# Focus Group on Unified Data Mining Engine (UDME 2010): Addressing Challenges

***2-Hour Focus Group Proposal***

**SHIVANSHU K. SINGH (PRIMARY CONTACT)**
Department of Computer Engineering,
Charles W. Davidson College of Engineering,
San Jose State University,
One Washington Square, San José, CA 95192-0180.
Ph: (408) 701-8294
E-mail: shivanshukumar@gmail.com, shivanshu@vrlsoft.com
URL: http://www.shivanshusingh.com

**VIJAY KUMER ERANTI**
Google Inc.,
Mountain View, CA, USA
evijaykumar@yahoo.com

**DR. M.E. FAYAD**
Professor of Computer Engineering
Computer Engineering Dept., College of Engineering
San José State University
One Washington Square, San José, CA 95192-0180
Ph: (408) 924-7364, Fax: (408) 924-4153
E-mail: m.fayad@sjsu.edu, mfayad@vrlsoft.com
http://www.engr.sjsu.edu/fayad
pattern.ijop.org – **Pattern Blog**
fayadsblog.vrlsoft.com – Fayad's Report

## INTRODUCTION

Data mining is the discovery of knowledge of analyzing enormous set of data, by extracting the meaning of the data, and then predicting the future trends. Data mining also helps to find out secret information from large databases, and also assists companies to take sound decisions, based on knowledge and information.

If we closely look into any data-mining tool, we can see that there are some common core logic, which are independent of the data and the applications, but most of existing implementations try to ignore that fact and concentrate on the specific problem, in that way the tool becomes limited to only to a particular set of data for specific application.

Data mining is finding interesting patterns in data. The main challenge of any data-mining engine is how to apply different algorithms or different techniques on different set of data to find an interesting pattern, which is useful to business. It is very extremely difficult to come with some standard way of analyzing the data. The enormous volume and the complexity of the data makes it impossible to run same algorithms on different dataset. Nowadays, there are different vendors, who are trying to solve this problem, but mostly they support a subset of different algorithms. None of them has come up with any stable engine, which can work in any data set and in any domain.

In the last decade, the improvement in storage and CPU speed has created a huge opportunity for different data mining applications, ranging from CRM to medical health care application. The evolution of data mining is shown in table 1.

Now it is very difficult to develop a single application, which can take care all of these problems. It's a dream to think of an application, which can iterate through any data and create pattern. Data mining deals with useful pattern and not just pattern, now whether a pattern is useful or not depend on the context where it is applied. Present day tools depend mainly on the expert about what kind of algorithms to apply and how to analyze the output, because most of them are generic and there is no context specific logic that is attached to the application.

| **Evolutionary Step** | **Business Question** | **Enabling Technologies** | **Product Providers** | **Cha** |
|---|---|---|---|---|
| Data Collection<br>(1960s) | "What was my total revenue in the last five years?" | Computers, tapes, disks | IBM, CDC | Retr e, sta de |
| Data Access<br>(1980s) | "What were unit sales in New England last March?" | Relational databases (RDBMS), Structured Query Language (SQL), ODBC | Oracle, Sybase, Informix, IBM, Microsoft | Retr e, d data at l |
| Data Warehousing & Decision Support (1990s) | "What were unit sales in New England last March? Drill down to Boston." | On-line analytic processing (OLAP), multidimensional databases, data warehouses | Pilot, Comshare, Arbor, Cognos, Microstrategy | Retr e, d data at n le |

| | | | | |
|---|---|---|---|---|
| Data Mining<br>(Emerging Today) | "What's likely to happen to Boston unit sales, next month? Why?" | Advanced algorithms, multiprocessor computers, massive databases | Pilot, Lockheed, IBM, SGI, numerous startups (nascent industry) | Pros prc info de |

Table 1. Steps in the Evolution of Data Mining [12].

Here is summary of the problems that we are facing  today in the existing data mining tools

1. **Difficult to use**– Existing data mining tools try to cover different data mining applications, thus become very difficult to configure and run.
2. **Needs Expert to run the tool** – No domain or problem specific logic is tied with the tool, therefore needs expert to run the to tool and analyze the result
3. **Difficult to add new functionality** - Because of the size and complexity of each tool, it is very difficult to add any new feature.
4. **Difficult to interface** -  There is no way  that the algorithms can be developed by some other companies that can be integrated with the tool easily
5. **Short Lifetime -**  There is no stable component in the tool and with time the tool becomes obsolete, as new tools take the market, and changing the exiting tool to incorporate new feature is difficult and require lot of changes.
6. **Limited Number of algorithms** – Existing tools only provide limited number of algorithm and sometimes use of multiple algorithms is quite limited.
7. **Need lot of resources**: Existing tools are not optimized for any specific application, therefore they need lot of resources, such as runtime memory, hard disk etc.

Thus, this focus group topic is driven forward by three main questions. First, "how can we develop a unified data mining engine {UDME)?" Second, "what kind of technologies and tools to build such an Engine?" and third, "how can we overcome the existing problems?"

## OBJECTIVE AND MOTIVATION

Building such an engine is not an easy exercise, specifically, when several factors can undermine their quality success, such as cost, time, and lack of systematic approaches. We would like to architect and develop a Unified Data Mining Engine (UDME), that has the some or all of the following properties:

1. **Ease of use**– Multiple tools can be developed easily mainly by focusing on specific problems, because they all can share the core services, that are provided by the UDME.
2. **No Need of Expert to run the tool** – Domain specific knowledge, such as verification, selection of tool etc can be implemented in the tool itself, while developing the tool.
3. **Easy to add new functionality** - – The application specific logic should be separate from the core logic, therefore new application specific functionality can be added quite easily, without making any change in the core logic.
4. **Easy to interface** - The design should be based on system of independent patterns, they can be developed by 3rd party vendors.
5. **Long Lifetime -** The engine should also be based on stable core logic, which has a long lifetime, the application logic should be loosely connected which can change over time.
6. **Multiple algorithms** – The engine must support any number of algorithms.
7. **Fewer resources**: The proposed engine should be developed by connecting several patterns or components. Depending on the application domain, the engine can use patterns or components, which are necessary and thus it needs less resources when compared to existing tools.
8. **Stable**: The engine should be stable over time, and provide a simple way to apply different data mining and data analysis algorithms on different sets of data in any domain.
9. **Isolation of Application logic**: We must also isolate the stable knowledge from any application specific logic, therefore different applications can use the same core knowledge, which need not be changed.
10. **Minimum Maintenance Cost** – Maintenance cost of such an engine should be very minimal.

The focus group topic will address the unified data mining engine' challenges, and also debate several issues that are related to the following questions. We also want researchers, framework developers, and application developers to discuss and debate the following questions that are related to:

I. **UDME Architecture**
   a. What is the best approach for building such an engine?
   b. What are the bases of creating the engine architecture?
   c. Are there any guidelines, methodologies, and/or processes for an engine architecture creation and development?
   d. What are the components of the unified data mining engine architecture?
   e. What kind of patterns or components that appear in UDME ?
   f. Show, how your engine architecture meets the above UDME properties?

II. **UDME Development**
   a. What is the ultimate way to develop such an engine?
   b. What are the techniques and tools that are used for developing such an engine?
   c. Show, how to extend your engine to new application logics?

## PARTICIPANTS

The projected audience will comprise of software engineering researchers, software analysts, software architects, software designers, software developers, Object-Oriented technologists, and software methodologists. Others, who may be interested, are also welcome to participate. Participants are expected to possess basic background on software modeling, object-oriented technology, and software patterns. Familiarity and association with UML is also preferred. The expected number of participants for this focus group is around 20+.

## SOLICITATION, SUBMISSION AND SELECTION PROCESS

We will have the following system of invitation:

1. General invitation: We will have a general call for papers, that will be publicized to over 10,000+ possible participants and an invitation will be sent to all the patterns groups.
2. Special invitation: We will have a large number of key people in the pattern community who will be receiving their call for papers.
3. Very special invitation: We will also send an invitation to some noteworthy people to participate and submit their papers to the focus group.

Each submission will be reviewed by at least three experienced reviewers. Based on the received reviews, the organizers will then select the accepted papers.

## FOCUS GROUP FORMAT

The two-hour focus group will consist of invited speakers and single-track presentation sessions. The overall action plan is to have an open seesion for presentations of position papers and discussion. The main theme of the sessions will be determined based on the position statements. All accepted position statements are expected to be presented in the focus group. A summary report on the focus group will also be posted on the web. (Please refer to the tentative agenda of the focus group in the preliminary call for papers given at the end of this proposal)

## PREVIOUS WORKSHOPS

Previous workshop on the same topic:

1. "Accomplishing Software Stability Workshop" with M. Laitinen, OOPSLA '99, Denver, Colorado, Nov 1999. The number of participants was 27.
2. "Software Stability: Timeless Architectures, Systems of Patterns, and Model-Based for Reuse" Mohamed E. Fayad and Haitham S. Hamza, IEEE IRI 03, Las Vegas, Nevada, October 2003.
3. **M.E. Fayad** and H.S. Hamza. AICCSA '05 First Workshop on Software Stability: Timeless Architectures and Systems of Patterns. AICCSA '05 Full day workshop, The 3rd ACS/IEEE International Conference on Computer Science Systems and Applications, January 3, 2005, Cairo, Egypt. http://www.engr.sjsu.edu/fayad/workshops/AICCSA-05/
4. **M.E. Fayad** and H.S. Hamza. IEEE '05 Second Workshop on Software Stability @ Work. IEEE IRI '05 Full day workshop, The 2005 IEEE International Conference on Information Reuse and Integration (IEEE IRI '05), August 15-17, 2005, Las Vegas, NV, USA.

5. **M.E. Fayad** and H.S. Hamza. ECOOP 2005 First Workshop on Building Systems Using Patterns: Examine the Illustrious Claim. Full day workshop, 2005 19th European Conference on Object-Oriented Programming (ECOOP 2005), July 25-29, 2005, Glasgow, Scotland .**http://2005.ecoop.org/workshops.html**, and **http://www.engr.sjsu.edu/fayad/workshops/ECOOP05/**
6. **M.E. Fayad, I.A. Zualkernan.** and H.S. Hamza. AICCSA06 5th International Workshop on Software Stability: Methodologies, Applications and Tools. Full day workshop, 2006 4th ACS/IEEE International Conference on Computer Science Systems and Applications, *March 8-11, 2006, Sharjah, United Arab Emirates* http://www.cs.utk.edu/aiccsa06/
7. **M.E. Fayad, H.S. Hamza, and E. Segura**. The First IEEE International Workshop on Software Pattern: Addressing Challenges (SPAC 07). In conjunction with COMPSAC 2007 -- Full day workshop, Beijing, July 24-27, 2OO7. **http://conferences.computer.org/compsac/2007/workshops/SPAC**
8. Licia Capra, Rami Bahsoon, Wolfgang Emmerich, **M.E. Fayad**: The international workshop on software architectures and mobility (SAM 2008). ICSE Companion 2008: 1033-1034 http://www.cs.bham.ac.uk/~rzb/sam.htm

## REQUESTED EQUIPMENT

There are no special requirements. Standard equipment for PowerPoint presentations and an overhead projector are just sufficient.

## ORGANIZERS' BIOGRAPHIES

**SHIVANSHU K. SINGH** is a Co-Founder and Senior Researcher at vrlSoft, Inc., Palo Alto, California, USA. His research is focused on the areas of Unified Software Engines, Software Architecture, Architectural and Stable patterns, Knowledge Maps, Spatiotemporal databases, Software Engineering processes, Requirements Engineering, Collaborative Systems and more. Shivanshu received his Bachelor of Technology degree in Information and Communication Technology from Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, India in 2007 and has 3+ years of professional experience in software engineering, research and development and teaching. He is involved in the development of some major journals in the field of software engineering and multiple new business developments. He has multiple journal, conference and column publications in his name. He is a columnist in the International Journal of Software Architecture (IJSA) and the Editor in Chief for the International Journal of Unified Software Engines (IJUSE). He is involved in writing several successful research and business proposals to various government agencies, institutions, public and private organizations, NSF, SBIR and industrial proposals etc. and is also in the process of writing two books, to be titled 'Knowledge Maps' and 'The Unified Software Engine'. Shivanshu is the Lead – Research and Development, of the research team, with more than 10 graduate students, working on several areas related to Software Engineering research, under Dr. M. E. Fayad at San Jose State University. He is an invited member of the Phi Kappa Phi Honor Society, for academic excellence; He is currently working towards his Master of Science degree in Software Engineering at San Jose State University, San Jose, California and towards a PhD degree thereafter.

**DR. M.E. FAYAD** is a Full Professor of Computer Engineering at San Jose State University from 2002 to present. He is one of the **founders and president** of ***Arab Computer Society (ACS)*** from April 04 to April 2007. Dr. Fayad is ***a known and well recognized authority*** in the domain of theory and the applications of software engineering. Fayad's publications are in the very core, archival journals and conferences in the software engineering field**.** Dr. Fayad has published more than **218** high quality papers, that includes more than 40 profound reports in reputed journals, and 90 advanced articles in refereed conferences, more than 20 journal columns, 16 blogged columns; 9 well-cited theme issues in prestigious journals and flagship magazines, 24 different workshops in very respected conferences, over 125 tutorials, seminars, and short presentations in 20+ different counties, a founder of 5 new online journals, NASA Red Team Review of QRAS and NSF-USA Research Delegations' Workshops to Argentina and Chili and four authoritative books, of which three of them are translated into different languages such as Chinese. **Dr. Fayad** received an MS and a Ph.D. in computer science from the University of Minnesota at Minneapolis. He is the lead author of several classic Wiley books**.**

==============================================================

## CALL FOR PAPERS

# Focus Group on Unified Data Mining Engine (UDME 2010): Addressing Challenges

*Reno/Tahoe, Nevada, USA,* ***October 17 - 21, 2010***

**(in conjunction with SPLASH 2010)**

## INTRODUCTION

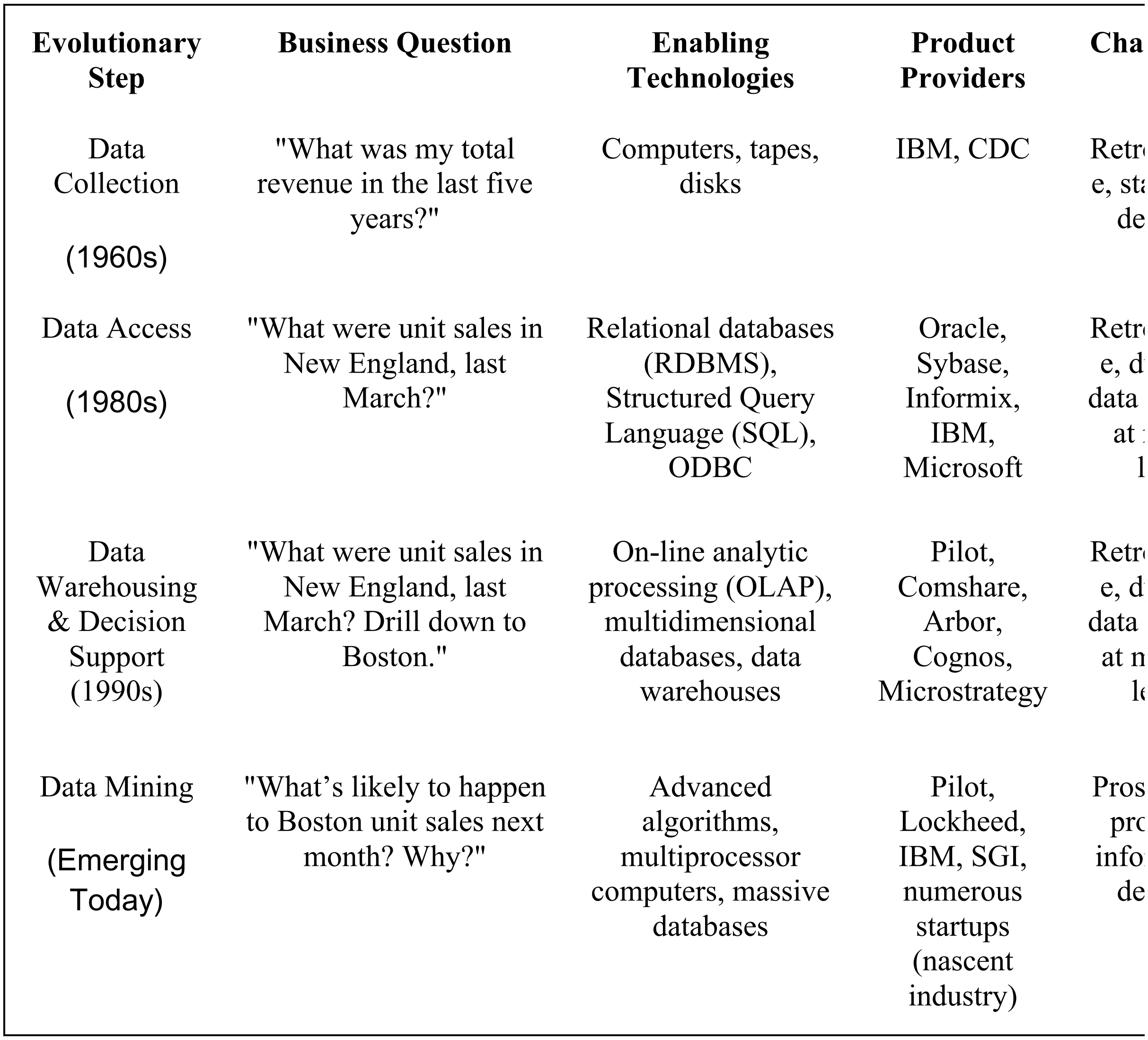
Data mining is the discovery of knowledge of analyzing enormous set of data, by extracting the meaning of the data and then predicting the future trends. Data mining helps us to find out secret information from large databases, and also helps companies to take sound decisions, based on knowledge and information.

If we closely take a look into any data-mining tool, we can see there are some common core logic, which are independent of the data and the applications, but most of existing implementations try to ignore that fact and concentrate on the specific problem, in that way the tool becomes limited to only to a particular set of data for specific application.

Data mining is also finding interesting patterns in data. The main challenge of any data-mining engine is how to apply different algorithms or different techniques, on different set of data, to find interesting pattern, which is very useful to business. It is extremely difficult to come with some standard way of analyzing the data. The enormous volume and the complexity of the data make it impossible to run same algorithms on different dataset. Nowadays, there are different vendors, who are trying to solve this problem, but mostly they support a subset of different algorithms. None of them has come up with any stable engine, which can work in any data set and in any domain.

In the last decade, the improvement in storage and CPU speed has created a huge opportunity for different data mining application, ranging from CRM to medical health care application. The evolution of data mining is shown in table 1.

Now it is very difficult to develop a single application, which can take care all of these problems. It's a dream even to think of an application, which can iterate through any data and will find pattern. Data mining also deals with useful pattern, not just patterns, now whether a pattern is useful or not, depends on the context where it is usually applied. Present day tools depend solely on the expert about what kind of algorithms to apply, and how to analyze the output, because most of them are generic, and there is no context specific logic is attached to the application.



| **Evolutionary Step** | **Business Question** | **Enabling Technologies** | **Product Providers** | **Cha** |
|---|---|---|---|---|
| Data Collection<br>(1960s) | "What was my total revenue in the last five years?" | Computers, tapes, disks | IBM, CDC | Retr<br>e, st<br>de |
| Data Access<br>(1980s) | "What were unit sales in New England, last March?" | Relational databases (RDBMS), Structured Query Language (SQL), ODBC | Oracle, Sybase, Informix, IBM, Microsoft | Retr<br>e, d<br>data<br>at<br>l |
| Data Warehousing & Decision Support (1990s) | "What were unit sales in New England, last March? Drill down to Boston." | On-line analytic processing (OLAP), multidimensional databases, data warehouses | Pilot, Comshare, Arbor, Cognos, Microstrategy | Retr<br>e, d<br>data<br>at n<br>l |
| Data Mining<br>(Emerging Today) | "What's likely to happen to Boston unit sales next month? Why?" | Advanced algorithms, multiprocessor computers, massive databases | Pilot, Lockheed, IBM, SGI, numerous startups (nascent industry) | Pros<br>pr<br>info<br>de |

Table 1. Steps in the Evolution of Data Mining [12].

Here is s summary of the problems that we face today in the existing data mining tools

8. **Difficult to use**– Existing data mining tools try to cover all different data mining applications, thus it becomes very difficult to configure and run.
9. **Needs Expert to run the tool** – No domain or problem specific logic is tied with the tool, therefore needs expert to run the to tool and analyze the result
10. **Difficult to add new functionality -** Because of the size and complexity of each tool, it is very difficult to add any new feature.
11. **Difficult to interface -** There is no way those algorithms developed by some other companies, can be integrated with the tool easily
12. **Short Lifetime -** There is no stable component in the tool and with time the tool become obsolete, as new tools take the market, changing the exiting tool to incorporate new feature is difficult and require lot of changes.
13. **Limited Number of algorithms** – Existing tool only provide limited number of algorithm and sometime use of multiple algorithms is very limited.
14. **Need lot of resources**: Existing tools are not optimized for any specific application, therefore they need lot of resources, such as runtime memory, hard disk etc.

Thus, this focus group topic is driven forward by three main questions. First, “how can we develop a unified data mining engine {UDME)?” Second, “what kind of technologies and tools to build such an Engine?” and third, “how can we overcome the existing problems?”

## OBJECTIVE AND MOTIVATION

Building such an engine is not an easy exercise, specifically, when several factors can undermine their quality success, such as cost, time, and lack of systematic approaches. We would like to architect and develop a Unified Data Mining Engine (UDME), that has the some or all of the following properties:

11. **Ease of use**– Multiple tools can be developed easily by focusing on specific problems, because they all can share the core services, that are provided by the UDME.
12. **No Need of Expert to run the tool** – Domain specific knowledge such as verification, selection of tool etc, can be implemented in the tool itself, while developing the tool.
13. **Easy to add new functionality - –** The application specific logic should be separate from the core logic, therefore new application specific functionality can be added easily, without making any change in the core logic.
14. **Easy to interface -** The design should be based on system of independent patterns, they can be developed by 3rd party vendors.
15. **Long Lifetime -** The engine should be based on stable core logic, which has a long lifetime, the application logic should be loosely connected which can change over time.
16. **Multiple algorithms** – The engine must support any number of algorithms.
17. **Fewer resources**: The proposed engine should be developed by connecting several patterns or components. Depending on the application, a domain the engine can use patterns or components, which are necessary therefore it needs less resources compare to existing tools.
18. **Stable**: The engine should be stable over time, and provide a simple way to apply different data mining and data analysis algorithms on different sets of data in any domain.

19. **Isolation of Application logic**: We must also isolate the stable knowledge from any application specific logic, therefore different applications can use the same core knowledge, which need not to be changed.
20. **Minimum Maintenance Cost** – Maintenance cost of such an engine should be very minimal.

The focus group will address the unified data mining engine challenges and debate several issues that are related to the following questions. We also want researchers, framework developers, and application developers to discuss and debate the following questions related to:

**III. UDME Architecture**

a. What is the best approach for building such an engine?
b. What are the bases of creating the engine architecture?
c. Are there any guidelines, methodologies, and/or processes for an engine architecture creation and development?
d. What are the components of the unified data mining engine architecture?
e. What kind of patterns or components that appear in UDME ?
f. Show how your engine architecture meets the above UDME properties.

**IV. UDME Development**

a. What is the ultimate way to develop such an engine?
b. What are the techniques and tools for developing such an engine?
c. Show how to extend your engine to the new application logics?

## PAPER FORMAT: SUBMISSION & PARTICIPATION

Developers and programmers, who are interested in participating in the workshop, are requested to submit a short position paper (**2-3 pages**), by representing views and experiences that are relevant to the given discussion topic. Please include full mailing address, e-mail address, phone number, fax number, and a designated contact author. Focus group papers will be selected depending on their relevance to the focus group theme. Papers should be submitted electronically by e-mailing it to the lead organizer of the focus group: Shivanshu K. Singh <shivanshukumar@gmail.com>. We also encourage authors to present novel and fresh ideas, critiques of existing work, and practical studies.

Each accepted focus group paper must be presented in the person, either by the author or by one of the co-authors. To foster and promote lively discussions, authors are encouraged to present open-ended questions and one or two main statements for the purpose of discussion at the focus group. Submissions must be made either in MS-Word or RTF formats (please, DO NOT compress files).

People who are interested in participating in the focus group, without making any submissions are requested to fill out the participation form and e-mail to any of the focus group organizers.

---------------------------------------------

PARTICIPATION FORM:
Name and Affiliation:
Position:
Address:
E-mail:
URL:
Areas of interest:
Reasons for Attending?
-----------------------------------------------

For more information please visit any of the following websites:

**http://www.hillside.net/plop/2010/**

You may also contact the organizers, either by e mail or by phone.

## PROPOSED AGENDA

1. Welcome and introduction of participants. The organizers will first provide a short overview of all open issues, and also of the main arguments arising out of the position papers. (Estimated time: 10-15 minutes)

2. Selected authors (who'll be representing the main trends) will be allotted 10-15 minutes, to explain and discuss their position paper with the audience. We are expecting about 5-10 position papers in this session. (Estimated time: 90 minutes)

3. The organizers will also propose an identification process of the major issues, and the participants will then discuss, choose and select what they perceive are the hottest issues to be examined and analyzed. (Estimated time: 10-15 minutes)

(Total estimated time: 120 minutes, i.e. about two hours +/- 5 minutes)

**IMPORTANT DATES** -- Will be decided based on acceptance process.

## ORGANIZERS: Point of Contacts:

**SHIVANSHU K. SINGH (PRIMARY CONTACT)**
Department of Computer Engineering,
Charles W. Davidson College of Engineering,
San Jose State University,
One Washington Square, San José, CA 95192-0180.
Ph: (408) 701-8294
E-mail: shivanshukumar@gmail.com, shivanshu@vrlsoft.com
URL: http://www.shivanshusingh.com

**VIJAY KUMER ERANTI**
Google Inc.,
Mountain View, CA, USA
evijaykumar@yahoo.com

**DR. M.E. FAYAD**
Professor of Computer Engineering
Computer Engineering Dept., College of Engineering
San José State University
One Washington Square, San José, CA 95192-0180
Ph: (408) 924-7364, Fax: (408) 924-4153
E-mail: m.fayad@sjsu.edu, mefayad@gmail.com
http://www.engr.sjsu.edu/fayad
pattern.ijop.org – **Pattern Blog**
fayadsblog.vrlsoft.com – Fayad's Report