Synopsis

Avery developed science is the result of the scrupulous analysis of a vast amount of data.

To advance science of education, researchers need (a) to establish reliable sources of data, which could later be (b) analyzed.

This project is to establish a reliable procedure for developing a large amount of reliable data using techniques developed in other field (physics) for high frequency "datamining".

The proposal has been developed by Pr. Plamen Ivanov and Dr. Valentin Voroshilov (valbu@bu.edu)

"Developing Strategies and Technology for Generating and Analysis of Longitudinal High Frequency Data Streams from Faculty and Students".

The structure of the document:

2-Page Summary of The Proposal (with *brief* part V: Further Development)

"Developing Strategies and Technology for Generating and Analysis of Longitudinal High Frequency Data Streams from Faculty and Students".

I. Introduction:

This project, when realized, has a potential to be awarded The $ 320 million Yidan Prize in Education (http://www.yidanprize.org/en/), because the ultimate result of the project is to transform science of education and, hence, education. The essence of the project is developing a revolutionary and science-based innovative approach to describing, structuring, analyzing, and assessing the teaching and learning process.

This 2-page presentation is to present the shortest version of the proposal, but the innovative nature of the proposed project demands more detailed representation, which is offered later.

The *big data analysis* has entered many important human practices. For example, one can point at such fields like: Human Genome Project (DNA v. health), healthcare and epidemiology (spread of diseases), particle physics, social and business networking (Facebook, Twitter, Snapchat, Instagram, cellphone communication, telemedicine, remote business communication), national security (trends in various networks), business network analysis (AirB&B, Uber, Lift, Netflix), trading stocks, currency exchange (live records of massive volume of transactions). Within all those fields, data scientists were able to: (1) establish protocols and procedures for quantifying data, for collecting, structuring, comparing and sharing vast amounts of data; and (2) for mining the large data bases for extracting valuable and reliable information on the correlations between multiple parameters based on various types and levels of data coming from multiple sources.

However, despite the fact that education represents one of the most vastly spread and one of the most important human practices, the methods developed in other fields for (1) collecting, and (2) mining BIG data have not found applications in the field of education. Current approaches do not provide understanding of the deep structure of teaching and learning processes, do not lead to development of quantitative measures of the quality of teaching, and development of quantitative measures of the trends in teaching (e.g. the measure of the improvement in teaching), and development of quantitative measures of the student progress correlated with student learning outcomes.

II. Description of the current state of the Educational Data Mining (a.k.a. EDM):

1. EDM is in the stage of an early development and rather represents Advanced Educational statistics (e.g. Educational Data Mining Society has been formed only five years ago: educationaldatamining.org).
2. Currently the following approaches are used to obtain various educational data:

- Observing school teachers or college faculty while teaching and assessing teacher's actions using various observation protocols (e.g. BOPR, COPUS, MarzanoOP, RTOP, GORP).
- Observing school and college students while being taught using various observation protocols (e.g. a "STEM class observation protocol").
- Collecting responses to various surveys (e.g. "National Survey of Student Engagement", "National Survey of College Faculty").
- Collecting data during various student-computer interactions when using various computer-based media (MOOCs, computer games, intelligent tutoring systems, online content delivery systems, online homework delivery systems).

It is important to stress that:

(A) When data collection methods are based on the use of surveys or observation protocols, they are typically used only ones or twice during a teaching period (a semester, or a year); these methods are typically used to observe of a small percentage of teachers and students.

(B) Data collected using computer-based media does not access the everyday reflection of students on the learning process (actions taken for absorbing information and developing skills, and following results and satisfaction); does not access the everyday reflection of teaching faculty on the teaching process and on the student progress; this data typically presents the aggregated student response on the course as a whole (ranking the difficulty of a course, ranking homework assignments, indicating relevance of a textbook and other resources, overall satisfaction); mostly present two-parametric correlations like "time used for homework" – "final grade".

Currently, educational data: is collected during isolated educational projects; does not represent longitudinal streams of high frequency data collected during the full term of learning; does not satisfy criteria for being "big data" (except few collected via student-computer interactions); does not involve data streams with a large number of parameters; does not allow cross analysis for searching stable correlations between multiple parameters. In its current state, EDM is rather Advanced Educational Statistics.

Currently, there is NO research which:
(1) regularly and *frequently* (e.g. several times a week) collects data *simultaneously* from teaching faculty *and* from students during the *whole* period of teaching a course (not just via observing one lecture);
(2) uses media technologies, including phone apps, to collect the desired sets of educational data incoming from *multiple* sources (faculty, disciplines, departments, institutions);
(3) uses technologies to mining data in searching for stable correlations between different factors affecting teaching-learning practices and student's performance using *multivariable* (multi-parametric) space.
Currently, there is no "brick-and-mortal" educational institution which collects from faculty and from students high frequency responses about multiple features of a teaching and learning processes. There is no institution which collects and cross-correlates multiple responses across various disciplines over a long period of time.

III. The scope and immediate goals of the proposed project:

The project will pioneer (A) the development of a new type of a big data base via collecting longitudinal streams of high frequency data in the field of education; (B) the development of the new methodology for mining new type of educational data and extracting valuable and reliable information on the correlations between various parameters of multiple data sources of different types and levels (faculty, departments, institutions).
Every day zillions of apps are being used by millions of people. People already have habits of tracking information every day (calories intake, calories burned, steps made, miles traveled, etc.). Why not harness the new technologies and the new habit to generate a stream of high frequency educational data?
The goals:

1. Establishing a set of measurable and universal (but modifiable) parameters which will be used for describing the state and structure of any teaching and learning processes (i.e. for any course).
2. Developing one questionnaire for teaching faculty and one questionnaire for students, which they will use during a course regularly and frequently for self-observation, for assessing students' actions and progress, for assessing faculty teaching actions and traits.
3. Developing an app for collecting the data provided by students and faculty.
4. Developing the strategy for analyzing the data coming from faculty and students in search for correlations.
5. Developing a web-site for collecting the data coming from faculty and students.
6. Piloting the program

We are proposing collecting *high frequency longitudinal responses* (from faculty and students: before the beginning of the course, then after each lecture, after each exam, summative responses after two weeks of a course about lectures, labs and all other features of the course, generalized responses after each third of a semester, and the accumulative responses just before and after the final examination). The goal is to develop procedures which will allow to visualize the structure of the responses, changes in the structure, trends in changes in the structure. This should allow to access regularly student reflection on the course and on his or her performance during the course (how do students assess the difficulty of various assignments, the clarity or helpfulness of lectures, workbooks, textbook, office hours, etc., helpful traits of a lecturer). This also should allow to access regularly the structured reflection of a faculty on teaching approach selected for the course, on students' readiness, behavior, performance, success. The next goal is to demonstrate the existence of stable trends in correlations between various parameters affecting learning process of students.

IV. Resources.

The project will leverage the existence of the expertise and resources allocated at the Boston University: including scientists who have deep expertise in developing and application of methods for collecting and organizing big data coming from multiple sources, for quantifying data, extracting information from big data on important correlations between multiple parameters describing functioning of various systems or subsystems, finding cross relations, describing information transfer between multiple sources. Using noise reduction methods, finding critical points and visualizing state transitions (PI Prof. Plamen Ivanov), and experienced teaching faculty (co-PI Dr. Valentin Voroshilov), and high computational facility (GHPCC).

V. Future development.

The proposed approach to educational data mining is pioneering the development of the new type of educational data, and the new methodology for collecting and mining that new type of educational data.
It has a potential to follow the history of the Human Genome Project (started at Boston University).

Extended 3-Page Summary of The Proposal (with *extended* part V: Further Development)

"Developing Strategies and Technology for Generating and Analysis of Longitudinal High Frequency Data Streams from Faculty and Students".

I. Introduction:

This project, when realized, has a potential to be awarded The $ 320 million Yidan Prize in Education (http://www.yidanprize.org/en/), because the ultimate result of the project is to transform science of education and, hence, education. The essence of the project is developing a revolutionary and science-based innovative approach to describing, structuring, analyzing, and assessing the teaching and learning process.

The *big data analysis* has entered many important human practices. For example, one can point at such fields like: Human Genome Project (DNA v. health), health care and epidemiology (spread of diseases), particle physics, social and business networking (Facebook, Twitter, Snapchat, Instagram, cellphone communication, telemedicine, remote business communication), national security (trends in various networks), business network analysis (AirB&B, Uber, Lift, Netflix), trading stocks, currency exchange (live records of massive volume of transactions). Within all those fields, data scientists were able to: (1) establish protocols and procedures for quantifying data, for collecting, structuring, comparing and sharing vast amounts of data; and (2) for mining the large data bases for extracting valuable and reliable information on the correlations between multiple parameters based on various types and levels of data coming from multiple sources.

However, despite the fact that education represents one of the most vastly spread and one of the most important human practices, the methods developed in other fields for (1) collecting, and (2) mining BIG data have not found applications in the field of education. Current approaches do not provide understanding of the deep structure of teaching and learning processes, do not lead to development of quantitative measures of the quality of teaching, and development of quantitative measures of the trends in teaching (e.g. the measure of the improvement in teaching), and development of quantitative measures of the student progress correlated with student learning outcomes.

II. Description of the current state of the Educational Data Mining (a.k.a. EDM):

1. EDM is in the stage of an early development and rather represents Advanced Educational statistics (e.g. Educational Data Mining Society has been formed only five years ago: educationaldatamining.org).

2. Currently the following approaches are used to obtain various educational data:

- Observing school teachers or college faculty while teaching and assessing teacher's actions using various observation protocols (e.g. BOPR, COPUS, MarzanoOP, RTOP, GORP).
- Observing school and college students while being taught using various observation protocols (e.g. a "STEM class observation protocol").
- Collecting responses to various surveys (e.g. "National Survey of Student Engagement", "National Survey of College Faculty").
- Collecting data during various student-computer interactions when using various computer-based media (MOOCs, computer games, intelligent tutoring systems, online content delivery systems, online homework delivery systems).

It is important to stress that:

(A) When data collection methods are based on the use of surveys or observation protocols, they are typically used only ones or twice during a teaching period (a semester, or a year); these methods are typically used to observe of a small percentage of teachers and students.

(B) Data collected using computer-based media does not access the everyday reflection of students on the learning process (actions taken for absorbing information and developing skills, and following results and satisfaction); does not access the everyday reflection of teaching faculty on the teaching process and on the student progress; this data typically presents the aggregated student response on the course as a whole (ranking the difficulty of a course, ranking homework assignments, indicating relevance of a textbook and other resources, overall satisfaction); mostly present two-parametric correlations like "time used for homework" – "final grade".

Currently, educational data: is collected during isolated educational projects; does not represent longitudinal streams of high frequency data collected during the full term of learning; does not satisfy criteria for being "big data" (except few collected via student-computer interactions); does not involve data streams with a large number of parameters; does not allow cross analysis for searching stable correlations between multiple parameters. In its current state, EDM is rather Advanced Educational Statistics.

Currently, there is NO research which:

(1) regularly and *frequently* (e.g. several times a week) collects data *simultaneously* from teaching faculty *and* from students during the *whole* period of teaching a course (not just via observing one lecture);

(2) uses media technologies, including phone apps, to collect the desired sets of educational data incoming from *multiple* sources (faculty, disciplines, departments, institutions);
(3) uses technologies to mining data in searching for stable correlations between different factors affecting teaching-learning practices and student's performance using *multivariable* (multi-parametric) space.
Currently, there is no "brick-and-mortal" educational institution which collects from faculty and from students high frequency responses about multiple features of a teaching and learning processes. There is no institution which collects and cross-correlates multiple responses across various disciplines over a long period of time.

III. The scope and immediate goals of the proposed project:

The project will pioneer (A) the development of a new type of a big data base via collecting longitudinal streams of high frequency data in the field of education; (B) the development of the new methodology for mining new type of educational data and extracting valuable and reliable information on the correlations between various parameters of multiple data sources of different types and levels (faculty, departments, institutions).
Every day zillions of apps are being used by millions of people. People already have habits of tracking information every day (calories intake, calories burned, steps made, miles traveled, etc.). Why not harness the new technologies and the new habit to generate a stream of high frequency educational data?
The goals:
1. Establishing a set of measurable and universal (but modifiable) parameters which will be used for describing the state and structure of any teaching and learning processes (i.e. for any course).
2. Developing one questionnaire for teaching faculty and one questionnaire for students, which they will use during a course regularly and frequently for self-observation, for assessing students' actions and progress, for assessing faculty teaching actions and traits.
3. Developing an app for collecting the data provided by students and faculty.
4. Developing the strategy for analyzing the data coming from faculty and students in search for correlations.
5. Developing a web-site for collecting the data coming from faculty and students.
6. Piloting the program
We are proposing collecting *high frequency longitudinal responses* (from faculty and students: before the beginning of the course, then after each lecture, after each exam, summative responses after two weeks of a course about lectures, labs and all other features of the course, generalized responses after each third of a semester, and the accumulative responses just before and after the final examination). The goal is to develop procedures which will allow to visualize the structure of the responses, changes in the structure, trends in changes in the structure. This should allow to access regularly student reflection on the course and on his or her performance during the course (how do students assess the difficulty of various assignments, the clarity or helpfulness of lectures, workbooks, textbook, office hours, etc., helpful traits of a lecturer). This also should allow to access regularly the structured reflection of a faculty on teaching approach selected for the course, on students' readiness, behavior, performance, success. The next goal is to demonstrate the existence of stable trends in correlations between various parameters affecting learning process of students.

IV. Resources.

The project will leverage the existence of the expertise and resources allocated at the Boston University: including scientists who have deep expertise in developing and application of methods for collecting and organizing big data coming from multiple sources, for quantifying data, extracting information from big data on important correlations between multiple parameters describing functioning of various systems or subsystems, finding cross relations, describing information transfer between multiple sources. Using noise reduction methods, finding critical points and visualizing state transitions (PI Prof. Plamen Ivanov), and experienced teaching faculty (co-PI Dr. Valentin Voroshilov), and high computational facility (GHPCC).

V. Future development.

The proposed approach to educational data mining is pioneering the development of the new type of educational data, and the new methodology for collecting and mining that new type of educational data (including high frequency prime series and network analysis, and the used of the modern methods of dynamics of cross interactions, establishing the information concert between different sources of data, and other).
The most important feature of the proposed project is its *scalability*.
Starting from one faculty teaching one course during one semester, it has a potential to evolve and involve multiple faculty teaching various disciplines at different departments, and even at different institutions. The existence of high frequency responses coming from different parts and different levels of the educational system will allow to cross analyze the data and to establish stable correlations, which do not depend – or on the contrary, heavily depend – on the type of a teaching faculty (e.g. a newly hired professor v. a Nobel Laureate), or a teaching discipline (physics or

chemistry), or a type of a course (elementary v. advanced), or a type of an institutional approach (traditional v. modified, blended, studio).

The first stage (one faculty, one course) will allow to establish basic principles for structuring, quantifying, collecting data, developing a methodology for building a data base, building a data base, developing methodology for mining the data, approbation of methods for collecting and mining data. The data will be collected using high frequency responses from the professor and from the students in a course. The goal is to see how the professor assess his performance and performance of the students, as well as to see how students assess the quality of the teaching and their own performance, and to correlate those data with the grades students receive via various grading procedures (homework, laboratory exercises, quizzes, midterms, final examination). The result of this stage will be demonstrating the viability of the approach via establishing various correlations between multiple features of a teaching process, of a learning process, and the learning outcomes of students.

Next stages would require scaling up the funding to broad the scale of the project including: prolonging teaching the original course, involving additional faculty from the same department, involving faculty from other departments, involving other institutions (this stage will leverage the existing collaborations in high computing and internet teaching between BU, MIT, and Harvard). At the later stages of the project BU may be able to establish a new department employing specialists from the physics department, department of engendering, BU IT personal, and designated for maintaining all technological processes related to functioning of the new data base (*when the seed grant will be supported by additional funds*, including NSF). Boston University also can become a hosting facility for hosting the centralized data base (at the Holyoke Green High Performance Computing Center).

This project has a potential to transform the science of education, and – hence – the education. Using quantification of teaching strategies based on high frequency responses gathering from multiple sources researches will be able to compare different teaching strategies, to find what different teaching strategies have in common and what makes them different in terms of affecting student learning. Having cross institutional, cross departmental, and cross disciplinary data analysis researches will be able to establish quantitative measures for making transparent the structure of teaching practices, for assessing the quality of teaching practices and their results; and to have a quantitative comparison of teaching strategies of different type, to see what type of strategies is more effective in what kind of teaching environment. This also will help to structure the learning process (absorbing information and skill development) by different types of students; to provide quantitative measures for the correlations between various factors influencing student learning and the everyday and the overall student performance assessed via various grading procedures.

We are first who are proposing this pioneering approach, moving the research in the new direction (collecting truly big data; building a shared and growing data base), which will begin from establishing principles of how to collect big educational data. At the same time, we will be developing and testing methods for dealing with large amounts of educational data coming in the form of various data streams from different sources and related to various aspect of teaching and learning processes. We will be developing methods for synchronizing and cross correlating the vast amount of educational data within and across departments and institutions, which is something that has not yet been done before.

This proposal has a potential to follow the history of *the Human Genome Project* (started at BU by Prof. Charles DeLisi). By merging contemporary methods developed for collecting and mining big data when streams of high frequency data coming from many different sources with educational practices happening at different time scales and locations, we will have an opportunity to initiate the process of reforming education.

We will be the first to start the development of the new type of quantitative measures for education.

This proposal is to lay the foundation, the base stone in the process of building the new system for collecting and mining big educational data.

Full Proposal: Textual Format (9 pages)

"Developing Strategies and Technology for Generating and Analysis of Longitudinal High Frequency Data Streams from Faculty and Students".

Key words:

Big data, big data analysis, high frequency data, instantaneous response data, cross correlation analysis, quantitative assessment of student performance, quantitative assessment of teacher performance, quantitative assessment of a teaching process, quantitative assessment of a teaching practice, quantitative assessment of the structure of a teaching process, quantitative assessment of the structure of a learning process (knowledge absorption, skill development), quantitative assessment of correlations between the structure of a teaching process and the structure of a learning process, quantitative assessment of the quality of teaching, quantitative assessment of the learning outcomes of students, quantitative assessment of the factors influencing the learning outcomes of students, longitudinal collection of high frequency responses from various sources.

I. Introduction:

This project, when realized, has a potential to be awarded The $ 320 million Yidan Prize in Education (http://www.yidanprize.org/en/), because the ultimate result of the project is to transform science of education and, hence, education. The essence of the project is developing a revolutionary and science-based innovative approach to describing, structuring, analyzing, and assessing the teaching and learning process.

The *big data analysis* has entered many important human practices. In every field the big data analysis stands on two pillars: (a) collecting vast amount of data (gigabytes or terabytes of data), and creating a large and shared database; and (b) mining the collected data and extracting information about important correlations.

For example, one can point at such fields like: Human Genome Project (the search for correlations between the structure of a DNA and health conditions), data mining in health care and epidemiology (analyzing spread of various diseases), particle physics (collecting data from particle accelerators and astronomical observations), social and business networking (Facebook, Twitter, Snapchat, Instagram, cellphone communication, telemedicine, remote business communication, etc.), national security (trends in various networks), business network analysis (AirB&B, Uber, Lift, Netflix), trading stocks and currency exchange (millions of transactions are recorded in real time).

Within all those fields, data scientists were able to: (1) establish protocols and procedures for quantifying data, for collecting, structuring, comparing and sharing vast amounts of data; and (2) for mining the large data bases for extracting valuable and reliable information on the correlations between various sources and types and levels of data. The amount of cross flow of data is growing every year.

However, despite the fact that the practice of education represents one of the most vastly spread and one of the most important human practices, the methods developed in other fields for (1) collecting, and (2) mining BIG data have not found applications in the field of education.

At the current state of educational data mining:

1) Projects are localized in space and time – limited by one institution, one course, one teaching term (no longitudinal research) (https://onlinelearningconsortium.org/read/online-learning-journal/).

2) Research is limited in its scope by a small number of quantified parameters (usually a pair of parameters: "blended learning" – "student confidence"; "inter- or outer- student interactions" – "sense of community"; "scripted roles" – "cognitive presence"; "cultural differences" – student satisfaction"; etc.).

3) Research is usually based on a selected theory of teaching or learning, which currently has rather a holistic or heuristic nature, i.e. represents a set of working rules (Kolb's experimental learning theory; Polya's Problem Solving Techniques; Zimmerman's Self-Regulated Learning; etc.).

4) The majority of studies do not satisfy criteria for being "big data". Rare exception represents one or two year studies related to MOOCs (https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2586847).

However, currently available research concentrates the study on searching for correlations between student overall satisfaction, or intensions for taking a course, or a drop rate and various socioeconomic factors (gender, race, age, income, geographical location, course payment, etc.). These studies do not probe the structure of a learning process of students taking MOCCs.

II. Brief description of the current structure of the Educational Data Mining (a.k.a. EDM):

1. Educational Data Mining is in the stage of an early development (e.g. Educational Data Mining Society has been formed only five years ago: educationaldatamining.org).

2. Currently the following approaches are used to obtain data:

- Observing school teachers or college faculty while teaching and assessing teacher's actions using various observation protocols (e.g. BOPR, COPUS, MarzanoOP, RTOP, GORP).
- Observing school and college students while being taught using various observation protocols (e.g. a "STEM class observation protocol").
- Collecting responses to various surveys (e.g. "National Survey of Student Engagement", "National Survey of College Faculty").
- Collecting data during various student-computer interactions when using various computer-based media (MOOCs, computer games, intelligent tutoring systems, online content delivery systems, online homework delivery systems).

It is important to stress that:

(A) When data collection methods are based on the use of surveys or observation protocols, they are typically used only ones or twice during a teaching period (a semester, or a year); these methods are typically used to observe of a small percentage of teachers and students.

(B) Data collected using computer-based media does not access the everyday reflection of students on the learning process (actions taken for absorbing information and developing skills, and following results and satisfaction); does not access the everyday reflection of teaching faculty on the teaching process and on the student progress; this data typically presents the aggregated student response on the course as a whole (ranking the difficulty of a course, ranking homework assignments, indicating relevance of a textbook and other resources, overall satisfaction); mostly present two-parametric correlations like "time used for homework" – "final grade".

To summarize:

Current research does not involve data streams with a large number of parameters and coming from various source; current educational data are collected during isolated educational projects; does not satisfy criteria for being "big data" (less by orders of magnitudes); does not represent longitudinal streams of high frequency data incoming during the full term of learning; does not involve data streams with a large number of parameters; does not allow cross analysis for searching stable correlations between multiple parameters. In its current state, Educational Data mining is rather Advance Educational Statistics.

Current approaches do not provide understanding of the deep structure of teaching and learning processes, do not lead to development of quantitative measures of the quality of teaching, and development of quantitative measures of the trends in teaching (e.g. the measure of the improvement in teaching), and development of quantitative measures of the student progress correlated with student learning outcomes.

3. Currently, there is NO research which:

(1) regularly and *frequently* (e.g. several times a week) collects data *simultaneously* from teaching faculty *and* from students during the *whole* period of teaching a course (not just via observing one lecture);

(2) uses media technologies, including phone apps, to collect the desired sets of educational data incoming from *multiple* sources (faculty, disciplines, departments, institutions);

(3) uses technologies to mining data in searching for stable correlations between different factors affecting teaching-learning practices and student's performance using *multivariable* (multi-parametric) space.

Currently, there is no "brick-and-mortal" educational institution which collects from faculty and from students high frequency responses about multiple features of a teaching and learning processes. There is no institution which collects and cross-correlates multiple responses across various disciplines over a long period of time. And there are no educational institutions which collaborate in collecting and mining big data.

Research in education is missing and needs to start collecting BIG data – in the same sense as this term is used in other fields, in core sciences.

But before starting collecting big educational data we need to establish underlining principles which should be used for collecting big educational data; i.e. principles, which should be used to organize big educational data, to describe the structure of that data, to collect the data, and to make the data available for a broad community of professionals and stakeholders involved in education.

Developing methods and methodologies for collecting, sharing, and analyzing big data in education is especially important for the current highly technological society. It is a known fact that when two courses of the same type are taught at different institutions educators often observe different results, and students experience a different success. However, every employer would expect that when students coming from different institutions present similar diploma with transcripts listing the similar set of core technological courses, those students would demonstrate the similar set of skills. Collecting, sharing, and analyzing big educational data generated within departments and institutions, and across different departments and institutions, using the same set of rules and protocols for coding, quantifying, representing, storing the data should help to achieve the better uniformity in documentation and representation the knowledge and skills of graduates.

Exchanging information between institutions will allow to structuring big data, and to establishing common principles on how to collect and structure bid data in education. Having big data with providing easy access to big data across multiple institutions will provide ample opportunities for researches to data mining and to extracting quantitative information on the quality of teaching and on the ways for improving curricula, and methods of teaching.

To generate big data, research cannot be based on collecting a single individual response recorded once or twice a year or a semester. Collecting big data which will be sufficient for quantitative analysis, requires a longitudinal practice of collecting highly frequent, repetitive, timely correlated responses from all important participants of a teaching-learning process (i.e. teaching faculty and students).

Collecting timely correlated high frequency responses from multiple sources can be done only through the use of electronic media. Collecting timely correlated high frequency responses will require the development of a number of quantitative measures related to different factors and parameters which reflect the quality of education, and which are important for assessing the quality of teaching, the level of student involvement and responsiveness to teaching, and correlate those factors with the student learning outcomes based on students' homework and test grades, and student performance at various content inventories (for example, the Force Concept Inventory).

This technology will allow to provide important insights to individual faculty and will be helpful for a gradually development of their teaching skills. In particular, having an immediate feedback from big data will be essential for newly appointed faculty. The feedback from big data mining will be essential also for sharing teaching experience between colleagues who teach similar courses, and to allow to exchange the quantitative information on how courses have been taught, or how courses are being taught, and what is the current response from students, and what is a trend in students' response to a course.

Similarly to collecting and mining big educational data within an educational institution, collecting such data across institutions also will allow to exchange information (coming from high frequency measures) and to develop common views on general principles of teaching (not related to a specific institution or subject), and on general approaches for evaluating teaching practices and the results of those practices (within specific disciplines, and beyond specific disciplines).

The big data analysis will allow to see those features of teaching practice which are not related to the type of an institution, as well as features of teaching practice related to that specific type of an institution (e.g. a big "Ivy-league" institution, or a small public college, or even a high school). That will become possible if the mining of big data will be scaled up from one institution introducing the approach, to a group of similar institutions, to groups of institutions of different types. It needs to be stressed, that this type of cross institutional informational exchange can be done only when all parties use the same protocols for collecting, storing, and interpreting data. This can be achieved if all the parties will be using one (the same for *all* institutions!) database.

In this proposal we introduce a very innovative (a "Blue Chip") type of approach, where high frequency and highly quantitative data is used to represent the state of education at different levels of education, at different systems and subsystems of education (individuals, departments, groups of individuals, disciplines, universities, colleges, schools)

The project will lead to the development of a technological infrastructure to query and to extract information necessary for giving regular and quantitative feedback to an individual faculty, a department, an institution, as well as across departments and institutions. Teacher organizations, and policy makers, and officials of all levels in the system of education will be able to use this information for evaluating and ranking different institutions (including formal procedures like auditing and accreditation), for assessing the performance of individual faculty, for ranking the performance of similar departments in different institutions, and form making all types of decisions.

III. The scope and immediate goals of the proposed project:

The project will pioneer (A) the development of a new type of a big data base via collecting longitudinal streams of high frequency data in the field of education; (B) the development of the new methodology for mining new type of educational data and extracting valuable and reliable information on the correlations between various parameters of multiple data sources of different types and levels (faculty, departments, institutions).

Every day zillions of apps are being used by millions of people. People already have habits of tracking information every day (calories intake, calories burned, steps made, miles traveled, etc.). Why not harness the new technologies and the new habit to generate a stream of high frequency educational data?

The goals:

1. Establishing a set of measurable and universal (but modifiable) parameters which will be used for describing the state and structure of any teaching and learning processes (i.e. for any course).

2. Developing one questionnaire for teaching faculty and one questionnaire for students, which they will use during a course regularly and frequently for self-observation, for assessing students' actions and progress, for assessing faculty teaching actions and traits.

3. Developing an app for collecting the data provided by students and faculty.

4. Developing the strategy for analyzing the data coming from faculty and students in search for correlations.

5. Developing a web-site for collecting the data coming from faculty and students.

6. Piloting the program

This proposal is to pioneer educational data mining within a framework of one department of one institutions – we propose Boston university physics department will pioneer this project. However, we envision a future extension of this project to sister disciplines, like chemistry, mathematics, engineering – *the pillars of STEM education*.

We are planning on collecting from faculty and students *high frequency longitudinal responses* of various types:

* before the beginning of the course;

* after each lecture;

* after each exam;

* summative responses after two weeks of a course about lectures, labs and all other features of the course;

* generalized responses after each third of a semester;

*accumulative responses just before and after the final examination.

Responses will be collected using all available media; the primary media will be a smart phone, but participants will be able to use tablets as well as regular computers (all media will require an Internet connections). To enter responses participants will be filling up specifically designed questionnaires.

To quantify the stream of information on a uniform basis all answers to each question in each questionnaire will be graded in a form of a number between 0 and 10. The standard grades for various courses assignments (homework, laboratory work, quizzes, exams) will be including in the collected data, which will allow to find important correlations between the overall student performance and various features of learning and teaching processes. The results of this work will be used for developing and testing different method for analysis and establishing stable correlations.

The general goal of the project is to develop procedures which will allow to visualize the structure of the responses, changes in the structure, trends in changes in the structure. This should allow to access regularly student reflection on the course and on his or her performance during the course (how do students assess the difficulty of various assignments, the clarity or helpfulness of lectures, workbooks, textbook, office hours, etc., helpful traits of a lecturer). This also should allow to access regularly the structured reflection of a faculty on teaching approach selected for the course, on students' readiness, behavior, performance, success.

Our expectation (a.k.a. a hypothesis) is that further analysis of the data will demonstrate the existence of stable correlations between various parameters affecting learning process of students; trends in changes in correlations between various parameters affecting learning process of students. Some of the correlations and trends will manifest itself at local levels of a teaching-learning process (one group of students, one course). Continuation and broadening of the approach presented in this proposal will uncover other correlations and trends which will manifest itself at higher levels of a teaching-learning process (a department, an institution, a discipline). Ultimately, we expect to prove the existence of universal trends in student behavior observed a semester by a semester, which do not depend on what faculty teaches the course; or on the prevailing type of students taking the course (this study requires longitudinal collection and mining data – at least within one institution).

Even within one department we will be able to developed methods for understanding network interactions (correlations between different courses, faculty, semesters). Those methods will present an opportunity for scaling the project up to including interdepartmental data analysis, and in the future across institutional analysis (the same course – different universities; the same faculty – different courses; the same courses – different textbooks or course organization, the same course – different teaching strategies, the same professor – different teaching strategies, the same courses and strategies – but institutions of different type, etc.).

IV. Resources.

The project will leverage the existence of the expertise and resources allocated at the Boston University: including scientists who have deep expertise in developing and application of methods for collecting and organizing big data coming from multiple sources, for quantifying data, extracting information from big data on important correlations between multiple parameters describing functioning of various systems or subsystems, finding cross relations, describing information transfer between multiple sources. Using noise reduction methods, finding critical points and visualizing state transitions (PI Prof. Plamen Ivanov), and experienced teaching faculty (co-PI Dr. Valentin Voroshilov), and high computational facility (GHPCC).

V. Future development.

The goal of this proposal is to seeks a seed funding from the university to develop a prove of a concept model which will be used for seeking funding from the NSF (https://www.nsf.gov/div/index.jsp?div=DUE or

https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504924&org=EHR&from=home).

The proposed approach to educational data mining is pioneering the development of the new type of educational data, and the new methodology for collecting and mining that new type of educational data (including high frequency prime series and network analysis, and the used of the modern methods of dynamics of cross interactions, establishing the information concert between different sources of data, and other).

The unique feature and one of the most important assets of the proposed project is its *scalability*.

Starting from one faculty teaching one course during one semester, it has a potential to evolve and involve multiple faculty teaching various disciplines at different departments, and even at different institutions. The existence of high frequency responses coming from different parts and different levels of the educational system will allow to cross analyze the data and to establish stable correlations, which do not depend – or on the contrary, heavily depend – on the type of a teaching faculty (e.g. a newly hired professor v. a Nobel Laureate), or a teaching discipline (physics or chemistry), or a type of a course (elementary v. advanced), or a type of an institutional approach (traditional v. modified, blended, studio).

The first stage (one faculty, one course) will allow to establish basic principles for structuring, quantifying, collecting data, developing a methodology for building a data base, building a data base, developing methodology for mining the data, approbation of methods for collecting and mining data. The data will be collected using high frequency responses from the professor and from the students in a course. The goal is to see how the professor assess his performance and performance of the students, as well as to see how students assess the quality of the teaching and their own performance, and to correlate those data with the grades students receive via various grading procedures (homework, laboratory exercises, quizzes, midterms, final examination). The result of this stage will be demonstrating the viability of the approach via establishing various correlations between multiple features of a teaching process, of a learning process, and the learning outcomes of students.

Next stages would require scaling up the funding to broad the scale of the project including: prolonging teaching the original course, involving additional faculty from the same department, involving faculty from other departments, involving other institutions (this stage will leverage the existing collaborations in high computing and internet teaching between BU, MIT, and Harvard). At the later stages of the project BU may be able to establish a new department employing specialists from the physics department, department of engendering, BU IT personal, and designated for maintaining all technological processes related to functioning of the new data base (*when the seed grant will be supported by additional funds*, including NSF). Boston University also can become a hosting facility for hosting the centralized data base (at the Holyoke Green High Performance Computing Center).

When big data mining will be including data from different institutions researchers will be able to establish correlations which do not depend on the type of a courses, or type of institutions, or experience of a faculty, etc. This will allow to establish the general (universal) characteristics of teaching. This will also help to conduct a quantitative comparison of teaching strategies of different type, to see what type of strategies are more effective in what kind of teaching environment. Using quantification of strategies based on student responses, and being able to compare different teaching strategies and to find what different teaching strategies have in common and what makes them different (in terms of affecting student learning), researchers will be able to establish general characteristics of the process of improving teaching.

This project has a potential to transform the science of education, and – hence – the education. Using quantification of teaching strategies based on high frequency responses gathering from multiple sources researches will be able to compare different teaching strategies, to find what different teaching strategies have in common and what makes them different in terms of affecting student learning.

Having cross institutional, cross departmental, and cross disciplinary data analysis researches will be able to establish quantitative measures for making transparent the structure of teaching practices, for assessing the quality of teaching practices and their results; and to have a quantitative comparison of teaching strategies of different type, to see what type of strategies is more effective in what kind of teaching environment. This also will help to structure the learning process (absorbing information and skill development) by different types of students; to provide quantitative measures for the correlations between various factors influencing student learning and the everyday and the overall student performance assessed via various grading procedures.

As one of the results, there will be developed a course taught at BU and other institutions (including the use of EdX resource) on methods for numerating educational data, collecting data, storing and sharing educational data in physics (at first) at the level of a department, and then at a level of an institution and at a cross institutional level, and at different departments, and how to cross-analyze data for various purposes, including ranking of individual faculty, departments, institutions for various decision makers (college or university officials, policy makers).

For the first time in a long history of science of education a new territory will be opened – the territory of the implementation in education of the new type of quantitative measures, which will allow the development of new methods for evaluating the quality of teaching.

We are first who are proposing this pioneering approach, moving the research in the new direction (collecting truly big data; building a shared and growing data base), which will begin from establishing principles of how to collect big educational data. At the same time, we will be developing and testing methods for dealing with large amounts of educational data coming in the form of various data streams from different sources and related to various aspect of teaching and learning processes. We will be developing methods for synchronizing and cross correlating the vast amount of educational data within and across departments and institutions, which is something that has not yet been done before.

This application for a grant is the first and small step in the direction of the development of the completely new area of cross disciplinary research which will combine new methods of collecting and mining big data with the research on the structure, effectiveness and quality of various teaching-learning processes.

By merging contemporary methods developed for collecting and mining big data when streams of high frequency data coming from many different sources with educational practices happening at different time scales and locations, we will have an opportunity to initiate the process of reforming science of education.

The collection of big data and formation of big database triggers the necessity for formation of new methods for analyzing the big data base and for extracting reliable information on various correlations between different parts or aspects of the processes generated the data. The existence of shared big data will lead to the development of new methodologies for describing various aspects of education and measuring the quality of education.

With the support of the NSF, this project has a potential to follow the history of *the Human Genome Project* (which started at Boston University by Prof. Charles DeLisi).

With this proposal we have an opportunity to play role in the development of a science of education a similar to the role the Human Genome Project plays in health sciences.

VI. Summative Conclusion.

This very innovative approach is based on two pillars:

1. Developing a new kind of data base that will be constantly growing accumulating data about an individual faculty, a specific course, and then accumulating data coming from different professors teaching different course within the same disciplines and different disciplines within the same institution and from different institutions. But we will develop the first building blocks of the data base and will develop the methodology for continuous growth of the data base. This project team will become the home-base for this new database suing the unique high computational facility GHPCC.

With the involvement of BU, a new department employing specialists from the physics department, department of engendering, BU IT and designated for maintaining all technological processes related to functioning of this data base can be created in the future (when the seed grant will be supported funds provided by the NSF and other stakeholders).

2. The second pillar of the proposal is the data mining of the data collected in the data base. The process of data mining will request development of new methodology for mining vast and various educational data. The ultimate goal will be establishing universal principles of teaching and learning supported by quantitative measures, as well as universal measures for assessing the quality of teaching, and methods for improving teaching.

Ultimately, this project falls into two general strategic development areas which every science founding agency wants to develop (including NSF): creating a big data base of new kind of data, and development of new methods for data mining and extracting information from this type of data.

This proposal is the response to the demands of a scientific development and policy defining agencies. Currently many decisions related to improving education made by official in the field of education and policy makers are based on such factors like: personal political or philosophical preference, connections with a certain group or institution, circumstantial or even anecdotal evidence.

This project is to address in a completely new way, a quantitative modern way, the questions of the search for the basic fundamental laws and principles underlining and governing process of educational at various levels in various systems within various time frames; the laws which explain the influence of various factors on the process of accumulating information and the quality of teaching; the laws which demonstrate correlations between various factors of a teaching process and the learning outcomes demonstrated by students.

This proposal will leverage the technologies already existing in multimedia, and we are going to develop protocols for quantifying data, and a set of applications for collecting data from faculty and students.

And also, this proposal will leverage the expertise of the faculty of Boston University who have spent years in the field of data mining large data bases (physiologic data, Human Genome Project data), particle physics data, network science, high frequency business data analysis, and other, and have developed the appropriate techniques to quantify data, to cross relate data from various sources, under various conditions of nose, under various conditions of diversity of data – and equally unique. And finally, within this proposal we are going to leverage the existing resources of computational power, high computing center available to BU, MIT, and Harvard. This collaboration also presents a leverage for moving the project beyond BU.

This will bring the completely revolutionary approach to introducing new quantitative high frequency longitudinal measures in the field of education.

Ultimately the project will generate big data related to different scales in education: one faculty scale, departmental scale, interdepartmental or institutional scale, interinstitutional scale. Correlation analysis in part will show how different elements of an educational process influence the learning outcomes of students. For example, in different institutions different professor can teach the same physics course using the same or different teaching materials. One professor maybe a Nobel Prize winner, and other one just recently hired. Hence we will have to find a method to quantify those differences.

With each new stage of the project researchers will have to find methods for quantifying differences between different faculty, or different departments, or different disciplines, or different institutions, as well as differences in student body. This cross correlational analysis will show features of educational process which depend on those differences, and which are universals within certain scale, or discipline, or other domain of teaching.

These correlations will have different level of universality, some will be very general (do not depend on the structure of teaching process), some will be more individual and less general (specific for a given discipline, for example observed only when physics is taught). Time analysis will show changes and trends in changes within a specific domain of teaching (for example when the same faculty teaches the same course over several years; this information can be used for teacher evaluation, making decision on a promotion or tenure).

It will allow to compare longitudinally the evolution of a teaching process, and even to make predictions about the effectiveness and result of a teaching process.

The uniqueness of the approach is (1) in collecting data at a high frequency (which currently does not exist: with the shortest period is the time interval between two consecutive lectures or two consecutive student-teacher or student-student interactions); (2) collecting high frequency data from all parties involved in teaching (professor, teaching assistants, students); (3) collecting data over long period of time (the study is longitudinal); (4) collecting data from a large number of different sources (different faculty, courses, disciplines, departments, institutions).

Students involved in the project will be expected to have developed a habit of using media for self-assessing the learning process, which to be expected in resulting of having a higher than average level of self-reflection. This statement is a hypothesis of the future research, in which students who participated in the project will continue generating data after their graduation. The project also can be extended beyond the institution of education and include business environment, hiring personal. That will lead to establishing quantitative correlations between learning experience of former students and their professional performance.

Our team will be the first to start the development of the new type of quantitative measures for education.

This proposal is to lay the foundation, the base stone in the process of building the new system for collecting and mining big educational data.

Full Proposal: Visual Presentation (4 pages)

"Developing Strategies and Technology for Generating and Analysis of Longitudinal High Frequency Data Streams from Faculty and Students".

Big Data
en.wikipedia.org/wiki/Big_data

Existing Areas of Research:
* Human Genome Project (DNA v. health correlations).
* Health care and epidemiology (telemedicine, spread of diseases).
* particle physics.
* social networking (Facebook, Twitter, Snapchat, Instagram, cellphone communication).
* Business networking and analytics (remote business communication, AirB&B, Uber, Lift, Netflix, trading stocks, currency exchange).
* National security (trends in various networks).

Fundamentals of the big data mining:
(1) Establish protocols and procedures for quantifying data, for collecting, structuring, comparing and sharing vast amounts of data.
(2) Developing methods for cross analysis of vast amount of data coming from many different sources.

Results:
Valuable and reliable information on correlations between multiple parameters influencing the state of a research system (a.k.a. laws).

One of the most vastly spread and one of the most important human practices:
EDUCATION – is NOT a part of BIG data mining research
(Educational Data Mining currently is rather Advanced Educational Statistics)

The Current State of Educational "Data Mining"

Educational "Data Mining"

http://www.educationaldatamining.org/

Live Teaching and Learning Observation Protocols, including using Phone App (GORP)

http://www.bu.edu/provost/planning/program-learning-outcomes-assessment/resources-for-assessment/generalized-observation-and-reflection-protocol-gorp-tool-pilot-2015/

http://ed.fnal.gov/trc_new/program_docs/instru/classroom_obs.pdf

http://tdop.wceruw.org/

http://www.iobservation.com/files/Marzano-Protocol-Using_Rounds1009.pdf/

http://www.serve.org/uploads/docs/Gen%20Documents/STEM_Classr_Observ_Protocol_Generic_10_27_14_.pdf

http://www.public.asu.edu/~anton1/AssessArticles/Assessments/Chemistry%20Assessments/RTOP%20Reference%20Manual.pdf

http://www.cwsei.ubc.ca/resources/files/COPUS_protocol.pdf

http://education.ufl.edu/325t/files/2013/06/QualityUrbanClassroomsSobelUCDenver.pdf

http://projectachieve.info/assets/files/pdfs/BEHAVIORAL_OBSERVATION_PROTOCOL_808.pdf)=

National Survey of Student Engagement

http://nsse.indiana.edu/

http://onlinelibrary.wiley.com/doi/10.1002/ir.283/abstract

National and other Surveys of Faculty

http://nces.ed.gov/surveys/nsopf/

https://nces.ed.gov/surveys/nsopf/survey_forms.asp

http://heri.ucla.edu/monographs/HERI-FAC2014-monograph.pdf

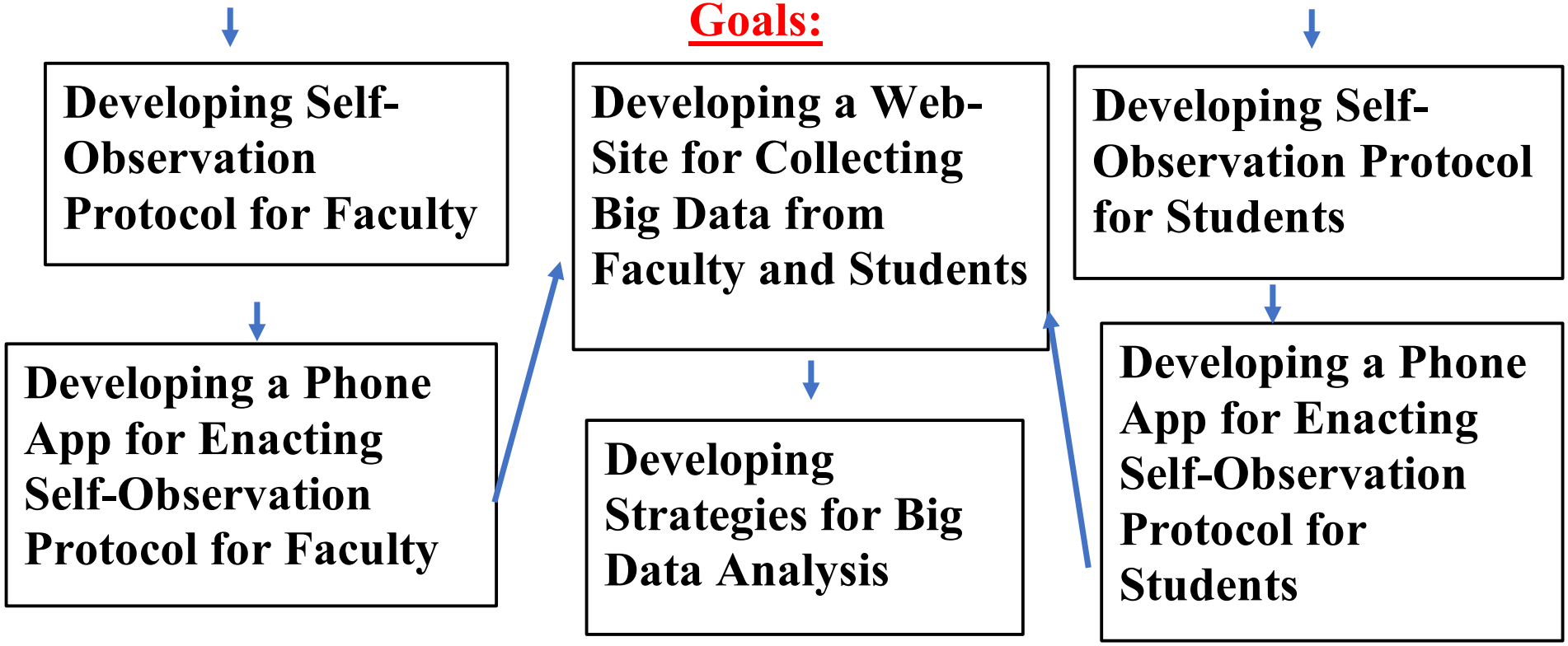
http://www.bu.edu/apfd/files/2009/10/Boston-University-COACHE2010-Executive-Summary1.pdf

Using technologies to analyze data coming from live interacting between students and computer-based media (MOOCs, computer games, online content managing systems, online homework delivery systems) http://www.educationaldatamining.org/proceedings

Conclusion:

Currently, there is NO research which: (1) regularly and frequently collects data from teaching faculty during the period of teaching a complete course (not just observing one lecture); (2) regularly collects data from students in the same course; (3) uses technologies (including phone apps) to collect the two sets of data; (4) uses technologies to analyze the data stream in searching for correlations which can help to understand how different factors affect students' performance.

Goals:



Project

Unique: (1) collecting high frequency responses; (2) simultaneous data streams from all participants (faculty, students); (3) longitudinal study (years); (4) data from multiple sources (faculty, departments, institutions).

General Stages of Growth:

1. One department, one course, one faculty (this proposal).
2. One department, several courses, several faculty.
3. One institution, several departments.
4. Several Institutions.

Boston University Physics Department

Leverage:

1. Researchers with deep expertise in big data mining.
2. History of longitudinal collaborative projects (Human Genome Project).
3. High Computing Facility.
4. Collaborative relations with MIT and Harvard University.

1st Stage Results:

1. Basic principles for structuring, quantifying, collecting data
2. Initial data base
3. Methodology for mining the new type of data
4. Approbation of methods for collecting and mining data.
5. Demonstration of the viability of the approach, and initiating further funding.

Ultimate Goal:

Establishing universal principles of teaching and learning supported by well-established quantitative measures, used to assess various aspects education.

Ultimate Results: 1) Providing a solid foundation for a science of education; 2) Providing decision makers with reliable quantitative measures of the quality of education.