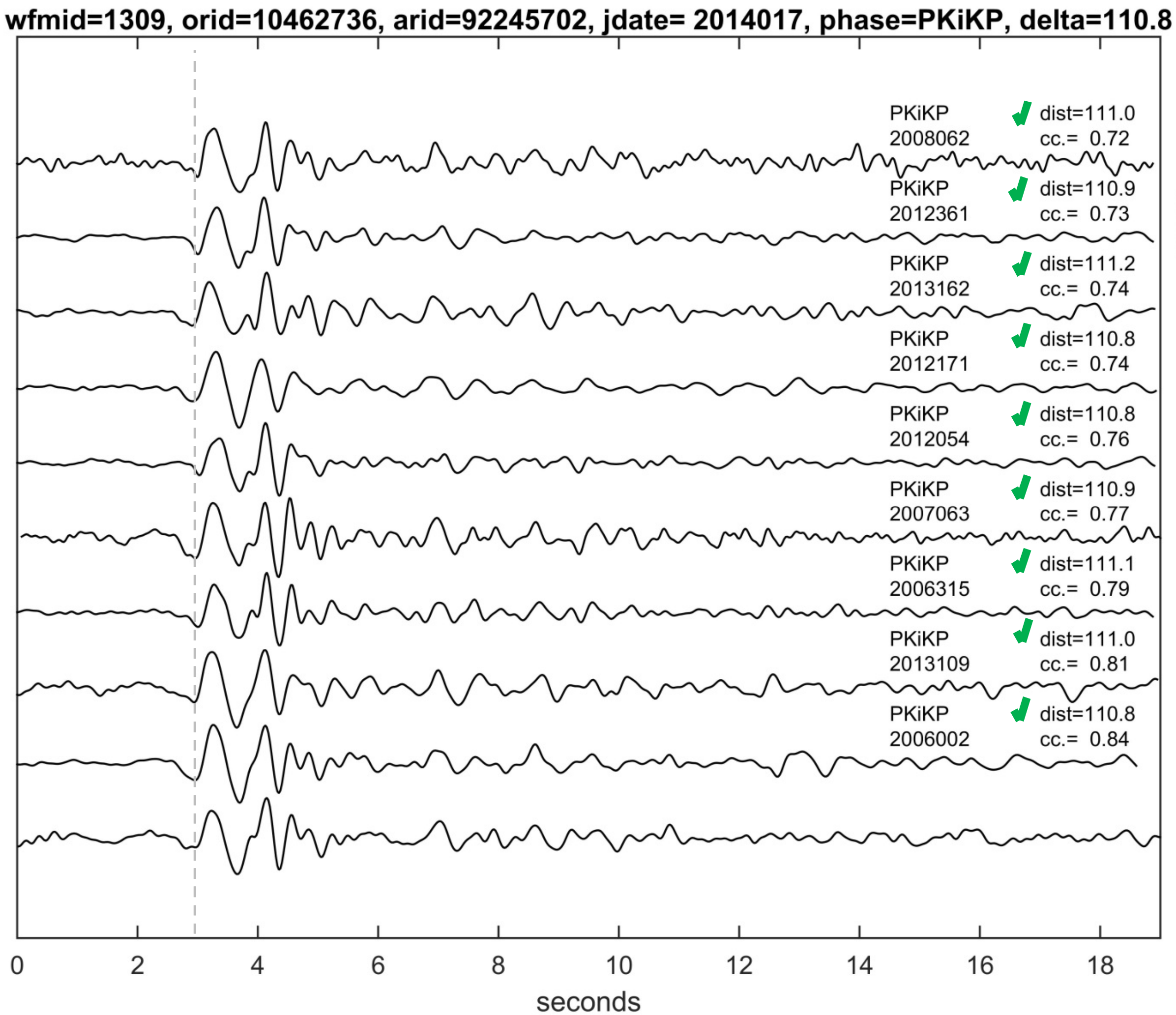
Teleseismic P Accepted Match Example #1

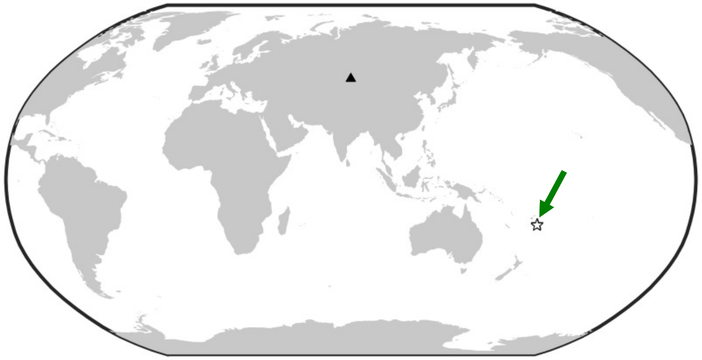




- Much shorter duration, simpler signal, but still matched!
- Pool of good matches much smaller (only 8): earthquakes have longer recurrence interval than mine explosions
- Match pool spans 2004-2013

Teleseismic PKiKP Accepted Match Example #1

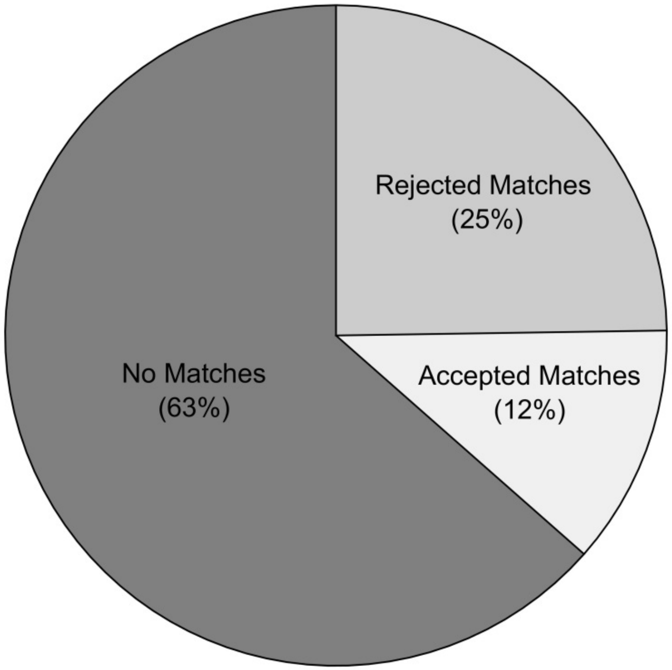




- Very complex signal is unmistakable (high confidence match)
- Match pool spans 2006-2013

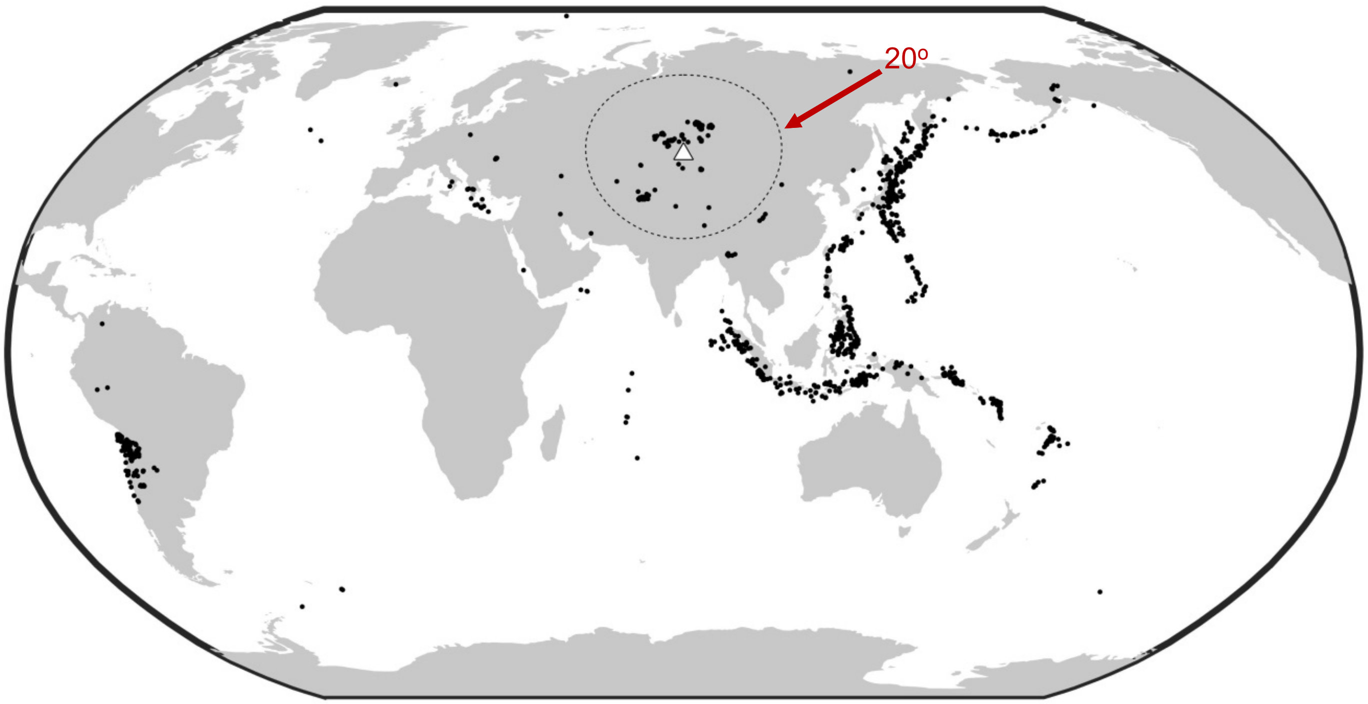
Match Results Summary (1)

- 63% of 2014 signal detections could not be identified (no matches or 1 match)
- Of remainder, 25% were rejected due to lack of event location consistency
- Analysis of 12% accepted matches:
  - 68% of those validated using LEB info
  - 93% of regional matches (Pn, Pg, Sn, Lg) validated
  - 67% of teleseismic first P validated, (important because 60% of all LEB associated detections are tele first P
  - 71% of core phases validated (PKPbc, PKB, PKiKP))
  - Poorer performance for secondary phases (42% PcP matches validated)



2014 signal identification

Match Results Summary (2)



- Majority of matches are at teleseismic distances because majority of events are at teleseismic distances
- Nearly all teleseismic matches are likely to be earthquakes
- Having a long time baseline for template archive is very important for matching earthquakes (much slower recurrence rate than mining events)

National Nuclear Security Administration

Sandia National Laboratories

Summary

- **Motivation** for work was observation that computation cost of waveform correlation was limiting size of signal archives that could be searched
- **Objective** was to see if ANN methods could facilitate searching of very large archives
- **Findings**
  - Our KPCA + KD Trees approach allows us to search archives at least as large as the total number of signals recorded by one of the most active stations in the IMS network (~300K, algorithm can easily handle much larger)
  - No specialized distributed computing system needed:
uses single COTS PC (albeit a fairly powerful one)
  - Adding a match pool screening step based on consistency of metadata was crucial for getting good results
  - Overall the proportion of detected signals that could be matched was modest (12% total, 8.6% validated), but impact on automatic processing could still be very significant
  - Long archive time baselines very important for matching earthquakes

Future work

- **Frequency dependence:** Build indexes for several overlapping narrower bands to try to better enhance the signals from particular event locations.
- **Array processing:** Build indexes for multiple channels and combine the results to take advantage of the coherency and move-out of signals across the array. This *should dramatically reduce the number of false event locations* in the ANN set of potential matches.
- **Signal Detector:** Apply directly as a signal detector rather than as an identifier of signals detected by another algorithm. Already the algorithm performs better than real-time, so could be used as an operational detector. We are currently working on implementing it in a distributed computing framework to make it fast enough to perform repeated testing on long time intervals (e.g. several months).

Backup slides

Archive Projection: Reducing the Size of the Kernel Matrix

- When the database is really large, we save time by building the kernel representation from *r* randomly sampled representatives.

r << n   (e.g. 500 out of 300,00)

$$K = \begin{bmatrix} k(x_1, x_1), k(x_1, x_2), \ldots, k(x_1, x_r) \\ k(x_2, x_1), k(x_2, x_2), \ldots, k(x_2, x_r) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ k(x_r, x_1), k(xr, x_2), \ldots, k(xr, x_r) \end{bmatrix}$$

r x r matrix

Archive Projection: Transforming the Kernel Matrix

- For next step, *K* must be symmetric, positive semi-definite

- *k=corr(x,y)* does not meet this requirement, but we can transform the correlation coefficients to fix this problem:

$$k = e^{(1 - corr(x,y))}$$

Archive Projection:
PCA in Kernel Space

- KPCA: Applying Principal Component Analysis to the Kernel matrix (K) creates a matrix (W) that can be used to project the archive to a form such that correlation scores have been mapped to Euclidean distance, i.e. correlated items can be found via searching for Euclidean nearest neighbors

ANN Search: KD-Tree Forests

- KD-Tree Forest:

  Build an ANN index from a group of randomly varied KD-Trees and query them collectively

Advantages	Disadvantages
Log(d) query time, easy to implement, memory efficient	**Only works for Euclidean similarity**

Our interest is Waveform Correlation

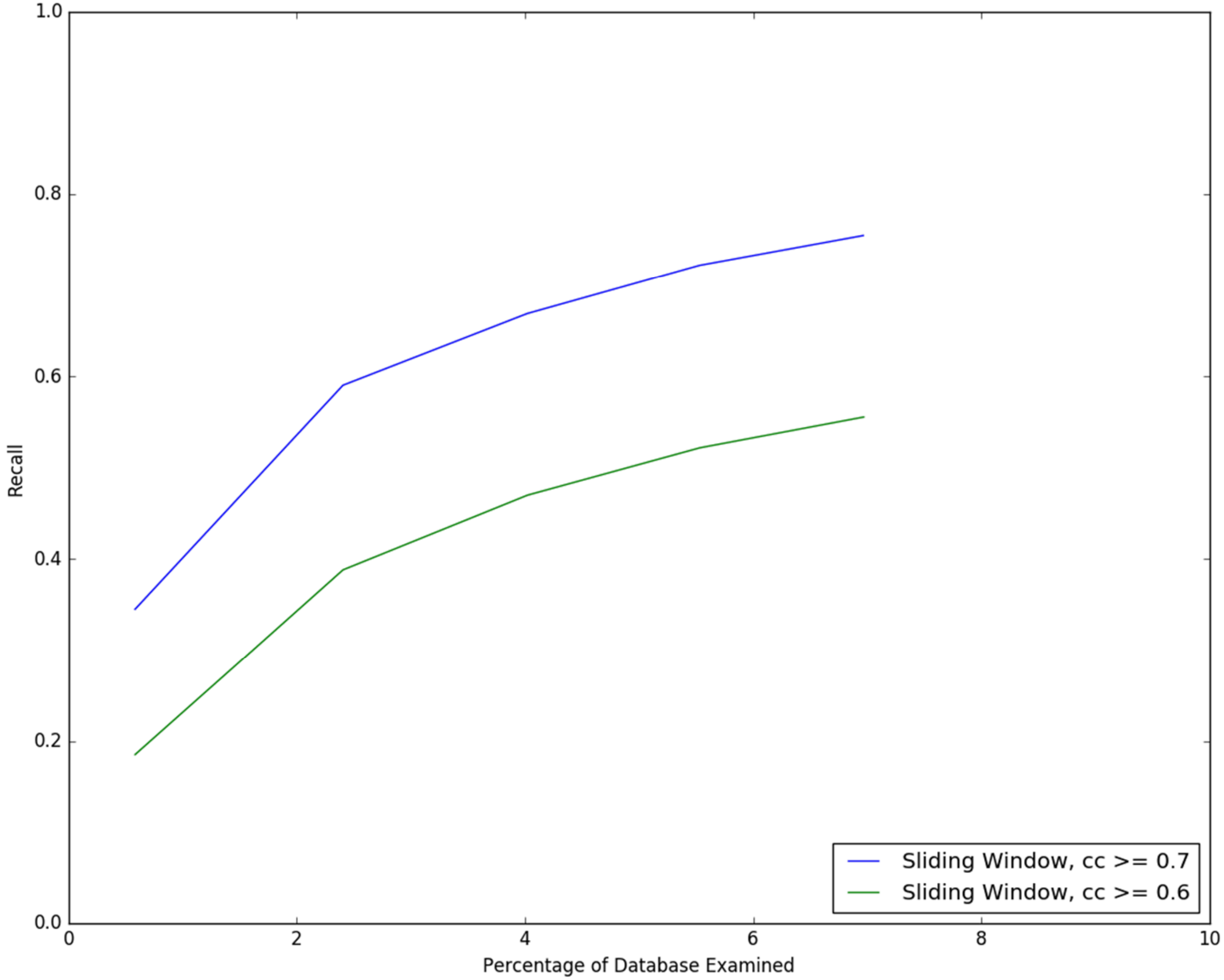
Archive Search: KD-Tree Forests

- How do we create a set of randomly varied trees?
  - During each split decision, randomly select from the top *n* highest variance dimensions

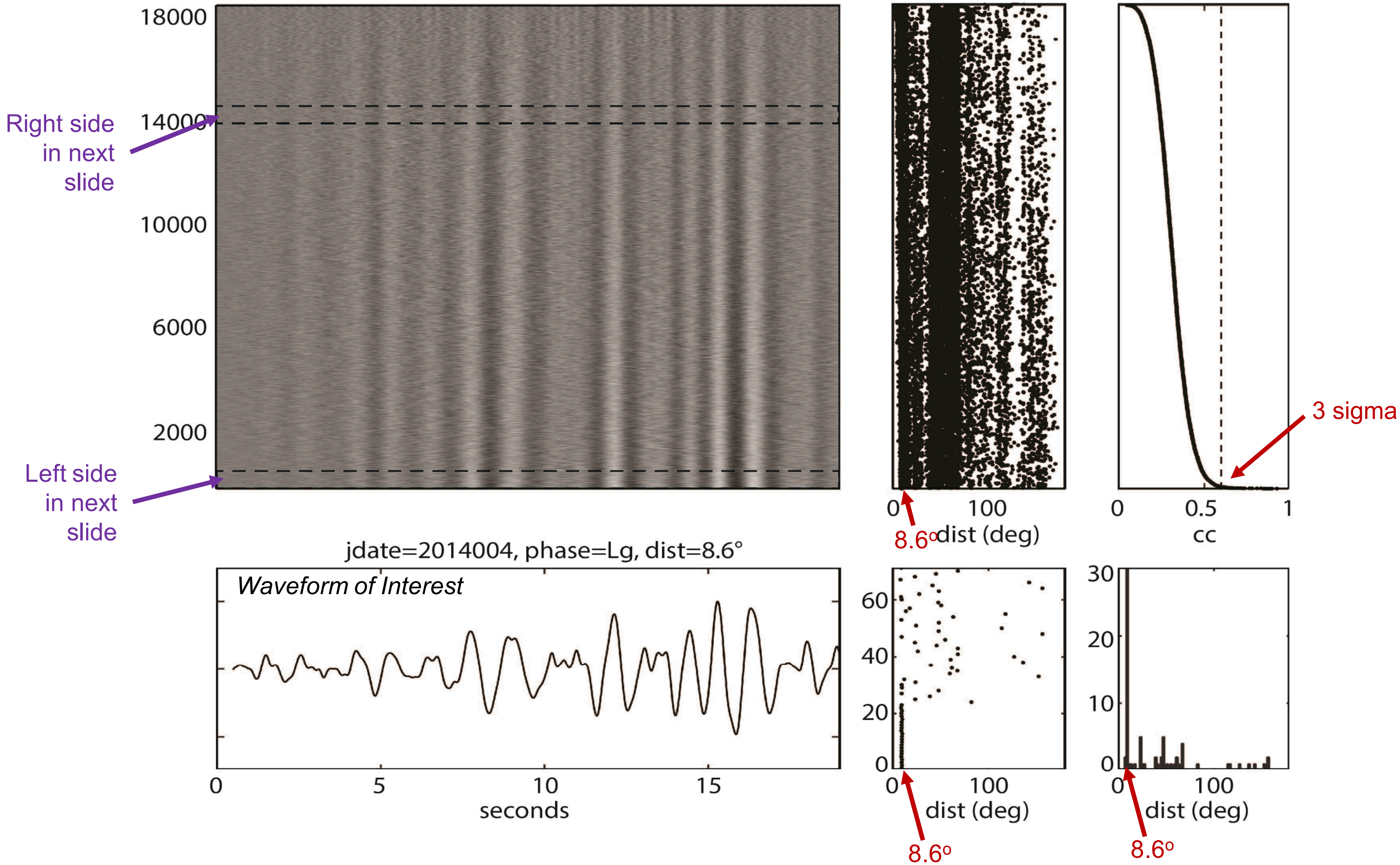
- How do we search the forest?
  - Search all trees until a leaf is encountered
  - During initial search, keep a sorted list of decision points based on how close the query was to the mean
  - Continue search by walking the list and taking the opposite path down to a leaf (re-visit most marginal decisions)

ANN Search: Performance Tests

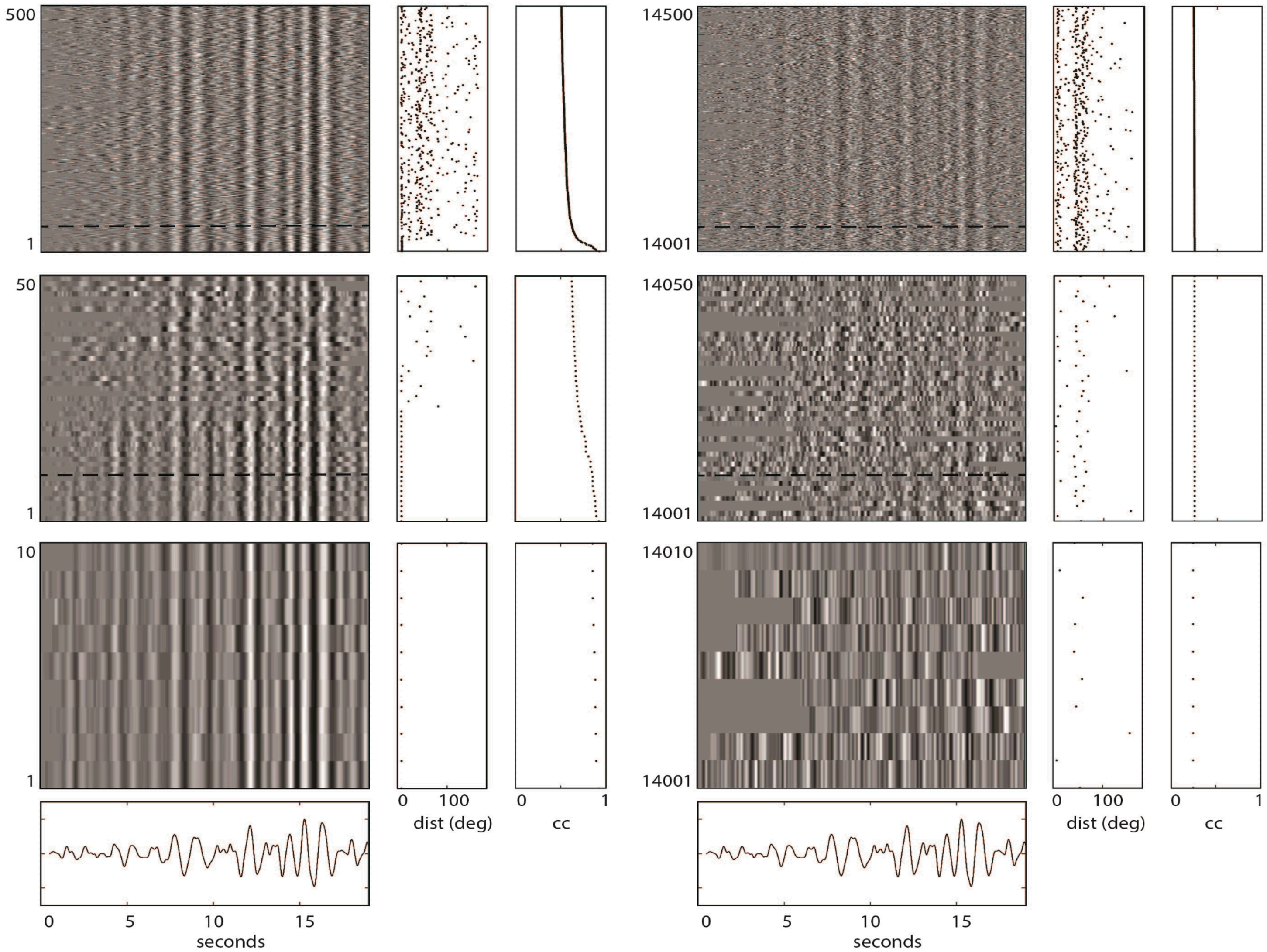
Sliding window performance based on quality of neighbors



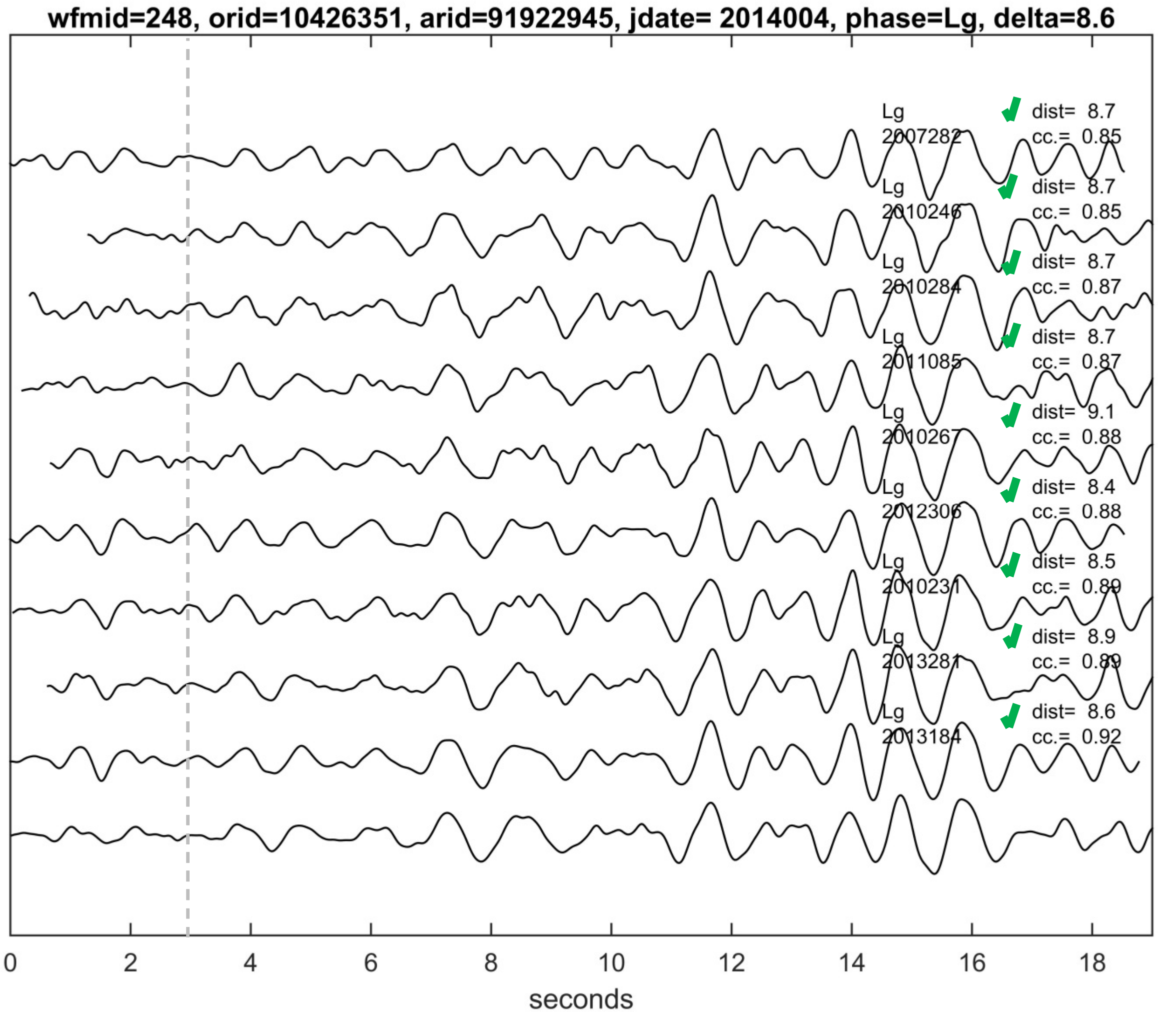
Example of KLSH results for a single waveform (Lg, 8.6°)

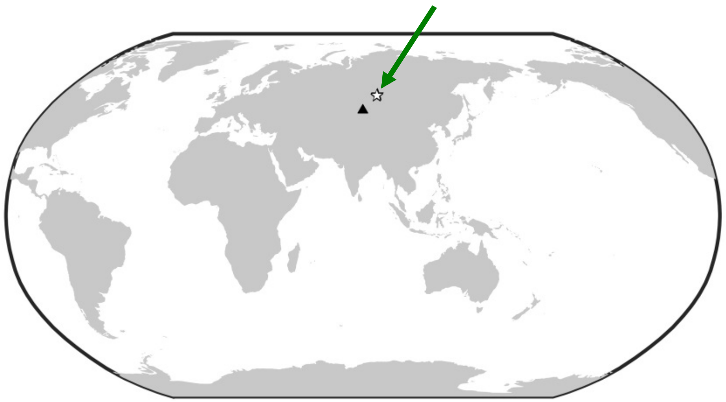


More on Lg Example:
Beware of “matches” with low correlation scores



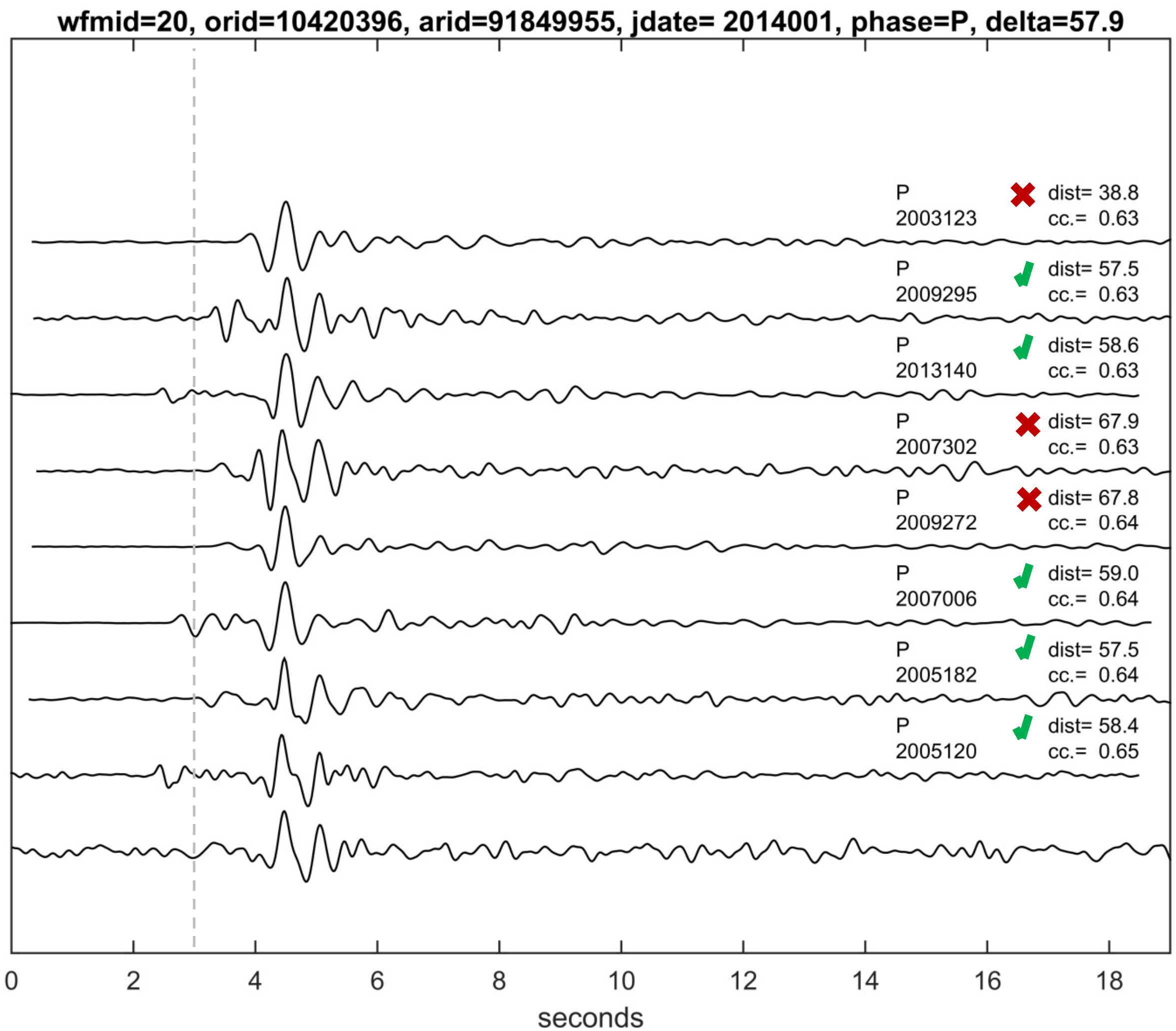
Regional Accepted Match Example #1 (Lg)

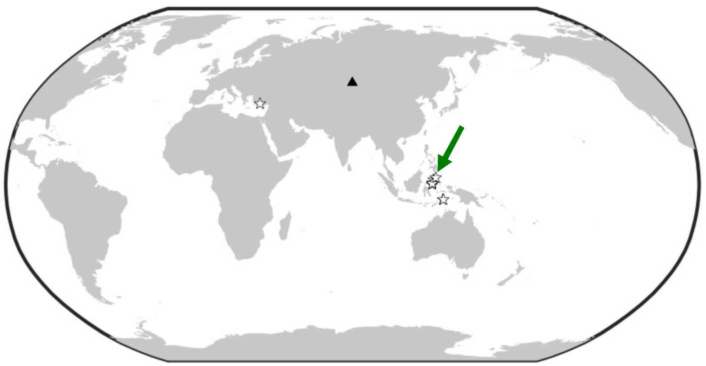




- Complex, long-duration signal, hence no disagreement in match pool
- Match pool span 2007-2013
- Mining explosions?

Teleseismic P Accepted Match Example #2



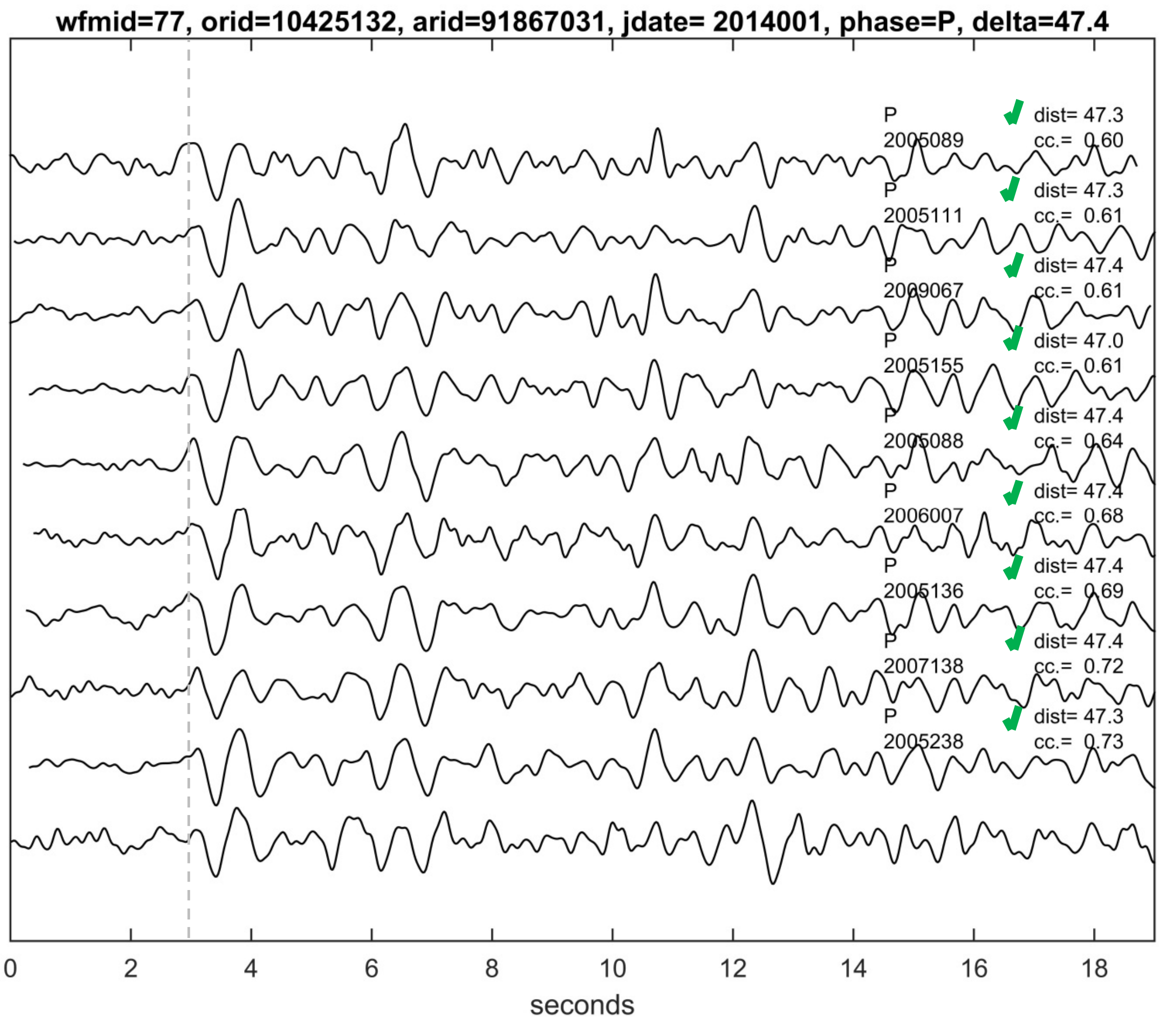


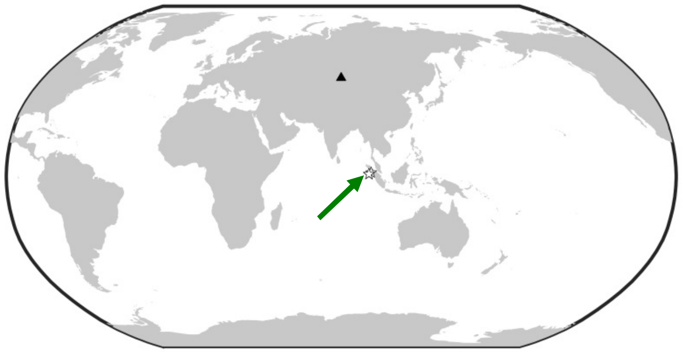
- Pool of matches again small
- 3 of 8 wrong, but still enough consistency in top 4 to accept the match

National Nuclear Security Administration

Sandia National Laboratories

Teleseismic P Accepted Match Example #3



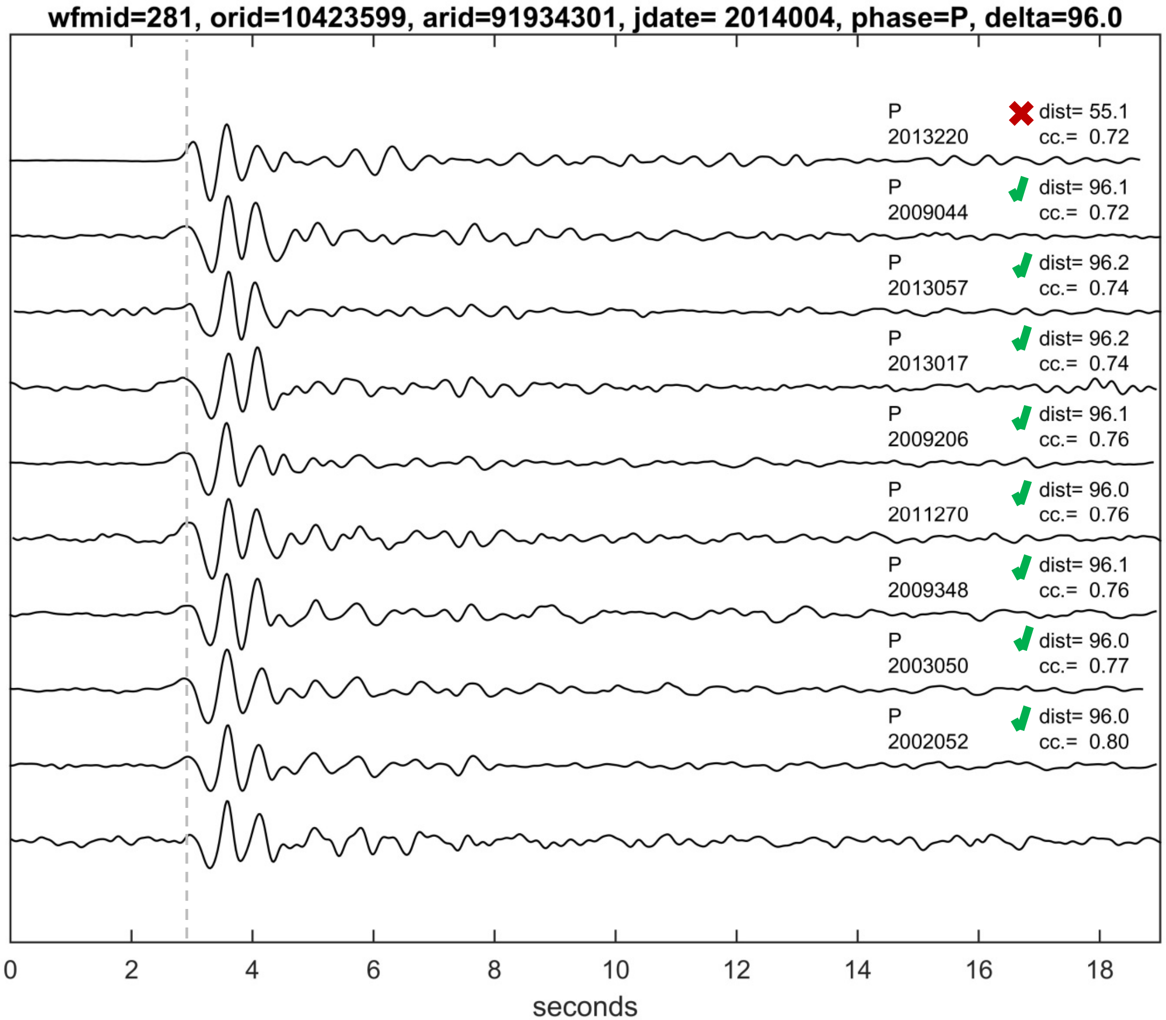


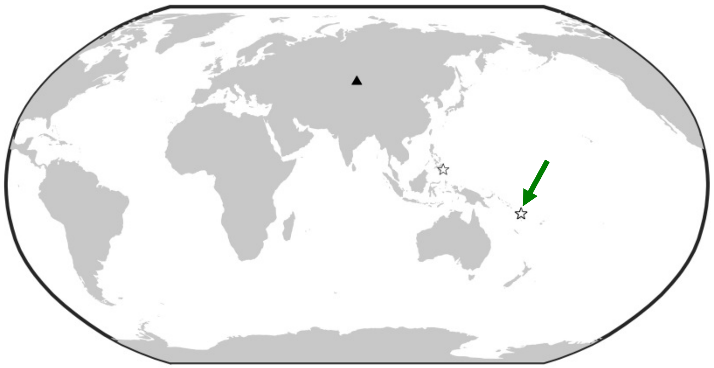
- Long duration, complex signal (path effect?) so perfect agreement in match pool

National Nuclear Security Administration

Sandia National Laboratories

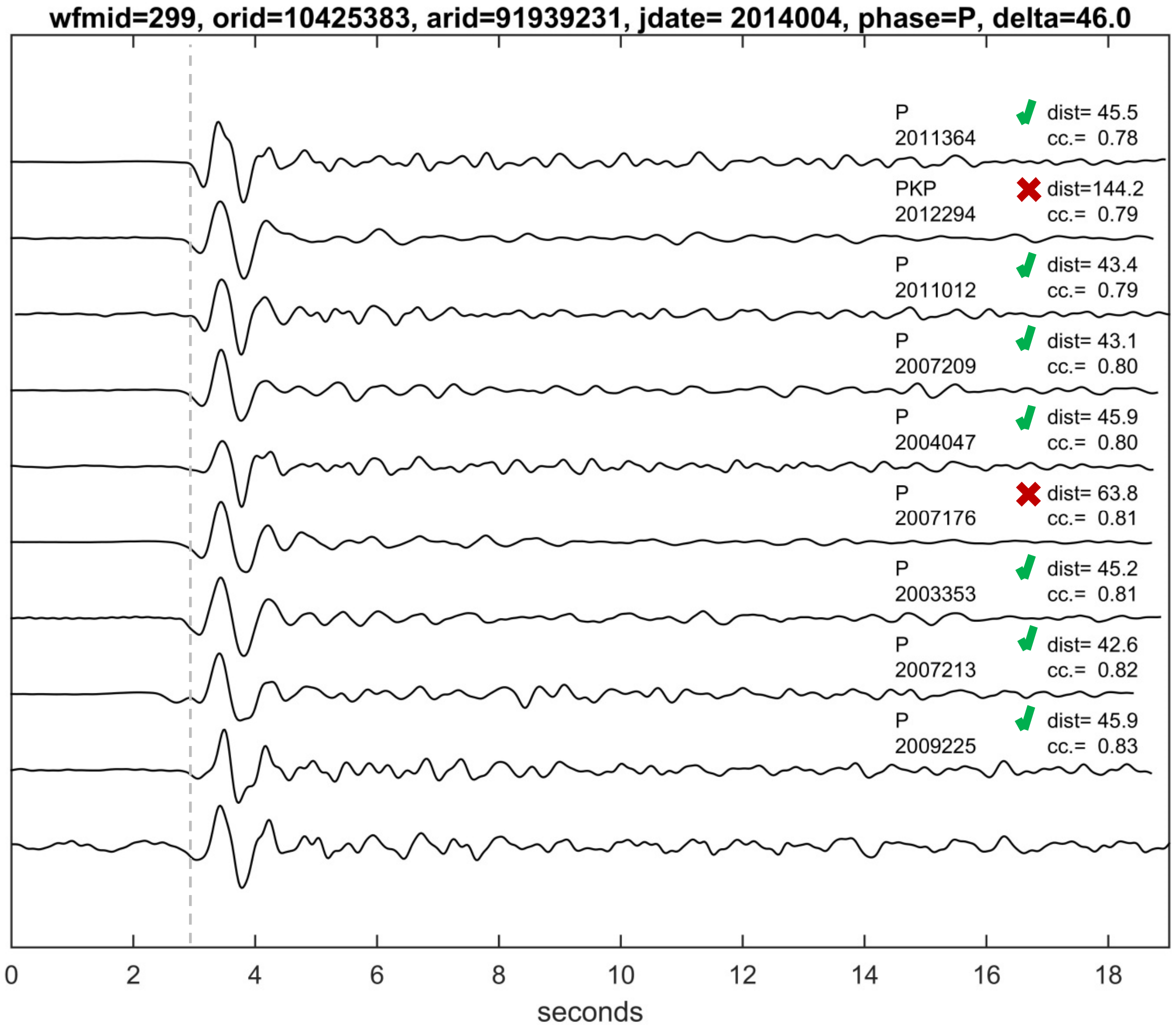
Teleseismic P Accepted Match Example #4

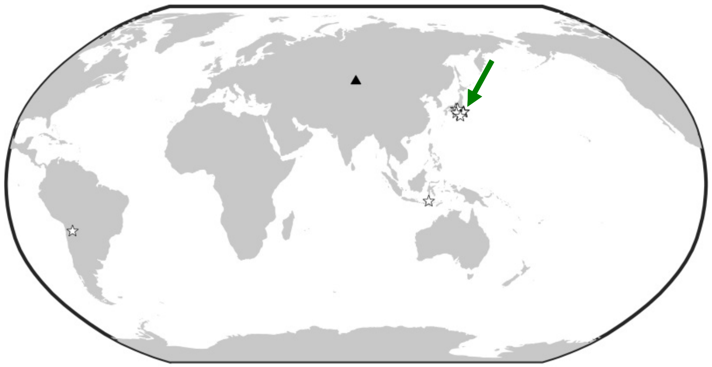




- Moderately complex signal, so good agreement among ANN set

Teleseismic P Accepted Match Example #5





- Signal seems neither long nor very complex, but match is accepted and is correct