

## 2009 IEEE ICDM PBC Brain Connectivity Challenge

2009 Institute of Electrical and Electronics Engineers International Conference on Data Mining  
Pittsburgh Brain Competition Brain Connectivity Challenge

### 1. General Description of the Problem

**Overview.** Mapping the connections of the human brain is a historic scientific challenge with important medical applications. The 2009 ICDM Pittsburgh Brain Competition (PBC) seeks to map the "cables of the brain." **The goal of this competition is to take the set of 300,000 streamlines of brain axon paths and classify them into 20-50 fiber tracts that provide a consistent segmentation across multiple scans of the same person and between persons.** The competition seeks to identify data mining methods that can perform useful automatic segmentation of the human brain fiber tracts. The methods of segmentation are valuable for basic neuroscience, medical applications, and reverse engineering of the brain.

**Background.** The human brain has over 150,000 km of fibers (see [wiki](#)) connecting a hundred billion neurons. The [axons](#) travel in identifiable **fiber tracts** where many neurons travel following a similar path as a bundle of fibers pass through the brain (see Figure 1). These tracts appear in atlases used by surgeons and neuroscientists. Computational segmentation methods are not available to do accurate segmentation at present.

**Data Set.** The competition data set was collected using Magnetic Resonance Imaging (MRI) based [diffusion weighted techniques](#) to image the diffusion "pipes" representing bundles of cortical axons (connections) non-invasively on live humans. The PBC staff created High Definition Fiber Tracking (HDFT) data sets for three individuals using advanced MRI based diffusion imaging (Diffusion Spectrum Imaging) and image reconstruction techniques (multi-shell Q-Ball) which is likely the highest resolution data set of its type ever provided. The data set includes about 300,000 fiber streamlines per brain scan per person scanned on two occasions. The data include one training data set and five test data sets.

**Tests and Data Submission.** There are two broad goals: 1) supervised learning to segment the brain and match the fiber tract mapping of an expert neuroanatomist; 2) unsupervised learning segmenting fiber tracts consistently within and between brains maximizing similarity of the tracts. For each brain an expert neuroanatomist has segmented the brain identifying eight major fiber tracts that have been generally agreed upon in the neuroanatomy scientific community. The quantitative tests include tests: 1) *Supervised Learning MatchExpert - segment the tracts identifying the eight major tracts; and Unsupervised Learning* 2A) *Within subject within scan segmentation of all tracts; 2B) Within subject between scan segmentation; 2C) Between subject segmentation.* Contestant researchers must segment/categorize six brain scans (3 individuals scanned twice) identifying 20-50 fiber tracts that account for over 80% of the fibers. The PBC site will compute a quality of segmentation score on submitted entries. Contestants can submit the training data Brain<sub>1</sub>Scan<sub>1</sub> once per day. The other brains (Brain<sub>1</sub>Scan<sub>2</sub> to Brain<sub>3</sub>Scan<sub>2</sub>) can be submitted at most 5 times and once per day. **Results must be uploaded by November 2, 2009 and method write-ups by November 6, 5PM EST.**

**Awards & Scoring.** Contestant researchers will be able to examine their scores on each test and compare their scores to all the other entries at the end of the competition. They will be able to download fiber graphic viewing tools to view their results and compare them to other entries (see Figure 1). The awards will be presented at the [IEEE ICDM meeting](#) in December 2009 in Florida, USA. The awards will include: **Test 1 Expert Match Award** (\$2000 (US)) and **Test 2 Within and Between Subject Award** (\$2000). In addition, we offer the **Contest Crown** to the group that scores top in both test, which includes the option for one team member to fly to Pittsburgh, have a full HDFT scan done and the data made available to the team, and presented at ICDM (Includes \$1000 for travel or just the cash). **Contact/Information/Support.** Information about the competition can be obtained from [braincompetition.org](http://braincompetition.org). For questions and comments email [pbcc\(at\)pitt\(dot\)edu](mailto:pbcc@pitt.edu). The competition is run by the Pittsburgh Brain Competition project of the University of Pittsburgh and is advised by the [PBC Board and staff](#).



**Figure 1. Major fiber tracts that are the target for the competition**

- 1.1. *Contest YouTube overview and tutorial.* There is a short overview (9 minute) YouTube video [http://www.youtube.com/watch?v=IIDv4FLRg\\_4](http://www.youtube.com/watch?v=IIDv4FLRg_4) detailing the competition goals and methods (to be released September 8). There will be a tutorial (~ 30 minutes) detailing the methods involved to be released September 14.
- 1.2. *Applied value.* The within subject localization is used to follow an individual during development and medical intervention (e.g., neurosurgery, traumatic brain injury, Alzheimer's). The between subject comparison allows between group assessment (e.g., normal versus individuals with autism).

- 1.3. *Version Number.* Draft 0.2 release date 9/8/2009, check web site (braincompetition.org) for updates. Note the competition is advised by the [PBC Board](#) and there are likely to be small changes to the procedures in order to improve the scientific benefit of the competition.
- 1.4. *Data availability.* The PBC has a policy to make the data available for all competitions for research purposes for at least a year after the competition closes. It has so far made the data available for past competitions as long as there is use of the data by researchers. Hence, data from the 2006-2009 competitions are still available online. Access to the scoring programs is available though we do not allow entries between the time that the submission process closes and the awards are announced (typically one month). After awards are announced the submission process is re-opened.

## **2. Important Dates 2009 IEEE ICDM PBC Brain Connectivity Challenge**

<i>September 8</i>	<i>Announcement of competition, allow sign up,</i>
<i>September 14</i>	<i>Release of training data and scoring tools - web cast detailing procedures and providing a review of successful methods utilized in processing fiber data</i>
<i>October 1</i>	<i>Final release of all data sets</i>
<i>October 15</i>	<i>Allow scoring of all data</i>
<i>October 26</i>	<i>Registration closes</i>
<i>November 2</i>	<i>Results must be submitted 5PM US Eastern Standard Time</i>
<i>November 6</i>	<i>Write up of the methods due 5PM US Eastern Standard Time</i>
<i>December 6-9</i>	<i>Results of competition announced</i>

## **3. Rules**

All contestant researchers must abide by rules relating to ethical use of the data and publication communication. These include:

- 3.1. All contestants will register for the competition (braincompetition.org) providing contract information and agree to the terms of use of the data. For groups there will be only one entry per group and the group leader will identify the names and emails of the other members of the team. These may be changed up until the final submission.
- 3.2. The submitted materials can be made available on the PBC website including the methods write up and the score of the method. (Note: if desired we can report the results without identifying the group if the group should wish to remain anonymous). The default rule is we report the names of the top 50% of the entries. However we will show the rank score of the entry and place the method description on the web. This enables researchers looking over the web site to determine what worked and did not work over the entries and the techniques employed.
- 3.3. The entrant may submit up to 5 submissions of their results for scoring of the data with a maximum of one submission per day.
- 3.4. If the contestant publishes results using the PBC data, they will reference the PBC website and where appropriate, they agree to cite publications that describe

how the data was collected and send PBC a copy of their publication. Note: we encourage publication.

- 3.5. Researchers must abide by the Institutional Review Board human subject restrictions relating to protecting the anonymity of the subjects. If the contestant utilizes the MRI structural images (available with special request), the contestant agrees they will not reconstruct the face of the subject (technically possible) and display that face in any form that would enable recognition of the individual.

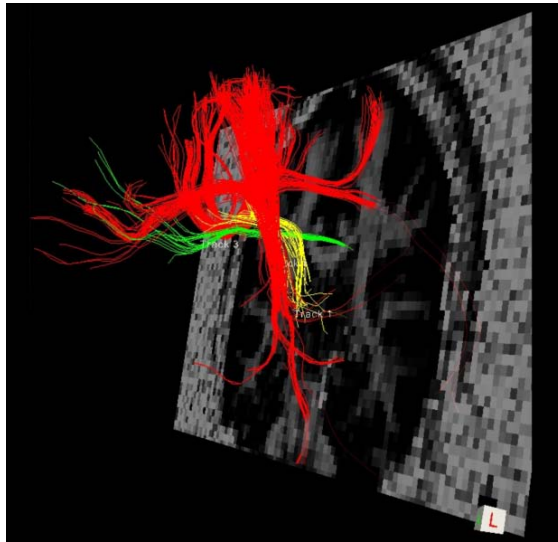
***Notice: The PBC Board reserves the right to modify competition rules in the event that new information is provided that indicates such changes would facilitate identifying more useful fiber brain connection mapping. All registered contestant researchers will be notified of any such change via their registered email address and notice will be placed on the web site.***

#### **4. Contest Challenges**

The contest involves mapping the cable pathways of the human brain. It will be done in a series of challenges. We do this through quantitative metrics of the quality of the segmentation.

##### **4.1. Supervised Learning Test 1 MatchExpert - segment the tracts identifying the 8 major tracts**

In this test the goal is to match the fiber tracts that our expert has identified in the data. Our expert, Juan C. Fernandez-Miranda, is a neurosurgeon and neuroanatomist with over a decade of experience in brain dissection and neurosurgery. He has identified tracts that have been recognized based on dissection work as the major tracts of the brain (see Figure 1). For each tract he identified three sets of fibers related to the tract in three states. We will refer to the sets as the green and red sets. The green fibers (the hit set) are the fibers that are clearly part of the tract. The red (the false alarm set) are fibers of other understood tracts or far distance fibers that could not be part of the tract. In some cases there may be yellow (fibers that are "too close to call"). They are ambiguous (e.g., following closely but make a turn that is not consistent with target tract and note scored). To use an analogy to times of the day, we have daylight (green), night (red), and dusk (yellow).



**Figure 2. Example three states coding of fiber data, green hit fibers, yellow ambiguous, red false alarms.**

The initial data set for our competition are the **diffusion stream lines** created via our **tractography** methods. For each brain there are about 300,000 such stream lines that we will refer to as **fibers**. You can think of these very much like you would a **fiber optic cable** with many strands taking a similar path and then at the end points stopping with different offsets.

Groups of individual fibers that travel a similar path are organized into fiber tracts. A **Fiber Tract** is a like a cable containing typically thousands of fibers with a coherent path. These fiber tracts take complex routes going from source to destination in the brain sometimes passing through other fiber tracts, sometimes turning to avoid crossing other tracts.

Our scoring metric for **Supervised Learning Test 1 MatchExpert** is the proportion of fibers (diffusion stream lines) that you matched that were classified by our expert as being part of a known tract. For each FiberTract you will get a **FiberTractScore** – the proportion of fibers that match the expert. The PBC scoring program will count the hit (green) fibers in your set, subtract the red fibers in your set, and divide by the total number of fibers our expert had for the tract (green).

$$\text{FiberTractScore} = (\text{GreenFibers} - \text{RedFibers}) / \text{GreenFibers}$$

Or

$$\text{FiberTractScore} = (\text{HitFibers} - \text{FalseAlarmFibers}) / \text{HitFibers}$$

Hence a perfect score would be 1.0. For any individual fiber tract we will restrict the range between 0-1 (note you could get a negative score if you had all red fibers for the target tract). For the **ContestExpertMatchFiberTractScore** we will take the average of the eight FiberTractScores for Brain<sub>2</sub> and Brain<sub>3</sub>. The Brain<sub>1</sub> data is used to develop the methods.

The green, yellow, red coding scheme was used to code the uncertainty that is present in the scanning and human classification methods. There are generally few or none yellow fibers so you will have to closely match the target to obtain a high score.

#### ***4.2. Unsupervised Learning - Full fiber tract segmentation within and between subject agreements.***

We seek to have a full brain segmentation that is reliable and consistent within and between subjects. In this test you must segment at least 80% of the fibers in the full test set into tracts. In dissection studies of the brain there are typically about 20-50 tracts that have been identified in various atlases of white matter. Depending on the researcher some tracks are divided into major levels and sub levels. There is no gold standard for tracts. There are atlases of fiber tracts (e.g., see [Schmahmann & Pandya 2006](#) and [Mori 2005](#) for an atlas). Also the major tracts can be divided into subtracts (e.g., a set of fibers with a specific known source and endpoint within a major tract) (e.g., in Figure 1 above the green fibers carry language related information). The competition will focus on machine learning/categorization to segment the major tracts.

The goal is to segment the brains using a consistent segmentation scheme within and between subjects and identify known tracts. In this case you will need to continue to categorize fibers into tracts until at least 80% of the fibers have been mapped and you have 20-50 identified tracts. The scoring will be based on the between and within subject accuracy of the tracts. We have scanned Brains 1-3 twice (data of the second scan will be released October 1).

The goal here is that you have good reliability within and between subjects. Contestants will submit a full fiber categorization scheme (e.g., 6 lists of about 300,000 fibers with the number 0-50 of the **TractNumber**. The tract numbering scheme must be consistent across all entries and the first 8 numbers must match the numbers provided in the Brain<sub>1</sub> dataset. TractNumber 0 can be used to indicate the fiber is not attributed to any tract). The PBC will score the between scan similarity of the fiber categorizations contestants have created. The concept is that a good segmentation scheme will create fiber bundles of high similarity within a fiber bundle and low similarity between bundles.

For each FiberTract the contestant researcher will identify the **ReferenceFiber** for that FiberTract. Contestants can choose their own ReferenceFibers or let the PBC scoring program choose the ReferenceFiber based on the geometric center of the bundle. Then the PBC will calculate the match of all the other fibers of the bundle to the appropriate reference fiber. This will continue starting with the largest to the smallest bundle until 80% of the fibers are accounted for. For all the fiber bundles for that subject we will have a list of the reference fibers (e.g., ReferenceFiber<sub>TractNumber</sub>) and number of fibers in the bundle (i.e., FiberCount<sub>TractNumber</sub>).

For each of 6 brains the contestant researcher will submit 2 files: 1) the **FiberTractReferencSummary** and 2) the **BrainCategorizedFibers** file.

**FiberTractReferencSummary** is an ASCII table similar to the columns identified and the text strings delimited by either a tab or space with an end of line (ASCII 015) at the end of the line (see below).

<b>FiberTractReferencSummary</b>		
<b>TractNumber</b>	<b>ReferenceFiber</b>	<b>FiberCount</b>
1	125221 (or -1 if PBC to choose)	421
...	...	...
50 (or less)	...	...

**BrainCategorizedFibers.** A list of an ASCII table with exactly the number or rows of the brain **FiberList** (see below). This is an ASCII table each line containing a number (0-50) and end or line (ASCII 015).

<b>BrainCategorizedFibers</b>	
<b>FiberNumber (not in table but offset)</b>	<b>BundleNumber</b>
1	1
2	1
...	...
297475 ( <i>example end line</i> )	50

A full submission of an entry will include six files:

<b>FileList for Submission</b>		
<b>Data set</b>	<b>FiberTractReferencSummary</b>	<b>BrainCategorizedFibers</b>
Brain1Scan1	FiberTractReferencSummaryBrain1Scan1	BrainCategorizedFibersBrain1Scan1
Brain1Scan2	FiberTractReferencSummaryBrain1Scan2	BrainCategorizedFibersBrain1Scan2
Brain2Scan1	FiberTractReferencSummaryBrain2Scan1	BrainCategorizedFibersBrain2Scan1
Brain2Scan2	FiberTractReferencSummaryBrain2Scan2	BrainCategorizedFibersBrain2Scan2
Brain3Scan1	FiberTractReferencSummaryBrain3Scan1	BrainCategorizedFibersBrain3Scan1
Brain3Scan2	FiberTractReferencSummaryBrain3Scan2	BrainCategorizedFibersBrain3Scan2

Note: the first 8 tracts 1-8, must match the tract numbers of the expert coding system. These are identified for Brain<sub>1</sub>Scan<sub>1</sub>.

#### **4.2.1.Test2A Within subject within scan segmentation**

For Brain<sub>i</sub>Scan<sub>1</sub> what is the similarity of all the other fibers within that fiber tract? The similarity metric is based on the weighted average of a series of fiber metrics including correlation of Brain<sub>i</sub>Scan<sub>1</sub> fiber ReferenceFiber<sub>TractNumber</sub> to all the fibers within the fibertract in that brain. We will publish the metric on the web page and the final weighting is awaiting final Board approval. The metric is basically the correlation of the geometric center of mass, end points, length, and correlation of the fiber points at each tenth of the fiber length. For Within subject within scan segmentation the calculation will determine the number of bundles needed in order of size from the largest to the

smallest bundle until at least 80% of the fibers are included. That is, for each fiber tract calculate the similarity of the reference fiber to all the other fibers within the fiber tract. The calculation of the reference fiber to itself is excluded. This is averaged for all the fibers up to 80% of the fiber streamlines and for all three Scan<sub>1</sub> (Brain<sub>1</sub>Scan<sub>1</sub>, Brain<sub>2</sub>Scan<sub>1</sub>, Brain<sub>3</sub>Scan<sub>1</sub>). Note the data for Brain<sub>1</sub>Scan<sub>1</sub> can be submitted at most once per day for an unlimited number of days. The data for the other brains can only be submitted once per day for up to 5 times for a competition entry after October 15. Entries that do not categorize at least 80% with 50 or fewer bundles will not be considered complete.

#### ***4.2.2. Test2B Within subject between scan segmentation***

This test is similar to the within subject within scan test but is now on two different scans for the same subject. For each Brain (1-3) and each fiber bundle the average similarity is based on match of Scan<sub>1</sub> to Scan<sub>2</sub> of the same subject. That is, using the reference fibers from Brain<sub>1</sub>Scan<sub>1</sub> score the fibers of Brain<sub>1</sub>Scan<sub>2</sub> and similarly for scan Brain<sub>2</sub>Scan<sub>1</sub> score the fibers of Brain<sub>2</sub>Scan<sub>2</sub> and Brain<sub>3</sub>Scan<sub>1</sub> score the fibers of Brain<sub>3</sub>Scan<sub>2</sub>. Note the data for Brain<sub>1</sub>Scan<sub>1</sub> can be submitted an at most once per day. The data for the other brains can only be submitted once per day for up to 5 times for a competition entry.

#### ***4.2.3. Test2C Between subject segmentation***

This now uses the reference of Brain<sub>1</sub>Scan<sub>1</sub> to calculate the similarity of fibers to Brain<sub>2</sub>Scan<sub>1</sub> and Brain<sub>3</sub>Scan<sub>2</sub>. Note at this point it is essential that the same Fiber Tract numbering scheme is used across all the data sets. This test is similar to the within subject between scan test but is now on different subjects and it is expected that the tracts will move around between subjects. This will be calculated using Brain<sub>1</sub>Scan<sub>1</sub> to Brain<sub>2</sub>Scan<sub>1</sub> and Brain<sub>3</sub>Scan<sub>1</sub>. The data for the other brains can only be submitted once per day for up to 5 times for a competition entry. The scoring program **requires that the same tract numbering scheme be used in all the brains**. If there is not a matching tract for Brain<sub>2</sub>Scan<sub>1</sub> or Brain<sub>3</sub>Scan<sub>1</sub> a match score of 0 will be added for the fibers in Brain<sub>1</sub>Scan<sub>1</sub> that do not have a match.

### ***5. Contestant/Researcher's Package***

The details of the data sets are provided in the document PBC 2009 IEEE ICDM Brain Connectivity Challenge Data Guide.pdf. Basically for each brain there is a **FiberList** for each fiber streamline (typically 300,000 per brain) with parameters for each (e.g, length, geometric center of mass, endpoints, xyz points at millimetre length). There is a field in the data for the TRACK\_ID that labels the 8 tracts of Brain 1 (or zero indicating not in any of scored tracts).

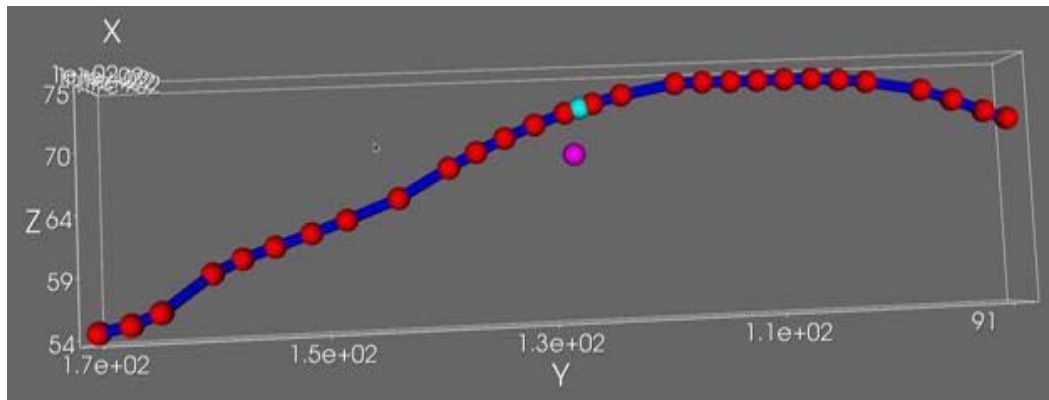
#### ***5.1. Description of Data Sets Provided***

Note there will be six directories for the six brain scans. Brain<sub>1</sub>Scan<sub>1</sub> to Brain<sub>3</sub>Scan<sub>2</sub>.

The files are:

- **FiberList** - lists all the fibers base features
- **FiberListExpanded** - full data with added features along the tract of the fiber that may such as the fractional anisotropy and Apparent Diffusion Coefficient at each point along the fiber. **Fractional Anisotropy (FA)** and **Apparent Diffusion Coefficient (ADC)**. Most contestants will likely not use the expanded data. It provides more information about the tract that might be useful. For example the FA measures when a fiber tract passes through another fiber tract and might be a useful feature to cluster fiber on.
- **RegionOfInterest FiberList**- This is a listing of the fibers for each brain making up a superset of the scored fiber set. This is provided to simplify testing of the Match. It lists for each expert coded FiberTract a set of fibers around that tract that might be close neighbours. A contestant might find it useful to use this set (about 30% of the full set) to develop their methods for the MatchExpert test.

The PBC 2009 IEEE ICDM Brain Connectivity Challenge Data Guide.pdf provides a detailed description of the data sets. The FiberList is a list of fiber lines. Each row of the table includes features describing the fiber. Figure 3 shows an example fiber. Note all location points are in a common brain space called [MNI space](#) in millimetres for the standard brain where x, y, z are right/left, forward/backwards, top/down and the 0,0,0 point that is about 4mm below a major fiber track called the [anterior commissure \(AC\)](#).



**Figure 3. Example fiber red points indicate fiber x,y,z position. Pink center of mass, light blue geometric center of mass.**

<b>FiberList</b> (one table per brain scan e.g., Brain <sub>1</sub> Scan <sub>1</sub> )		
<i>ColumnName</i>	<i>Description</i>	<i>Example</i>
<i>FiberID</i>	<i>Unique number</i> 1-300,000	2531

<i>FiberLenght</i>	<i>mm</i>	35.6
<i>NumberPoints</i>	<i>Number of points on fiber at 1mm spacing</i>	36
<i>CenterOfMass X,Y,Z</i>	<i>x,y,z in mm NMI</i>	40<tab>37<tab>5
<i>GeodesicCenterOfMass X,Y,Z</i>	<i>x,y,z in mm NMI</i>	39<tab>35<tab>5
<i>ListofPoints</i>	<i>NumberPoints list of fiber points x,y,z in mm NMI</i>	7<tab>11<tab>-5 ...

<b>FiberListExpanded</b> (one table per brain scan e.g., Brain <sub>1</sub> Scan <sub>1</sub> )		
<i>ColumnName</i>	<i>Description</i>	<i>Example</i>
<i>FiberID</i>	<i>Unique number 1-300,000</i>	2531
<i>NumberPoints</i>	<i>Number of points on fiber at 1mm spacing</i>	36
<i>FA values for each point on fiber</i>	<i>FA</i>	0.35<tab>...
<i>ADC values value for each point on fiber</i>	<i>ADC list</i>	41.3<tab> ...

## 6. Contest Papers (Draft Materials due one week after submission of data)

Each entry to be considered for a prize must submit a short abstract description for the entry.

### 6.1. Project abstract

**6.1.1. Title** of less than 100 characters

**6.1.2. Abstract** of less than 500 words

**6.1.3. Organization name and team description**

**6.1.4. Optional Links to team related web sites**

**6.1.5. Publication permission** and whether to disclose the team identity

**6.2. Project description** – PDF file 2-10 pages describing the methods used and what was learned as to what worked and did not work. We encourage submitters to do sensitivity analysis of their methods to characterize the marginal utility of various components of the approach they took.

**6.3. Visualization of tracts** – The PBC will provide freely available for academic/research/non-commercial use tools to load and visualize the tract files so contestants can pull up and examine their solutions graphically and optionally include graphics in their submission. The program will use a standard coloring scheme for the eight target tracts to allow easy comparison across entries. After submissions are complete tools will allow visually comparing your solution to that of other entrants.

## 7. Contest prizes and scoring

**7.1. Reporting.** Each team that has submitted a full entry will receive a score of their submission including how they did on each metric with plots of the range of all

the scores on the submissions on each BrainScan and in total. This scoring data will be available on the web so contestants can determine what groups did better on each metric. A query data base will be available to determine which method did best on each of the eight scored tracts, tracts of a given size, and particular brains. Contestants will be able to graphically look at your data can compare it to the data of the winning entries.

## **7.2. Contest Awards**

7.2.1. The awards will be presented at the [IEEE ICDM meeting](#) in December 2009 in Florida, USA. The awards will include: **Test 1 Expert Match Award** (\$2000) and **Test 2 Within and Between Subject Award** (\$2000 US) plus the **Contest Crown** special offer, if a group scores tops in both, we will offer the option for member of team to fly to Pittsburgh and have a full HDFT scan done, the data made available to the team and presented at ICDM (Includes \$1000 for travel or they can take the cash). **Contact/Information/Support.** Information about the competition can be obtained from [braincompetition.org](http://braincompetition.org). **Contest awards.** There will be a listing of the top three scores on each of the 4 tests.

7.2.2. **Board Choice awards.** The [Board members](#) will review the top entries and provide an award for the entry that appears to make the greatest contribution and benefit in two categories based on subjective evaluation of the entries with some reference to the quantitative scores:

Note the Board has the option to not give awards if there are insufficient entries to judge or to split prizes in multiple entries reach and equivalent level of performance. Write ups that do not detail the methods sufficiently to interpret the quality of the submission may be excluded. The PBC staff may direct specific requests for additional information regarding the methods used or request additional brain data sets to be processed.

## **8. Contest Board & Sponsoring Organization, Contact Information**

The Pittsburgh Brain Competition is run under the direction of Walter Schneider, Professor of Psychology at the [University of Pittsburgh](#) as a research activity. Decisions are made by the [PBC Board](#) of thirteen experts of brain imaging and computational methods and staff at the University of Pittsburgh. This is a non-profit scientific competition. For questions and comments send note to [pbcc@pitt.edu](mailto:pbcc@pitt.edu). This project is funded through grants from University of Pittsburgh. The scanning was done at the [University of Pittsburgh Medical Center](#), MR Research Center which provided support for this project.

*Draft September 8, 2009, see [braincompetition.org](http://braincompetition.org) for updates*