HHS Public Access

Author manuscript

Procedia Comput Sci. Author manuscript; available in PMC 2016 July 29.

Published in final edited form as:

Procedia Comput Sci. 2016 June ; 80: 1791–1800. doi:10.1016/j.procs.2016.05.454.

Biomedical Big Data Training Collaborative (BBDTC): An effort to bridge the talent gap in biomedical science and research

Shweta Purawat[1], **Charles Cowart**[1], **Rommie E. Amaro**[2], and **Ilkay Altintas**[1]

Shweta Purawat: shpurawat@sdsc.edu; Charles Cowart: charliec@sdsc.edu; Rommie E. Amaro: ramaro@ucsd.edu; Ilkay Altintas: altintas@sdsc.edu

[1]San Diego Supercomputer Center, University of California, San Diego, USA

[2]Department of Chemistry and Biochemistry, University of California, San Diego, USA

Abstract

The BBDTC (https://biobigdata.ucsd.edu) is a community-oriented platform to encourage high-quality knowledge dissemination with the aim of growing a well-informed biomedical big data community through collaborative efforts on training and education. The BBDTC collaborative is an e-learning platform that supports the biomedical community to access, develop and deploy open training materials. The BBDTC supports Big Data skill training for biomedical scientists at all levels, and from varied backgrounds. The natural hierarchy of courses allows them to be broken into and handled as *modules*. Modules can be reused in the context of multiple courses and reshuffled, producing a new and different, dynamic course called a *playlist*. Users may create playlists to suit their learning requirements and share it with individual users or the wider public. BBDTC leverages the maturity and design of the HUBzero content-management platform for delivering educational content. To facilitate the migration of existing content, the BBDTC supports importing and exporting course material from the edX platform. Migration tools will be extended in the future to support other platforms. Hands-on training software packages, i.e., *toolboxes*, are supported through Amazon EC2 and Virtualbox virtualization technologies, and they are available as: (*i*) downloadable lightweight Virtualbox Images providing a standardized software tool environment with software packages and test data on their personal machines, and (*ii*) remotely accessible Amazon EC2 Virtual Machines for accessing biomedical big data tools and scalable big data experiments. At the moment, the BBDTC site contains three open Biomedical big data training courses with lecture contents, videos and hands-on training utilizing VM toolboxes, covering diverse topics. The courses have enhanced the hands-on learning environment by providing structured content that users can use at their own pace. A four course biomedical big data series is planned for development in 2016.

Keywords

e-learning; biomedical; collaborative; big data; education; toolbox

1 Introduction

Researchers increasingly rely on Big Data, Computational Science, and High Performance Computing (HPC) to solve problems within scientific domains and explore them in new ways (Ryabinin & Chuprina, 2015). However, it is often difficult to effectively apply these approaches in practice. They draw heavily from domains including computer science and applied mathematics, and most researchers are not greatly invested in them (Chapman, et al., 2014). Moreover, petascale and exascale datasets require techniques to process and manage that many practioners have not encountered in their studies or work (Ryabinin & Chuprina, 2015).

University training is largely offered through a multidisciplinary Computational Science discipline, or as specializations in existing domains such as computational chemistry, computational physics, etc. (Chapman, et al., 2014). These offerings cannot keep up with current demand, or the rapid changes in technologies and practices. Students are not entering the field with enough real-world skills, and there aren't enough programs to rapidly affect a change in the talent gap.

To overcome this challenge, our vision is to cultivate an online community focused narrowly on data science and computing in biomedicine, and fostering high quality, well informed, and freely accessible knowledge. Our community, the Biomedical Big Data Training Collaborative (BBDTC), targets the development of technical skills as well as education. Our effort will assist educators by developing and communicating best practices for content development and deployment, along with adaptive learning, assessment metrics, and testing practices.

Our educational material not only includes traditional video and text, but all of the data and tools required to participate in our courses or learn independently. Tools come preinstalled and preconfigured on virtual machines that can be run on a local workstation, or in Amazon EC2 cloud. Users can adapt existing material to their own needs, by assembling parts of multiple courses into their own playlists. Our training material is accessible to a broad community of biomedical researchers and students, including those without access to high performance computing facilities.

By focusing on technical training as well as education, providing real-world datasets, and presenting packaged toolkits that are simple enough to learn yet powerful enough to perform real work, BBDTC differentiates itself from other online learning providers.

Unlike university courses, BBDTC's online presence makes it available to more content providers as well as more students. Being a community-driven resource allows it to evolve quickly and be more responsive to user needs.

The rest of this paper is organized as follows: In section 2, we describe the software architectures of our framework, its user interface, and how it integrates with other important software platforms. In section 3, we cover our current usage statistics. In section 4, we outline related work. In section 5 we conclude, and in section 6 we discuss future work. we describe the results of our efforts. In section 6, we will cover future work.

2 Approach: BBDTC Framework

BBDTC framework utilizes and extends HUBzero (McLennan, 2010), an open-source software platform that facilitates course creation and content management. BBDTC utilizes HUBzero's LAMP stack; a web-portal comprising the Joomla content management system, Apache web server, and MySQL database. BBDTC extends HUBzero modules whenever required to meet demands unique to the Biomedical Big Data research community. The platform is extended to Amazon EC2 Cloud infrastructure in order to enable scalable Big Data experiments. To facilitate offline learning, users can download complete packages as Virtualbox images from the BBDTC platform. The HUBzero platform is further enhanced through the implementation of cross platform integration module. Cross Platform Integration enables instructors to rapidly build content by leveraging their already existing courses on other platforms such as edX in a seamless manner.

The BBDTC framework comprises three main components: (*i*) BBDTC User Interface; (*ii*) BBDTC VM Toolbox; and (*iii*) Cross platform integration module. The components are explained in further detail in successive sub-sections.

2.1 BBDTC User Interface

BBDTC User Interface is designed with simplicity and usability at the center. It consists of four major sub-components: (*i*) course building interface for instructors and course managers; (*ii*) course viewing interface for learners; (*iii*) a repository with playlist functionality for every BBDTC contributor; and (*iv*) social media interface for creating an involved learning atmosphere.

Course Building Interface—The platform offers several features to create an online interactive class through a simple and intuitive interface. A course can be tailored to create multiple '*offerings*' with a subset of the content changed or updated. An offering is a collection of '*units*'. A unit can be further broken down into one or more '*lessons*'. Instructors can upload materials, assignments, assessments, tools using all or any existing options for embedding a video, uploading files, including a wiki page and attaching web-links. Videos are hosted on BBDTC's YouTube channel[1] and streamed on a BBDTC course page on user's computer. Instructors can track the progress of each learner through course progress bubbles, the real-time updates on quizzes taken, and the breakdown of scores and grades. Figure 2: BBDTC User Interface Screenshots demonstrations a course outline page, an embedded video lecture, an online quiz, learner's grade reporting and progress tracking dashboard.

Course Viewer Interface—Course viewer interface is for learners. Learners can only view the course contents but cannot upload or edit materials. They can track their personal progress and grades.

BBDTC Repository and Playlist Builder—Each learner is unique and needs a course structure that matches her personal learning requirements. The requirements include

[1]The BBDTC YouTube Channel: https://www.youtube.com/channel/UCj9gpCafVV23WcP_b66Wkdw

attributes such as different learning pace, prior knowledge of a subject etc. The playlist feature promotes personalization of content by enabling customization to meet an individual's requirements. BBDTC offers a central repository where users contribute by uploading course modules, tools, and datasets. BBDTC allows researchers to create their own customized catalogue for personalized education by selectively adding content from the repository. The playlist feature enables users to add materials from the repository and/or upload their own content, and arrange them in the order that aids their personal learning requirements. The user interface provides drag and drop function to add contents to the playlist.

Social Media Interface—Learners can utilize functionalities such as forums, discussion boards, blogs and groups to enhance their knowledge through community involvement. A forum provides an online discussion area where people can exchange queries, difficulties, concepts and solutions in form of messages. People with common interest can form peer groups. The group can share contents and exchange conversations privately or publicly. People in a group can form their own forums for discussions. Research has shown that collaborative techniques such as Pair Programming (Williams, 2002) results in improved learning outcomes. BBDTC promotes a connected learning atmosphere by providing multiple channels to collaborate and enhance the learning experience through peer engagement.

2.2 Toolbox Integration: Cloud and Local

In line with our goal of maximizing users' time for learning rather than installing tools, we introduce tailored pre-installed software environments that enhance the learning experience. Users can access these toolboxes over the cloud through Amazon EC2 and locally through Oracle Virtualbox virtualization technologies.

Users with access to the internet can remotely connect to pre-built Amazon EC2 Virtual Machines for accessing biomedical big data tools and perform scalable big data experiments. We also support local execution in equally simple style via downloadable Virtualbox Images. Every downloadable VM contains a standardized software tool environment with all required software packages and test data. The BBDTC provides an efficient software distribution mechanism using ready-to-go Virtual Machine Toolboxes. These toolboxes free the user of installation and configuration hassles. Users can directly get to the task of problem solving using specially customized virtual machines. The notable features of the two categories is as follows:

- **Downloadable VM toolbox on Local Machine:** In this category, the software tools are shared as a downloadable Virtualbox Image. The VM toolbox allows users to get a standardized software tool environment with their required software dependencies and the necessary test data on their personal machines.
- **Amazon EC2 VM Toolbox on the Cloud:** This category provides a ready-to-go solution for scalable big data experiments, saving users from setting up a local machine with necessary hardware specifications. Instead,

the users can directly access an Amazon EC2 Virtual Machine through the BBDTC website using a light-client machine. The system utilizes noVNC (Martin, 2011), an HTML5 based VNC client. The noVNC web-client communicates with a remote VNC server on Amazon EC2 Cloud infrastructure through websockets. Figure 3: Amazon EC2 VM Toolbox Integration Architecture demonstrates Amazon EC2 toolbox integration with BBDTC platform.

We have verified and tested the features of the BBDTC course-building interface. The toolbox framework is live and linked to an existing online-course on "*Scalable Bioinformatics Bootcamp*" that is available to learners through the BBDTC portal.

2.3 Cross Platform Integration

Organizing text, video, and other prepared materials into a new online course can be a significant effort. Once the new course has been designed, the metadata capturing this organization cannot always be migrated outside of the online application. The time taken to develop a course online discourages researchers and educators from redeveloping the same course across several MOOCs, encouraging 'vendor lock-in' to a particular provider.

The MOOC provider edX has taken steps to remedy this by defining a specification for *XBlocks*, hierarchical components of text, video, assessments, etc., that can be used like building blocks to construct a new course. Courses constructed using edX's *edX Studio* can be exported as an XBlocks blueprint to an author's local computer, merged and modified by third-party tools, and imported back into edX.

BBDTC has developed a 'back-end' tool to process this exported data from edX Studio, and import it into BBDTC as new courses and resources. This has enabled biomedical big data educators and researchers, who made early investments into edX, the ability to also provide that content in the centralized resource that is BBDTC. Similarly, BBDTC is currently working on a tool to perform the reverse operation, and export courses and resources from BBDTC as XBlocks blueprint; this will enable courses created using BBDTC's course building interface to imported into edX as new courses, or as a library of content to make new courses from.

3 Usage Statistics

In the first year of the development for BBDTC, we offered three complete and open biomedical training courses during the alpha release of the framework. We received an excellent response with a total of 162 users registered on the BBDTC website in the year 2015.

The National Biomedical Computation Resource (NBCR)[2] launched two biomedical training courses on the BBDTC platform in 2015: NBCR Summer Training Program - "Data to Structural Models" and NBCR & TCBG Joint Training Program - "Simulation-Based Drug Discovery". These week-long intensive training programs introduced new principles,

[2]National Biomedical Computation Resource (NBCR) website: http://nbcr.ucsd.edu

methods and NBCR tools to the biomedical community. The major goals of launching the courses online via BBDTC were to outreach a broader user base of biomedical research community. It also enabled us to test the platform capabilities and get feedback from the active users for further improvements. Following are the data analysis reports from the two courses:

Course 1, NBCR Summer Training Program - “Data to Structural Models”

We received a total of 31 enrollments for this course. BBDTC provided interface to track learners course progress and grades. We observed 6 students accessed full course material and 9 completed the quiz. Figure 5: NBCR Summer Training Program - “Data to Structural Models” represents user profile distribution according to Career level, Location and Continent.

Course 2, NBCR & TCBG Joint Training Program - “Simulation-Based Drug Discovery”

We received a total of 42 enrollments for this course. Figure 6: NBCR & TCBG Joint Training Program - “Simulation Based Drug Discovery” represents user profile distribution according to Career level, Location and Continent.

Course statistics information delivered by the BBDTC enables content creators to understand our learners and to provide a learning experience that goes beyond their expectations. It promotes course providers to experiment with new learning techniques and recognize how the learners respond. This structured feedback mechanism of BBDTC accelerates high quality content creation by iterative upgradation of techniques.

4 Related Work

A series of short tutorial videos, of the kind commonly used in Coursera and other MOOCs, was recently used to teach two graduate-level biostatistics courses at the Public Health Program at the University of New Mexico (Hund & Getrich, 2015). Students met twice a week in a computer lab, with access to the instructor as well as required software. The instructor also held office hours twice a week. In this setting, students’ overall perception of videos was positive, with 87.5% of the students stating that the videos were advantageous to their studying. Specifically, students found that shorter videos were easier to ‘re-watch’ and ‘skim’. “The ability to use the videos in a self-paced and repetitive manner was consistently cited as a major advantage by the students (Hund & Getrich, 2015).”

Students didn’t view the videos as a replacement for in-class instruction, but as a supplement; instructors found that the availability of the videos allowed for more time in class to be spent on discussion. Most importantly, it was found that using videos promoted social support between classmates (Hund & Getrich, 2015). Although students would watch videos individually, they would reference and recommend specific videos when communicating with other classmates.

In another study, conducted across six different public institutions, students took the same introductory course on statistics (Bowen, Chingos, Lack, & Nygren, 2012). At each institution, a control group took the course in a classroom-based setting, while a second

group took a *hybrid course*. This hybrid course used software developed at Carnegie Mellon University to guide students' mode of instruction, with one face-to-face meeting with an instructor, per week. Results from this study showed that students in the hybrid course performed as well as students in the classroom-based setting, in terms of course completion, course grades, and performance on standardized tests.

Recently, a group of researchers conducted a series of surveys with over five hundred students taking a select group of online courses, over a period of three years (Cole, Shelleu, & Swartz, 2014). Students found their experience with online instruction only moderately satisfactory, and hybrid courses were rated with higher satisfaction over purely online courses; lack of interaction with others was the most often reason cited for this dissatisfaction. The above research suggests that engagement with an instructor and/or a community is key to a student's overall satisfaction and performance when taking a course. Compared with older online learning paradigms, modern MOOCs are augmented with social networks that provide community and support through message boards and discussion groups (Duderstadt & Colloquium, 2013).

Today's MOOCs are primarily targeted at serving individuals, rather than communities (Duderstadt & Colloquium, 2013). BBDTC instead focuses on cultivating a community, by offering multiple social networking tools, and focusing on the biomedical big data community's particular needs. Rather than supplanting traditional education, the BBDTC expects to be successful by supplementing it.

5 Conclusion

Biomedical Big Data Science is a thriving and an exponentially growing community. Scientists need a quick-access gateway to high-quality knowledge content on topics that interest them. BBDTC is a community-oriented platform with an aim to promote knowledge sharing and accelerate information transfer rate from the field experts to an inquisitive learner. Smooth information dissemination will encourage new approaches and enable community members to combine tools from diverse domains to solve critical problems.

The collaborative provides essential tools to facilitate content development and deployment of Biomedical Big Data training and education. The simple and intuitive platform enables the community to contribute at many different levels such as by adding course modules and playlists, adding new tools or working with the development team to introduce customized VM packages. BBDTC's fact based progress tracking dashboard and statistical reports enable content creators to understand learners and to provide a learning experience that goes beyond their expectations.

6 Future Work

There are several areas in which we plan to expand our current work. To give an experiential view of a scientific facility, we plan to add Virtual Field Trips to BBDTC. Using this feature, community members can gain fundamental understanding of each other's strengths to better collaborate. This feature will enable labs to achieve higher visibility among their target community members.

To further facilitate easy dissemination of read-to-use tools, we plan to automate the VM toolbox generation process. In addition, we will enhance BBDTC's integration with Amazon AWS, by launching virtual machines using AMI id information from the BBDTC website.

For fast exchange of MOOC course structure, we plan to extend the ability to import and export feature to additional platforms that are gaining traction in the community. A user-interface to allow educators to individually import and export data will also be explored in the future. In the playlist interface, we plan to include access to course repository and resources from the same interface.

Finally, a tool is only as good as its utility to the stakeholders. We will build incentive for biomedical researchers to share materials. We will aggressively increase community engagement with the key stockholders, such as, R25 awardees, BD2K Centers of Excellence, NIH P41 BTRRs by serving their training and educational material through BBDTC, and acting on feedback at every step.

Acknowledgments

This work is supported by NIH R25 GM114821-01 for BBDTC and P41 GM103426 for NBCR.

References

Bowen, WG.; Chingos, MM.; Lack, KA.; Nygren, TI. Interactive Learning Online at Public Universities: Evidence from Randomized Trials. Ithaka S+R; 2012.

Chapman, B.; Calandra, H.; Crivelli, S.; Dongarra, J.; Hittenger, J.; Lathrop, S., et al. DOE Advanced Scientific Advisory Committee (ASCAC): Workforce Subcommittee Letter. Washington, DC: USDOE Office of Science; 2014.

Cole MT, Shelleu DJ, Swartz LB. Online Instruction, E-Learning, and Student Satisfaction: A Three Year Study. The International Review of Research in Open and Distance Learning. 2014; 15(6): 111–131.

Duderstadt JJ, Colloquium IG. The Impact of Technology on Discovery and Learning in Research Universities. Ninth Glion Colloquium. 2013 Jun.:5–9.

Hund L, Getrich C. A Pilot Study of Short Computing Video Tutorials in a Graduate Public Health Biostatistics Course. Journal of Statistics Education. 2015; 23(2)

Martin, J. noVNC project website. 2011. Retrieved from https://kanaka.github.io/noVNC/

McLennan, Ma. HUBzero: a platform for dissemination and collaboration in computational science and engineering. Computing in Science & Engineering. 2010:48–53.

Ryabinin, K.; Chuprina, S. ICCS 2015 International Conference On Computational Science. Vol. 51. Amsterdam: Elsevier; 2015. Using Scientific Visualization Tools to Bridge the Talent Gap; p. 1734-1741.

Williams, La. In support of pair programming in the introductory computer science course. Computer Science Education. 2002:197–212.

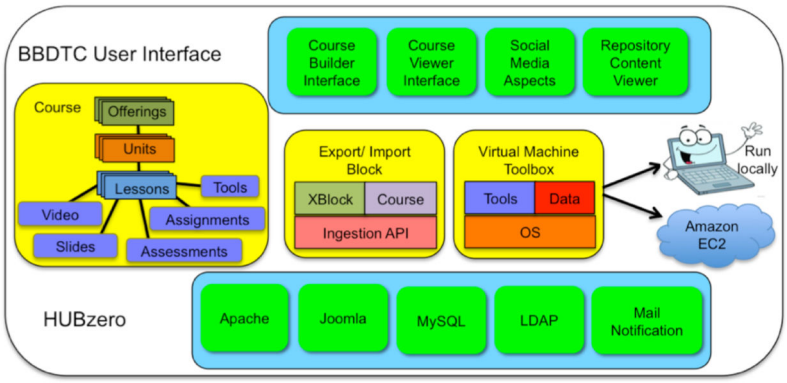

Figure 1.
BBDTC Framework

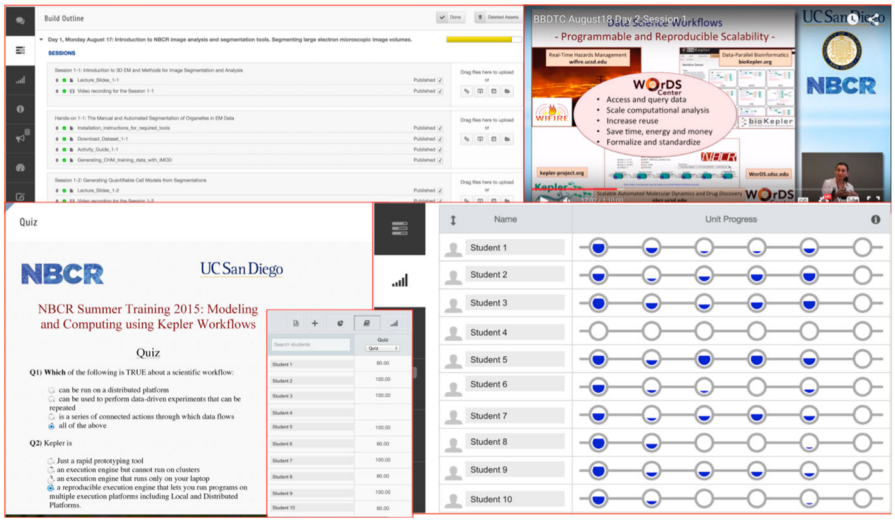

Figure 2.
BBDTC User Interface Screenshots

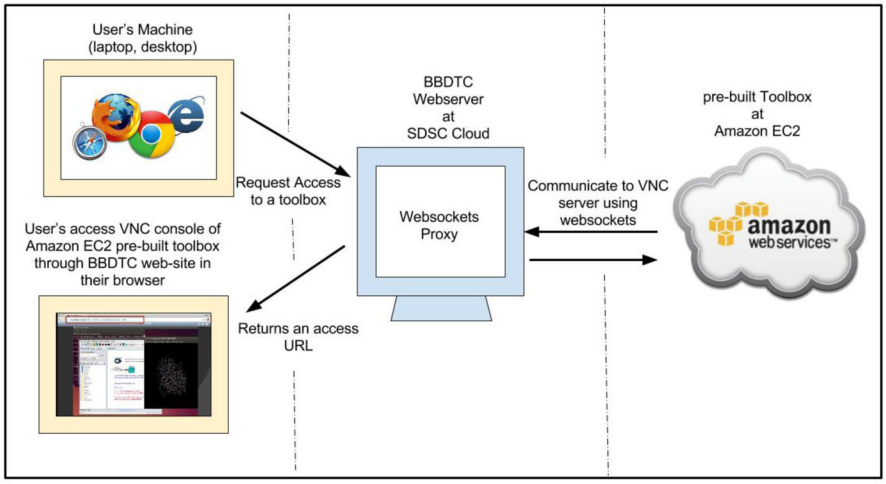

Figure 3.
Amazon EC2 VM Toolbox Integration Architecture

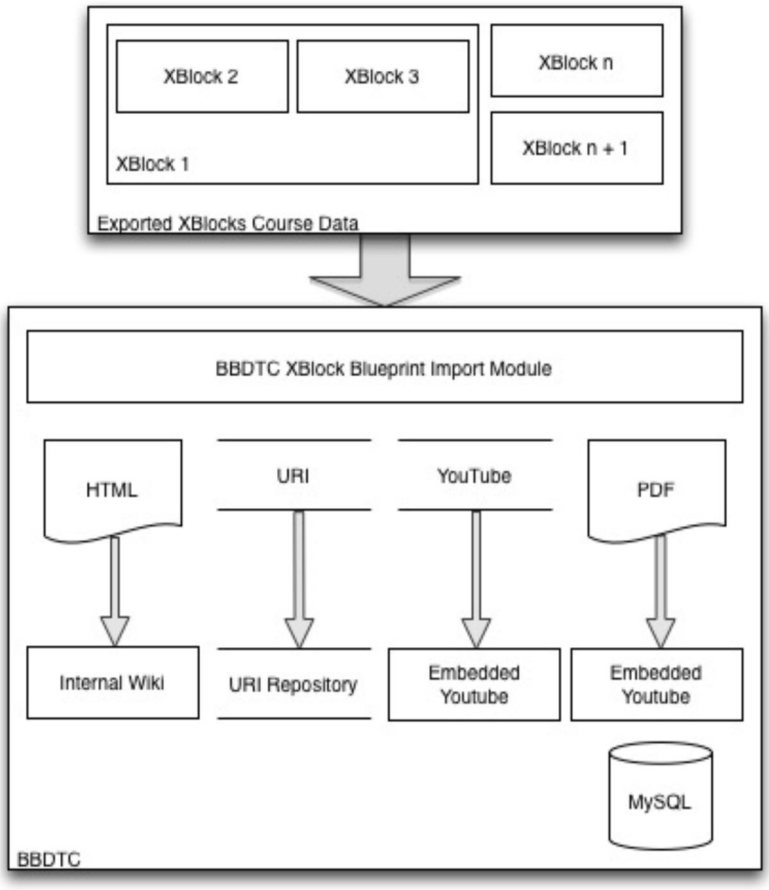

Figure 4.
Importing exported edX Studio data

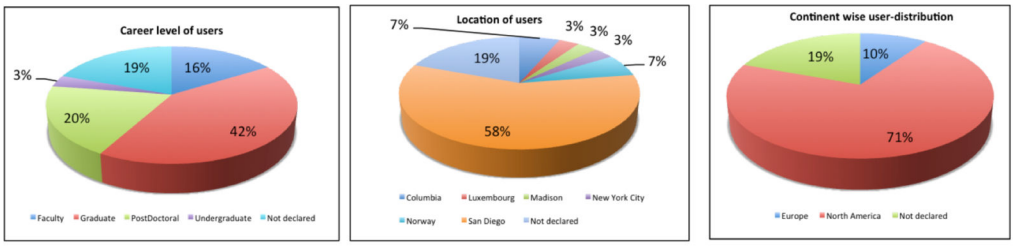

Figure 5.
NBCR Summer Training Program - “Data to Structural Models”

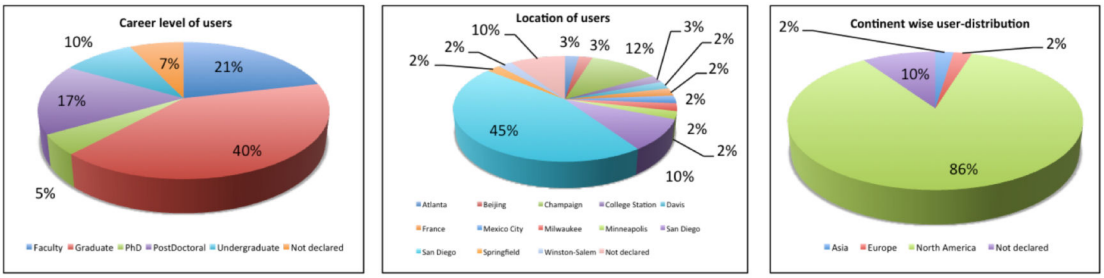

Figure 6.
NBCR & TCBG Joint Training Program - “Simulation Based Drug Discovery”