

# UCR Insect Classification Contest



Organized by:

Yanping Chen

Eamonn Keogh

Gustavo E. A. P. A. Batista

## Briefing Document

[www.cs.ucr.edu/~eamonn/CE/contest.htm](http://www.cs.ucr.edu/~eamonn/CE/contest.htm)

# There are two planned phases to our contests

**Phase I:** July to November 16<sup>th</sup> 2012

(this contest)

- The task is to produce the best distance (similarity) measure for insect flight sounds.
- The contest will be scored by 1-nearest neighbor classification.
- The prizes include \$500 cash and engraved trophies.

**Phase II:** Spring 2013 to Fall 2013 (tentatively)

(future contest)

- Possibly two tasks:
  - A more general insect flight sound contest (your classifier does not have to be distance based, you can use *any* classifier).
  - Clustering, or anomaly detection or... of insect sounds
- Contest may be co-located with a ML/DM conference.
- The prizes may include a larger cash prize, engraved trophies, invited paper to a journal etc.

# Background to the Task: I



The history of humankind is intimately connected to insects. Insect borne diseases kill a million people<sup>1</sup> and destroy tens of billions of dollars worth of crops annually<sup>2</sup>. However, at the same time, *beneficial* insects pollinate the majority of crop species, and it has been estimated that approximately one third of all food consumed by humans is directly pollinated by bees alone.

Given the importance of insects in human affairs, it is somewhat surprising that computer science has not had a larger impact in entomology. We believe that recent advances in sensor technology are beginning change this, and a new field of *Computational Entomology* will emerge.

If we could inexpensively count and classify insects, we could plan *interventions* more accurately, thus *saving lives* in the case of insect vectored disease, and *growing more food* in the case of insect crop pests.



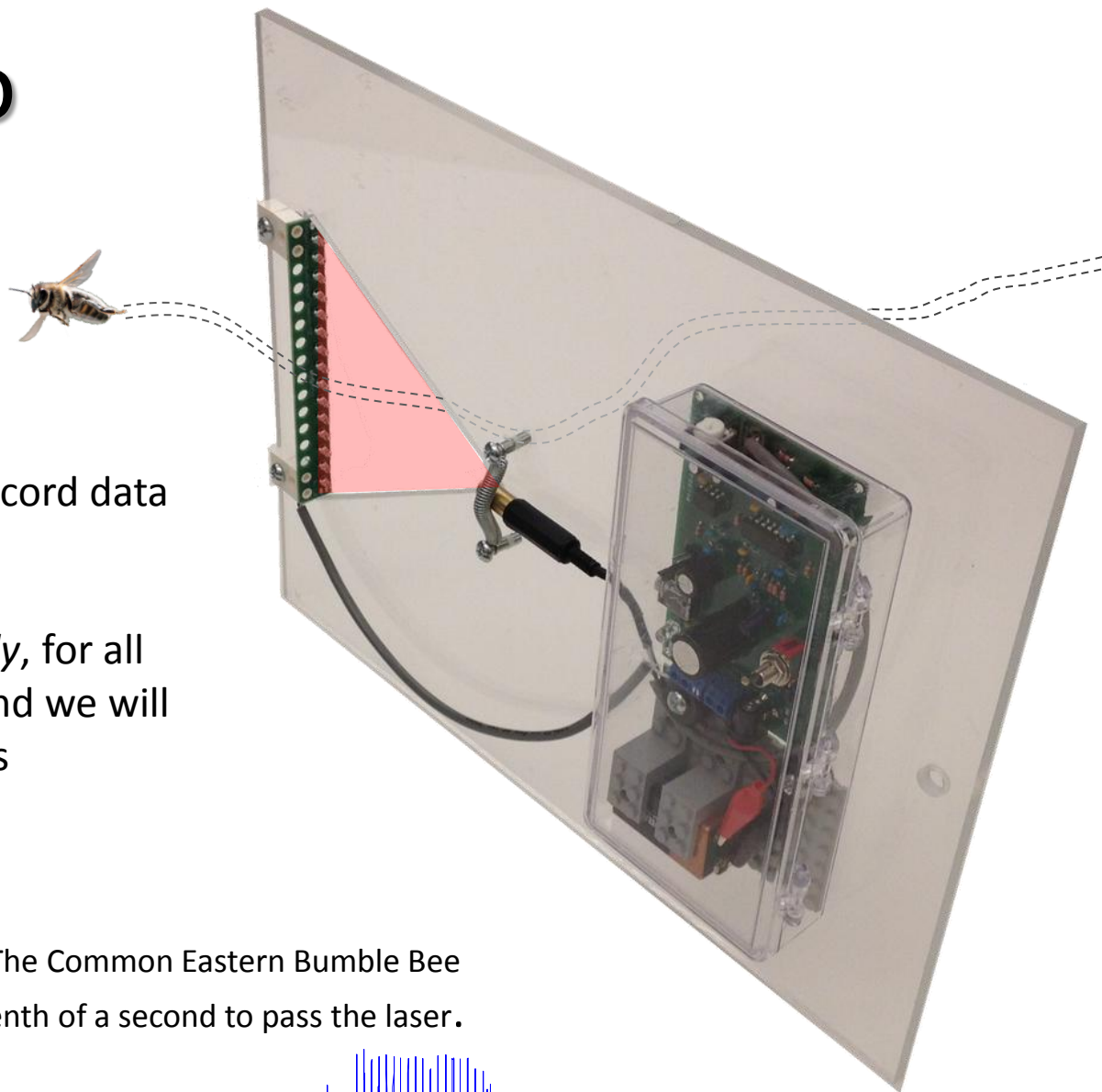
1: Malaria is the first insect vectored disease that comes to mind, but there is also West Nile disease, African trypanosomiasis, Dengue fever, Pogosta disease etc.

2: Aphids, caterpillars, grasshoppers, leafhoppers and crickets all cause damage to crop plants. Currently, insects alone consume or damage sufficient food to feed 1 billion people (Oerke EC. 2006. *Crop losses to pests*. The Journal of Agricultural Science 144(01): 31-43.)

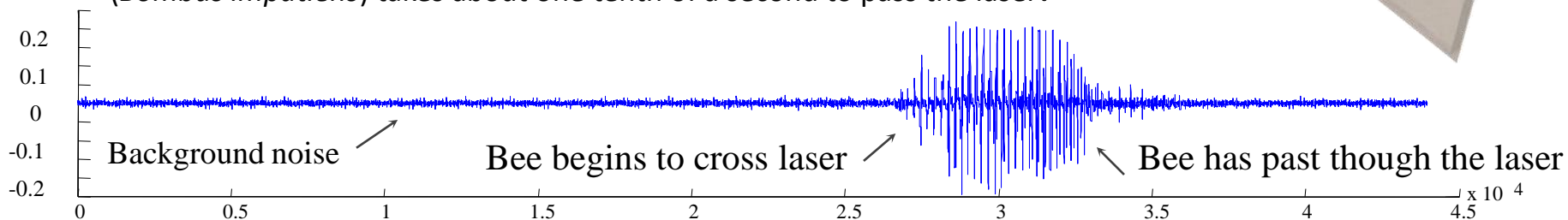
# Background to the Task: II

At UCR we have built sensors to record data from flying insects.

While the data is collected *optically*, for all intents and purposes it is *audio*, and we will refer to it as such in the rest of this document.

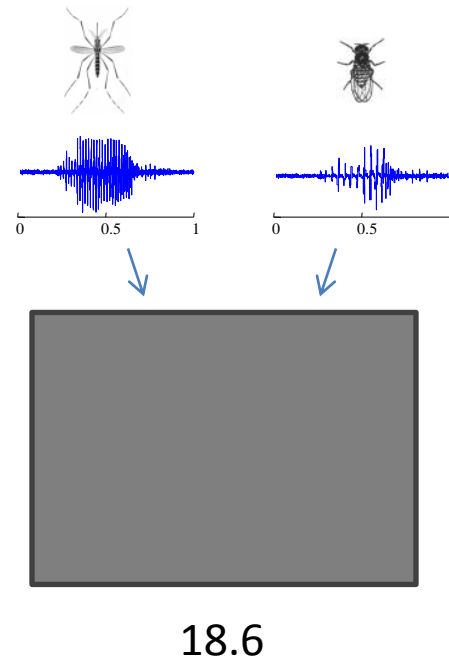
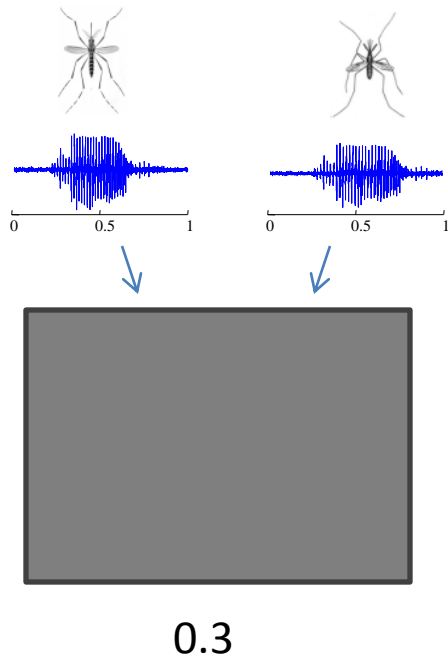


One second of audio from our sensor. The Common Eastern Bumble Bee (*Bombus impatiens*) takes about one tenth of a second to pass the laser.



# The Task

- The task is to build a distance function (i.e. a computer program) that takes in two audio snippets and calculates their similarity.



# Prizes

- **Overall winner:** Whoever gets the highest  $E$  Score
  - \$500 prize
  - An engraved trophy
- **(Most weeks on) Top of the Leaderboard** (if tie, highest *final E* accuracy wins)
  - \$100 prize
  - An engraved trophy
- **Judges Prize:** (optional, given at the discretion of the judges. Could be more than one)
  - \$250 prize(s)

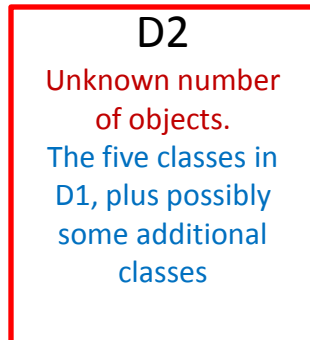
The judges prize is design to reward some team/individual that produces a great idea within this contest or more generally within the area of computational entomology, but does not (necessarily) win the contest. For example:

- Suppose the winning team scores 90%, but has a 1,000 lines of code. A team that scores 89.9% with just *two* lines of code might win this prize.
- Supposes a team figures out how to cheat. For example they note that class 1 is only on stereo left, and class 2 in only on stereo right. By telling us how they might be able to cheat, we can make the Phase II contest better. This would deserve a prize.
- Suggesting an interesting task for Phase II could be worth a prize.

If you want to *explicitly* be considered for this prize, send your idea to the judges.

# Evaluation

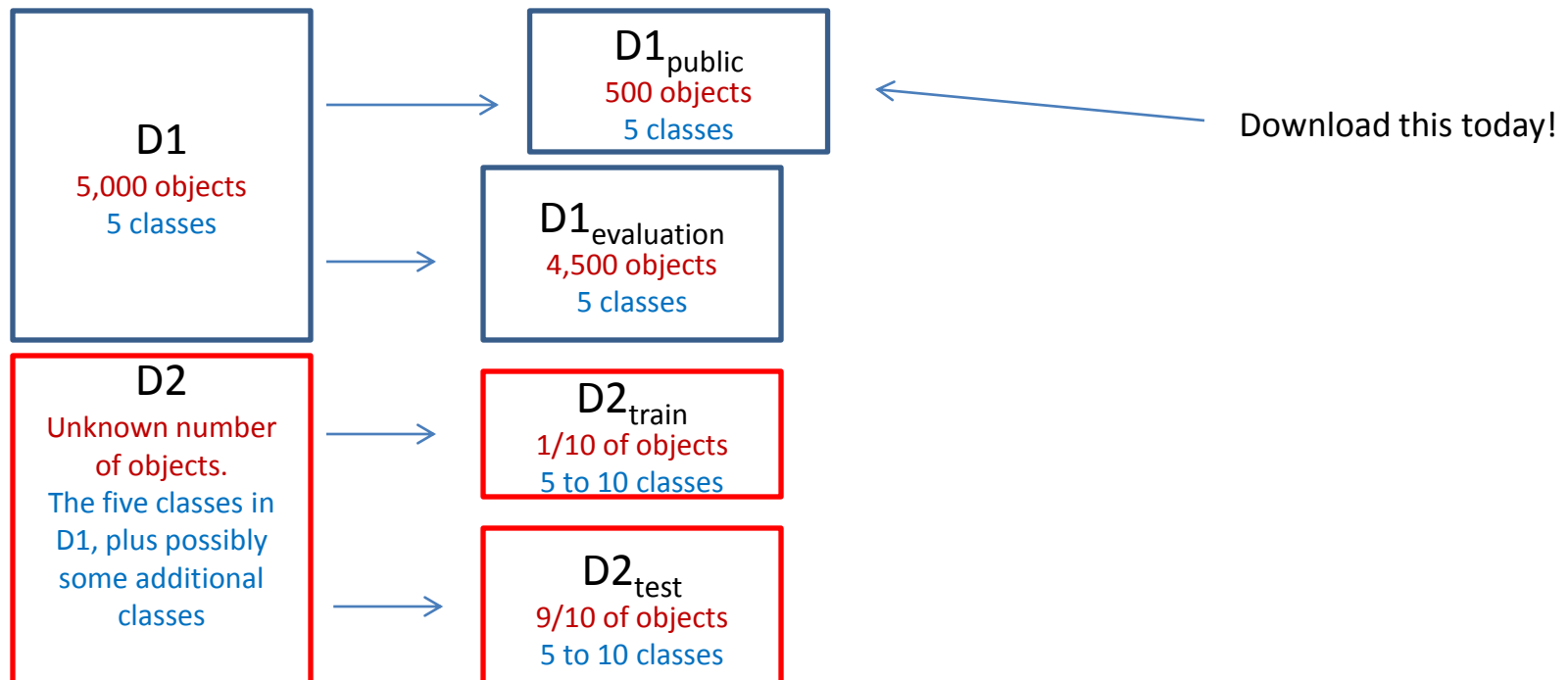
- Evaluation will be done with one-nearest neighbor classification.
- We have 5,000 exemplars in five classes. Call this D1.
- We are currently collecting more data, including possibly new (but similar) classes, in an identical format with the same sensors<sup>1</sup>. Call this D2.



<sup>1</sup>: We reserve the right to slightly modify the sensors, as we try to improve reduce noise, reduce the power needed etc. However the new data will be essentially the same as the D1.

# Evaluation

- We have split D1 into two sets  $D1_{\text{public}}$  and  $D1_{\text{evaluation}}$ . We used a random shuffle, but preserved the class ratios.
- You have access to  $D1_{\text{public}}$  today. You can download it from [www.cs.ucr.edu/~eamonn/CE/contest.htm](http://www.cs.ucr.edu/~eamonn/CE/contest.htm)
- When we finish collecting D2, we will split it the same way.





# Evaluation

- Use  $D1_{\text{public}}$  to test your distance function.
- We suggest using the provided code to test the leave one-out-accuracy of the one-nearest-neighbor algorithm using your distance function. But you can do anything you want.



- By November 16<sup>th</sup> 2012, submit your distance function. We will publicly announce the results within a week.
- We will only announce the names of teams in the top 50% or in the top 10, whichever is larger.

# Evaluation

- The evaluation score  $E$  is the (unweighted) mean of your two scores:
  - Using  $D1_{\text{public}}$  to classify  $D1_{\text{evaluation}}$
  - Using  $D2_{\text{train}}$  to classify  $D2_{\text{test}}$
- Thus  $E$  is defined as:

$$E = ( \text{accuracy}(D1_{\text{public}} | D1_{\text{evaluation}}) + \text{accuracy}(D2_{\text{train}} | D2_{\text{test}}) ) / 2$$

The code we will use to do the evaluation is near identical to the code we have distributed with the data, except instead of leave-one-out we will use the train/test split

Read  $\text{accuracy}(Y | X)$ , as the accuracy of  $X$  using the model  $Y$

# Evaluation

- Note that we *may* make part or all of  $D2_{\text{train}}$  available before the contest ends.
- We will decide later based on the interest in our contest, and the number of entries etc.
- If we do so, we will not give you feedback about your results on it, but *you* will be able to do cross validation on it ( that is to say, it will be labeled)
- If this happens, we will simply create a link at [www.cs.ucr.edu/~eamonn/CE/contest.htm](http://www.cs.ucr.edu/~eamonn/CE/contest.htm) one Friday before the contest ends, we will not broadcast the release.

$D2_{\text{train}}$   
1/10 of objects  
5 to 10 classes

← *May* be available before the contest ends.

# Leaderboard

- While you can try to predict your accuracy by doing cross-validation on  $D1_{\text{public}}$ , you can also gain some feedback by asking us to test your current distance measure with  $D1_{\text{evaluation}}$
- You can only do this *once a week*.
- Send your matlab distance function (named for your team leader/organization, i.e **MIT\_Smith.m**) to [UCR.insect.contest2012@cs.ucr.edu](mailto:UCR.insect.contest2012@cs.ucr.edu) before any Friday at noon (PST). We will run your code to test  $\text{accuracy}(D1_{\text{public}} | D1_{\text{evaluation}})$ , and post your accuracy on the leaderboard (in most cases within a few days).
- We will only tell you your accuracy, not which examples you got wrong etc.
- We strongly recommend that every team does this at least once before the final scoring. That way, we can all be pretty sure your final code will run for the final evaluation.

# FAQ I

- Q) The prize money is not a lot...
- A) True, we are hoping that the socially noble nature of the research, and the fun of the challenge will be enough incentive. Note that while the building of the sensors and the data collection was funded by the Bill and Melinda Gates Foundation and the Vodafone America, their funds *cannot* be used to pay the prize. Thus the prizes for the Phase I are coming out of Dr. Keogh's pocket.
- Q) Tell me more about the data...
- A) The data instances are one second long sound files. However the insect signal is typically only a few hundredths of a second long, and *approximately* centered in about the middle of the file. The data before and after the insect sound is just noise from the sensor. In most cases if you listen to the files you can hear the distinctive buzz of the insects at about the halfway point. Note that the "sound" is measured with an *optical* sensor, rather than an *acoustic* one. This is done for various pragmatic reasons, however we don't believe it makes any difference to the task at hand. The sampling rate is 16000 Hz
- Q) Are the time stamps relevant?
- A) It is true that some insects are more active at certain times of the day. Thus, if you know the time (and the date and longitude) this could change the prior probabilities. However we want to focus just on the signal-processing here, so we have changed time/date information to remove any such clues.
- Q) Could the data be mislabeled?
- A) We are almost certain that no data is mislabeled in the sense that we might have mistakenly listed an instance as class A, when it is actually class B. However, it is possible that an instance has two or more insects flying past the sensors at once, and thus the sound is a mixture of two insects (of the same type). It is also possible that there is no insect sound in the file, just a noise "blip". However, we expect such instances to be vanishingly rare.

# FAQ II

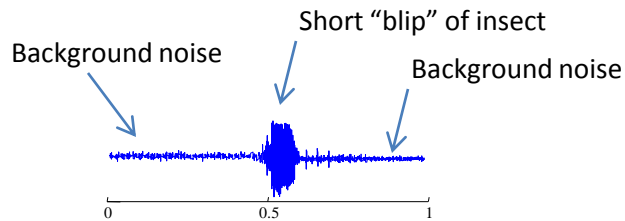
- Q) If I enter the contest, can I use my idea(s) for my own papers/patents/commercial products?
- A) Yes, your ideas belong to you. As discussed elsewhere, in order to enter the contest, you must share your code. And we will share the code/methodology of the winner (and possibly other entries) with the entire world, **after the contest is over**. However, any additional avenues/commercial venues you wish to peruse is your business. If for some reason you do not want your code/methodology be possibly shared with the world, please don't enter the contest.
  
- Q) Anything else I should know about the data?
- A) Some insects are sexually dimorphic (the males/females are different sizes/shapes). It is possible that one or more classes could have sexually dimorphic insects, however, we are not sure if this makes a difference that is important/exploitable. It also is possible that one class is male  $X$  and another class is female  $X$ , or that one class is juvenile  $Y$  and another class is adult  $Y$  etc. In every case we do believe that it is possible to differentiate the classes.
  
- Q) Are the two phases independent?
- A) Yes, you can enter either or both, you can use the same or different ideas in both. The only reason why we are having two phases is because, we will be still in the process of collecting data as the first phase is underway, and we want to gain experience in hosting a smaller contest before a larger version at a conference.
  
- Q) What are the tax implications of the prize?
- A) Sorry, we cannot give tax advice. Our university will report this income to US government. Talk to a tax professional.
  
- Q) Can you answer *my* question?
- A) Maybe, but if we do, the (edited for clarity/length) question and answer will be posted online for all to read.

# FAQ III

- Q) Can you recommend any papers I should read?
- A) This paper gives some more background to the *motivation* etc: *G. Batista, E. Keogh, A. Mafra-Neto, E. Rowton. Sensors and Software to allow Computational Entomology, an Emerging Application of Data Mining. SIGKDD 2011 Demo Paper.* Itai Cohen's amazing videos may be worth watching ( <http://vimeo.com/22997241#> ). However, you are mostly on your own.
- How fast does my code have to be?
- We don't really care about speed. However, if your code is so slow that it takes days or weeks to evaluate it, we would have a problem. The two exemplars you will be comparing are 1 second long each, and almost all matlab sound processing algorithms are *much* faster than real time. Thus, we are setting a comfortable 10-second maximum per comparison limit (amortized over the entire evaluation).
- Can I be on two teams?
- No, a person may only be on a single team. However a university/company may have multiple teams. If you are a professor and two subsets of your students want to compete, you must be on **only one** of those teams (or neither). The team list needs to be in the comments of the m-file submitted.
- How many people can be on a team?
- We don't care, just be sure to list them *all* in the comments of your code.
- Do I have to be in at university to compete?
- No, the contest is open to all. Companies, private individuals, high schools etc.

# FAQ IIII

- Q) What are the species in the sound files?
- A) We will tell you after contest is over. We don't think it makes any difference to your work.
  
- Q) Why not use Kaggle?
- A) We will probably use Kaggle for the Phase II of the contest. We wanted to have hands-on experience first.
  
- Q) Is anyone barred from the contest?
- A) To prevent any apparent or actual COI, students at UCR, past or current students of Dr. Keogh or Dr. Batista should not enter the main contest (they could enter the Judges Prize part of the contest, by submitting an idea).
  
- Q) Does the entire one second sound file contain insect sounds?
- A) No, as mentioned above, in every case, the insect sound is much shorter, about 1/10 to 1/1000 of a second, and *approximately* centered in the middle of the second.





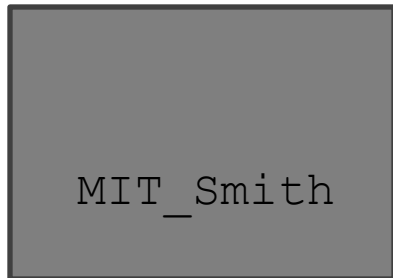
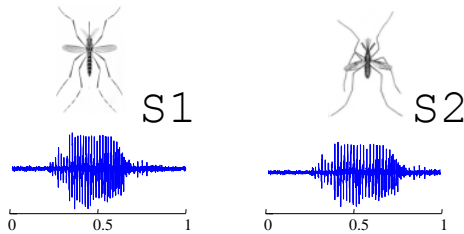
- Your distance function must be written in Matlab, version 7 or later.
- It must be a single m-file. However, within the m-file you can do anything you want, except access the web (assume zero network connectivity)
- Suppose your team leaders name is **Smith**, and your team is from **MIT**. Then your function should be called **MIT\_Smith.m**
- The first line of your function should be:

```
function dist = MIT_Smith(S1,S2)
```

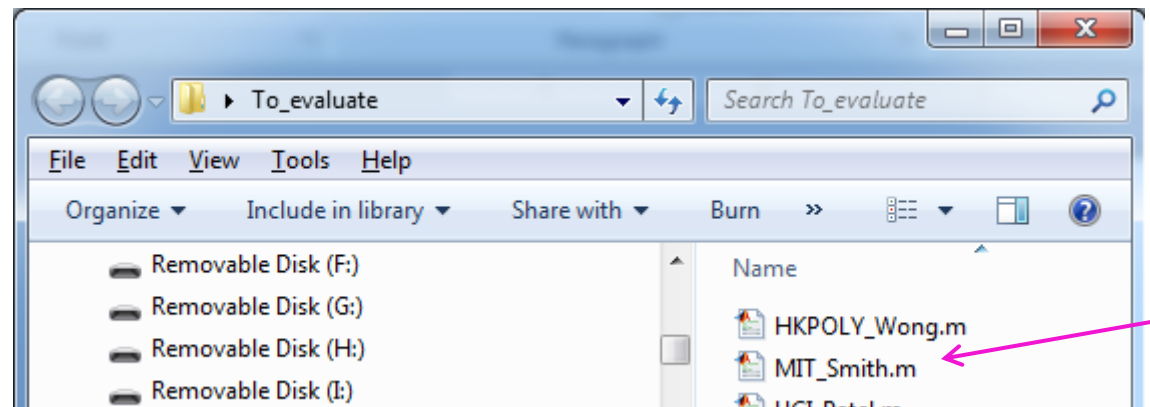
```
% This is the entry of Sue Smiths team
% to the UCR insect classification contest
% Team is Sue Smith and Joe Patel
% Contact info is sue@MIT.edu
```

```
...
```

The file name and function name must match



0.3



# Getting Started: I

For simplicity, we assume you have deleted all wav files from your default matlab directory. Download the `UCR_Contest.zip` file from [www.cs.ucr.edu/~eamonn/CE/](http://www.cs.ucr.edu/~eamonn/CE/) Unzip the file into your default matlab directory. Type `>> edit UCR_insect_classification_contest`

You can examine this file, which is what we suggest you use to test your function. It is just a simple 35-line leave-one-out nearest neighbor classifier.

Note the sub-function marked in pink. This is a sample entry, by a team lead by Prof John Doe from UCR. Hence he named the function `UCR_JohnDoe`. The function takes in two vectors (which are sound files) and returns their distance.

Simple 35-line leave-one-out nearest neighbor classifier.

```
% Here is a sample distance measure algorithm, rewrite and rename.
function this_distance = UCR_JohnDoe(unknown_object, compare_to_this_object)

% This is the entry of Prof John Does team to the UCR insect classification contest
% Team is John Doe and Quintin Qoe. Contact info is doe@UCR.edu

this_distance = rand; % This is a dummy distance function, write your own!
```

```
Editor: C:\Users\eamonn\Documents\MATLAB\UCR_insect_classification_contest.m
File Edit Insert Tools Debug Desktop Window Help
1 function UCR_insect_classification_contest(yourFunctionName) % Yanning Chen, Eamonn Keogh, Gustavo Batista
2
3 Distance_Algorithm = str2func(yourFunctionName);
4 TEST_DATA = dir('*.wav');
5 TEST_class_labels = importdata('classLabel.mat');
6 correct = 0; % Initialize the number we got correct
7
8 for i = 1 : length(TEST_class_labels) % Loop over every instance in the test set
9 classify_this_object = wavread(TEST_DATA(i).name); % Get object to test
10 this_object_actual_class = TEST_class_labels(i,2); % Find ground truth
11 best_distance = inf;
12 for j = 1 : length(TEST_DATA)
13 if [j == i] % Don't compare to self
14 object_to_compare = wavread(TEST_DATA(j).name);
15 this_distance = Distance_Algorithm(classify_this_object,object_to_compare);
16 if (this_distance < best_distance)
17 best_distance = this_distance;
18 predicted_class = TEST_class_labels(j,2);
19 end
20 end
21 end
22 if predicted_class == this_object_actual_class
23 correct = correct + 1;
24 disp([int2str(i), ' out of ', int2str(length(TEST_class_labels)), ' done, correctly classified ']) % Report progress
25 else
26 disp([int2str(i), ' out of ', int2str(length(TEST_class_labels)), ' done, misclassified ']) % Report progress
27 end;
28 end;
29
30 % Create Report
31 disp(['Evaluation results for ', yourFunctionName]);
32 disp(['The dataset you tested has ', int2str(length(TEST_class_labels(:,2))), ' classes']);
33 disp(['The data set is of size ', int2str(length(TEST_class_labels(:,1)))]);
34 disp(['The error rate was ', num2str((length(TEST_class_labels)-correct)/length(TEST_class_labels))]);
35 % End Report
36
37 % Here is a sample distance measure algorithm, rewrite and rename.
38 function this_distance = UCR_JohnDoe(unknown_object, compare_to_this_object)
39
40 % This is the entry of Prof John Does team to the UCR insect classification contest
41 % Team is John Doe and Quintin Qoe. Contact info is doe@UCR.edu
42
43 this_distance = rand; % This is a dummy distance function, write your own!
44
```

# Getting Started: II

Let us test this file, type `>> UCR_insect_classification_contest('UCR_JohnDoe')`

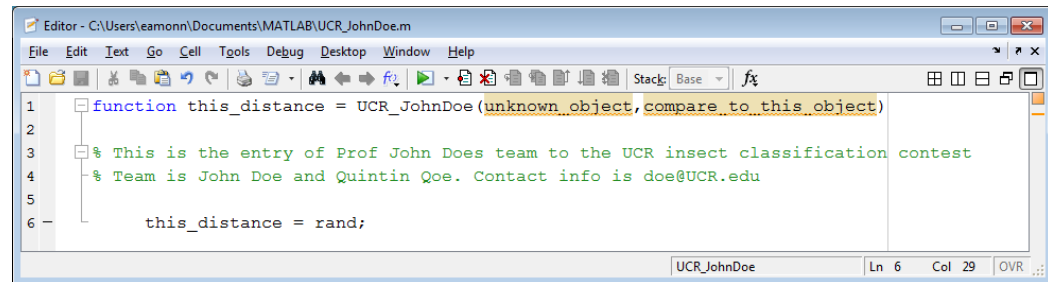
Note that we had to pass in the name of our function. The result is:

```
EDU>> UCR_insect_classification_contest('UCR_JohnDoe')
1 out of 500 done, misclassified
2 out of 500 done, misclassified
3 out of 500 done, correctly classified
...
498 out of 500 done, misclassified
499 out of 500 done, misclassified
500 out of 500 done, misclassified
Evaluation results for UCR_JohnDoe:
The dataset you tested has 5 classes
The data set is of size 500.
The error rate was 0.786
```

Suppose that Prof. Doe is happy with his result and he wants to submit it. He will send an email to [UCR.insect.contest2012@cs.ucr.edu](mailto:UCR.insect.contest2012@cs.ucr.edu) with **just** the m-file, like this:

Note:

- The function name and the file name must be the same.
- The function name must be in the form `<organization>_<team leaders name>`
- The comments must list the team, and an email contact.
- Do **not** include our 35-line leave-one-out nearest neighbor classifier! Send *only* your distance function.



```
Editor - C:\Users\eamonn\Documents\MATLAB\UCR_JohnDoe.m
File Edit Text Go Cell Tools Debug Desktop Window Help
Stack: Base fx
1 function this_distance = UCR_JohnDoe(unknown_object,compare_to_this_object)
2
3 % This is the entry of Prof John Does team to the UCR insect classification contest
4 % Team is John Doe and Quintin Qoe. Contact info is doe@UCR.edu
5
6     this_distance = rand;
```

- We don't care if you create a distance function or a similarity function\*.
- However our evaluation system assumes a *distance* function.
- Thus, if you are creating a similarity function, in the last line of your code, convert it to a distance function using:

$$x = 1 - x \quad \text{or} \quad x = 1/(x+\text{eps})$$

(or whatever is the appropriate transformation, this is *your* decision)

\* In general, distance functions range from 0 to inf, with smaller being more similar. Similarity functions range from 0 to 1, with larger being more similar. We don't care if you have a measure or metric or ultrametric etc

- You can assume the following toolboxes are on our machine.

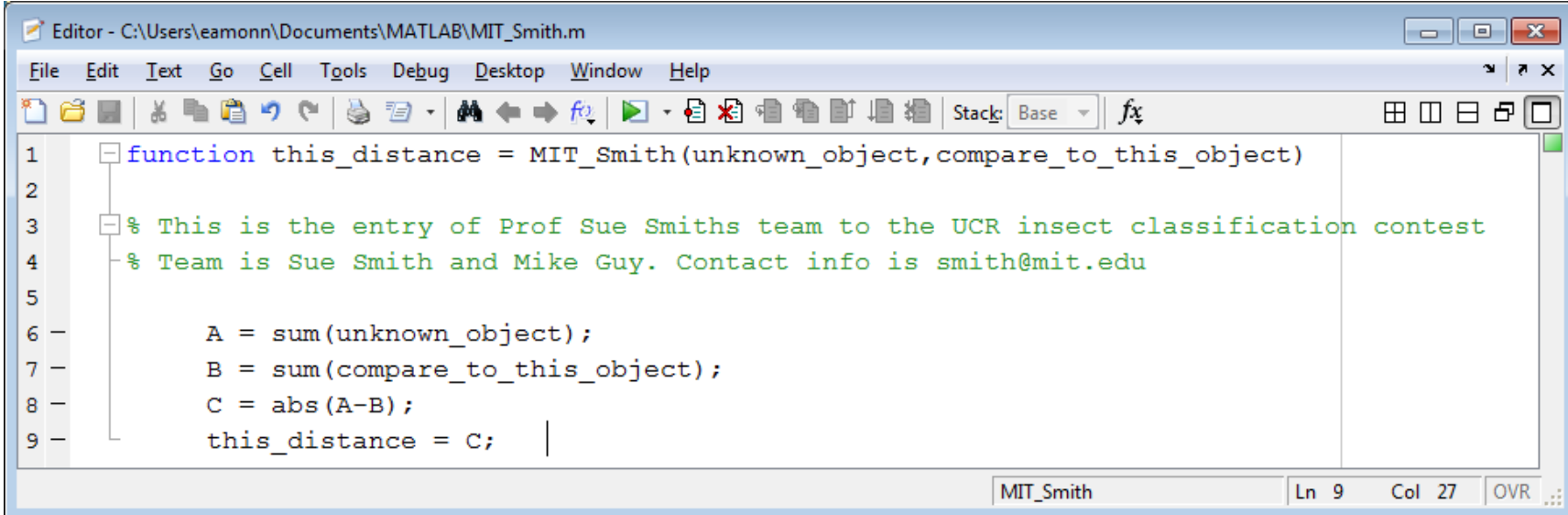
Simulink	Version 7.5	(R2010a)
Control System Toolbox	Version 8.5	(R2010a)
Image Processing Toolbox	Version 7.0	(R2010a)
Optimization Toolbox	Version 5.0	(R2010a)
Signal Processing Blockset	Version 7.0	(R2010a)
Signal Processing Toolbox	Version 6.13	(R2010a)
Statistics Toolbox	Version 7.3	(R2010a)
Symbolic Math Toolbox	Version 5.4	(R2010a)

- If your code requires additional toolboxes, you will need to provide them to us (at your expense) with instructions. We will spend up to one hour trying to installing the provided toolbox(s), after that we disqualify your entry.

# Checklist: Please check before submitting

- Did you send your entry to [UCR.insect.contest2012@cs.ucr.edu](mailto:UCR.insect.contest2012@cs.ucr.edu) ?
- Is your entry in the format below?
  - Filename and function name are the same
  - Filename is *<name of your institution> <underscore> <name of team leader>*
  - Function is in the sample format. It takes in **just two equal length vectors**, and returns a **single** number.
  - The comments list the team name and a contact email.
- Did you send **only** your single m-file (like the sample below), we don't want any other wrapper code, we don't want you to send your m-file embedded in our UCR\_insect\_classification\_contest.m etc

A sample entry looks like this (but will probably be longer, have more code etc)



```
Editor - C:\Users\eamonn\Documents\MATLAB\MIT_Smith.m
File Edit Text Go Cell Tools Debug Desktop Window Help
[Icons] Stack: Base fx
1 function this_distance = MIT_Smith(unknown_object,compare_to_this_object)
2
3 % This is the entry of Prof Sue Smiths team to the UCR insect classification contest
4 % Team is Sue Smith and Mike Guy. Contact info is smith@mit.edu
5
6     A = sum(unknown_object);
7     B = sum(compare_to_this_object);
8     C = abs(A-B);
9     this_distance = C;
```

MIT\_Smith Ln 9 Col 27 OVR