345 Sunset Drive
Selinsgrove, PA 17870

December 14, 2015

SRC, Inc.
7502 Round Pond Road
North Syracuse, New York 13212-2510

Dear Human Resources, Hiring Manager, and Hiring Team:

Thank you for taking the time to interview me for SRC's senior data analyst position. The purpose of this document is to provide details about my capabilities and potentials. The content of this binder is largely driven by the job responsibilities and hiring qualifications indicated in the job posting.

> Disclaimer: I am far removed from your needs and organizations. It seems likely that I may have misinterpreted your requirements or your interests. If I did so, please forgive me. My goal is to make sure that it is clear to you who I am, what I can offer, and what my weaknesses and strengths are.

It is my understanding that the successful applicant will have assured SRC that s/he can contribute significantly to the United States of America's or its allies' military, cyber, and critical infrastructure defenses. To do this, s/he will illustrate abilities to:

- Propose products and services (e.g. via whitepapers)
- Team with subject matter experts (e.g. in SRC's product lines)
- Harmonize with DoD researchers (e.g. by bridging through PEOs)
- Collaborate with industry partners (e.g. with a partner that complements SRC)
- Continuously expand their analysis skills (e.g. using probabilistic graphs)

Necessarily, the hired candidate's responsibilities will be executed as part of structured, developing, and rolling plan, that has broad SRC support. As a substantive contributor, the hired applicant will need to be able to relate to other individuals.

The table of contents provides a list of the identified responsibilities and qualifications and this binder's affiliated section's letter. Each section is reached by using PDF bookmarks and/or in-document hyperlinks.

To be very clear, I am a slightly experienced researcher who is aggressively refining his ability to analyze data using modern tools and techniques. I am nowhere near the competency level in data analysis that I desire. I am also an engineer who enjoys bringing new ideas to life and arguing for needed products. In closing, I am very serious about discovering using data. Again, thank you so much for interviewing me.

All the very best,

[original signature provided on hardcopy]

Russell W. Robbins

Back to Cover Letter

Table of Contents

Section A - Robbins' Initial Cover Letter (copy)

This section contains my initial cover letter requesting that SRC consider my application for employment as a senior data analyst.

(Please turn the page.)

Russell W. Robbins, Ph.D.

345 Sunset Drive
Selinsgrove, PA 17870

November 30, 2015

SRC, Inc.
7502 Round Pond Road
North Syracuse, New York 13212-2510

Dear Human Resources, Hiring Manager, and Hiring Team (if appropriate):

I would like to explore with you how I might deliver as a senior data analyst in the Advanced Analytics group at SRC. I am thrilled about this potential opportunity. Each of the following main bullets is a qualification or duty for the analyst role. Each of the checks indicates how my skills, experiences, or knowledge correspond to a specific qualification or duty.

- Master's and/or Ph.D. in data science, math, statistics, operations research, computer science or a related field
    - ✓ PhD (Computational Modeling) from Rensselaer Polytechnic Institute
    - ✓ Big Data Certificate from UC Berkeley Algorithms, People, & Machines Lab
    - ✓ Data Science Certificate from Johns Hopkins Biostatistics Department
    - ✓ Earning Machine Learning Certificate from University of Washington
    - ✓ Earning Business Analytics Certificate from Penn/Wharton
    - ✓ Planned Big Data Certificate from UC San Diego
    - ✓ Planned Data Mining Certificate from U of Illinois at Urbana-Champaign
- Experience cleaning and manipulating data
    - ✓ Built program that merges six mishmash files and outputs 562 orderly files.
    - ✓ Reshaped, bridged, & merged US college enrollment files from 2002-2012.
    - ✓ ETLed 1.4 M records from non-relational data store to Oracle.
    - ✓ Building program that writes parser programs for any condensed ASCII files.
- Experience with predictive modeling
    - ✓ Predicted actions of persons using health activity trackers (e.g., fitbit)
    - ✓ Predicted image categories based on deep features.
    - ✓ Predicted release year of songs based on audio characteristics.
    - ✓ Predicting Medicaid insurable drug abuse treatment patients.
- Experience with developing and implementing complex algorithms
    - ✓ Built assessable optimizing algorithm for forecasting, purchasing, developing new products, producing, marketing, and pricing.
    - ✓ Built executable integration of all combinations for computing binary classification performance probabilities and derivative ratios.
    - ✓ Extracted numerical features from categorical data using one-hot-encoding, then reduced dimensionality of features with hashing, to train logit.
- Experience using statistical analysis tools such as R and RapidMiner
    - ✓ I have substantively used R, RapidMiner, Python, Stata, SPSS, and MATLAB.
    - ✓ I have substantively used several IDEs related to R and Python.
    - ✓ I have substantively used many R libraries and Python packages.

- Exceptional creative, critical thinking and problem solving skills

  - ✓ Building modular, interactive, generative agent-based model of English.
    - o Built most comprehensive, in memory/disk, declarative, grammatical, semantic, non-proprietary model of English language.
    - o Identified method for automatically loading software agent memories with declarative information.
    - o Identified linkable and executable reasoning, decision making, judgment, and problem solving, and communication theories.
    - o Researching methods to derive and load procedural knowledge into software agents.
    - o Researching methods for building software agent specific knowledge using machine learning inside software agents.
    - o Note Oracle 12c supports graph data structures and text indexes.
    - o Note distributed, parallel, scalable processing is almost mainstream.

  - ✓ Built first executable model of ethical problem solving.
  - ✓ Hypothesized how software agents could be ethics advisors.
  - ✓ Built first immersive case based learning environment with NPCs.
  - ✓ Built first scenario building software centering on supporting diversity, perspectives, and pluralism.

- Experience with machine learning (natural language processing a plus)

  - ✓ Predicted unigrams, bigrams, trigrams, 4-grams, and 5-grams.
  - ✓ Found most similar Wikipedia articles using bag of words model.
  - ✓ Identified distinct neuron firing patterns induced by different stimuli.
  - ✓ Predicting prices using feature extraction, Ridge, Lasso, and KNN.
  - ✓ Merging classic probabilistic NLP approaches with functionalism.

- Excellent interpersonal and communication skills

  - ✓ Led software projects at the IRS, GE, IBM, NXP Semiconductors, & MapInfo.
  - ✓ Led four emerging technology projects at the University of Pittsburgh.
  - ✓ Wrote 19 peer reviewed accepted manuscripts, including 12 publications.
  - ✓ Wrote seven grant proposals, four awarded.
  - ✓ Presented at 13 conferences.

- Three or more years of experience (including internship experience)

  - ✓ Two years of military experience
  - ✓ Seven years of professional non-academic experience
  - ✓ Sixteen years of teaching experience
  - ✓ Ten years of research experience after PhD

- Experience with analyzing large volumes of data using distributed processing architectures such as Hadoop

  - ✓ Substantively used Apache Spark, RDDs, as well as map, reduce, lambdas.
  - ✓ Substantively used Cassandra, MongoDB, AllegroGraph.
  - ✓ Evaluated Virtuoso, MarkLogic, OrientDB.
  - ✓ Taking Hadoop Platform and Application Framework (UC San Diego).

- Active Top Secret clearance

  - ✓ I am sorry. I do not have an active Top Secret clearance.
  - ✓ I had a secret clearance from 1985 to 1987.
  - ✓ I do not have any skeletons in my closet.

- Data Mining

  - ✓ I am aware of the types of algorithms and their underlying approaches.
  - ✓ I am studying Data Mining & Analysis, Zaki and Meira, 2014.
  - ✓ I am taking a clustering course in Machine Learning Certificate in January.
  - ✓ There is a clustering course in my planned Data Mining Certificate.
  - ✓ There is a clustering course in my planned (second) Big Data Certificate.

- Trend, Pattern, and Statistical Analysis

  - ✓ Covered in two Masters
  - ✓ Covered in PhD
  - ✓ Covered in Data Science Certificate (earned)
  - ✓ Covered in Big Data Certificate (earned)
  - ✓ Covered in Machine Learning Certificate (underway)
  - ✓ Covered in planned certificates
  - ✓ Studying Elements of Statistical Learning, Hastie et al. 2009
  - ✓ Studying Pattern Recognition and Machine Learning, Bishop, 2006
  - ✓ Studying Probabilistic Graphical Models, Koller & Friedman, 2009

- Analysis of structured and unstructured data

  - ✓ I am very comfortable assessing "unstructured" data.
  - ✓ I do not experience unstructured data (e.g., text) as unstructured.
  - ✓ My experiences with images and audio have been positive.
  - ✓ I have no fundamental issues studying classically quantitative or pseudo-quantitative data.

- Communicate complex machine learning concepts to non-specialists

  - ✓ I have not done this yet.
  - ✓ I have taught clustering to unskilled undergraduates.
  - ✓ I have taught technical skills to undergraduates.

- Develop white papers to cultivate new business opportunities

  - ✓ I wrote seven proposals, of which four were successful.
  - ✓ I was awarded two NSF grants.
  - ✓ I teamed with the IRS, GE, IBM, NXP Semiconductors, and MapInfo.

- Temporary travel (up to 10%) of the time

  - ✓ I love to travel.

At this point I will change the direction of this cover letter significantly.

I worked as an entry level professor at three institutions over nine years, after my PhD. Seeing this information in a resume may naturally lead to the question "Why didn't you make tenure or at least move up in the organizations?" While answering this question for myself, I was able to identify several factors. For conciseness and clarity, I will only focus only two interacting factors:

- I taught rigorous or very rigorous courses.
- I had Attention Deficit Hyperactivity Disorder (ADHD), but did not realize I had ADHD.

Continued on page 4 …

ADHD symptoms I (and my unfortunate students) experienced as I taught included being easily distracted and switching from one activity to another activity quickly and frequently. In particular, these two ADHD symptoms played a significant role in hampering my students' learning. You can imagine the frustration of a student held to a very high bar in their learning, and who should be able to learn in a carefully scaffolded way; However, this student's professor cannot consistently provide a lecture that has the form A->B->C->D->E. Understandably, this student would be very frustrated; Again, understandably, this student should share this information via student evaluations (or if necessary, more timely, means).

Because of this kind of problem, usually, I was not able to consistently achieve above a "4" on a scale of "1 to 5 with 5 as the best score", across three sections, in a semester. In teaching roles, as I am sure you are aware, excellent teaching is required. To fulfill this requirement, I attempted to address the less than excellent ratings students shared in evaluations. I sought and applied anonymous feedback from students, evaluations from expert instructors, and teaching centers' staffs suggestions. I found, learned, and applied educational theory prescriptions. I learned my deficiencies; I also learned the necessary solutions.

However, despite how hard I tried, I was not able to completely stop the symptomatic ADHD behaviors that frustrated my students. Fortunately, I am now treated. However, after 16 years of that uphill battle, I have no energy left to teach up to 155 student customers each semester. Now, as I close this letter, I quickly discuss my personal interest in the geographic location of Syracuse, New York, which complements my professional interest in a data analyst role in a non-profit research organization.

I lived in upstate New York for 16 years and found it the best place I have lived. This is largely due to the fact that my wife is from a small town near Utica, but it is also because of the Adirondacks, the snow, our friends, etc. Thus, pursuing this prospective professional role is also personally appropriate for my wife, our children, and myself. However, this interest does not preclude my family and I from considering living in SRC's non-headquarter locations, if I were fortunate enough to join your organization and the need arises.

My portfolio, which only includes "well framed" work products that I am interested in sharing publically, is at: https://robbinsr.squarespace.com. Less well-framed research or analyses could be provided if you are interested. I would very much appreciate the opportunity to interview. My phone number is 570-884-3647. My email address is russ.robbins@outlook.com. Thank you so much for considering me as a potential contributor to your teams, your customers, and your organization.

All the very best,

Russell W. Robbins

Section B – Robbins' Résumé (copy)

This section contains a one-page summary of my work experiences.

(Please turn the page.)

Russell W. Robbins, Ph.D.

570-884-3647
russ.robbins (Skype)
russ.robbins@outlook.com

345 Sunset Drive
Selinsgrove, PA 17870
https://robbinsr.squarespace.com

EDUCATION

Exp. 03/16	Business Analytics Certificate	University of Pennsylvania	100 to date
Exp. 03/16	Machine Learning Certificate	University of Washington	100 to date
8/15	Big Data Certificate	Univ. of California, Berkeley	94.5/100
8/15	Data Science Certificate	Johns Hopkins University	95.5/100
12/05	Ph.D. Engineering Science	Rensselaer Polytechnic Inst.	3.86/4.00
12/04	M.S. Information Technology	Rensselaer Polytechnic Inst.	3.86/4.00
05/97	M.S. Accounting	Binghamton University	3.70/4.00
12/90	B.S.B.A. Finance	University of Missouri	3.20/4.00

EXPERIENCE

09/14 - 11/15
Student

- Learned broad set of tools and techniques to capture, represent, manipulate, analyze, and share data.

09/13 - 09/14
Assistant Professor
Susquehanna Univ.

- Designed framework that integrates software development lifecycle and project management responsibilities.
- Taught 63 students. Wrote one paper. Built two systems.

08/08 – 08/13
Visiting Asst. Prof.
Univ. of Pittsburgh

- Led four NSF or dean funded technology projects.
- Taught analytics by using SAP ERP competitively.
- Taught 1925 students. Wrote six papers. Coached 15 students.
- Wrote four proposals; three awarded. Built four systems.

08/05 – 08/08
Assistant Professor
Marist College

- Led software projects at IBM and NXP Semiconductors.
- Led two self-assessment efforts for five faculty.
- Taught 780 students. Coached ten students.
- Wrote three papers. Wrote two proposals; one awarded.

09/97 – 06/05
Clinical Asst. Prof.
Adjunct Instructor
Systems Analyst
Rensselaer Polytechnic

- Led software projects at GE Specialty Materials and MapInfo.
- Co-led student data ETL from legacy to relational database.
- Built prototype data warehouse. Built ethics decision aid.
- Built computational model of ethical decision making.
- Taught 1080 students. Coached 16. Wrote dissertation.

01/98 – 12/99
Project Manager
Achaean Technology

- Prototyped and marketed first enterprise information system for a particular type of social service agency.
- Led ten employees.

05/90 – 05/94
Clerk
IBM

- Distributed $3B annually to governments and suppliers.
- Handled “hot”, “large”, “electronic” payments.
- Led 5 people. Improved business processes.

08/85 – 08/87
U.S. Army

- Chauffeured and learned from mid-level officers.
- Trained as front line radio operator. Honorable Discharge.

Section C – Robbins' References Contact Information (new; added 4th reference)

This section contains contact information for and background information about my relationships with my references.

Please note that I have added Ms. Sharon Kunkel to the original list that I submitted to SRC. Sharon is the Registrar at Rensselaer Polytechnic Institute. I teamed with her associate registrar, who was a subject matter expert in information that Rensselear Polytechnic Institute collected and kept about its students and their relationships with the university.

Sharon's associate registrar and I worked together to extract, transform, and load information from a System 32, RPG accessible, non-relational data store, to an Oracle database that underlay SCT Banner, which is an enterprise system for post-secondary educational institutions. Sharon can characterize my nonacademic professional work when I was in a non-research environment.

Since the employment application was limited to three references, I have included Sharon's information here.

(Please turn the page.)

Russell W. Robbins, Ph.D.

REFERENCES

DR. BRIAN S. BUTLER
Professor and Dean of the College of Information Studies
University of Maryland
301-405-2033
bsbutler@umd.edu

Brian was my supervisor at the University of Pittsburgh.

DR. WILLIAM E. HEFLEY
Clinical Professor of Information Systems
Naveen Jindal School of Management
The University of Texas at Dallas
972-883-5006
William.Hefley@utdallas.edu

Bill was a senior colleague of mine at the University of Pittsburgh.

MS. SHARON L. KUNKEL
Registrar
Office of the Registrar
Rensselaer Polytechnic Institute
518-276-6028
kunkes@rpi.edu

Sharon was my "dotted-line" supervisor and customer at Rensselaer.

DR. WILLIAM A. WALLACE
Yamada Corporation Professor
Professor of Industrial and Systems Engineering
Professor of Civil and Environmental Engineering
Professor of Cognitive Sciences
Member, Faculty of Information Technology
518-276-6854
wallaw@rpi.edu

Al has been my teacher, mentor, patron, colleague, and friend.

Section D – Robbins' Curriculum Vitae

This section has a detailed summary of activities I have been fortunate to perform professionally.

(Please turn the page.)

Russell William Robbins, Ph.D.

Curriculum Vitae

WORK EXPERIENCE:

9/2014 - 10/2015
Student
Selinsgrove, PA
Supervisor: Self

Summary: 15 months of intensive training in quantitative analysis.

1. If appropriate please also see portfolio at http://robbinsr.squarespace.com .

2. Earned Big Data certificate from UC Berkeley AMPLab and EdX. Certificate attached.

3. Earned Data Science certificate from Johns Hopkins University Biostatistics and Coursera.

4. Participating in a five course Business Analytics sequence from the University of Pennsylvania Wharton School and Coursera.

5. Participating in a six course Machine Learning sequence from the University of Washington statistics and computer science schools as well as Coursera.

6. Built a program that writes parser programs for the National Health Interview Survey Data.

7. Developed broad understanding of current state of a particular portion of the online education market by evaluating or using over twenty online education vendors.

8. Studied neurons as they fired in a zebrafish.

9. Learned how to reduce dimensions.

10. Predicted website click through rates.

11. Predicted words used by bloggers.

12. Assessed the accuracy of statistical functions in R.

13. Built and documented repeatable data processing pipelines.

STUDENT continued on next page

STUDENT continued from page 1

Fundamental Machine Learning Skills

- Classification (basic)
- Regression (basic)
- Resampling (rudimentary)
- Model Selection (rudimentary)
- Regularization (rudimentary)
- Non-linear Models (rudimentary)
- Tree based Methods (rudimentary)
- Support Vector Machines (rudimentary)
- Clustering (rudimentary)

Fundamental Statistics Skills

- Descriptive Statistics (proficient)
- Distributions (basic)
- Probability Theory (proficient)
- Bayes Theorem (basic)
- Hypothesis Testing (between basic and proficient)
- Simple & Multiple Linear Regression (proficient)
- One way & Multifactor ANOVA (basic)
- Logistic & Ordinal Regression (proficient)
- Binomial Test (basic)
- Chi square Contingency Tables (basic)
- Non-parametric Alternatives (proficient)

Statistical Programming Toolboxes

- MATLAB (evaluated)
- Octave (used)
- Minitab (used)
- Python (used)
- R (used)
- Rattle (learned)
- Revolution R (used)
- SAS (evaluated)
- Stata (used)

Databases

- Cassandra (formal training)
- Ontotext GraphDB (used)
- MongoDB (formal training)
- Neo4j (evaluated)
- Stardog (used)
- Virtuoso (evaluated)

STUDENT continued on next page

STUDENT continued from page 2

Development Environments

• Anaconda (evaluated)
• Databricks (used)
• Enthought Canopy (evaluated)
• IDLE (evaluated)
• iPython interpreter (evaluated)
• iPython notebook(used)
• Komodo (evaluated)
• Oracle SQL Developer (used)
• Oracle Applications (used)
• Pycharm (used)
• Spyder (used)
• Stanford Protege (used)
• Teradata Studio Express (evaluated very lightly)
• TopBraid Composer (used)
• Visual Studio (evaluated)
• Web Storm (used)
• Wing (evaluated)
• WinPython (used)

Declarative/Procedural Programming Languages

• CSS (rudimentary)
• HTML5 (basic)
• Javascript (rudimentary)
• JSON (basic)
• Markdown (basic)
• Pandoc (basic)
• Python (familiar)
• OWL (basic)
• R (proficient)
• RDF (basic)
• RDFS (basic)
• Regular Expressions (between basic and proficient)
• Spark (familiar)
• SPARQL (basic)
• XML (basic)

Sample R Libraries

• caret (used)
• ggplot2 (used)
• data.table (used)
• knitr (used)
• lattice (used)
• regex (used)
• plyr (used)
• rCharts (used)

STUDENT continued on next page

STUDENT continued from page 3

Sample Python Packages

• rPython (evaluated)
• Beautiful Soup (used)
• Bottle (used)
• Core NLP (evaluated)
• iPython (used)
• MatPlotLib (evaluated)
• NumPy (used)
• Pandas (used)
• PyMongo (used)
• pyR (evaluated)
• PySpark (used)
• Re (used)

08/2013 - 08/2014
Susquehanna University
Sigmund Weis School of Business
Selinsgrove, PA
Assistant Professor of Information Systems
Supervisor: Dr. Barbara McElroy (570-372-4242)

Summary: One year of experience teaching, researching, writing, and programming.

1. Designed, built, used, and assessed information technology based instructional system that allowed users to practice project management and software engineering skills, observe and refine their developing knowledge, and build a portfolio of their experiences and results. Instructional system aided users learning how to generate and assure system requirements. Instructional design of system is based upon problem based learning. Application is based upon a project management oriented software engineering approach described below.

2. Instructional system focused on linking concept of operations, requirements analysis, design, development and/or purchase, testing, customer assurance, and people, using the best from the Project Management Body of Knowledge from the Project Management Institute as well as the Software Engineering Body of Knowledge and its many explicit and implicit standards promulgated by the IEEE, and other concepts.

3. Planning was risk centered and began with identifying and beginning the tracking of and consensus building among sponsors, customers, managers, super users, and the current situation, identifying problems, effects of the planned system on the current environment as well as installed systems, and users. Planning continued through identifying project drivers, constraints, and known issues and measures. It also included focuses on measures, monitoring, and planned uses for human, financial, and physical capital.

4. Information system solutions built by users included improving the process of moving passengers through the business processes of air flight passenger ticketing, baggage checking, and boarding as well as scheduling salt trucks based upon integrating roadside weather stations data.

SUSQUEHANNA continued on next page

SUSQUEHANNA continued from page 4

5. Project management oriented software engineering then moved on to requirements analysis. In requirements analysis, first pertinent business events were captured, then as is use cases, as is data models, and as is process models were developed to clarify the business events. In each case these models enabled capturing a low level of granularity and helped analysts identify potential measures.

6. Process then focused on Design. Design was forced to map to the Analysis as Analysis was forced to map to Concept of Operations and initial Planning. Design included to be use cases including preconditions, minimal guarantees, success guarantees, triggers, primary scenario script, as well as extensions, exceptions, misuses. Design used some of the same modeling techniques as Analysis but had three layers, including working prototypes as well as the documentation of evolved but approved requirements.

7. Project management tasks, risks, and quality concerns were identified by using the evolved requirements. These tasks, risks, and quality concerns, when integrated with human resources, physical plant, and constraints, then drove the schedule, costs, and which of the prioritized requirements could be fulfilled after considering interdependencies of the requirements.

8. Project management tasks included standard development or purchase methodologies as well as testing, installation, verification and validation, development of documentation and training materials, as well as customer and other stakeholder activities.

9. Developed, used, and analyzed results from online but learning science theory driven surveys of customer satisfaction. Used the assessment features in Blackboard® Learn to control quality in my blended courses or electronic learning environments and to align organizational goals with student needs. I have also used Web 2.0 and mobile learning tools such as Socrative™, PollEverywhere™, and Qualtrics™. Socrative was particularly helpful for understanding how well students were learning, during a workshop- oriented lecture, where students and I practiced skills together.

10. Continued to establish contacts and maintain active involvement in instructional design and/or technology and related areas through participation in professional activities, by publishing and sharing "Clarifying the SAP ERPsim Experience." Please see: http://russrobbins.info/assets/clarifying.pdf.

11. Engaged in general and focused market research into existing and emerging technologies, with an emphasis on balancing the use of commercial products with open source, collaborative web technologies, and open educational content, especially in the context of the semantic web.

12. Completed courses for Data Science Specialization at Johns Hopkins University.

13. Wrote and published paper. See: http://robbinsr.github.io/assets/papers/clarifying.pdf

---

VITA continues on next page

08/2008 - 08/2013
University of Pittsburgh
Joseph M. Katz Graduate School of Business
Pittsburgh, PA
Visiting Assistant Professor of Information Systems
Supervisor: Dean Brian S. Butler (301-405-2033)
Note: Brian is now at the University of Maryland, College Park.

Summary: Five years of experience managing and co-managing all aspects of multiple technology training programs for five years. I describe four examples below. Taught 1925 students project management, data analysis, and information systems management.

1. In the INNOVATE project we investigated collaboration technologies, researched, compared, and purchased software. Used collaboration best practices and software discovered in projects to evaluate team technologies, product comparisons, and leadership/member competency assessments. Described how results provided activities which were prescribed in model curriculum developed by professional association. Prescribed topics that were taught included business processes, emerging technologies, globalization, human–computer interactions, and the impacts of digitization. I managed a $23,000 budget, short schedules, and 7 stakeholders, risks driven by customers' values, and requirements driven by learning goals. See: http://robbinsr.github.io/assets/teaching_.pdf

2. In the SAP ERPSIM project, we built blended learning to help students learn business process optimization. I anticipated, eliminated, and mitigated risks driven by a lack of documentation and an insufficient wireless network. Students experientially learned business process optimization as they practiced globalism collaboratively by running a German muesli manufacturing company in an innovative simulation using actual enterprise software. Analyzed electronic transaction data to assess students' abilities to optimize business processes. Taught and supported instructors. I managed a $5,200 budget, short schedules, 1000+ students, and 100+ requirements. See: http://robbinsr.github.io/assets/papers/clarifying.pdf

3. In the SIMULATE project we developed NSF-sponsored curriculum for learning ethics. Under a colleagues' and my management, we built 2 courses, 10 cases, and two systems. Used learning gain tests. Students used innovative curriculum, method, and software to experientially and collaboratively learn ethical decision making in a globally diverse world. Analyzed criteria, decision making processes, and decisions within software. Described how curriculum is grounded in ethics, knowledge, and cognitive theories. Described assessment of curriculum using theories. Reported findings that students learned about the importance of diversity, multiple perspectives, values, and pluralism. I individually managed $101,491 budget, a 3-year schedule, 10 staff, 200+ students, and 100+ requirements. Our primary risks were driven by a general lack of knowledge in ethics education. See: http://robbinsr.github.io/assets/papers/information.ethics.pdf

4. In the VIRTUALVERSITY we integrated media rich collaborative environment (3D ICC TERF®), instructional design theories, and case based learning. Instructional system helped students project management. Analyzed data from electronic instructional system's whiteboards, documents, text chats, and audio/video records. Students learned project management experientially as they collaboratively recommended project solutions. The embedded case protagonists were actors and reflected the diverse and global modern corporation. Described project objectives and how product achieved objectives. Reported theory grounded study.

UNIVERSITY OF PITTSBURGH continued on next page

UNIVERSITY OF PITTSBURGH continued from page 6

Published assessment methods and theory for selecting instructional technology. I managed $20,000 budget, a two-year schedule, three contracts, and coordinated 10+ stakeholders and 100+ requirements. Risks were caused because we did not manage scope. See: http://robbinsr.github.io/assets/papers/virtual.teaching.pdf

5. Collected and analyzed quantitative and qualitative research data to support study of ethical decision making. Data was collected using software built by a team I led. Data was also collected using audio/video as well as concurrent/retrospective reports. Data was analyzed using theory grounded coding schemes.

6. Evaluated, proposed, and implemented vendors' software products and services, coordinated vendors' professional services.

7. Envisioned, proposed, designed, managed programmers, assessed, and shared information systems.

8. Led, contributed to, and assessed student learning.

9. Wrote and published seven refereed journal articles, conference proceedings, or book chapters.

10. Helped students develop the skills to predict a market's behavior quantitatively. I first became familiar with the set of commercial software available in the market. I then became certified as an instructor user as well as a trainer of instructor users. Following these efforts, I argued successfully for pilot funding, ran pilots, and implemented the simulation throughout my courses as well as several other courses provided by other instructors.

08/2005 - 08/2008
Marist College
School of Computer Science and Mathematics
Poughkeepsie, NY
Assistant Professor of Information Systems
Supervisor: Dr. Roger Norton (845-575-3610)

Summary: Three years of experience teaching, researching, and leading.

1. Information Systems Faculty Coordinator.

2. Library Committee Chair.

3. Led projects at NXP Semiconductors and IBM corporation. Teams developed (for example) databases to support human resource management understanding employees' responsibilities.

4. Taught and applied software design. User interfaces followed usability principles. Pseudo code built upon design patterns. Architecture leveraged customers' infrastructure.

MARIST continued on next page

MARIST continued from page 7

5. Taught and applied software quality assurance. Teams assessed (for example) security, usability, and reliability across units, components, modules, internal/external interfaces, and system.

6. Used UML Use Case, Activity, Class, Sequence, Communication, State, Component diagrams

7. Used IEEE Standards for software quality assurance, quality metrics, test documentation, unit testing, verification and validation, reviews, user documentation, and configuration management.

8. Met with stakeholders to discuss user needs and reach consensus on product needs.

9. Analyzed user needs across diverse stakeholder groups to identify common solutions.

10. Translated broad concepts into specific system requirements to ensure customer needs are met.

11. Identified analytics needs and elicited requirements with customers.

12. Prioritized lists of requested functionality, reports, or data points for solutions.

13. Captured and used data to identify issues with processes.

14. Developed standard data nomenclature, definitions, and valid values for existing data elements.

15. Advocated for and supported data driven decision making.

16. Created and reviewed functional requirements and conducted quality assurance on software.

17. Performed software life cycle management and acceptance testing.

18. Oversaw the design and development of data queries and reports.

19. Collected and analyzed data to understand ethical decision making.

20. Orally presented to students daily.

21. Wrote and published two refereed journal articles and conference papers.

22. Wrote proposal and was awarded National Science Foundation grant.

23. Developed written curricula for students.

24. Led two information systems programs self-review in preparation for Middle States Accreditation.

25. Built with small team the MS in Technology Management

VITA continues on next page

01/1999 - 06/2005
Rensselaer Polytechnic Institute
Lally School of Management and Technology
Troy, NY
Adjunct Instructor/Clinical Assistant Professor of Information Systems
Supervisor: Dr. Joseph Ecker (518-276-6383)

Summary: Six and one half years of experience teaching based on students doing to learn.

1. Taught and used of IEEE standards (e.g., requirements specification) to build (for example) database- driven systems such as 1) an integrated projects binder, 2) an employee proposal development aide, 3) an employee training site locator, 4) an employee training registration system, 5) a human resources training calendar, and 6) a human resources training evaluation system at GE Specialty Materials and MapInfo.

2. Met with stakeholders to discuss user needs and reach consensus on product needs.

3. Analyzed user needs across diverse stakeholder groups to identify common solutions.

4. Identified analytics needs and elicited requirements with customers.

5. Prioritized a list of requested functionality, reports, or data points for BI solutions.

6. Captured and used human resources and other data to identify issues with a process.

7. Provided data that allowed troubleshooting of customer issues with human resources.

8. Oversaw the design and development of human resource data queries and reports.

9. Translated broad concepts into specific system requirements to ensure customer needs are met.

10. Analyzed user needs across diverse stakeholder groups to identify common solutions.

11. Facilitated discussions with customers to identify analytics needs and determine priorities.

12. Advocated for and supported data-driven decision making.

13. Participated on teams to develop software and data solutions.

14. Analyzed user needs to inform system requirements for information technology or services.

15. Led teams that developed software or data solutions.

16. Coordinated with stakeholders to discuss user needs / reach consensus on product development.

17. Created and reviewed functional requirements and conducted quality assurance on software.

18. Performed software life cycle management and acceptance testing.

RENSSELAER continued on next page

RENSSELAER continued from page 9

01/1999 - 06/2005
Rensselaer Polytechnic Institute
Lally School of Management and Technology
Troy, NY
Adjunct Instructor/Clinical Assistant Professor of Information Systems

19. Used IEEE Guide for Developing Software Life Cycle Processes.

20. Used IEEE Standard for Software Project Management Plans.

21. Used IEEE Guide for System Definition-Concept of Operations.

22. Used IEEE Standard - Recommended Practice for Software Requirements Specifications.

23. Used IEEE Guide for Developing System Requirements Specifications.

24. Used IEEE Standard - Recommended Practice for Software Design Descriptions.

01/1998 - 12/2000
Achaean Technology
Watervliet, NY
Co-Founder / Project Manager / Salesperson
Supervisor: Self

Summary: Developed/marketed state-of-the-art and only enterprise-wide database driven software to support operations (including human resources) of agencies providing care to the intellectually disabled.

1. Analyzed user needs across diverse stakeholder groups to identify common solutions.

2. Facilitated discussions with customers to identify analytics needs and determine priorities.

3. Developed standard nomenclature, data definitions, and valid values for new data elements.

4. Advocated for and supported data-driven decision making.

5. Maintained list of planned or requested functionality, reports, and data points for BI solutions.

6. Participated on teams to develop software and data solutions.

7. Analyzed user needs to inform system requirements for information technology or services.

8. Led teams that developed software or data solutions.

9. Coordinated with stakeholders to discuss user needs / reach consensus on product development.

ACHAEAN continued on next page

ACHAEAN continued from page 10

10. Created and reviewed functional requirements.

11. Conducted quality assurance on software.

12. Performed software life cycle management and acceptance testing.

13. Planned and marketed emerging technology solutions.

09/1997 - 12/2000
Rensselaer Polytechnic Institute
Administrative Information Services
Troy, NY
Business Analyst
Supervisor: John Wilder (unknown whereabouts)
Rensselaer Polytechnic Human Resources: (518-276-6302)

Summary: 27 months of experience as systems analyst.

1. Coordinated student records extraction, transformation, and loading from legacy to ERP. 1.4 million person/course units converted. 68,000 student records converted. 44,000 degrees.

2. Developed application providing sample student records for conversion audit.

3. Created application to show registered students' detail given various student characteristics.

4. Created application to show degrees/honors per student given various student characteristics.

5. Developed proof of concept data warehouse. Application provided registration counts, credits, gross and net tuition by many criteria.

6. Replicated relevant production environment in data warehouse.

7. Met with stakeholders to discuss user needs and reach consensus on product needs.

8. Prioritized a list of requested functionality, reports, or data points for BI solutions.

9. Identified analytics needs, elicited requirements, and determined priorities with customers.

10. Oversaw the design and development of data queries and reports about personnel.

11. Analyzed user needs across diverse stakeholder groups to identify common solutions.

12. Designed queries and reports using business intelligence software based on customer needs.

RENSSELAER BUSINESS ANALYST continued on next page

RENSSELAER BUSINESS ANAYST continued from page 11

13. Translated broad concepts into specific system requirements to ensure customer needs are met.

14. Analyzed user needs across diverse stakeholder groups to identify common solutions.

15. Facilitated discussions with customers to identify analytics needs and determine priorities.

16. Developed standard data definitions, and valid values for new and existing data elements.

17. Advocated for and supported data driven decision making.

18. Maintained list of planned or requested functionality, reports, and data points for BI solutions.

19. Participated on teams to develop software and data solutions.

20. Analyzed user needs to inform system requirements for information technology or services.

21. Designed and developed workforce (on faculty usage) data queries and reports using BI solutions.

22. Coordinated with stakeholders to discuss user needs / reach consensus on product development.

23. Created and reviewed functional requirements.

24. Conducted quality assurance on software.

25. Performed acceptance testing.

26. Used Oracle, Brio, Informatica, SCT Banner, and Sequiter, etc.

05/1990 - 05/1994
IBM
Corporate Accounts Payable
Endicott, NY
Clerk
Supervisor: Charles Costantino (unknown whereabouts)
IBM Human Resources: (800 -426-4968)

1. Distributed $3,000,000,000 annually.
2. Analyzed business processes.
3. Proposed business process improvements orally and in writing.
4. Implemented business process improvements.

VITA continued on next page

08/1985 - 08/1987
United States Army
40th Signal Battalion
Tindall and Hatfield
Fort Huachuca, AZ
31K - Radio Operator / Chauffeur
Supervisor: Lieutenant Colonel (at the time) John D. Hartman (unknown whereabouts)
US Army Human Resources: (888-276-9472)

EDUCATION:

Job Related Training:

Data Science Certificate (Complete)
- Johns Hopkins University
- Biostatistics Department and Coursera
- 9 difficult courses + 1 significantly difficult capstone.
- 95.5 / 100

Big Data Certificate (Complete)
- University of California, Berkeley
- Computer Science Department and EdX
- 2 significantly difficult courses with 9 projects
- 94.5 / 100

Business Analytics Certificate / Wharton School / In Process

Machine Learning Certificate / University of Washington / In Process

Courses:

- Customer Analytics (Current)
- Machine Learning Foundations (Current)
- Scalable Machine Learning (August 2015)
- Data Science Capstone (August 2015)
- Introduction to Big Data with Apache Spark (July 2015)
- Cassandra Operations and Performance Tuning (July 2015)
- Cassandra Core Concepts (June 2015)
- MongoDB for Developers (May 2015)
- Try Git (May 2015)
- JavaScript Road Trip Part 2 (April 2015)
- Python Fundamentals (March 2015)
- JavaScript Road Trip Part 1 (February 2015)
- Front End Formations (January 2015)
- Front End Foundations (December 2014)
- Developing Data Products (December 2014)
- Statistical Inference (December 2014)
- Practical Machine Learning (November 2014)
- Reproducible Research (November 2014)
- Regression Models (November 2014)
- Exploratory Data Analysis (June 2014)

- Getting and Cleaning Data (June 2014)
- R Programming (June 2014)
- Data Scientist's Toolbox (June 2014)
- Using R for Data Mining (Summer 2012)
- Using R for Programming and Simulation (Summer 2012)
- Participant Centered Learning Seminar (April 2012)
- SAP ERPsim: Instructor Training Level 1 (August 2011)
- SAP ERPsim: Train the Trainer Training Level 2 (August 2011)
- Introduction to SAP Business ByDesign™ (2011)
- Using R for Generalized Linear and Additive Models (2011)
- Using R for PLS Path Modeling Using R (2011)
- Using R for Statistical Research Analyses II (2011)
- Using R for Statistical Research Analysis I (2011)
- Introduction to SAP ECC 6.0 ERP Using Global Bike Inc. (2010)
- Introduction to SAP ECC 6.0 ERP course (2010)
- Federal Agencies Sponsored Prospective Funding Briefing (2009)
- Understanding Islamic Frameworks in a Global Context Symposium (2009)
- Invited Participant, NSF Sponsored Building an Educational Technology Research Agenda Early Career Symposium (2008)
- EPIC Cognitive Architecture Workshop (2008)
- CLARION Cognitive Architecture Workshop (2008)
- COGNET Cognitive Task Analysis and Modeling Workshop (2003)
- Computational Analysis of Social & Organizational Systems (CASOS) Summer Institute Carnegie Mellon University (2002)
- RePast (Java-based Agents) Workshop, University of Chicago (2002)

Rensselaer Polytechnic Institute
Doctorate 12/2005
GPA: 3.86 of a maximum 4.00
Credits Earned: 91 Semester hours
Major: Engineering Science
Minor: Ethics

1. Performed literature reviews.

2. Collected and analyzed data using observation, surveys, video/audio recording, content analysis, verbal protocol analysis.

3. Built, verified and validated, and used computational model of ethics based upon earlier analysis and experimented using computational model.

4. Coursework included the following:

    a. Calculus (Math Department, School of Science)

    b. Advanced Behavioral Statistics (Psychology Department, School of Science)

    c. Research Methods 2 (Decision Sciences and Engineering Systems Department, School of Engineering)

RENSSELAER DOCTORATE continued on next page

RENSSELAER DOCTORATE continued from page 14

d. Discrete Structures (Computer Science Department, School of Science)

e. Data Structures and Algorithms (Computer Science Department, School of Science)

f. Database Systems (Computer Science Department, School of Science)

g. Decision Support and Expert Systems (Decision Science and Engineering Systems Department, School of Engineering)

h. Software Engineering (Electrical and Computer Systems Department, School of Engineering)

i. Cognition (Psychology Department, School of Science)

j. Cognitive Architecture Development (Psychology Department, School of Science)

k. Statistics and Operations Management (Management Department, School of Management)

l. Business Economics, (Management Department, School of Management)

Rensselaer Polytechnic Institute
Master's Degree 12/2004
GPA: 3.86 of a maximum 4.00
Credits Earned: 60
Major: Information Technology
Relevant Coursework, Licenses and Certifications:

1. Built ethical decision support information system.

2. Evaluated information system using experiment.

3. Results indicated preliminary support for the hypothesis that information technology can be used to aid individuals considering ethical dilemmas.

Binghamton University
Master's Degree 05/1997
GPA: 3.7 of a maximum 4.0
Credits Earned: 70 semester hours
Major: Accounting

1. Coursework included:

a. Auditing 3
b. Auditing 2
c. Auditing 1

BINGHAMTON MASTERS continued on next page

BINGHAMTON MASTERS continued from page 15

d. Legal Environment 2
e. Legal Environment 1
f. (Advanced) Financial Accounting Theory
g. Intermediate Accounting Theory
h. Financial Accounting
i. Managerial Accounting Theory
j. Cost Accounting
k. Statistical Analysis for Management
l. Managerial Finance
m. Financial Management
n. Business Economics
o. Total Quality Management
p. Federal Income Tax 1
q. Computer Tools
o. Management Information Systems
m. Project Management

University of Missouri, Columbia
Columbia, MO
Bachelor's Degree
12/1990
GPA: 3.2 of a maximum 4.0
Credits Earned: 120 semester hours
Major: Finance and Banking

JOURNAL ARTICLES

1. Fleischmann, K.R., Robbins, R.W., and Wallace, W.A. (Winter 2011). "Information Ethics Education for a Multicultural World" Journal of Information Systems Education. Special Issue: Special Issue on Ethics & Social Responsibility 22(3): 191-202.

2. Robbins, R.W. and Butler, B.S. (Summer 2009). "Selecting a Virtual World Platform." Journal of Information Systems Education. Special Issue: Impacts of Web 2.0 and Virtual World Technologies on IS Education 20(2): 199-210.

3. Fleischmann, K.R., Robbins, R.W., and Wallace, W.A. (Jan 2009). "Designing Educational Cases for Intercultural Information Ethics: The Importance of Diversity, Perspectives, Values, and Pluralism." Journal of Education for Library and Information Science 50(1): 4-14.

4. Robbins, R.W., and Wallace, W.A. (August 2007). "Decision Support for Ethical Problem Solving: A Multi-agent Approach." Decision Support Systems 43(4): 1571-1587.

VITA continues on next page

CONFERENCE PUBLICATIONS

5. Robbins, R. W. (August 2014). “Clarifying the SAP ERPsim Experience.” Proc. 2014 Americas Conference on Information Systems. Association for Information Systems. Savannah, GA.

6. Fleischmann, K.R., Robbins, R.W., and Wallace, W.A. (January 2011). “Collaborative Learning of Ethical Decision-Making via Simulated Cases.” Proc 2011 i-Conference. Seattle, WA. Available in ACM Digital Library.

7. Robbins, R.W. and Butler, B.S. (December 2010). “Virtual Teaching Cases? An Exploratory Study.” Proc. 2010 International Conference on Information Systems. Association for Information Systems. Saint Louis, MO.

8. Robbins, R.W. and Butler, B.S. (August 2009). “Teaching and Learning Collaboratively and Virtually Proc. 2009 Americas Conference on Information Systems. Association for Information Systems. San Francisco, CA. Paper No. 655.

9. Robbins, R.W. and Hall, D.J. (August 2007). “Decision Support for Individuals, Groups, and Organizations: Ethics and Values in the Context of Complex Problem Solving.” Proc. 2007 Americas Conference on Information Systems. Association for Information Systems. Keystone, Colorado.

10. Robbins, R.W., Wallace, W.A., and B. Puka, (April 2004). "Supporting Ethical Problem Solving: An Exploratory Investigation." Proc. 2004 ACM SIGMIS CPR, pp. 134-143. ACM Press.

BOOK CHAPTER

11. Robbins, R.W., Fleischmann, K.R., and Wallace, W.A. (2009). “Computing and Information Ethics Education Research.” Handbook of Research on Technoethics. Luppicini, R. and Adell, R. (Eds.). pp. 391-408. Information Science Reference. New York.

DISSERTATION

12. Robbins, R.W. (2005). "Understanding Individual and Group Ethical Problem Solving: A Computational Ethics Approach. Rensselaer Polytechnic Institute.

OTHER ACCEPTED REFEREED MANUSCRIPTS

13. Fleischmann, K.R., Koepfler, J.A., Robbins, R.W., and Wallace, W.A. (October 2011). “CaseBuilder: A GUI Web App for Building Interactive Teaching Cases.” 74th Annual Meeting of the American Society for Information Science and Technology. New Orleans, LA.

14. Robbins, R.W., Wallace, W.A., and Gao, L. (October 2009). “Cognitive Agents for Ethical Problem Solving.” 2009 North American Association for Computational Social and Organization Sciences Annual Conference. Phoenix, AZ.

OTHER ACCEPTED REFEREED MANUSCRIPTS continued on next page

OTHER ACCEPTED REFEREED MANUSCRIPTS continued from page 17

15. Robbins, R.W. and Wallace, W.A. (July 2008). "Understanding Complex Problem Solving: The Case of Ethics Decision Making." CogSci 2008, Washington, DC.

16. Fleischmann, K.R., Robbins, R.W., and Wallace, W.A. (January 2008). "Education Simulation for Information Ethics: Connecting Education with Practice." Association for Library and Information Science Education Annual Conference 2008. Philadelphia.

17. Robbins, R.W. (December 2006). "Towards Developing Descriptive Ethics Theories for Management Science: Using Interdisciplinary Research and Information Systems." 2006 International Federation for Information Processing Working Group 8.2 Organizations and Society in Information Systems Pre-ICIS Workshop, Milwaukee.

18. Robbins, R.W. and Wallace, W.A. (November 2006). "A Computational Model of a Group of Individuals Resolving an Ethical Dilemma: Virtual Experiments." 2006 Institute for Operations Research and the Management Sciences Annual Meeting, Pittsburgh.

19. Robbins, R.W. and Wallace, W.A. (October 2005). "Describing Ethical Problem Solving Dynamically: A Computational Modeling Approach." Ethics: The Guiding Light - The 12th Annual International Conference Promoting Business Ethics, St. Johns University, New York.

20. Robbins, R.W. and Wallace, W.A. (December 2004). "Towards Supporting Ethical Problem Solving in Individuals and Groups." AIS SIGDSS workshop "Expanding the Boundaries for Decision Support Systems" pre-2004 International Conference on Information Systems. Association for Information Systems. Washington, D.C.

GRANTS and AWARDS

Principal Investigator: $300,000, National Science Foundation, Educational Simulation for Computing and Information Ethics. Collaboration with colleagues at University of Maryland College Park and Rensselaer Polytechnic Institute. 2007-2010.

Principal Investigator: $1,200, Experience Based Learning Grant, Joseph M. Graduate School of Business, University of Pittsburgh, August 2011.

Principal Investigator: $11,494, National Science Foundation, Research for Undergraduate Education. May 2011.

Coinvestigator: $23,000, Educational Technology Innovation Grant, The Virtual Firm: An Interactive Environment for Teaching IT Opportunity Recognition. March 2011.

Coinvestigator: $20,000, Collaborative Technology Innovation Grant. 2009-2010.

Finalist: Excellence in Ethics Dissertation Proposal Competition at the University of Notre Dame. 2004.

Senior Personnel: $287,557 as component of $5,000,000 NSF proposal for research ethics education commons. One of two of twenty proposals deemed very competitive; the other was awarded. March 2010. Not funded.

SERVICE TO INSTITUTION

2008-2011: Operational Lead, Virtual Katz 2.0 project
Joseph M. Katz Graduate School of Business
2007-2008: Chair, Library Development Committee, Marist College
2006-2008: Member, Library Development Committee, Marist College
2006-2008: Co-lead, Information Systems Area self-assessment
(for re-accreditation)
2006-2008: Coordinator, Information Systems Area
2006-2007: Member, Information Literacy Teacher Search Committee
2005-2006: Member, Assistant Professor Search Committee
2005: Member, MS in Technology Management Curriculum Committee
School of Computer Science and Mathematics, Marist College
2003-2005: Faculty Intervention Program Mentor, Rensselaer Polytechnic

SERVICE TO COMMUNITY

2012: Faculty Mentor, SAP Student Dashboard Competition
2011-2013: Faculty Residence Hall Mentor, University of Pittsburgh
2011-2013: Kan Jam Faculty Sponsor, University of Pittsburgh
2011-2013: Ascend Faculty Sponsor, University of Pittsburgh
2010: Faculty Mentor, International Project Management Competition (Team won two first prizes in three possible categories.)
2010-2013: Hip Hop Dance Club Adviser, University of Pittsburgh 2001-2004: Board Member, Singles Outreach Services, Inc.

SERVICE TO ACADEMIA AT LARGE

International Conference on Information Systems 2010
Associate Editor, Decision Support and Knowledge Management

International Conference on Information Systems 2010
Associate Editor, IS Philosophy

Academy of Management 2009
Facilitator, Stakeholder Perspectives

Americas Conference on Information Systems
Co-chair, Human Characteristics and Decisions

PROFESSIONAL REFERENCES:

Available upon request.

Section E - Data: Relating to other individuals

This section shares quantitative and qualitative information about how I have related to people. It does so using data from my work over the past fifteen years. My references (Section D) can provide additional uncensored information.

(Please turn the page.)

Please find below summary quantitative information from 54 undergraduate students at Susquehanna University and 101 undergraduate students at the University of Pittsburgh.

Questions and results from Susquehanna University course and instructor evaluations which speak to my ability to relate to individual students follow:

"To what degree did the instructor inspire you to set and achieve goals which really challenged you as a student?"

RESULTS	SECTION 1 (N=19)	SECTION 2 (N=19)	SECTION 3 (N=16)
AVERAGE (1 TO 5)	4	4.3	4.5
% INDICATING 4 OR 5	74%	84%	94%

Table 1: First form of a quantitative "Russ can relate to other individuals" measure.

"To what degree did the instructor display a personal interest in your learning?"

RESULTS	SECTION 1 (N=19)	SECTION 2 (N=19)	SECTION 3 (N=16)
AVERAGE (1 TO 5)	4.7	4.7	4.7
% INDICATING 4 OR 5	100%	95%	100%

Table 2: Second form of quantitative "Russ can relate to other individuals" measure.

The following is a course / instructor evaluation question that may summarize all my interactions with each respondent during a semester at the University of Pittsburgh.

"Would you recommend this course to other students?"

ANSWER	SECTION 1 (N=32)	SECTION 2 (N=35)	SECTION 3 (N=29)
DEFINITELY YES	37.5%	40%	48.3%
PROBABLY YES	53.1%	45.7%	48.3%
PROBABLY NO	9.4%	5.7%	3.4%
DEFINITELY NO	0%	8.6%	0%

Table 3: Third form of a quantitative "Russ can relate to other individuals" measure.

The next two pages share two letters which qualitatively describe my relationships with two working graduate students. One of these students and I interacted when he was a student in a classroom setting. The other student interacted with me when she was a remote student of mine. Both Marist College students.

After those letters I have placed the full evaluations that are the source of Tables 1, 2, and 3 above.

Special Note:
To protect privacy this testimonial has been anonymized.

Date removed.

To Whom It May Concern:

My name is [Name Removed]. I would like to take a moment to tell you about my experience at Marist College during the [date removed] semester as a student in MSIS 647 "Information Analysis." I am a second year MBA student at Western Connecticut State University in Danbury, Connecticut. I chose to take "Information Analysis" at Marist College in order to graduate on time.

"Information Analysis" was the greatest educational experience in my collegiate career to date. This was due to the class professor: Dr. Russ Robbins. Dr. Robbins was the most informative educator I have ever encountered. Information Systems is a field in which I am unfamiliar. When he went over the course outline during the first class meeting, I was nothing short of scared. I was going to quit and graduate later than planned – anything – to avoid his class. However, Dr. Robbins informed me that it would be in my best interest to stay. He assured me that even with no information systems or computer science experience I would be an asset to the class. He indicated that my managerial background would be insightful. And above all, he indicated that if I needed help, he would be there for me. It was clear that Dr. Robbins did something no other professor had: he believed in me!

Well I am glad I stayed. There were times when I struggled. But when the going was tough, Dr. Robbins helped me grasp the concepts at hand. If we did not meet face-to-face, we spoke on the phone, for extended periods. He stopped at nothing to help me succeed. And the end result was astonishing; I now understand how to perform system analysis.

I may never take another course in information systems. And after this course, I say that with regret. However my one experience was worth a lifetime. Dr. Robbins is a magnificent individual who is an asset to his institution. Dr. Robbins helped me re-gain confidence I had lost long ago. He does one thing professors do not do much of any more: he makes sure students learn! If you would like to discuss Dr. Robbins and his teaching please do not hesitate to contact me at [phone number removed.]

Closing and signature removed.

Name removed.

Special Note:
To protect privacy this testimonial has been anonymized.

Name removed.
Address removed.
Address removed.

Date removed.

To Whom It May Concern,

I am a student working toward a Masters Degree in Information Systems Management at Marist College. This semester I had the privilege of taking my sixth course, Information Analysis (MSIS 647) as independent study with Dr. Russ Robbins. I am writing this to thank Dr. Robbins for providing me with an exceptional learning experience. I originally chose Marist College because the on-line course content is the same high quality as courses offered on campus. Because I take courses exclusively on-line, I would not have been able to take a course this semester if Dr. Robbins had not been available. There were no other required IS courses being offered on-line for which I met the prerequisites.

Dr. Robbins made this course a valuable learning experience in many ways. First, he selected a textbook that would challenge me (I have over ten years experience working in IS). Next, I was concerned because I study at home and don't have access to case tools or UML drawing tools. Dr. Robbins arranged for me to use IBM's Rational software. I had never before used case tools so this was a great introduction that I would not have had otherwise. In addition, he arranged for me to use Visio for UML drawings, which allowed me to efficiently create and deliver my project assignments. Dr. Robbins initiated telephone calls to give assignments and homework feedback and actually took time to teach some of the concepts. I had more interaction and feedback from Dr. Robbins in this one course than I've had with all my other instructors combined. I also appreciated Dr. Robbins' overall approach which emphasized learning by doing, which is especially valuable to me with prospective employers. In summary, Dr. Robbins' generous time and effort greatly enhanced learning in this course. I would definitely recommend his course to other students.

Sincerely,

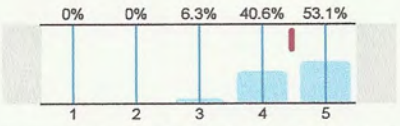
Name and signature removed.

Professor Russell Robbins

INTRO TO INFORMATION SYSTEMS(BUSMIS-1060)13011-2134
Spring 2013
RESPONDENTS = 68% OF NUMBER REGISTERED

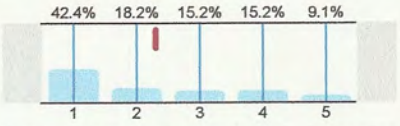
1. TEACHING EVALUATION

1.1) The course was intellectually challenging.



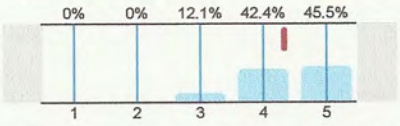
n=32
av.=4.47
dev.=0.62

1.2) The text and materials were helpful in learning concepts and techniques.



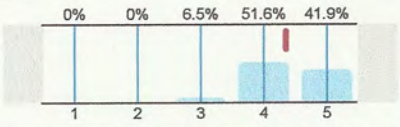
n=33
av.=2.3
dev.=1.4

1.3) The course increased my knowledge.



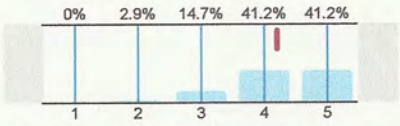
n=33
av.=4.33
dev.=0.69

1.4) The instructor was well prepared for class activities.



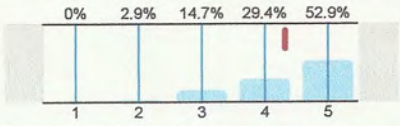
n=31
av.=4.35
dev.=0.61

1.5) The instructor effectively interpreted difficult or abstract ideas.



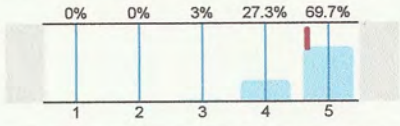
n=34
av.=4.21
dev.=0.81

1.6) The instructor provided appropriate feedback.



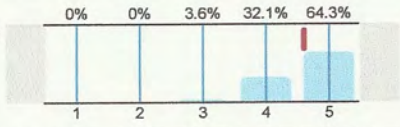
n=34
av.=4.32
dev.=0.84

1.7) The instructor created an atmosphere conducive to learning.



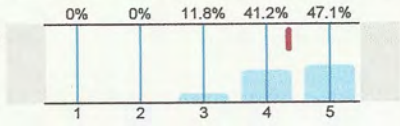
n=33
av.=4.67
dev.=0.54

1.8) The instructor was accessible to students. (Do not answer if no basis to judge)



n=28
av.=4.61
dev.=0.57

1.9) Express your judgment of the instructor's **overall teaching effectiveness:**



n=34
av.=4.35
dev.=0.69

1.10) Would you recommend this instructor to other students?

n=32

Probably not	9.4%
Probably yes	53.1%
Definitely yes	37.5%

ROBBINS, R W

Susquehanna University

Business Information & Data Processing Services 174
MWF 10:00
Fall 2013

IDEA Diagnostic Form Report

To learn more, see the Interpretive Guide: www.theideacenter.org/diagnosticguide.pdf

Of the 21 students enrolled, 19 responded (90%). Feedback from individual classes is always useful to guide improvement efforts. Typically, multiple classes should be used for evaluation, using more classes when they are small (fewer than 10) or when they have low response rates (less than 60%) (see www.theideacenter.org/AdminDecisions).

Summary Evaluation of Teaching Effectiveness

Teaching effectiveness is assessed in two ways: **A. Progress on Relevant Objectives**, a weighted average of student ratings of the progress they reported on objectives selected as "Important" or "Essential" (double weighted) and **B. Overall Ratings**, the average student agreement with statements that the teacher and the course were excellent. The **SUMMARY EVALUATION** is the average of these two measures. Individual institutions may prefer to combine these measures in some other manner to arrive at a summary judgment.

Converted Averages are standardized scores that take into account the fact that the average ratings for items on the IDEA form are not equal; students report more progress on some objectives than on others. Converted scores all have the same average (50) and the same variability (a standard deviation of 10); about 40% of them will be between 45 and 55. Because measures are not perfectly reliable, it is best to regard the "true score" as lying within plus or minus 3 of the reported score.

For comparative purposes, use converted averages. Your converted averages are compared with those from all classes in the IDEA database. If enough classes are available, comparisons are also made with classes in the same broad *discipline* as this class and/or with all classes that used IDEA at your *institution*. The *Interpretive Guide* offers some suggestions for using comparative results; **some institutions may prefer to establish their own "standards" based on raw or adjusted scores rather than on comparative standing**.

Both unadjusted (raw) and adjusted averages are reported. The latter makes classes more comparable by considering factors that influence student ratings, yet are beyond the instructor's control. Scores are adjusted to take into account student desire to take the course regardless of who taught it (item 39), student work habits (item 43), instructor reported class size, and two multiple item measures (student effort not attributable to the instructor and course difficulty not attributable to the instructor).

Your Average Scores

	Your Average (5–point scale)	
	Raw	Adj.
A. Progress on Relevant Objectives [1] **Four** objectives were selected as relevant (Important or Essential –see page 2)	4.2	4.4
Overall Ratings		
B. Excellent Teacher	3.9	4.2
C. Excellent Course	3.6	4.1
D. Average of B & C	3.8	4.2
Summary Evaluation (Average of A & D) [1]	**4.0**	**4.3**

[1] If you are comparing Progress on Relevant Objectives from one instructor to another, use the converted average.

[2] The process for computing Progress on Relevant Objectives for the Discipline and Institution was modified on May 1, 2006. Do not compare these results with reports generated prior to this date.

Your Converted Average When Compared to All Classes in the IDEA Database

Comparison Category	A. Progress on Relevant Objectives		Overall Ratings: B. Excellent Teacher		Overall Ratings: C. Excellent Course		Overall Ratings: D. Average of B & C		Summary Evaluation (Average of A & D)	
	Raw	Adj.	Raw	Adj.	Raw	Adj.	Raw	Adj.	Raw	Adj.
Much Higher Highest 10% (63 or higher)										
Higher Next 20% (56–62)		59								56
Similar Middle 40% (45–55)	55		46	50		53	45	52	50	
Lower Next 20% (38–44)					44					
Much Lower Lowest 10% (37 or lower)										
Your Converted Average When Compared to Your:[2]										
Discipline (IDEA Data)	55	62	47	52	43	55	45	54	50	58
Institution	51	60	45	51	42	57	44	54	48	57

IDEA Discipline used for comparison:
Business Information & Data Processing Services

Student Ratings of Learning on Relevant (Important and Essential) Objectives

Average unadjusted (raw) and adjusted progress ratings are shown below for those objectives you identified as "Important" or "Essential." **Progress on Relevant Objectives** (also shown on page 1) is a weighted average of student ratings of the progress they reported on objectives selected as "Important" or "Essential" (double weighted). The percent of students rating each as "1" or "2" (either "no" or "slight" progress) and as "4" or "5" ("substantial" or "exceptional" progress) is also reported. These results should help you identify objectives where improvement efforts might best be focused. Page 3 contains suggestions about the types of changes you might consider to obtain more satisfactory results. Also, refer to the **POD–IDEA Center *Learning* Notes** (www.theideacenter.org/podidea/PODNotesLearning.html).

	Importance Rating	Your Average (5–point scale)		Percent of Students Rating		Your Converted Average When Compared to Group Averages					
						IDEA Database		IDEA Discipline[1]		Your Institution[1]	
		Raw	Adj.	1 or 2	4 or 5	Raw	Adjusted	Raw	Adjusted	Raw	Adjusted
21. Gaining factual knowledge (terminology, classifications, methods, trends)	Minor/None										
22. Learning fundamental principles, generalizations, or theories	Important	4.2	4.2	5%	79%	54 Similar	56 Higher	53 Similar	57 Higher	49 Similar	55 Similar
23. Learning to *apply* course material (to improve thinking, problem solving, and decisions)	**Essential**	**4.3**	**4.5**	**5%**	**84%**	**55 Similar**	**61 Higher**	**54 Similar**	**62 Higher**	**52 Similar**	**63 Much Higher**
24. Developing specific skills, competencies, and points of view needed by professionals in the field most closely related to this course	Minor/None										
25. Acquiring skills in working with others as a member of a team	Important	4.3	4.7	0%	84%	56 Higher	62 Higher	58 Higher	66 Much Higher	54 Similar	63 Much Higher
26. Developing creative capacities (writing, inventing, designing, performing in art, music, drama, etc.)	Minor/None										
27. Gaining a broader understanding and appreciation of intellectual/cultural activity (music, science, literature, etc.)	Minor/None										
28. Developing skill in expressing myself orally or in writing	Minor/None										
29. Learning how to find and use resources for answering questions or solving problems	Minor/None										
30. Developing a clearer understanding of, and commitment to, personal values	Minor/None										
31. Learning to *analyze* and *critically evaluate* ideas, arguments, and points of view	Important	4.1	4.2	5%	74%	53 Similar	56 Higher	57 Higher	64 Much Higher	50 Similar	56 Higher
32. Acquiring an interest in learning more by asking my own questions and seeking answers	Minor/None										
Progress on Relevant Objectives		**4.2**	**4.4**			**55**	**59**	**55**	**62**	**51**	**60**

[1] The process for computing Progress on Relevant Objectives for the Discipline and Institution was modified on May 1, 2006. Do not compare these results with reports generated prior to this date.

Much Higher = Highest 10% of classes (63 or higher)
Higher = Next 20% (56–62)
Similar = Middle 40% (45–55)
Lower = Next 20% (38–44)
Much Lower = Lowest 10% (37 or lower)

Description of Course and Students

Students described the course by rating three items related to "level of academic challenge." Results cannot be interpreted as "good" or "bad"; in general, these ratings have a slight positive relationship with measures of academic achievement. The three items describing your students relate to their academic motivation and work habits and are key factors in developing adjusted ratings.

Course Description	Your Average (5–point scale)	Your Converted Average When Compared to Group Averages					
		IDEA Database		IDEA Discipline		Your Institution	
33. Amount of reading	2.9	46	Similar	47	Similar	45	Similar
34. Amount of work in other (non–reading) assignments	3.9	58	Higher	57	Higher	54	Similar
35. Difficulty of subject matter	4.3	64	Much Higher	69	Much Higher	60	Higher
Student Description							
37. I worked harder on this course than on most courses I have taken.	3.6	51	Similar	54	Similar	48	Similar
39. I really wanted to take this course regardless of who taught it.	2.4	34	Much Lower	30	Much Lower	29	Much Lower
43. As a rule, I put forth more effort than other students on academic work.	3.8	56	Higher	49	Similar	45	Similar

Much Higher = Highest 10% of classes (63 or higher)
Higher = Next 20% (56–62)
Similar = Middle 40% (45–55)
Lower = Next 20% (38–44)
Much Lower = Lowest 10% (37 or lower)

Improving Teaching Effectiveness

One way to improve teaching effectiveness is to make more use of the teaching methods closely related to learning on specific objectives.

- Review page 2 to identify the objective(s) where improvements are most desirable.
- Use the first column to answer the question, "Which of the 20 teaching methods are most related to learning on these objective(s)?"
- Review the next two columns to answer the question, "How did students rate my use of these important methods?"
- Read the last column to answer the question, "What changes should I consider in my teaching methods?"
- Beyond specific methods, do the results suggest a general area (e.g., Stimulating Student Interest) where improvement efforts should be focused?

Suggested Actions are based on comparisons with ratings for classes of similar size and level of student motivation. **Consider increasing use** means you employed the method less frequently than those teaching similar classes. **Retain current use or consider increasing** means you employed the method with typical frequency. **Strength to retain** means you employed the method more frequently than those teaching similar classes. More detailed suggestions are in the **Interpretive Guide** (**www.theideacenter.org/diagnosticguide.pdf**), **POD–IDEA Center Notes** (**www.theideacenter.org/podidea**), and **POD–IDEA Center *Learning* Notes** (**www.theideacenter.org/podidea/PODNotesLearning.html**).

Teaching Methods and Styles

	Relevant to Objectives: (see page 2)	Your Average (5–point scale)	Percent of Students Rating 4 or 5	Suggested Action
Stimulating Student Interest				
8. Stimulated students to intellectual effort beyond that required by most courses	**All selected objectives**	**4.2**	**74%**	**Strength to retain**
15. Inspired students to set and achieve goals which really challenged them	**All selected objectives**	**4.0**	**74%**	**Strength to retain**
13. Introduced stimulating ideas about the subject	**22, 23, 31**	**4.2**	**78%**	**Strength to retain**
4. Demonstrated the importance and significance of the subject matter	**22, 23**	**4.4**	**95%**	**Strength to retain**
Fostering Student Collaboration				
16. Asked students to share ideas and experiences with others whose backgrounds and viewpoints differ from their own	**25, 31**	**4.1**	**79%**	**Strength to retain**
18. Asked students to help each other understand ideas or concepts	**25, 31**	**4.4**	**95%**	**Strength to retain**
5. Formed "teams" or "discussion groups" to facilitate learning	**25**	**4.5**	**84%**	**Strength to retain**
Establishing Rapport				
2. Found ways to help students answer their own questions	All selected objectives	4.1	74%	Retain current use or consider increasing
7. Explained the reasons for criticisms of students' academic performance	**23, 31**	**4.2**	**84%**	**Strength to retain**
1. Displayed a personal interest in students and their learning	**23**	**4.7**	**100%**	**Strength to retain**
20. Encouraged student–faculty interaction outside of class (office visits, phone calls, e–mails, etc.)	Not relevant to objectives selected	4.6	95%	
Encouraging Student Involvement				
11. Related course material to real life situations	23	3.9	74%	Retain current use or consider increasing
19. Gave projects, tests, or assignments that required original or creative thinking	**25, 31**	**4.4**	**89%**	**Strength to retain**
14. Involved students in "hands on" projects such as research, case studies, or "real life" activities	**25**	**4.5**	**84%**	**Strength to retain**
9. Encouraged students to use multiple resources (e.g. data banks, library holdings, outside experts) to improve understanding	Not relevant to objectives selected	3.7	58%	
Structuring Classroom Experiences				
6. Made it clear how each topic fit into the course	22, 23	4.1	74%	Retain current use or consider increasing
10. Explained course material clearly and concisely	22, 23	3.7	58%	Retain current use or consider increasing
12. Gave tests, projects, etc. that covered the most important points of the course	22	4.3	84%	Retain current use or consider increasing
3. Scheduled course work (class activities, tests, projects) in ways which encouraged students to stay up–to–date in their work	Not relevant to objectives selected	4.3	84%	
17. Provided timely and frequent feedback on tests, reports, projects, etc. to help students improve	Not relevant to objectives selected	4.5	84%	

5–point Scale: 1 = Hardly Ever 2 = Occasionally 3 = Sometimes 4 = Frequently 5 = Almost Always

Statistical Detail

	Number Responding							
	1	2	3	4	5	Omit	Avg.	s.d.
1. Displayed a personal interest in students and their learning	0	0	0	6	13	0	4.7	0.5
2. Found ways to help students answer their own questions	0	2	3	6	8	0	4.1	1.0
3. Scheduled course work (class activities, tests, projects) in ways...	0	0	3	7	9	0	4.3	0.7
4. Demonstrated the importance and significance of the subject matter	0	1	0	8	10	0	4.4	0.8
5. Formed "teams" or "discussion groups" to facilitate learning	0	0	3	4	12	0	4.5	0.8
6. Made it clear how each topic fit into the course	1	0	4	6	8	0	4.1	1.1
7. Explained the reasons for criticisms of students' academic...	1	0	2	8	8	0	4.2	1.0
8. Stimulated students to intellectual effort beyond that required by...	0	1	4	4	10	0	4.2	1.0
9. Encouraged students to use multiple resources (e.g. data banks,...	1	1	6	5	6	0	3.7	1.1
10. Explained course material clearly and concisely	1	2	5	5	6	0	3.7	1.2
11. Related course material to real life situations	1	2	2	7	7	0	3.9	1.2
12. Gave tests, projects, etc. that covered the most important points...	0	1	2	7	9	0	4.3	0.9
13. Introduced stimulating ideas about the subject	0	0	4	7	7	1	4.2	0.8
14. Involved students in "hands on" projects such as research, case...	0	0	3	3	13	0	4.5	0.8
15. Inspired students to set and achieve goals which really...	0	2	3	8	6	0	3.9	1.0
16. Asked students to share ideas and experiences with others...	0	1	3	9	6	0	4.1	0.8
17. Provided timely and frequent feedback on tests, reports,...	0	1	2	3	13	0	4.5	0.9
18. Asked students to help each other understand ideas or concepts	0	0	1	9	9	0	4.4	0.6
19. Gave projects, tests, or assignments that required original or...	0	1	1	6	11	0	4.4	0.8
20. Encouraged student–faculty interaction outside of class (office...	0	0	1	5	13	0	4.6	0.6

Key: 1 = Hardly Ever 2 = Occasionally 3 = Sometimes 4 = Frequently 5 = Almost Always

The details on this page are of interest primarily to those who want to confirm scores reported on pages 1–3 or who want to determine if responses to some items were distributed in an unusual manner.

Converted Averages are reported only for relevant learning objectives (Important or Essential –see page 2) and other items for which comparisons were provided.

Notes:
Discipline code selected on FIF: 5212
Discipline code used for comparison: 5212

	Number Responding								Converted Avg.		Comparison Group Average		
	1	2	3	4	5	Omit	Avg.	s.d.	Raw	Adj.	IDEA	Discipline	Institution
21. Gaining factual knowledge (terminology, classifications,...	1	0	3	6	9	0	4.2	1.1	NA	NA	4.0	4.1	4.3
22. Learning fundamental principles, generalizations, or...	**1**	**0**	**3**	**6**	**9**	**0**	**4.2**	**1.1**	**54**	**56**	**3.9**	**4.0**	**4.2**
23. Learning to *apply* course material (to improve thinking,...	**1**	**0**	**2**	**6**	**10**	**0**	**4.3**	**1.0**	**55**	**61**	**4.0**	**4.1**	**4.2**
24. Developing specific skills, competencies, and points of view...	0	2	2	7	8	0	4.1	1.0	NA	NA	4.0	4.1	4.2
25. Acquiring skills in working with others as a member of...	**0**	**0**	**3**	**7**	**9**	**0**	**4.3**	**0.7**	**56**	**62**	**3.9**	**3.9**	**4.1**
26. Developing creative capacities (writing, inventing, designing,...	2	1	3	6	7	0	3.8	1.3	NA	NA	3.9	3.5	4.2
27. Gaining a broader understanding and appreciation of...	2	2	0	9	6	0	3.8	1.3	NA	NA	3.7	3.4	4.1
28. Developing skill in expressing myself orally or in writing	2	2	1	5	9	0	3.9	1.4	NA	NA	3.8	3.5	4.0
29. Learning how to find and use resources for answering questions...	2	0	2	7	8	0	4.0	1.2	NA	NA	3.7	3.9	3.9
30. Developing a clearer understanding of, and commitment to,...	2	2	0	5	10	0	4.0	1.4	NA	NA	3.8	3.6	3.9
31. Learning to *analyze* and *critically evaluate* ideas,...	**1**	**0**	**4**	**6**	**8**	**0**	**4.1**	**1.1**	**53**	**56**	**3.8**	**3.7**	**4.0**
32. Acquiring an interest in learning more by asking my own...	1	0	2	5	11	0	4.3	1.1	NA	NA	3.8	3.8	3.9

Key: 1 = No apparent progress 2 = Slight progress 3 = Moderate progress 4 = Substantial progress 5 = Exceptional progress Bold = Selected as Important or Essential

33. Amount of reading	2	3	9	5	0	0	2.9	0.9	46	NA	3.2	3.1	3.3
34. Amount of work in other (non–reading) assignments	1	0	6	5	7	0	3.9	1.1	58	NA	3.4	3.6	3.6
35. Difficulty of subject matter	1	0	1	8	9	0	4.3	1.0	64	NA	3.4	3.3	3.6

Key: 1 = Much Less than Most 2 = Less than Most 3 = About Average 4 = More than Most 5 = Much More than Most

36. I had a strong desire to take this course.	5	2	8	3	1	0	2.6	1.2	NA	NA	3.7	3.5	3.7
37. I worked harder on this course than on most courses I have taken.	0	0	8	10	1	0	3.6	0.6	51	NA	3.6	3.5	3.7
38. I really wanted to take a course from this instructor.	2	3	9	5	0	0	2.9	0.9	NA	NA	3.4	3.4	3.6
39. I really wanted to take this course regardless of who taught it.	3	5	11	0	0	0	2.4	0.8	34	NA	3.3	3.3	3.7
40. As a result of taking this course, I have more positive feelings...	2	1	7	7	2	0	3.3	1.1	41	50	3.9	3.8	4.0
41. Overall, I rate this instructor an excellent teacher.	0	1	6	5	7	0	3.9	1.0	46	50	4.2	4.1	4.3
42. Overall, I rate this course as excellent.	1	2	5	6	4	1	3.6	1.1	44	53	3.9	3.9	4.0
43. As a rule, I put forth more effort than other students on...	0	0	5	12	2	0	3.8	0.6	56	NA	3.6	3.9	4.0

Key: 1 = Definitely False 2 = More False than True 3 = In Between 4 = More True than False 5 = Definitely True

No Additional Questions.

ROBBINS, R W

Susquehanna University

Business Information & Data Processing Services 174
MWF 1:45
Fall 2013

IDEA Diagnostic Form Report

To learn more, see the Interpretive Guide: www.theideacenter.org/diagnosticguide.pdf

Of the 23 students enrolled, 19 responded (83%). Feedback from individual classes is always useful to guide improvement efforts. Typically, multiple classes should be used for evaluation, using more classes when they are small (fewer than 10) or when they have low response rates (less than 60%) (see www.theideacenter.org/AdminDecisions).

Summary Evaluation of Teaching Effectiveness

Teaching effectiveness is assessed in two ways: **A. Progress on Relevant Objectives**, a weighted average of student ratings of the progress they reported on objectives selected as "Important" or "Essential" (double weighted) and **B. Overall Ratings**, the average student agreement with statements that the teacher and the course were excellent. The **SUMMARY EVALUATION** is the average of these two measures. Individual institutions may prefer to combine these measures in some other manner to arrive at a summary judgment.

Converted Averages are standardized scores that take into account the fact that the average ratings for items on the IDEA form are not equal; students report more progress on some objectives than on others. Converted scores all have the same average (50) and the same variability (a standard deviation of 10); about 40% of them will be between 45 and 55. Because measures are not perfectly reliable, it is best to regard the "true score" as lying within plus or minus 3 of the reported score.

For comparative purposes, use converted averages. Your converted averages are compared with those from all classes in the IDEA database. If enough classes are available, comparisons are also made with classes in the same broad *discipline* as this class and/or with all classes that used IDEA at your *institution*. The *Interpretive Guide* offers some suggestions for using comparative results; **some institutions may prefer to establish their own "standards" based on raw or adjusted scores rather than on comparative standing**.

Both unadjusted (raw) and adjusted averages are reported. The latter makes classes more comparable by considering factors that influence student ratings, yet are beyond the instructor's control. Scores are adjusted to take into account student desire to take the course regardless of who taught it (item 39), student work habits (item 43), instructor reported class size, and two multiple item measures (student effort not attributable to the instructor and course difficulty not attributable to the instructor).

Your Average Scores

	Your Average (5-point scale)	
	Raw	Adj.
A. Progress on Relevant Objectives [1] **Four** objectives were selected as relevant (Important or Essential –see page 2)	4.0	3.9
Overall Ratings		
B. Excellent Teacher	4.1	4.1
C. Excellent Course	3.7	3.8
D. Average of B & C	3.9	3.9
Summary Evaluation (Average of A & D) [1]	**4.0**	**3.9**

[1] If you are comparing Progress on Relevant Objectives from one instructor to another, use the converted average.

[2] The process for computing Progress on Relevant Objectives for the Discipline and Institution was modified on May 1, 2006. Do not compare these results with reports generated prior to this date.

Your Converted Average When Compared to All Classes in the IDEA Database

Comparison Category	A. Progress on Relevant Objectives		Overall Ratings: B. Excellent Teacher		Overall Ratings: C. Excellent Course		Overall Ratings: D. Average of B & C		Summary Evaluation (Average of A & D)	
	Raw	Adj.	Raw	Adj.	Raw	Adj.	Raw	Adj.	Raw	Adj.
Much Higher Highest 10% (63 or higher)										
Higher Next 20% (56–62)										
Similar Middle 40% (45–55)	51	49	49	48	46	48	48	48	50	49
Lower Next 20% (38–44)										
Much Lower Lowest 10% (37 or lower)										
Your Converted Average When Compared to Your: [2]										
Discipline (IDEA Data)	50	51	50	51	46	50	48	51	49	51
Institution	47	49	48	50	44	52	46	51	47	50

IDEA Discipline used for comparison:
Business Information & Data Processing Services

Student Ratings of Learning on Relevant (Important and Essential) Objectives

Average unadjusted (raw) and adjusted progress ratings are shown below for those objectives you identified as "Important" or "Essential." **Progress on Relevant Objectives** (also shown on page 1) is a weighted average of student ratings of the progress they reported on objectives selected as "Important" or "Essential" (double weighted). The percent of students rating each as "1" or "2" (either "no" or "slight" progress) and as "4" or "5" ("substantial" or "exceptional" progress) is also reported. These results should help you identify objectives where improvement efforts might best be focused. Page 3 contains suggestions about the types of changes you might consider to obtain more satisfactory results. Also, refer to the **POD–IDEA Center *Learning* Notes** (www.theideacenter.org/podidea/PODNotesLearning.html).

	Importance Rating	Your Average (5–point scale)		Percent of Students Rating		Your Converted Average When Compared to Group Averages					
						IDEA Database		IDEA Discipline[1]		Your Institution[1]	
		Raw	Adj.	1 or 2	4 or 5	Raw	Adjusted	Raw	Adjusted	Raw	Adjusted
21. Gaining factual knowledge (terminology, classifications, methods, trends)	Minor/None										
22. Learning fundamental principles, generalizations, or theories	Important	4.0	3.8	0%	68%	51 Similar	48 Similar	49 Similar	48 Similar	45 Similar	46 Similar
23. Learning to *apply* course material (to improve thinking, problem solving, and decisions)	**Essential**	**3.9**	**3.8**	**11%**	**74%**	**49 Similar**	**47 Similar**	**47 Similar**	**47 Similar**	**46 Similar**	**49 Similar**
24. Developing specific skills, competencies, and points of view needed by professionals in the field most closely related to this course	Minor/None										
25. Acquiring skills in working with others as a member of a team	Important	4.3	4.3	6%	89%	56 Higher	56 Higher	57 Higher	59 Higher	53 Similar	57 Higher
26. Developing creative capacities (writing, inventing, designing, performing in art, music, drama, etc.)	Minor/None										
27. Gaining a broader understanding and appreciation of intellectual/cultural activity (music, science, literature, etc.)	Minor/None										
28. Developing skill in expressing myself orally or in writing	Minor/None										
29. Learning how to find and use resources for answering questions or solving problems	Minor/None										
30. Developing a clearer understanding of, and commitment to, personal values	Minor/None										
31. Learning to *analyze* and *critically evaluate* ideas, arguments, and points of view	Important	3.8	3.7	6%	61%	49 Similar	47 Similar	52 Similar	53 Similar	45 Similar	46 Similar
32. Acquiring an interest in learning more by asking my own questions and seeking answers	Minor/None										
Progress on Relevant Objectives		**4.0**	**3.9**			**51**	**49**	**50**	**51**	**47**	**49**

[1] The process for computing Progress on Relevant Objectives for the Discipline and Institution was modified on May 1, 2006. Do not compare these results with reports generated prior to this date.

Much Higher = Highest 10% of classes (63 or higher)
Higher = Next 20% (56–62)
Similar = Middle 40% (45–55)
Lower = Next 20% (38–44)
Much Lower = Lowest 10% (37 or lower)

Description of Course and Students

Students described the course by rating three items related to "level of academic challenge." Results cannot be interpreted as "good" or "bad"; in general, these ratings have a slight positive relationship with measures of academic achievement. The three items describing your students relate to their academic motivation and work habits and are key factors in developing adjusted ratings.

Course Description	Your Average (5–point scale)	Your Converted Average When Compared to Group Averages					
		IDEA Database		IDEA Discipline		Your Institution	
33. Amount of reading	2.9	47	Similar	48	Similar	46	Similar
34. Amount of work in other (non–reading) assignments	4.0	60	Higher	59	Higher	56	Higher
35. Difficulty of subject matter	4.1	61	Higher	65	Much Higher	57	Higher
Student Description							
37. I worked harder on this course than on most courses I have taken.	3.6	51	Similar	54	Similar	48	Similar
39. I really wanted to take this course regardless of who taught it.	2.9	43	Lower	43	Lower	38	Lower
43. As a rule, I put forth more effort than other students on academic work.	4.1	63	Much Higher	57	Higher	52	Similar

Much Higher = Highest 10% of classes (63 or higher)
Higher = Next 20% (56–62)
Similar = Middle 40% (45–55)
Lower = Next 20% (38–44)
Much Lower = Lowest 10% (37 or lower)

Improving Teaching Effectiveness

One way to improve teaching effectiveness is to make more use of the teaching methods closely related to learning on specific objectives.

- Review page 2 to identify the objective(s) where improvements are most desirable.
- Use the first column to answer the question, "Which of the 20 teaching methods are most related to learning on these objective(s)?"
- Review the next two columns to answer the question, "How did students rate my use of these important methods?"
- Read the last column to answer the question, "What changes should I consider in my teaching methods?"
- Beyond specific methods, do the results suggest a general area (e.g., Stimulating Student Interest) where improvement efforts should be focused?

Suggested Actions are based on comparisons with ratings for classes of similar size and level of student motivation. **Consider increasing use** means you employed the method less frequently than those teaching similar classes. **Retain current use or consider increasing** means you employed the method with typical frequency. **Strength to retain** means you employed the method more frequently than those teaching similar classes. More detailed suggestions are in the **Interpretive Guide** (**www.theideacenter.org/diagnosticguide.pdf**), **POD–IDEA Center Notes** (**www.theideacenter.org/podidea**), and **POD–IDEA Center *Learning* Notes** (**www.theideacenter.org/podidea/PODNotesLearning.html**).

Teaching Methods and Styles

	Relevant to Objectives: (see page 2)	Your Average (5–point scale)	Percent of Students Rating 4 or 5	Suggested Action
Stimulating Student Interest				
13. Introduced stimulating ideas about the subject	**22, 23, 31**	**3.6**	**58%**	**Consider increasing use**
8. Stimulated students to intellectual effort beyond that required by most courses	**All selected objectives**	**4.3**	**84%**	**Strength to retain**
15. Inspired students to set and achieve goals which really challenged them	**All selected objectives**	**4.0**	**74%**	**Strength to retain**
4. Demonstrated the importance and significance of the subject matter	**22, 23**	**4.4**	**89%**	**Strength to retain**
Fostering Student Collaboration				
16. Asked students to share ideas and experiences with others whose backgrounds and viewpoints differ from their own	25, 31	3.6	53%	Retain current use or consider increasing
18. Asked students to help each other understand ideas or concepts	**25, 31**	**4.1**	**79%**	**Strength to retain**
5. Formed "teams" or "discussion groups" to facilitate learning	**25**	**4.5**	**89%**	**Strength to retain**
Establishing Rapport				
7. Explained the reasons for criticisms of students' academic performance	23, 31	4.0	74%	Retain current use or consider increasing
2. Found ways to help students answer their own questions	**All selected objectives**	**4.3**	**89%**	**Strength to retain**
1. Displayed a personal interest in students and their learning	**23**	**4.7**	**95%**	**Strength to retain**
20. Encouraged student–faculty interaction outside of class (office visits, phone calls, e–mails, etc.)	Not relevant to objectives selected	4.6	95%	
Encouraging Student Involvement				
11. Related course material to real life situations	23	4.1	84%	Retain current use or consider increasing
19. Gave projects, tests, or assignments that required original or creative thinking	**25, 31**	**4.4**	**84%**	**Strength to retain**
14. Involved students in "hands on" projects such as research, case studies, or "real life" activities	**25**	**4.3**	**84%**	**Strength to retain**
9. Encouraged students to use multiple resources (e.g. data banks, library holdings, outside experts) to improve understanding	Not relevant to objectives selected	3.4	50%	
Structuring Classroom Experiences				
10. Explained course material clearly and concisely	**22, 23**	**3.5**	**53%**	**Consider increasing use**
6. Made it clear how each topic fit into the course	22, 23	3.9	72%	Retain current use or consider increasing
12. Gave tests, projects, etc. that covered the most important points of the course	**22**	**4.4**	**89%**	**Strength to retain**
3. Scheduled course work (class activities, tests, projects) in ways which encouraged students to stay up–to–date in their work	Not relevant to objectives selected	4.5	89%	
17. Provided timely and frequent feedback on tests, reports, projects, etc. to help students improve	Not relevant to objectives selected	3.9	72%	

5–point Scale: 1 = Hardly Ever 2 = Occasionally 3 = Sometimes 4 = Frequently 5 = Almost Always

Statistical Detail

	Number Responding							
	1	2	3	4	5	Omit	Avg.	s.d.
1. Displayed a personal interest in students and their learning	0	0	1	3	15	0	4.7	0.6
2. Found ways to help students answer their own questions	0	1	1	9	8	0	4.3	0.8
3. Scheduled course work (class activities, tests, projects) in ways...	0	1	1	5	12	0	4.5	0.8
4. Demonstrated the importance and significance of the subject matter	0	1	1	7	10	0	4.4	0.8
5. Formed "teams" or "discussion groups" to facilitate learning	0	1	1	4	13	0	4.5	0.8
6. Made it clear how each topic fit into the course	0	2	3	7	6	1	3.9	1.0
7. Explained the reasons for criticisms of students' academic...	0	1	4	8	6	0	4.0	0.9
8. Stimulated students to intellectual effort beyond that required by...	0	1	2	7	9	0	4.3	0.9
9. Encouraged students to use multiple resources (e.g. data banks,...	3	2	4	2	7	1	3.4	1.5
10. Explained course material clearly and concisely	0	2	7	8	2	0	3.5	0.8
11. Related course material to real life situations	0	2	1	9	7	0	4.1	0.9
12. Gave tests, projects, etc. that covered the most important points...	0	1	1	7	10	0	4.4	0.8
13. Introduced stimulating ideas about the subject	0	3	5	8	3	0	3.6	1.0
14. Involved students in "hands on" projects such as research, case...	0	1	2	7	9	0	4.3	0.9
15. Inspired students to set and achieve goals which really...	1	1	3	6	8	0	4.0	1.2
16. Asked students to share ideas and experiences with others...	0	2	7	6	4	0	3.6	1.0
17. Provided timely and frequent feedback on tests, reports,...	0	2	3	7	6	1	3.9	1.0
18. Asked students to help each other understand ideas or concepts	0	1	3	8	7	0	4.1	0.9
19. Gave projects, tests, or assignments that required original or...	0	1	2	4	12	0	4.4	0.9
20. Encouraged student–faculty interaction outside of class (office...	0	1	0	5	13	0	4.6	0.8

Key: 1 = Hardly Ever 2 = Occasionally 3 = Sometimes 4 = Frequently 5 = Almost Always

The details on this page are of interest primarily to those who want to confirm scores reported on pages 1–3 or who want to determine if responses to some items were distributed in an unusual manner.

Converted Averages are reported only for relevant learning objectives (Important or Essential –see page 2) and other items for which comparisons were provided.

Notes:
Discipline code selected on FIF: 5212
Discipline code used for comparison: 5212

									Converted Avg.		Comparison Group Average		
									Raw	Adj.	IDEA	Discipline	Institution
21. Gaining factual knowledge (terminology, classifications,...	0	0	3	10	6	0	4.2	0.7	NA	NA	4.0	4.1	4.3
22. Learning fundamental principles, generalizations, or...	**0**	**0**	**6**	**7**	**6**	**0**	**4.0**	**0.8**	**51**	**48**	**3.9**	**4.0**	**4.2**
23. Learning to *apply* course material (to improve thinking,...	**0**	**2**	**3**	**8**	**6**	**0**	**3.9**	**1.0**	**49**	**47**	**4.0**	**4.1**	**4.2**
24. Developing specific skills, competencies, and points of view...	0	0	6	6	7	0	4.1	0.8	NA	NA	4.0	4.1	4.2
25. Acquiring skills in working with others as a member of...	**0**	**1**	**1**	**8**	**8**	**1**	**4.3**	**0.8**	**56**	**56**	**3.9**	**3.9**	**4.1**
26. Developing creative capacities (writing, inventing, designing,...	1	2	3	8	5	0	3.7	1.1	NA	NA	3.9	3.5	4.2
27. Gaining a broader understanding and appreciation of...	2	6	2	5	4	0	3.2	1.4	NA	NA	3.7	3.4	4.1
28. Developing skill in expressing myself orally or in writing	2	1	6	8	2	0	3.4	1.1	NA	NA	3.8	3.5	4.0
29. Learning how to find and use resources for answering questions...	0	4	8	4	3	0	3.3	1.0	NA	NA	3.7	3.9	3.9
30. Developing a clearer understanding of, and commitment to,...	3	3	4	5	3	1	3.1	1.4	NA	NA	3.8	3.6	3.9
31. Learning to *analyze* and *critically evaluate* ideas,...	**0**	**1**	**6**	**7**	**4**	**1**	**3.8**	**0.9**	**49**	**47**	**3.8**	**3.7**	**4.0**
32. Acquiring an interest in learning more by asking my own...	0	2	4	6	7	0	3.9	1.0	NA	NA	3.8	3.8	3.9

Key: 1 = No apparent progress 2 = Slight progress 3 = Moderate progress 4 = Substantial progress 5 = Exceptional progress **Bold = Selected as Important or Essential**

33. Amount of reading	0	7	6	6	0	0	2.9	0.8	47	NA	3.2	3.1	3.3
34. Amount of work in other (non–reading) assignments	0	0	3	13	3	0	4.0	0.6	60	NA	3.4	3.6	3.6
35. Difficulty of subject matter	0	1	2	11	5	0	4.1	0.8	61	NA	3.4	3.3	3.6

Key: 1 = Much Less than Most 2 = Less than Most 3 = About Average 4 = More than Most 5 = Much More than Most

36. I had a strong desire to take this course.	5	3	5	4	2	0	2.7	1.4	NA	NA	3.7	3.5	3.7
37. I worked harder on this course than on most courses I have taken.	0	2	6	8	3	0	3.6	0.9	51	NA	3.6	3.5	3.7
38. I really wanted to take a course from this instructor.	1	3	14	1	0	0	2.8	0.6	NA	NA	3.4	3.4	3.6
39. I really wanted to take this course regardless of who taught it.	2	4	7	5	1	0	2.9	1.1	43	NA	3.3	3.3	3.7
40. As a result of taking this course, I have more positive feelings...	2	1	3	7	6	0	3.7	1.3	48	51	3.9	3.8	4.0
41. Overall, I rate this instructor an excellent teacher.	1	0	3	7	8	0	4.1	1.0	49	48	4.2	4.1	4.3
42. Overall, I rate this course as excellent.	0	1	8	6	4	0	3.7	0.9	46	48	3.9	3.9	4.0
43. As a rule, I put forth more effort than other students on...	0	0	5	8	6	0	4.1	0.8	63	NA	3.6	3.9	4.0

Key: 1 = Definitely False 2 = More False than True 3 = In Between 4 = More True than False 5 = Definitely True

No Additional Questions.

ROBBINS, R W

Susquehanna University

Business Information & Data Processing Services 174
MWF 3:00
Fall 2013

IDEA Diagnostic Form Report

To learn more, see the Interpretive Guide: www.theideacenter.org/diagnosticguide.pdf

Of the 19 students enrolled, 16 responded (84%). Feedback from individual classes is always useful to guide improvement efforts. Typically, multiple classes should be used for evaluation, using more classes when they are small (fewer than 10) or when they have low response rates (less than 60%) (see www.theideacenter.org/AdminDecisions).

Summary Evaluation of Teaching Effectiveness

Teaching effectiveness is assessed in two ways: **A. Progress on Relevant Objectives**, a weighted average of student ratings of the progress they reported on objectives selected as "Important" or "Essential" (double weighted) and **B. Overall Ratings**, the average student agreement with statements that the teacher and the course were excellent. The **SUMMARY EVALUATION** is the average of these two measures. Individual institutions may prefer to combine these measures in some other manner to arrive at a summary judgment.

Converted Averages are standardized scores that take into account the fact that the average ratings for items on the IDEA form are not equal; students report more progress on some objectives than on others. Converted scores all have the same average (50) and the same variability (a standard deviation of 10); about 40% of them will be between 45 and 55. Because measures are not perfectly reliable, it is best to regard the "true score" as lying within plus or minus 3 of the reported score.

For comparative purposes, use converted averages. Your converted averages are compared with those from all classes in the IDEA database. If enough classes are available, comparisons are also made with classes in the same broad *discipline* as this class and/or with all classes that used IDEA at your *institution*. The *Interpretive Guide* offers some suggestions for using comparative results; **some institutions may prefer to establish their own "standards" based on raw or adjusted scores rather than on comparative standing**.

Both unadjusted (raw) and adjusted averages are reported. The latter makes classes more comparable by considering factors that influence student ratings, yet are beyond the instructor's control. Scores are adjusted to take into account student desire to take the course regardless of who taught it (item 39), student work habits (item 43), instructor reported class size, and two multiple item measures (student effort not attributable to the instructor and course difficulty not attributable to the instructor).

Your Average Scores

	Your Average (5-point scale)	
	Raw	Adj.
A. Progress on Relevant Objectives [1] **Four** objectives were selected as relevant (Important or Essential –see page 2)	4.2	4.5
Overall Ratings		
B. Excellent Teacher	4.3	4.6
C. Excellent Course	3.4	4.1
D. Average of B & C	3.8	4.4
Summary Evaluation (Average of A & D) [1]	**4.0**	**4.5**

[1] If you are comparing Progress on Relevant Objectives from one instructor to another, use the converted average.

[2] The process for computing Progress on Relevant Objectives for the Discipline and Institution was modified on May 1, 2006. Do not compare these results with reports generated prior to this date.

Your Converted Average When Compared to All Classes in the IDEA Database

Comparison Category	A. Progress on Relevant Objectives		Overall Ratings: B. Excellent Teacher		C. Excellent Course		D. Average of B & C		Summary Evaluation (Average of A & D)	
	Raw	Adj.	Raw	Adj.	Raw	Adj.	Raw	Adj.	Raw	Adj.
Much Higher Highest 10% (63 or higher)										
Higher Next 20% (56–62)		61		57						58
Similar Middle 40% (45–55)	55		51			52	47	55	51	
Lower Next 20% (38–44)					42					
Much Lower Lowest 10% (37 or lower)										

Your Converted Average When Compared to Your:[2]

Discipline (IDEA Data)	54	64	52	61	41	54	47	58	51	61
Institution	51	62	50	59	40	56	45	58	48	60

IDEA Discipline used for comparison:
Business Information & Data Processing Services

Student Ratings of Learning on Relevant (Important and Essential) Objectives

Average unadjusted (raw) and adjusted progress ratings are shown below for those objectives you identified as "Important" or "Essential." **Progress on Relevant Objectives** (also shown on page 1) is a weighted average of student ratings of the progress they reported on objectives selected as "Important" or "Essential" (double weighted). The percent of students rating each as "1" or "2" (either "no" or "slight" progress) and as "4" or "5" ("substantial" or "exceptional" progress) is also reported. These results should help you identify objectives where improvement efforts might best be focused. Page 3 contains suggestions about the types of changes you might consider to obtain more satisfactory results. Also, refer to the **POD–IDEA Center *Learning* Notes** (**www.theideacenter.org/podidea/PODNotesLearning.html**).

	Importance Rating	Your Average (5–point scale)		Percent of Students Rating		Your Converted Average When Compared to Group Averages					
						IDEA Database		IDEA Discipline[1]		Your Institution[1]	
		Raw	Adj.	1 or 2	4 or 5	Raw	Adjusted	Raw	Adjusted	Raw	Adjusted
21. Gaining factual knowledge (terminology, classifications, methods, trends)	Minor/None										
22. Learning fundamental principles, generalizations, or theories	Important	4.3	4.5	6%	88%	58 Higher	61 Higher	56 Higher	63 Much Higher	52 Similar	60 Higher
23. Learning to *apply* course material (to improve thinking, problem solving, and decisions)	**Essential**	**4.2**	**4.6**	**6%**	**81%**	**54 Similar**	**62 Higher**	**52 Similar**	**63 Much Higher**	**51 Similar**	**64 Much Higher**
24. Developing specific skills, competencies, and points of view needed by professionals in the field most closely related to this course	Minor/None										
25. Acquiring skills in working with others as a member of a team	Important	4.3	4.8	0%	69%	56 Higher	64 Much Higher	57 Higher	69 Much Higher	53 Similar	66 Much Higher
26. Developing creative capacities (writing, inventing, designing, performing in art, music, drama, etc.)	Minor/None										
27. Gaining a broader understanding and appreciation of intellectual/cultural activity (music, science, literature, etc.)	Minor/None										
28. Developing skill in expressing myself orally or in writing	Minor/None										
29. Learning how to find and use resources for answering questions or solving problems	Minor/None										
30. Developing a clearer understanding of, and commitment to, personal values	Minor/None										
31. Learning to *analyze* and *critically evaluate* ideas, arguments, and points of view	Important	3.9	4.1	13%	63%	51 Similar	55 Similar	54 Similar	63 Much Higher	47 Similar	55 Similar
32. Acquiring an interest in learning more by asking my own questions and seeking answers	Minor/None										
Progress on Relevant Objectives		**4.2**	**4.5**			**55**	**61**	**54**	**64**	**51**	**62**

[1] The process for computing Progress on Relevant Objectives for the Discipline and Institution was modified on May 1, 2006. Do not compare these results with reports generated prior to this date.

Much Higher = Highest 10% of classes (63 or higher)
Higher = Next 20% (56–62)
Similar = Middle 40% (45–55)
Lower = Next 20% (38–44)
Much Lower = Lowest 10% (37 or lower)

Description of Course and Students

Students described the course by rating three items related to "level of academic challenge." Results cannot be interpreted as "good" or "bad"; in general, these ratings have a slight positive relationship with measures of academic achievement. The three items describing your students relate to their academic motivation and work habits and are key factors in developing adjusted ratings.

Course Description	Your Average (5–point scale)	Your Converted Average When Compared to Group Averages					
		IDEA Database		IDEA Discipline		Your Institution	
33. Amount of reading	2.9	46	Similar	47	Similar	45	Similar
34. Amount of work in other (non–reading) assignments	3.6	54	Similar	51	Similar	51	Similar
35. Difficulty of subject matter	4.4	67	Much Higher	72	Much Higher	62	Higher
Student Description							
37. I worked harder on this course than on most courses I have taken.	4.0	58	Higher	61	Higher	55	Similar
39. I really wanted to take this course regardless of who taught it.	2.3	32	Much Lower	28	Much Lower	28	Much Lower
43. As a rule, I put forth more effort than other students on academic work.	3.8	56	Higher	48	Similar	44	Lower

Much Higher = Highest 10% of classes (63 or higher)
Higher = Next 20% (56–62)
Similar = Middle 40% (45–55)
Lower = Next 20% (38–44)
Much Lower = Lowest 10% (37 or lower)

Improving Teaching Effectiveness

One way to improve teaching effectiveness is to make more use of the teaching methods closely related to learning on specific objectives.

- Review page 2 to identify the objective(s) where improvements are most desirable.
- Use the first column to answer the question, "Which of the 20 teaching methods are most related to learning on these objective(s)?"
- Review the next two columns to answer the question, "How did students rate my use of these important methods?"
- Read the last column to answer the question, "What changes should I consider in my teaching methods?"
- Beyond specific methods, do the results suggest a general area (e.g., Stimulating Student Interest) where improvement efforts should be focused?

Suggested Actions are based on comparisons with ratings for classes of similar size and level of student motivation. **Consider increasing use** means you employed the method less frequently than those teaching similar classes. **Retain current use or consider increasing** means you employed the method with typical frequency. **Strength to retain** means you employed the method more frequently than those teaching similar classes. More detailed suggestions are in the **Interpretive Guide** (**www.theideacenter.org/diagnosticguide.pdf**), **POD–IDEA Center Notes** (**www.theideacenter.org/podidea**), and **POD–IDEA Center *Learning* Notes** (**www.theideacenter.org/podidea/PODNotesLearning.html**).

Teaching Methods and Styles

	Relevant to Objectives: (see page 2)	Your Average (5–point scale)	Percent of Students Rating 4 or 5	Suggested Action
Stimulating Student Interest				
13. Introduced stimulating ideas about the subject	22, 23, 31	3.8	75%	Retain current use or consider increasing
4. Demonstrated the importance and significance of the subject matter	22, 23	4.1	75%	Retain current use or consider increasing
8. Stimulated students to intellectual effort beyond that required by most courses	**All selected objectives**	**4.5**	**94%**	**Strength to retain**
15. Inspired students to set and achieve goals which really challenged them	**All selected objectives**	**3.9**	**69%**	**Strength to retain**
Fostering Student Collaboration				
16. Asked students to share ideas and experiences with others whose backgrounds and viewpoints differ from their own	**25, 31**	**3.9**	**56%**	**Strength to retain**
18. Asked students to help each other understand ideas or concepts	**25, 31**	**4.2**	**75%**	**Strength to retain**
5. Formed "teams" or "discussion groups" to facilitate learning	**25**	**4.5**	**88%**	**Strength to retain**
Establishing Rapport				
2. Found ways to help students answer their own questions	**All selected objectives**	**4.4**	**88%**	**Strength to retain**
7. Explained the reasons for criticisms of students' academic performance	**23, 31**	**4.2**	**75%**	**Strength to retain**
1. Displayed a personal interest in students and their learning	**23**	**4.7**	**100%**	**Strength to retain**
20. Encouraged student–faculty interaction outside of class (office visits, phone calls, e–mails, etc.)	Not relevant to objectives selected	4.9	100%	
Encouraging Student Involvement				
11. Related course material to real life situations	23	3.9	75%	Retain current use or consider increasing
19. Gave projects, tests, or assignments that required original or creative thinking	**25, 31**	**4.6**	**88%**	**Strength to retain**
14. Involved students in "hands on" projects such as research, case studies, or "real life" activities	**25**	**4.4**	**81%**	**Strength to retain**
9. Encouraged students to use multiple resources (e.g. data banks, library holdings, outside experts) to improve understanding	Not relevant to objectives selected	3.2	31%	
Structuring Classroom Experiences				
6. Made it clear how each topic fit into the course	22, 23	4.1	81%	Retain current use or consider increasing
10. Explained course material clearly and concisely	22, 23	3.9	75%	Retain current use or consider increasing
12. Gave tests, projects, etc. that covered the most important points of the course	22	4.3	81%	Retain current use or consider increasing
3. Scheduled course work (class activities, tests, projects) in ways which encouraged students to stay up–to–date in their work	Not relevant to objectives selected	4.4	88%	
17. Provided timely and frequent feedback on tests, reports, projects, etc. to help students improve	Not relevant to objectives selected	4.3	88%	

5–point Scale: 1 = Hardly Ever 2 = Occasionally 3 = Sometimes 4 = Frequently 5 = Almost Always

Statistical Detail

	Number Responding							
	1	2	3	4	5	Omit	Avg.	s.d.
1. Displayed a personal interest in students and their learning	0	0	0	5	11	0	4.7	0.5
2. Found ways to help students answer their own questions	0	1	1	5	9	0	4.4	0.9
3. Scheduled course work (class activities, tests, projects) in ways...	0	2	0	4	10	0	4.4	1.0
4. Demonstrated the importance and significance of the subject matter	1	1	2	3	9	0	4.1	1.3
5. Formed "teams" or "discussion groups" to facilitate learning	0	1	1	3	11	0	4.5	0.9
6. Made it clear how each topic fit into the course	2	0	1	5	8	0	4.1	1.3
7. Explained the reasons for criticisms of students' academic...	0	1	3	4	8	0	4.2	1.0
8. Stimulated students to intellectual effort beyond that required by...	1	0	0	4	11	0	4.5	1.0
9. Encouraged students to use multiple resources (e.g. data banks,...	1	3	7	2	3	0	3.2	1.2
10. Explained course material clearly and concisely	2	0	2	5	7	0	3.9	1.3
11. Related course material to real life situations	2	2	0	4	8	0	3.9	1.5
12. Gave tests, projects, etc. that covered the most important points...	2	0	1	2	11	0	4.3	1.4
13. Introduced stimulating ideas about the subject	2	0	2	7	5	0	3.8	1.3
14. Involved students in "hands on" projects such as research, case...	0	0	3	4	9	0	4.4	0.8
15. Inspired students to set and achieve goals which really...	0	2	3	5	6	0	3.9	1.1
16. Asked students to share ideas and experiences with others...	0	2	5	2	7	0	3.9	1.1
17. Provided timely and frequent feedback on tests, reports,...	1	0	1	6	8	0	4.3	1.1
18. Asked students to help each other understand ideas or concepts	0	1	3	4	8	0	4.2	1.0
19. Gave projects, tests, or assignments that required original or...	1	0	1	1	13	0	4.6	1.1
20. Encouraged student–faculty interaction outside of class (office...	0	0	0	1	14	1	4.9	0.3

Key: 1 = Hardly Ever 2 = Occasionally 3 = Sometimes 4 = Frequently 5 = Almost Always

The details on this page are of interest primarily t those who want to confirm scores reported on pages 1–3 or who want to determine if responses to some items were distributed in an unusual manner.

Converted Averages are reported only for relevan learning objectives (Important or Essential –see page 2) and other items for which comparisons were provided.

Notes:
Discipline code selected on FIF: 5212
Discipline code used for comparison: 5212

	Number Responding								Converted Avg.		Comparison Group Average		
	1	2	3	4	5	Omit	Avg.	s.d.	Raw	Adj.	IDEA	Discipline	Institution
21. Gaining factual knowledge (terminology, classifications,...	1	0	1	5	9	0	4.3	1.1	NA	NA	4.0	4.1	4.3
22. Learning fundamental principles, generalizations, or...	**1**	**0**	**1**	**5**	**9**	**0**	**4.3**	**1.1**	**58**	**61**	**3.9**	**4.0**	**4.2**
23. Learning to *apply* course material (to improve thinking,...	**1**	**0**	**2**	**5**	**8**	**0**	**4.2**	**1.1**	**54**	**62**	**4.0**	**4.1**	**4.2**
24. Developing specific skills, competencies, and points of view...	1	0	2	5	8	0	4.2	1.1	NA	NA	4.0	4.1	4.2
25. Acquiring skills in working with others as a member of...	**0**	**0**	**5**	**1**	**10**	**0**	**4.3**	**0.9**	**56**	**64**	**3.9**	**3.9**	**4.1**
26. Developing creative capacities (writing, inventing, designing,...	1	1	4	2	8	0	3.9	1.3	NA	NA	3.9	3.5	4.2
27. Gaining a broader understanding and appreciation of...	3	3	4	2	4	0	3.1	1.5	NA	NA	3.7	3.4	4.1
28. Developing skill in expressing myself orally or in writing	1	1	7	2	5	0	3.6	1.2	NA	NA	3.8	3.5	4.0
29. Learning how to find and use resources for answering questions...	0	2	5	3	6	0	3.8	1.1	NA	NA	3.7	3.9	3.9
30. Developing a clearer understanding of, and commitment to,...	3	1	5	3	4	0	3.3	1.4	NA	NA	3.8	3.6	3.9
31. Learning to *analyze* and *critically evaluate* ideas,...	**0**	**2**	**4**	**4**	**6**	**0**	**3.9**	**1.1**	**51**	**55**	**3.8**	**3.7**	**4.0**
32. Acquiring an interest in learning more by asking my own...	0	1	4	4	7	0	4.1	1.0	NA	NA	3.8	3.8	3.9

Key: 1 = No apparent progress 2 = Slight progress 3 = Moderate progress 4 = Substantial progress 5 = Exceptional progress Bold = Selected as Important or Essential

33. Amount of reading	2	3	7	3	1	0	2.9	1.1	46	NA	3.2	3.1	3.3
34. Amount of work in other (non–reading) assignments	0	0	8	6	2	0	3.6	0.7	54	NA	3.4	3.6	3.6
35. Difficulty of subject matter	0	0	2	6	8	0	4.4	0.7	67	NA	3.4	3.3	3.6

Key: 1 = Much Less than Most 2 = Less than Most 3 = About Average 4 = More than Most 5 = Much More than Most

36. I had a strong desire to take this course.	5	2	4	4	1	0	2.6	1.4	NA	NA	3.7	3.5	3.7
37. I worked harder on this course than on most courses I have taken.	0	2	3	4	7	0	4.0	1.1	58	NA	3.6	3.5	3.7
38. I really wanted to take a course from this instructor.	3	0	8	1	4	0	3.2	1.4	NA	NA	3.4	3.4	3.6
39. I really wanted to take this course regardless of who taught it.	6	3	4	2	1	0	2.3	1.3	32	NA	3.3	3.3	3.7
40. As a result of taking this course, I have more positive feelings...	2	3	1	3	7	0	3.6	1.5	46	59	3.9	3.8	4.0
41. Overall, I rate this instructor an excellent teacher.	2	0	2	0	12	0	4.3	1.4	51	57	4.2	4.1	4.3
42. Overall, I rate this course as excellent.	3	1	2	6	4	0	3.4	1.5	42	52	3.9	3.9	4.0
43. As a rule, I put forth more effort than other students on...	0	0	4	11	1	0	3.8	0.5	56	NA	3.6	3.9	4.0

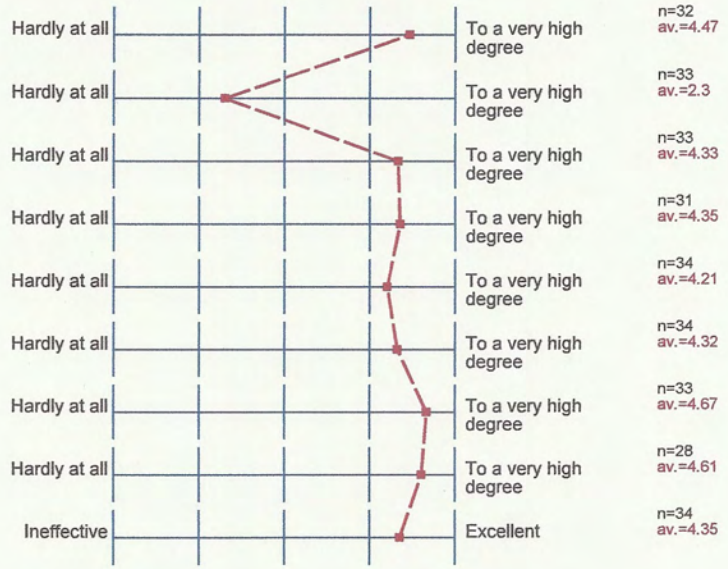
Key: 1 = Definitely False 2 = More False than True 3 = In Between 4 = More True than False 5 = Definitely True

No Additional Questions.

Profile

Subunit: BUSINESS
Name of the instructor: Professor Russell Robbins,
Name of the course: (Name of the survey) INTRO TO INFORMATION SYSTEMS(BUSMIS-1060) (13011-2134)

1. TEACHING EVALUATION

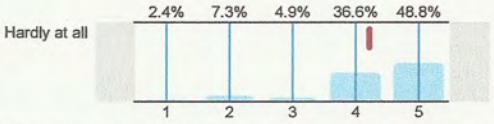


Professor Russell Robbins

INTRO TO INFORMATION SYSTEMS(BUSMIS-1060)12898-2134
Spring 2013
RESPONDENTS = 86% OF NUMBER REGISTERED

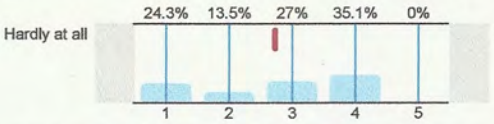
1. TEACHING EVALUATION

1.1) The course was intellectually challenging.



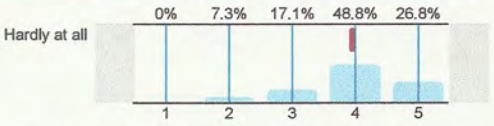
n=41
av.=4.22
dev.=1.01

1.2) The text and materials were helpful in learning concepts and techniques.



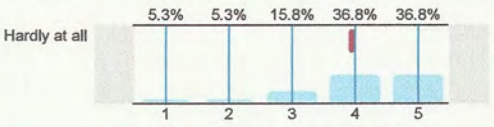
n=37
av.=2.73
dev.=1.19

1.3) The course increased my knowledge.



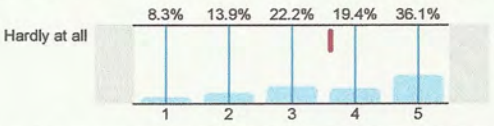
n=41
av.=3.95
dev.=0.86

1.4) The instructor was well prepared for class activities.



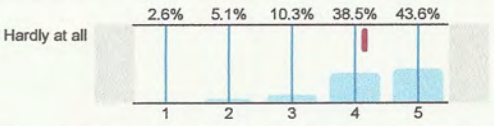
n=38
av.=3.95
dev.=1.11

1.5) The instructor effectively interpreted difficult or abstract ideas.



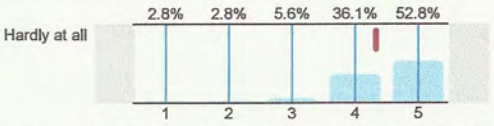
n=36
av.=3.61
dev.=1.34

1.6) The instructor provided appropriate feedback.



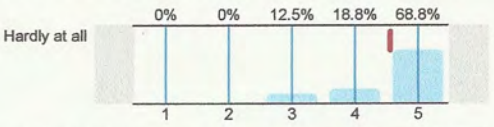
n=39
av.=4.15
dev.=0.99

1.7) The instructor created an atmosphere conducive to learning.



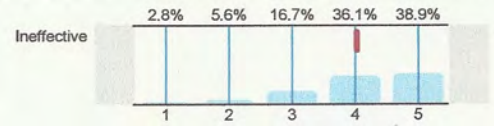
n=36
av.=4.33
dev.=0.93

1.8) The instructor was accessible to students. (Do not answer if no basis to judge)



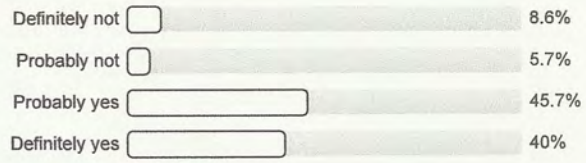
n=32
av.=4.56
dev.=0.72

1.9) Express your judgment of the instructor's **overall teaching effectiveness:**



n=36
av.=4.03
dev.=1.03

1.10) Would you recommend this instructor to other students?

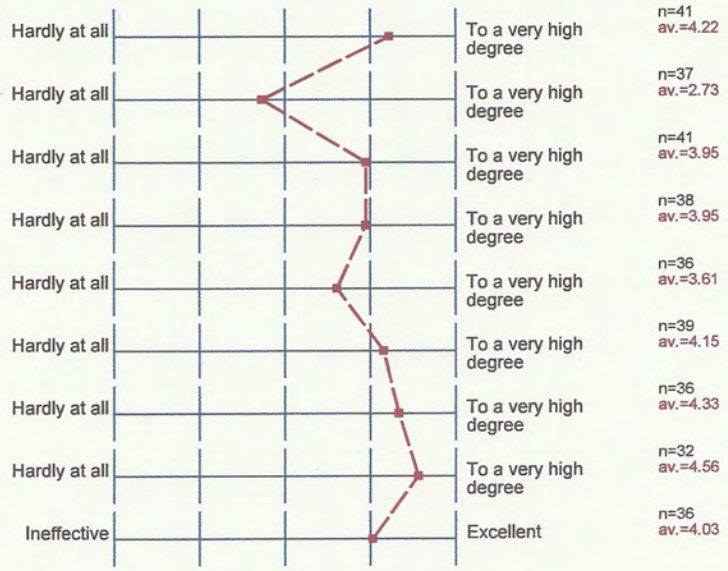


n=35

Profile

Subunit: BUSINESS
Name of the instructor: Professor Russell Robbins,
Name of the course: (Name of the survey) INTRO TO INFORMATION SYSTEMS(BUSMIS-1060) (12898-2134)

1. TEACHING EVALUATION

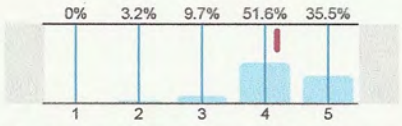


Professor Russell Robbins

INTRO TO INFORMATION SYSTEMS(BUSMIS-1060)/12839-2134
Spring 2013
RESPONDENTS = 66.67% OF NUMBER REGISTERED

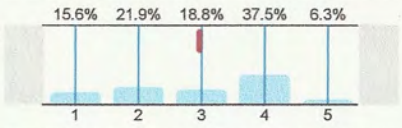
1. TEACHING EVALUATION

1.1) The course was intellectually challenging.



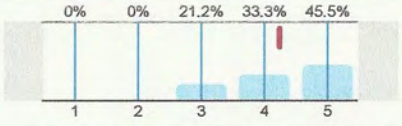
n=31
av.=4.19
dev.=0.75

1.2) The text and materials were helpful in learning concepts and techniques.



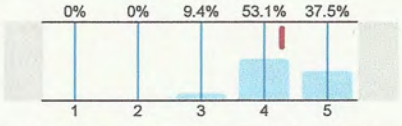
n=32
av.=2.97
dev.=1.23

1.3) The course increased my knowledge.



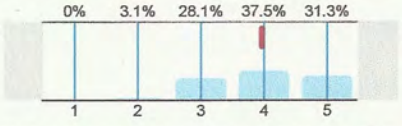
n=33
av.=4.24
dev.=0.79

1.4) The instructor was well prepared for class activities.



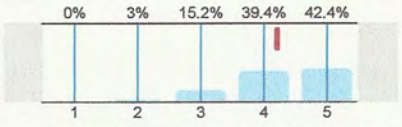
n=32
av.=4.28
dev.=0.63

1.5) The instructor effectively interpreted difficult or abstract ideas.



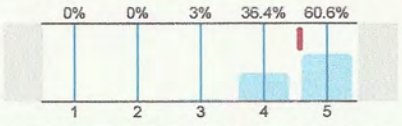
n=32
av.=3.97
dev.=0.86

1.6) The instructor provided appropriate feedback.



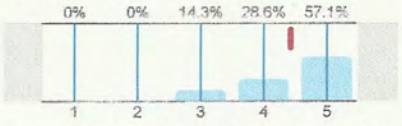
n=33
av.=4.21
dev.=0.82

1.7) The instructor created an atmosphere conducive to learning.



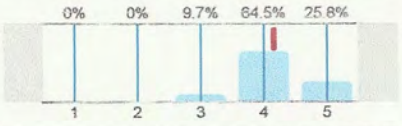
n=33
av.=4.58
dev.=0.56

1.8) The instructor was accessible to students. (Do not answer if no basis to judge)



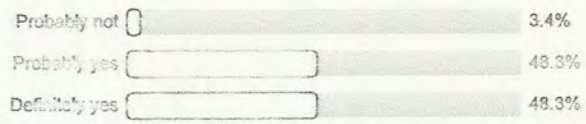
n=28
av.=4.43
dev.=0.74

1.9) Express your judgment of the instructor's overall teaching effectiveness:



n=31
av.=4.16
dev.=0.58

1.10) Would you recommend this instructor to other students?



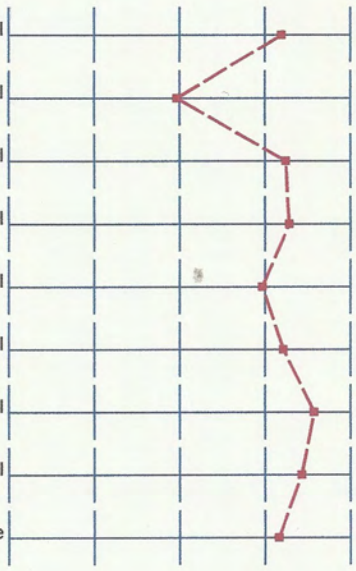
n=29

Profile

Subunit: BUSINESS
Name of the instructor: Professor Russell Robbins,
Name of the course: (Name of the survey) INTRO TO INFORMATION SYSTEMS(BUSMIS-1060) (12899-2134)

1. TEACHING EVALUATION

1.1)	The course was intellectually challenging.	Hardly at all	To a very high degree	n=31 av.=4.19
1.2)	The text and materials were helpful in learning concepts and techniques.	Hardly at all	To a very high degree	n=32 av.=2.97
1.3)	The course increased my knowledge.	Hardly at all	To a very high degree	n=33 av.=4.24
1.4)	The instructor was well prepared for class activities.	Hardly at all	To a very high degree	n=32 av.=4.28
1.5)	The instructor effectively interpreted difficult or abstract ideas.	Hardly at all	To a very high degree	n=32 av.=3.97
1.6)	The instructor provided appropriate feedback.	Hardly at all	To a very high degree	n=33 av.=4.21
1.7)	The instructor created an atmosphere conducive to learning.	Hardly at all	To a very high degree	n=33 av.=4.58
1.8)	The instructor was accessible to students. (Do not answer if no basis to judge)	Hardly at all	To a very high degree	n=28 av.=4.43
1.9)	Express your judgment of the instructor's **overall teaching effectiveness:**	Ineffective	Excellent	n=31 av.=4.16



Section F - Example: Successfully proposing products and services
(and designing, developing, implementing, and assessing a complex algorithm)

This section indicates my ability to bring a product's potential future versions into focus and how I can express that vision.

(Please turn the page.)

I was a professor. In that context I asked myself “What is the most valuable thing that I can do with my business students? My answer was: “Imparting a sense for the value of data analysis and helping students build a base they could then build upon.”

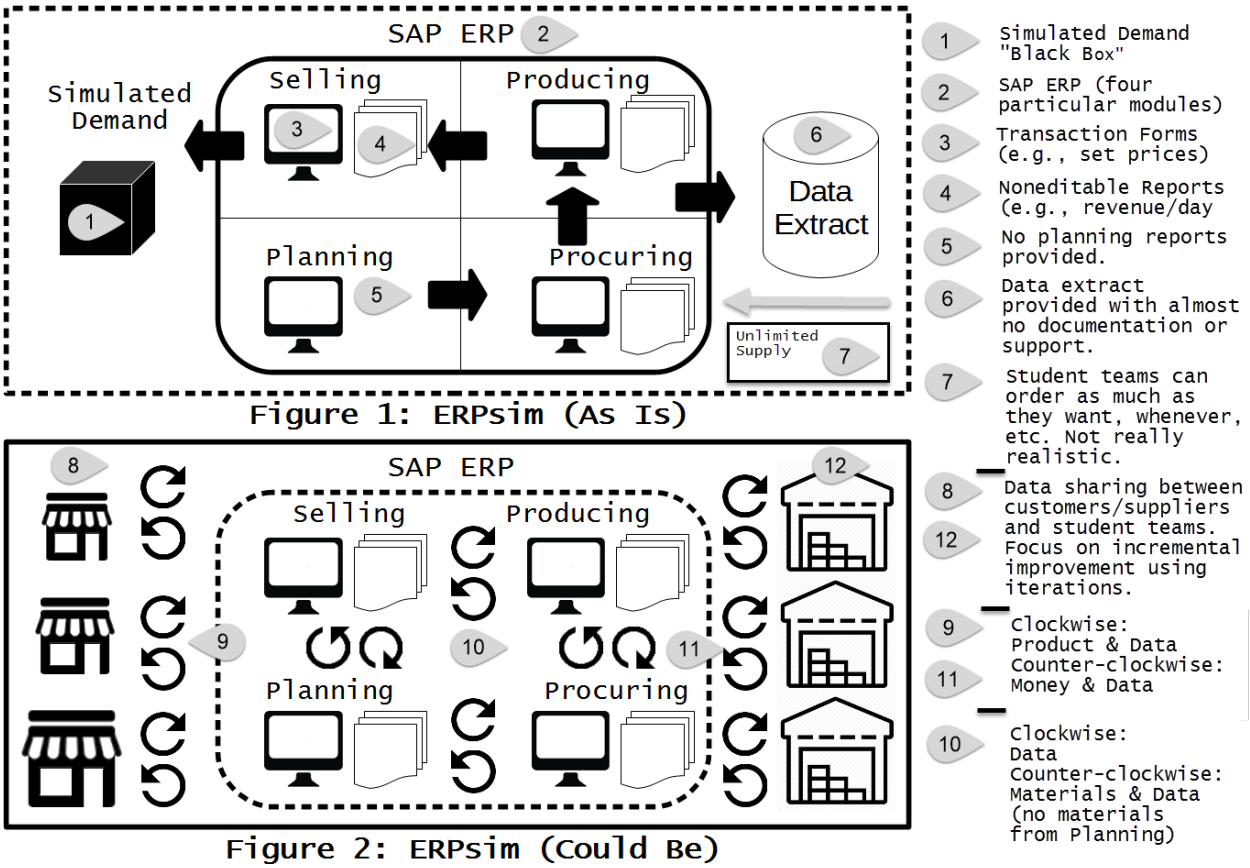


Figure 1: ERPsim (As Is)

Figure 2: ERPsim (Could Be)

Many individuals understand that what s/he values is what s/he strives to do well. In this context I sought to find an experience that would help students value analytics while learning analytics. I found part of this experience in the product known as SAP ERPsim. SAP ERPsim is a product in which students “run” companies using core transactions in four SAP ERP modules. A background simulation runs which simulates demand and purchases units of product. Please see Figure 1.

As sold, the SAP ERPsim focuses on the transactions part of the transactions / analytics cycle that businesses use. While a data extract option is provided, all other materials provided with the ERPsim simulation focus on students learning how to “push” product through a manufacturing organization. If a professor delivers the ERPsim product/service to students as advised by the ERPsim product/service providers, s/he gives students (at most) one half of the story.

A more complete ERPsim product (Figure 2) would have an analysis focus that stepped through descriptive analytical (what is), predictive analytical (what is likely), and prescriptive analytical (what should be) lenses. SAP ERPsim materials currently used to train students about performing transactions would need few changes.

However, additional materials would need to be developed to help students practice the analytics part of the transactions/analytics cycle that all businesses use.

In an ideal world, the additional materials would be embedded into the SAP ERP interface in ERPsim. A more realistic approach would be using another commercial off the shelf analytics software package that could be integrated with the existing ERPsim.

The attached paper “Clarifying the ERPsim Experience” shares some of my ideas regarding how the ERPsim product could be used better. Note that this paper was framed for an Information Systems Educator conference, as opposed to the providers of the ERPsim product.

Americas Conference on Information Systems

SPECIAL NOTE:

While this paper was accepted and published by the AMCIS conference, the paper was removed because I was not able to attend. I include this paper in my portfolio because it provides evidence of my potential for successfully proposing projects and services.

Clarifying the SAP ERPsim Experience

Journal:	*20th Americas Conference on Information Systems*
Manuscript ID:	AMCIS-0777-2014.R1
Submission Type:	Paper
Track:	General Track for IS Education < IS in Education, IS Curriculum, Education and Teaching Cases (SIGED)

Clarifying the SAP ERPsim Experience

Research-in-Progress

Russell W. Robbins
Susquehanna University

Abstract

This paper provides readers not familiar with the ERPsim games a simple overview. It also provides a detailed description of the ERPsim Manufacturing Game. Further, it provides a detailed prescription (a set of maxims and a roadmap) for using the simulation. A research plan is framed to answer the research question: "Will students that are taught using the maxims and the road map understand how information systems support business processes better?" If the planned experiment begins to show that the maxims and roadmap are efficacious, the researcher's contribution will be pedagogy that instructors can apply with non-MIS (and MIS) major students to help these students understand the value of MIS, and prospectively, to help these same students engage their MIS (e.g., data analytics) course more fully. The author also seeks advice from experienced ERPsim instructors as to how to improve the maxims/roadmap or the research plan.

Introduction

MIS faculty frequently find themselves in the challenging circumstances of teaching MIS courses to students who have not developed an appreciation for IS. Despite undergraduates living in a world that has ubiquitous computing, part-time MBA students interacting with information systems during the day prior to coming to class at night, and full-time MBA students using systems prior to starting graduate school, many of these students arrive to our classes believing they already understand IS, that IS are tangential to their developing core competencies, or both. These evaluations by students keep them from engaging an MIS course fully. In order to address this lackluster evaluation of MIS by non-MIS major students and its effects, the author clarified his students' experiences with the ERPsim learning experience (Leger 2006; Leger et al. 2007; Leger et al. 2012-2013), which asks students to use SAP ERP software as they "run" organizations.

ERPsim is a simulation game suite that helps students study ERP concepts by using an ERP system. There are three ERPsim simulation games: Distribution, Manufacturing, and Logistics. This paper discusses the Manufacturing Game. The ERPsim Manufacturing Game asks students to, in teams, run a company that makes and sells muesli (e.g., granola). If coached, students can develop their skills to optimize and synchronize the planning, procuring, manufacturing, and selling business processes. Figure 1 graphically shows major aspects of the game, including creating a forecast of sales which drives production, changing prices, and managing marketing expenses.

In order to use an ERPsim simulation game, an instructor's institution needs to be a member of the University Alliance. Information about how to become a member can be found at http://scn.sap.com/docs/DOC-7876. Further, since the games are complex, instructors must become certified. S/he can attend training and take a certification exam at the HEC Montreal ERPsim Lab or other venues. More information can be found at http://erpsim.ca or by emailing erpsim@hec.ca. Finally, licenses for learning materials must be purchased to run the Manufacturing and Logistics Games. The Distribution Game is free. The outline of the paper is thus:

1. ERPsim Manufacturing Game Overview
2. ERPsim Manufacturing Game in Detail
3. Hypothesized Maxims and Prescription for the ERPsim Manufacturing Game
4. Research Plan
5. Conclusion and Acknowledgements

ERPsim Manufacturing Game Overview

The ERPsim Manufacturing Game (hence, Game) asks student teams to "run" a German for-profit enterprise that makes and sells muesli to German independent grocers (e.g., convenience stores), grocery stores, and hypermarkets (e.g., Sam's Club) in three marketing regions (North, West, South). In addition to providing experiences that help students understand that transactions are at the foundation of today's commerce, it helps students understand the concept of a business process and that business processes are interdependent.

Key skills students can begin to appreciate are the abilities to plan purchases of raw materials, keep a manufacturing line running and fully utilized, and selling finished goods so as to maximize profit. The Introductory version of the Game is designed so that students are carefully introduced to SAP ERP functionality over the first three rounds. Rounds last at least 20 days but can be longer. Each "day" is roughly one minute in length. In the first round, students focus on learning how to use the SAP ERP pricing transaction form to set prices for muesli product distribution channels and a marketing expense transaction form which asks students to set marketing (e.g., advertising) expenses for each muesli product by day by sales region (hence, area).

During the second round of the introductory Game students learn to access reports (e.g., inventory levels, market prices). The second round also requires students to become familiar with and execute a transaction that converts planned production orders to actual production orders (which drive the muesli assembly line).

In the third round of the Game, students learn how to place their forecasts for sales (and thus, production) in the SAP ERP system. Further they learn how to initiate material requirements planning (MRP) using a SAP ERP transaction. MRP determines the raw materials that are necessary to create the muesli products; it also creates planned production orders. Students then learn to execute a transaction that creates all the purchase orders (PO) that are necessary to obtain the raw materials. Finally, students then learn how to convert planned production orders to actual production orders, and by doing so, placing production runs in the production schedule. Students also manage sales as they have in the first two rounds.

The three previous paragraphs describe the Introductory version of the game, the Extended version of the Game requires students to use all of the transactions across the planning, purchasing, producing, and selling business processes beginning in the first round. It also can include depreciation, fixed overhead, interest and loan management, and inventory storage cost features. The Advanced version of the Game includes an additional process—allocating finished goods muesli products to the three sales areas prior to their sale.

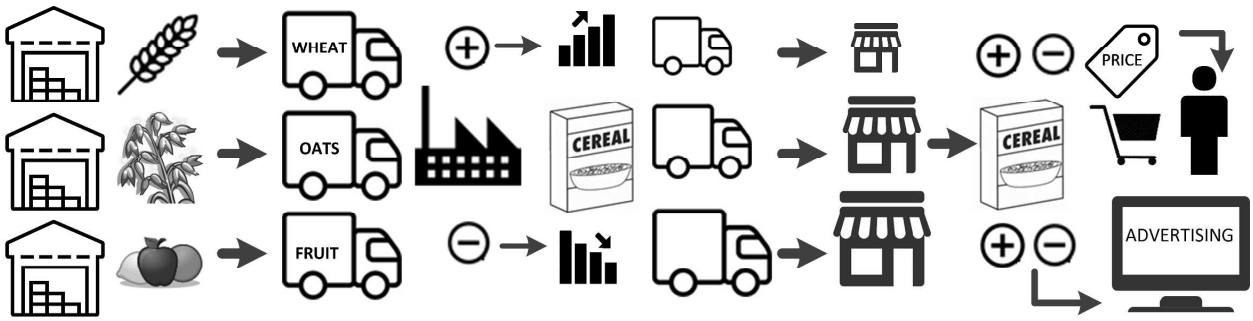


Figure 1: ERPsim Manufacturing Game Overview

ERPsim Manufacturing Game in Detail

Planning, Procuring, Manufacturing, and Selling Business Processes

The business cycle encompasses four business processes: Planning, Procuring, Manufacturing, and Selling. During Planning, students indicate the sales of the muesli products that they expect, and therefore need to manufacture by using the MD61 transaction form. Further, the required amounts of each raw material (e.g., wheat, oats, strawberries) are calculated when they use the MD01 transaction form to perform MRP. Process outputs are purchase requisitions and planned production orders.

During the Procuring business process, student teams use the ME59N transaction to create purchase orders (po) for the raw materials. POs are then "sent" to suppliers and the raw materials arrive at the student team's production facility in three to five days. Accounting transactions (e.g., goods receipt) are automated. After the raw materials arrive student teams convert planned production orders to production orders and start production runs using the CO41 transaction form. Upon production of muesli, the muesli products are available for sale at the prices set by student teams using VK32 transaction and as influenced by marketing (e.g., advertising) expenditures managed via the ZADS transaction. See Figure 2.

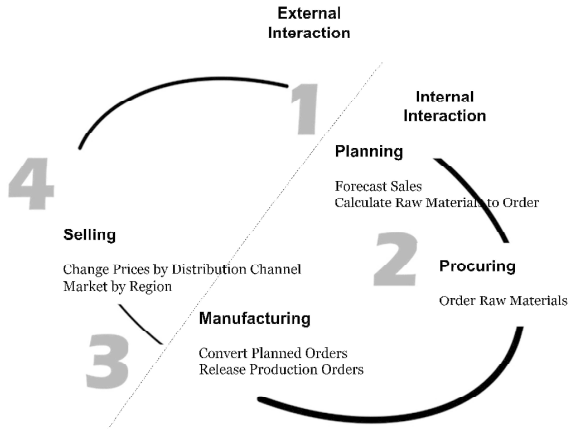


Figure 2: Actions to be Taken by Students in Every Manufacturing Game

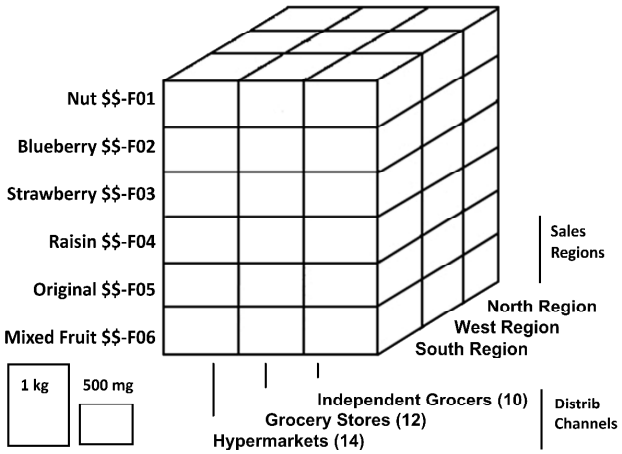


Figure 3: Products, Distribution Channels, and Sales Areas

Products, Distribution Channels, and Sales Areas

The Products that are by default "manufactured and sold" are Nut (Product Number $$-F01), Blueberry ($$-F02), Strawberry ($$-F03), Raisin ($$-F04), Original ($$-F05), and Mixed Fruit ($$-F06) Muesli in 1 Kg boxes. The $ in the product numbers represents a variable that is replaced with a teams' identifying letter. Each of the 1 kg boxes of muesli are primarily wheat, then oats, and sometimes, a smaller amount of each of the special ingredients of raisins, nuts, berries, etc. In the Extended and Advanced versions of the game, one to six of these six products can be replaced by new products which can be combinations of wheat, oats, nuts, and varying berries and can be in 500 mg boxes.

The Distribution Channels that are by default the organizations that the students sell through are hypermarkets (Distribution Channel 10), grocery stores (12), and independent grocers (14). Student teams can use the price change transaction form (VK32) to set and change prices for each of the six products for each of the three distribution channels. Prices changed on one day become effective the next day. There are 12 hypermarkets, 59 grocery stores, and 123 conveniences stores.

The Sales Areas are the North, the West and the South of Germany. These areas are primarily important to students as they determine, set, and use marketing (e.g., advertising) expenses via transaction form ZADS. The amounts that students enter for each product/sales area combination are the amounts of money (Euros) to be spent on marketing each day. These values can be changed each day and are effective

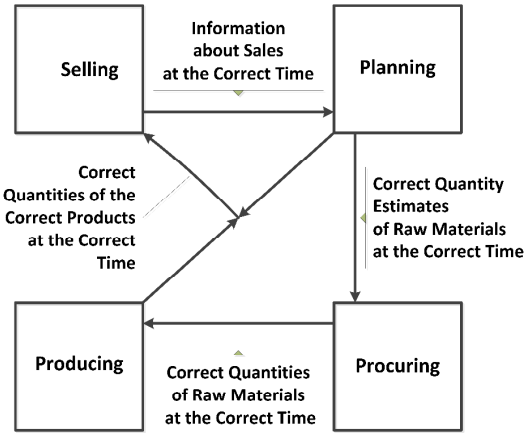


Figure 4: Optimal Flows of Information, Raw Materials, and Products

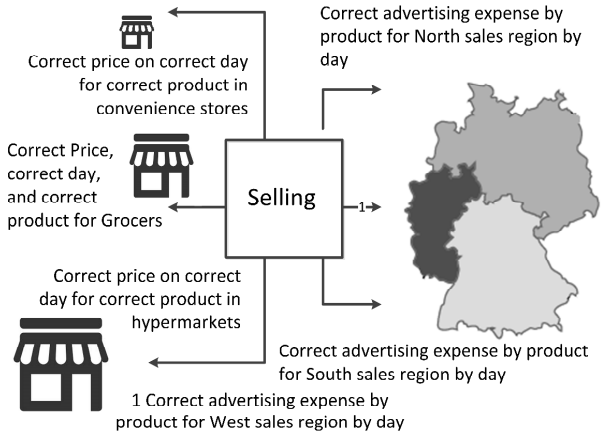


Figure 5: Optimal Flows of Information and Capital outside of Company

Business Process	Pivot Table	Fields
Planning	None	Not applicable.
Procuring	Muesli Purchase	Company (Student team), Product Name, Purchase Order Number, Purchase Order Date, Expected Delivery Time in Days, Actual Delivery Delay in Days, Receipt Date (in Quarter/Day form), Raw Material Cost
Manufacturing	Muesli Daily Stock	Company, Product Name, Product Number, Quarter, Week, Day, Quantity on Hand, Warehouse Location
Selling	Muesli Sales	Company, Customer Number, Product Number, Product Name, Price per Sales Order, Quantity Sold, Revenue per Sales Order, Distribution Channel, Sales Area Name, Quarter, Week, Day, Product Standard Cost, Warehouse Location
	Muesli Market Share	Product Name, Sales Area Name, Market Quantity Sold, Market Revenue, Company Quantity Sold, Company Revenue, Daily Quantity Sold, Daily Revenue, Week
	Product Profitability	Company, Product Name, Account (Revenue, Cost), Net Profit, Sales Area, Week
Financial Accounting	Muesli Financial	Account Number, Account Name, Financial Statement Level 1 Category (e.g., Liabilities), Financial Statement Level 2 Category (Current Liabilities), Financial Statement Level 3 Category (Bank), Credit/Debit, Euro Amount

Table 1: SAP ERPsim Reports

Business Process	Report	Information
Planning	None	Not applicable.
Procuring	Purchase Order Tracking Report	Products ordered, the quantity ordered of each, expected receipt date, whether goods have been received, and the expected payment date for each PO.
Manufacturing	Inventory Report	Amounts of raw materials, packing materials (boxes and bags), and finished goods for the day that has just transpired.
	Current Material Overview	Amount of a finished goods on hand. Amount of finished goods in process.
	Production Schedule Report	Production orders for each day. Each production batch/run includes when it was placed on the schedule, when it will start/finish, how much time it took to setup the line for that batch/run, how many units (boxes) should be produced, how many were produced, and the unit cost for each box.
	Production Cost Analysis Report	Variable and fixed costs for each product as well as the prices for each distribution channel.
	Raw Material Cost Per Order Report	Raw Materials and Costs by Production Order
Selling	Sales Order Report	Product, price, quantity sold in boxes, total revenue, customer number, and quarter, day, distribution channel, and sales area.
	Sales Summary Report	Product, quantity sold in boxes, and total revenue for each quarter and day.
	Price Market Report	Product, quantity sold, average price, every five days.
Financial Accounting	Financial	Revenue, expenses, assets, and liabilities for a quarter.

Table 2: Business Intelligence Excel Pivot Tables/Charts Reports

the next day. See Figures 2 and 3. Figures 4 and 5 indicate optimal flows with regards to internal and external flows of information and product.

SAP ERPsim Reports

See Table 1.

Excel Pivot Tables

To extend the information available to students as they play the Game as well as to facilitate drill-down analysis, the ERPsim Lab has developed a spreadsheet that can be populated with data from a specific run of a Game. See Table 2.

Data Extraction Tools

Business Process	SAP tables	ERPsim tables
Planning	None	None
Procuring	Purchase Order Header Purchase Order Detail	Purchase Order by Quarter/Day
Manufacturing- Bill of Material	Bill of Material Header Bill of Material Detail	None
Manufacturing- Production	Production Order Header Production Order Detail	Production Order by Quarter/Day
Manufacturing- Finishing	Production Confirmation Inventory Movement	None
Selling	Sales Order Header Sales Order Detail Customer	Sales Order by Quarter/Day Sales by Sales Area, Distribution Channel, and Quarter/Day Sales Area Contribution Margin by Quarter/Day
Financial Accounting	Transaction Header Transaction Detail Account Numbers Account Names	Income Statement Accounts Marketing Expense Accounts by Quarter, Day, Sales Area, and Product Balance Sheet Accounts
Materials	Material Material Type Material Location	Finished Goods by Quarter/Day
ERPsim Timing	None	Quarter/Day Occurrence Counting

Table 3: Table Data Available through ERPsim Simulation Data Extraction Tool

The Game has three related data extraction tools, which are Microsoft Database files which establish connections with SAP database instances and query data: Simulation, Cash-to-Cash, and Configuration. This paper discusses the Simulation data extraction tool (hence, Tool).

The Tool contains eighteen "vanilla" tables that exist in most SAP instances. The Tool has another eleven tables that are specific to the ERPsim games. Further, it contains fourteen queries. Table 3 describes the tables. The queries in the Tool are named clearly so are not described here. In effect the ERPsim tables accessible via the Tool largely link the SAP tables to a concept of time in the Game which is rounds and days (e.g., minutes). Further, three other Tool tables store information about Game sales while another three tables store Income Statement and Balance Sheet Account information. See Table 3.

Hypothesized Maxims and Roadmap for the Game

The Game is an excellent learning environment for non-MIS major business/management students. If the Game is used carefully, students can:

1. Develop an intuitive understanding about what a business process is.
2. Learn that business processes are interconnected.
3. Learn that business processes cannot be performed without information systems.
4. Learn that modern corporations use expansive information systems called enterprise systems.

5. Develop an appreciation for the complexity of enterprise systems.
6. Learn that even a very small decision made by one person in an organization can have significant financial effects (either positive or negative).
7. Develop intuitive understanding of the model of a manufacturing corporation (i.e., exchange money for inputs, transform those inputs into something new and valuable using people and equipment, exchange outputs for money).
8. See the inter-relationship of accounting and information systems.
9. See an inter-relationship of sales and information systems.
10. See an inter-relationship of operations management and information systems.

When using the Game as a proxy for a corporation and its use of data, the instructor asked student teams to use SAP ERP to first "operate" their organizations. Then the author asked the student teams to analyze their internal performance using data extracted from the ERPsim SAP ERP instance for their team. The author asked the students to develop an understanding of their external performance in the marketplace. He then asked the students to identify problems, develop, and implement action operational plans. Finally, the author asked the students to develop and implement strategies to dominate the marketplace. Again, students were to gather data, analyze the data, adjust strategy, and implement. The roadmap below is a result of knowledge the author gained during those runs of the Game.

As the Game is provided, it takes four class sessions of at least 60 minutes to run the game. This assumes students have already learnt the transactions and understand the four business processes and their contribution to the business cycle. Given its length and associated use of class time, the hypothesized maxims and prescription provided here is probably more appropriate as part of an introductory module of a data analytics course, thereby providing tangible motivation for students as they complete the rest of the course and tackle learning more difficult skills. Importantly, the author is interested in other experienced ERPsim instructors' thoughts about improving the hypothesized roadmap and about the provided experimental design that could be used to assess the roadmap as a whole or in parts. These maxims and the prescription are with regards to an Extended Game with pre-game capitalization, depreciation, fixed overhead, interest and loan management, and inventory storage cost features enabled, and that has eight rounds of 30 days each. The maxims identified below are ordered by degree of importance.

Hypothesized Maxims:

1. The production line runs continuously and production line setups are minimized but flexibility to change products is developed and then maintained. Thus the total amount of boxes of muesli that should be created across all production cycles (MD61 through CO41) in a round is represented by the equation below.

   $\Sigma = \alpha\beta - \gamma\delta\varepsilon$

   where

   Σ : the total amount of boxes to create in a month (e.g., 419,136)

   α : amount of boxes that can be made in a day (e.g., 21,000)

   β : the number of days in the round (e.g., 20)

   γ : the number of boxes that can be produced in an hour

   (e.g., 875, if production runs 5 days a week 24 hours a day)

   δ : the number of hours it takes to setup a new production configuration (e.g., 12)

   ε : the number of setups that are necessary in the month (e.g., 3)

   Note: If setup times (see below) are decreased via investment then 12 should be replaced with the new setup time. Note if production in a day increases past 21,000 (e.g., to 25000), that number should replace the 21,000 in the equation.

2. Price should be increased, decreased or held stable based upon the arc elasticity at the next price. If the arc elasticity is less than -1 (i.e., -1.01 to -∞) then prices should be increased. If the arc elasticity is greater than -1 (i.e., -.99 to +∞) prices should be decreased. In the case below price should not be changed, unless quantities sold are decreasing at the same offer price. (See number 2 below.) Arc elasticity should be computed thus:

   $E_p = \Delta Q / \Delta P * (P_2 - P_1)/(Q_2 - Q_1)$ (e.g., (2527 /.05) * (€1.80 – €1.75) / (21,346 – 23,873) = -1

   Note: If competitors lower their price on that product, and your quantities sold begin decreasing, you should decrease your price until your quantities sold stabilize. This price is referred to as the adjusted theoretical price below.

3. Product sales (and therefore, production), as much as possible, should occur in ranked order descending. The ranking of products in terms of selling preference should be based upon contribution margin at unit elasticity and as influenced by competitors' prices, with each product that has a higher contribution margin sold at the maximum quantity prior to moving to produce and sell another product.

   Note: if the product that has the highest contribution margin and which has a price set according to 1 and 2 above, then the next production runs should create that same product, until sales stop. Then the focus should move to the second highest contribution margin set using numbers 1 and 2 above, while monitoring for lack of sales or sales of the highest ranked product.

4. New products should be designed and assessed with regards to profitability, particularly in the .5 kg size in the distribution channel that supplies independent grocers. These new products should replace the poorest selling products that you start a game with.

5. Setup times should be decreased incrementally if sales keep increasing and price is at the adjusted theoretical price on all products, and the total profit in a round attributable to the setup time reduction is greater than its affiliated depreciation expense (.008 * the capacity improvement cost), and cash on hand is greater than the mean cash on hand over the rounds that have transpired. When the total profit – depreciation expense approaches the cost of interest on the loan or setup time has reached 2.5 hours, setup time reductions should be discontinued. An estimate of the increment to use is 1% of cash on hand. Teams should adjust from 1% based upon their risk tolerance.

   Note: If return on investment in capital for the previous falls below .0046 per month, then cash available greater than the mean cash on hand over the rounds that have transpired should be used to pay down the loan.

6. Production capacity should be increased incrementally if sales keep increasing and price is at the adjusted theoretical price, and the total profit in a round attributable to the capacity increase is greater than its affiliated depreciation expense (.008 * the capacity improvement cost), and cash on hand is greater than the mean cash on hand over the rounds that have transpired. When the total profit – depreciation expense approaches €0 production capacity improvements should be discontinued. An estimate of the increment to use is 1% of cash on hand. Teams should adjust from 1% based upon their risk tolerance.

   Note: If return on investment in capital for the previous falls below .0046 per month, then cash available greater than the mean cash on hand over the rounds that have transpired should be used to pay down the loan.

7. As prices are managed, so should marketing expenses. Marketing expenses for each product / sales area combination should maximize the effectiveness of Euros, in terms of quantity of sales increased per Euro spent on marketing in a Sales Area as measured via marketing expense elasticity. The equation below can be used to set optimal marketing spending. However, if quantity sold decreases (when price is relatively stable) and marketing expense is at the optimal amount given elasticity, then marketing spending should be eliminated for that product in that sales area.

   $$E_m = \Delta Q / \Delta M * (M_2 - M_1)/(Q_2 - Q_1)$$

Hypothesized Roadmap:

Round 1

- As per Maxim 1:
    - Do not make more product than can be made in 30 days.
    - Be sure to keep all products to be sold stocked so that you sell all products on all days and so that you have five days of each product at the end of the round.
    - Set lot size at 21,000.

- To fulfill Maxim 2:
    - Decrease prices on three products $$-F01, $$-F02, and $$-F03 in all three distribution channels from the start-of-the-game price + € .75 using decrements of € .05. If products begin selling quickly, and you will run out of product before the end of the round, change your decrement to € .025.
    - Increase prices on the other three product $$-F04, $$-F05, $$-F06 from the starting price - € .75 using increments of € .05. If products begin selling quickly, and you will run out of product before the end of the round, change your decrement to € .025.

Round 2

- As per Maxim 1:
    - Do not make more than 15 days of product in a production cycle.
    - Run business cycle twice. Run the second cycle no later than the ninth day of the round.
    - Be sure to keep all products to be sold stocked so that you sell all products on all days and so that you have five days of each product at the end of the round.
    - Set lot size at 22,000 at end of round.

- To fulfill Maxim 2:
    - Increase prices on three products $$-F01, $$-F02, and $$-F03 from the start-of-the-game price + € .75 using increments of € .05.
    - Increase prices on the other three product $$-F04, $$-F05, $$-F06 from the start-of-the-game price - € .75 using increments of € .05.

Round 3

- As per Maxim 1:
    - Do not make more than ten days of product in a production cycle.
    - Run the production cycle thrice. Run the second business cycle on fourth day of the round and the third no later than the fourteenth day of the round.
    - Be sure to keep all products to be sold stocked so that you sell all products on all days and so that you have five days of each product at the end of the round.
    - Set lot size at 23,000 at end of round.

- As per Maxim 2, compute arc elasticity for prices on the six products and price according to the theoretically correct price. Move to adjusted theoretically correct price if appropriate.

- As per Maxim 3, rank products by profitability.
- As per Maxim 4:
    - Design products to replace the products ranked 5th and 6th most profitable.
    - Eliminate inventory of products ranked 5 and 6 as quickly as possible while still making the most profit possible on each sale, so that the products you are designing can be sold in the next round.

Round 4

- As per Maxim 1:
    - Do not make more than 8 days of product in a production cycle.
    - Run the production cycle four times. Run the second business cycle on fourth day of the round and the third no later than the ninth day of the round, and the fourth no later than the fourteenth of the month.
    - Set lot size at 24,000 at end of round.
- As per Maxim 3, re-rank all six products.
- As per Maxim 4:
    - Manufacture the newly designed products in first business cycle.
    - Sell off least profitable product so that you are making the most profit possible as you sell it off. (Use a price is the highest possible where you expect to sell off all of the product but be sure product is out of stock by the end of round.)
    - Design new product.

Round 5

- As per Maxim 1:
    - Do not make more than 8 days of product in a production cycle.
    - Run the production cycle four times. Run the second business cycle on fourth day of the round and the third no later than the ninth day of the round, and the fourth no later than the fourteenth of the month.
    - Set lot size at 25,000 at end of round.
- As per Maxim 3:
    - Re-rank all six products.
- As per Maxim 4,
    - Manufacture the newly designed product in first production cycle.
    - Sell off least profitable product so that you are making the most profit possible as you sell it off. (Use a price is the highest possible where you expect to sell off all the inventory but be sure product is out of stock by the end of round.)
    - Design new product.
- As per Maxim 5, invest in setup time reductions or loan payment daily.
- To fulfill Maxim 7, increase marketing expenses on .5 kg products in the West and South sales regions in areas at increments of € 5 from € 25.

Round 6

- As per Maxim 1:
  - Do not make more than 8 days of product in a production cycle.
  - Run the production cycle four times. Run the second business cycle on fourth day of the round and the third no later than the ninth day of the round, and the fourth no later than the fourteenth of the month.
  - Set lot size at 25,000 at end of round.
- As per Maxim 3:
  - Re-rank all six products.
- As per Maxim 4,
  - Manufacture the newly designed product in first production cycle.
  - Sell off least profitable product so that you are making the most profit possible as you sell it off. (Use a price is the highest possible where you expect to sell off all the inventory but be sure product is out of stock by the end of round.)
  - Design new product.
- As per Maxim 6, invest in capacity increases or loan payment daily.
- To fulfill Maxim 7, decrease marketing expenses on .5 kg products in the West and South sales regions in areas at increments of € 5 from € 175.

Round 7

- As per Maxim 1:
  - Do not make more than 8 days of product in a production cycle.
  - Run the production cycle four times. Run the second business cycle on fourth day of the round and the third no later than the ninth day of the round, and the fourth no later than the fourteenth of the month.
  - Set lot size at 25,000 at end of round.
- As per Maxim 3:
  - Re-rank all six products.
- As per Maxim 4,
  - Manufacture the newly designed product in first production cycle.
  - Sell off least profitable product so that you are making the most profit possible as you sell it off. (Use a price is the highest possible where you expect to sell off all the inventory but be sure product is out of stock by the end of round.)
  - Design new product.
- As per Maxims 5 and 6, invest in capacity increases, setup time reductions, or loan payment daily.
- As per Maxim 7, compute marketing expense elasticities and set Marketing expenses accordingly.

Round 8

Follow Maxims 1, 3, 4, 5, and 6.

Research Plan

This is pedagogical action research seeking to help instructors use ERPsim more effectively. It seeks to begin to argue for a detailed roadmap for using ERPsim in the classroom by experimentally showing the difference between the learning outcomes achieved when using maxims and a prescribed approach as compared to not. The research question is: Will students that are taught using the maxims and the prescription above (under the auspices of the hypotheses) understand how information systems support business processes better?

One experimenter/instructor will concurrently teach two sections with at least thirty students in each section. One course will use the maxims (treatment). The other course will not (control). Each student will be asked the following quantitative and qualitative questions during the weeks directly before and after the simulation using an online survey given in class. The experimenter/instructor and a proctor will assure that students are not using the Internet as research tools as they complete the survey either via their laptops or phones. Students will not be remunerated for the time/effort they spend on the survey. The survey and the experiment will be approved by the Institute Review Board.

1. To what degree are information systems valuable for organizations?
2. To what degree are information systems helpful for companies?
3. To what degree are information systems useful to managers and employees?
4. To what degree are information systems necessary to support business processes?
5. To what degree are information systems relevant to businesses?1.
6. Why is a business process important?
7. How do information systems enable business processes?

Questions numbered 1 through 5 will be answerable via a six-point Likert scale (no neutral response). If there are 35 to 40 students in each section the author will review for normality using bar/boxplot/stem-leaf charts, analyzing skewedness and kurtosis, as well as the Shapiro-Wilk (1965) and Anderson-Darling (1954) tests. If normality is suggested, the author will compute, compare, and report parametric statistics (e.g., $\bar{x}$, σ, t, r, X^2). The non-parametric ordinal tests applied will be the Spearman rank-order coefficient ρ, Mann-Whitney U, and the Wilcoxon signed rank and rank sum tests.

The data from the Likert questions will also be dichotomized (lower three scale values vs. higher three scale values). The binomial test will be applied. The pre/post treatment/control contingency table will be analyzed using Fisher's exact test, Φ, and Cramer's V. Questions 6 and 7 will be analyzed using the codes below. The dichotomous nominal data will be studied as just described. Inter-coder agreement will be reported using Cohen's Kappa, Krippendorf's Alpha, and Kendall's W.

Question 6: (Why is a business process important?)

- Code 1. Did the student mention that a business process helps an organization provide value to the customer/patron?
- Code 2. Did the student mention that a business process helps an organization make a profit (or more revenue than expenses)?
- Code 3. Did the student mention that a business process helps an organization (or person) do something that is important for the business?
- Code 4. Did the student mention that business process frequently provides an input to or is dependent upon another business process?

Question 7: (How do information systems enable business processes?)

- Code 5. Did the student mention that information systems organize large amounts of data?
- Code 6. Did the student mention that information systems allow employees to do work in the business process that they would not be able to do otherwise?

- Code 7. Did the student mention that information systems help employees do much more work than they could otherwise?
- Code 8. Did the student mention that information systems allow employees to assess and improve their performance?

Conclusion

This paper provides readers not familiar with the ERPsim games a simple overview. It provides a detailed description of the ERPsim Manufacturing Game. Further, it provides a detailed prescription (a set of maxims and a roadmap) for using the simulation. A research plan is framed to answer the research question: "Will students that are taught using the maxims and the roadmap above understand how information systems support business processes better?" If the experiment begins to show that the maxims and roadmap are efficacious, the researcher's contribution will be pedagogy that instructors can apply with non-MIS students to help these students understand the value of MIS, and prospectively, to help these same students engage their MIS course fully.

Acknowledgements

The author thanks Dean John T. Delaney of the Katz Graduate School of Business at the University of Pittsburgh for providing monies to attend ERPsim training at HEC Montreal and the Innovation in Education Committee (of the Katz School) that paid for an initial 30 licenses to pilot ERPsim. Additional thanks go to the staff at the SAP University Alliance Competency Center at the University of Wisconsin-Milwaukee that provides SAP to for university professors.

REFERENCES

Leger, P.-M. 2006. "Using a Simulation Game Approach to Teach Enterprise Resource Planning Concepts," Journal of Information Systems Education (17:4), pp. 441-448.

Leger, P.-M., Babin, R. J., Pellerin, R., and Wagner, B. 2007. ""Erpsim"." Montreal: Montreal HEC, ERPsim Lab.

Leger, P.-M., Robert, J., Babin, G., Pellerin, R., and Wagner, B. 2012-2013. Participant's Guide. Montreal: HEC Montreal.

Shapiro, S. S., Wilk, M. B. 1965. "An analysis of variance test for normality (complete samples). Biometrika. (52:3-4) p. 593.

Anderson, T.W., and Darling, D.A. (1954) "A Test of Goodness-of-Fit." Journal of the American Statistical Assocation. (49) pp. 765-769.

Section G – Sketch: Designing a complex algorithm

This section describes a sketch that describes designing a complex algorithm. Sections F and I provide examples of designing, developing, implementing, and assessing an algorithm (which complement their content as indicated in their section headings).

I am clarifying this section and will resubmit under separate cover later during the work week beginning 12/14/15.

Section H - Example: Harmonizing with researchers in the Department of Defense

This section describes an instance of when I harmonized with a Department of Defense (DoD) researcher.

(Please turn the page.)

Note while this is my only example of interacting with a DoD party, I have interacted with five National Science Foundation supported researchers substantively, and a very large number of NSF supported scientists in less substantive ways.

In 2010, I brought and demonstrated an extraordinarily crude computational linguistic model of complex problem solving to the North American Association for the Computational Social and Organization Sciences (NAACSOS) annual meeting. Almost immediately after the demonstration I was approached by a program officer at the United States Office of Naval Research. We had lunch. Despite my grievances at the lack of fidelity in the model, he strongly encouraged me to pursue the linguistic/semantic/cognitive approach, because, in his (paraphrased) words, “he did not know of a researcher that had chosen a path even vaguely familiar” to mine. He also indicated his belief in the generalizability of the approach. Unfortunately, I could not follow through on his suggestion given the positions I held and the responsibilities they entailed.

Note that the sketch of candidate theories and tools described in Section H is a tremendously more thoughtful design than what underlay the implementation I demonstrated at NAACSOS 2010. Note also that the implementation I presented at NAACSOS 2010, in some respects, was much more meaningful than the simulation I shared in my Ph.D. dissertation.

Section I - Example: Teaming with subject matter experts
(and designing, developing, implementing, and assessing
a complex algorithm)

This section focuses on explaining how I have teamed with subject matter experts, in a particular project where, we, collaboratively, designed, built, assessed, and reported an immersive learning experience.

Another point where I teamed with subject matter experts is when I collaborated with the associate registrar (and others) at Rensselaer Polytechnic Institute to extract, transform, and load 1.4 million student/course records. A person who can discuss my ability to team with subject matter experts is Ms. Sharon Kunkel, the Registrar at Rensselear Polytechnic Institute, and who was the supervisor of the associate registrar. I also had a "dotted line" reporting relationship to Sharon. Sharon has agreed to field an email or a phone call to describe my work ethics. If necessary and appropriate, I encourage you to reach out to Sharon. She can be reached by email at kunkes@rpi.edu or by phone at 518-276-6028.

(Please turn the page.)

In the "business" of teaching, a complex problem is helping students who have little real world business experience understand or imagine the environments where business opportunities or problems occur, as these students learn a new skill. I co-led a team that built the first immersive[1] business case[2].

By teaming with subject matter experts[3], I was able to learn about:

- medical device manufacturers' operational processes and constraints
- computer games and how they can be used to help learning
- processes used by actors before, during, and after acting
- techniques used by graphic artists
- methods used to measure learning as students solve complex problems

I thought critically as I:

- evaluated products and services by installing, configuring, and using these
- embedded constraints that directed future unknown activities of learners
- provided structure to actors while encouraging careful improvisation
- framed my requests to consultants to make results tangible and reusable
- built highly detailed measures to provide transparency during assessment

I solved a complex problem, in carefully separated pieces, when I:

- constructed a 3D model of a medical device manufacturer's facilities
- imagined a series of activities that students would experience
- made "in game" gadgets that could guide students
- devised "in game" items that students would need to transform
- proposed answers actors could provide to anticipated student questions when actors played antagonist and protagonist roles "in game"

[1] "Immersive" in this sense refers to the concept of "being immersed in the task of learning."

[2] A "business case" is a multipage document which business school instructors ask students to read. Students then are expected to create solutions and share these in class. The purpose of the case is to help students practice a skill recently expressed in class. A "fundamental flaw" of business cases is the fact that they are abstracted real life situations. Further, the discussion of the abstract business cases, in the classroom is abstract.

[3] This example also indicates my ability to create, think critically, and solve complex, difficult to understand problems. It also shows how I have brought innovative and cutting edge technology and techniques to the table.

1-1-2010

VIRTUAL TEACHING CASES?AN EXPLORATORY STUDY

Russell W. Robbins
University of Pittsburgh, rrobbins@katz.pitt.edu

Brian S. Butler
Univrsity of Pittsburgh, bbutler@katz.pitt.edu

VIRTUAL TEACHING CASES? AN EXPLORATORY STUDY

Research-in-Progress

Russell W. Robbins
Brian S. Butler
Katz Graduate School of Business
University of Pittsburgh
rrobbins@katz.pitt.edu, bbutler@katz.pitt.edu

Abstract

This research, when complete, will represent a prototype of the development of a virtual teaching case and the use and assessment of the initial versions of research instruments whose aim is the assessment of this new form of teaching case, or any type of teaching case, with regards to learning efficacy, gains, satisfaction, and environment. The purpose of this virtual teaching case (that is, a teaching case, embedded within a virtual world) is to leverage the rich heritage of case-based teaching while helping today's students to learn by providing a more engaging environment where these students (experienced with multiplayer computer games and the Internet) can collaboratively practice project management skills such as planning scopes of work, schedules, and budgets—skills they have already learnt in class. In a virtual teaching case, students can experience the challenges of discovering problems; collaboratively creating, judging, and transforming resolutions; and reacting to changing circumstances.

Keywords: Teaching case, learning, evaluation methods and criteria, IS education, virtual world, project management, problem solving

Introduction

In the virtual teaching case research and development project reported here and elsewhere we have designed and built a new platform for learning ill-structured problem solving generally, and project management planning specifically, using a 3D immersive virtual world (Robbins and Butler 2009a, 2009b). In this section we summarize the virtual teaching case. In the next section we describe the theoretical justification of our approach. In the research, development, and teaching model section we first describe our desired learning outcomes and then our approach, with concrete examples at each step. Finally, in the research measures for teaching case quality section we describe how we plan to answer four research questions that focus on providing transparency into the virtual teaching case's learning efficacy and student satisfaction, as well as how traditional versus virtual teaching cases compare from a student's point of view.

The purpose of this virtual teaching case (that is, a teaching case, embedded within a virtual world) is to help students learn by providing an engaging environment where students can collaboratively practice project management skills such as planning scopes of work, schedules, and budgets—skills they have already learnt in class. Virtual worlds (VW) allow users to represent themselves as 3D animations (known as avatars); communicate with other users' avatars; and build, change, and travel within their 3D computer-game-like environments (Messinger et al. 2009). VWs show promise as a method for "enhancing, motivating, and stimulating learners' understanding of certain events, especially those for which the traditional notion of instructional learning have proven inappropriate or difficult," such as the teaching and learning of project management (Bares et al., 1998; Malone and Lepper, 1987; Pan et al., 2006; Phang and Kankanhalli 2009). The goal of our overall program is to enhance educational capabilities by developing technology and techniques for using VWs to provide realistic but safe practice scenarios for collaborative learning that can be used in combination with traditional methods such as classroom discussions, textbook reading, and traditional homework. In a VW, students can experience the challenges of discovering

problems that need addressing, collaboratively creating, judging, and transforming potential resolutions, and reacting to changing circumstances.

Although the scope of this project has primarily been limited to the MIS foci in the graduate and undergraduate business schools at the authors' university and one other university, the potential uses of interactive VW technologies and associated techniques as an educational platform is much broader. For example, while organizational problem solving (such as IT project management or IT opportunity identification) is central to the educational programs offered by business schools, there is increasing interest in other areas, including law, social work, government, and medicine to develop their students' capabilities to recognize, engage, and collaboratively address corollary complex organizational problems. In addition, VWs can support various modes of distance learning such as synchronous text/voice/video-based discussions, interactive teamwork of geographically distributed students using shared software applications, or embedded scenarios that allow teams of students to discover, react, or plan in context, and to some extent, in situ (Bronack et al., 2008; Lamont, 2007; Virtual World News, 2007). The development of interactive VWs with these foci and abilities to support education open up the possibility of interactions among students in schools not usually bridged or integrated (e.g., leadership and philosophy, information systems and computer science students, or between students at similar schools at different universities). In fact, in this project, during one of three pilots, IT project management students at the authors' mid-Atlantic United States' university and at a Southeastern US university collaborated. Therefore, there is broader potential impact for this innovative project which seeks to complement other IS education research (Borrajo et al., 2010; Dreher et al., 2009; Harris and Rea, 2009; Gupta and Bostrom, 2009; Keller, 2009; Law, 2007; Shen and Eder, 2009; Topi et al., 2010; Wagner and Ip, 2009; Wang and Brahman, 2009; Wu et al., 2010) and develop a framework for developing, using, and assessing VWs as learning environments in which students may practice their collaborative organization problem solving skills in safe but realistic contexts.

The virtual teaching case reported here asks students to role play, in teams, inside a virtual world, and to be part of a project management consulting firm that has been hired by a sterile disposable medical device manufacturing firm (The Trilleum Corporation) to restart their manufacturing operations relocation project and assure that their stalled project is completed successfully. By the end the virtual teaching case, the students, in teams, develop and present a plan within the virtual world that shows their client organization (represented by characters/avatars that are "brought to life" by professors, working professionals that are alumni, or professional actors) that they (the students acting as consultants) can help the client organization manage the scope, schedule, and budget of the manufacturing operations relocation project. The case begins when the client organization's managers indicate to the student consultants that the Trilleum Corporation, which manufactures intravenous catheters, blood transfusion kits, wound drainage sets, etc., was recently sanctioned by the United States Food and Drug Administration (FDA) and the corollary organization in the EU, the European Medicines Agency (EMEA), for not using Good Manufacturing Processes (GMP). Further, by mandate of the EMEA and FDA, the Trilleum Corporation must geographically transfer and integrate three manufacturing processes per an agreed upon and validated design by July 4th in order to avoid fines, avoid an otherwise mandatory shutdown, and continue to sell sterile medical disposables. Note we created a non-IT-centric case in order to make this case usable to all kinds of business students, not just MIS students. Also note that while some of the goals of the project are clear, for example—the geographical transfer and physical integration of three manufacturing lines into one line that is EMEA and FDA GMP-compliant by end of the July 4th shutdown week, there are many goals or activities within the problem that have to be identified, considered, ranked, and integrated by the students, such as addressing sales demand concurrent to operations integration, balancing competing stakeholder interests, and managing outsourcing contracts for parties involved in the move. Finally, note that since the problem is not fully defined, the planning process begins with a complex problem that must be structured.

Within the VW, the students collaborate virtually with others using shared applications, voice and text chat, and physical interactions (e.g., focused attention, waving). The VW contains virtual buildings, furniture, files, etc. One of two virtual buildings provides student consultants with a virtual consulting practice while the other is the student consultants' client's offices and manufacturing facilities. The virtual buildings are within a virtual city and students can move from one building to the other more traditionally, by "walking," or by taking advantage of features only available in virtual worlds such as teleporting from location to location. The students' virtual consulting practice contains a working conference room, a "war room," individual problem solving spaces, a lobby, and the consulting partners' office. While working in the virtual teaching case as consultants, students are able to, as necessary, obtain advice from a senior consulting partner—a character in the virtual case—that is played by the instructor of the

students that are using the Trilleum case. When the student consultants "arrive" at the Trilleum Corporation for the first time, they are brought to the Trilleum conference room and Trillieum's ill-defined problem is presented. Students and their avatars are introduced to avatars that represent Trilleum managers—played by the course instructor or other professors, alumni—especially those with operations or project management experience, or professional actors. Following an initial presentation, the students are shown the current and new manufacturing "spaces" at the virtual Trilleum Corporation.

After this first interaction with the client, the students move back to their virtual consulting practice and determine who, they, as a team, will interview at the client's location and what questions to ask while they are interviewing any particular Trilleum manager. The Trilleum managers that the student consultants are able to interview are Estella Hernandez—VP of Operations, Bill Rapinalo—Director of Manufacturing, Jeff Goldstein—Supply Chain Manager, Jorge Gonzalez—Maintenance Manager, Consuela Rodriguez—Production Manager, Sam Weyland—Facilities Manager, and Steve Gordon—the Quality and Safety Manager. Each of these virtual client managers has a virtual office at the Trilleum Corporation where s/he can meet with student consultants, and each office may have editable word processing, spreadsheet, or project management files, such as the planned manufacturing operations relocation design, information about what manufacturing processes have been moved already, as well as information about particular sterile disposables' demand and line capacity.

While the case is about applying and practicing project management planning skills, the students are provided information about the Trilleum Corporation's operations. We included this additional, extraneous information so that the students would experience a real, and complicated, context. In fact as a result of our second pilot, we learned that students can become significantly focused on understanding the operations when what is important for their project planning purposes is their focus on the logistical transfer of machines per an already agreed upon operations design, identifying persons qualified to validate installations, assuring additional movers, etc. Note that these kind of student experiences (e.g., where students believe they need certain information when they actually do not) provide an opportunity for a post-case discussion among the instructor and students about focusing upon the correct information in a problem or the need to refine the problem statement (planning the project) prior to starting other problem solving activities, and that this kind of discussion is not piqued by traditional, snapshot-of-the-past, paper-based cases. After interviewing Trilleum managers, student consultants then retreat to their consulting practice, develop early shared mental models of the problem context, re-interview clients at the client's site as necessary, and use educational scaffolds (such as samples of work breakdown structures (WBS), schedules, and budgets). At the end of the case, the students present their plans and the client asks the students to rework the plan on specific points that the students did not consider – often because the students, did not interview the client managers (again, played by professors, alumni, and professional actors) effectively. At the close of the case, the students' instructor shares an ideal (but non-unique) solution to the problem and leads a virtual discussion.

Theoretical Justification

The foundation for this project is the case method of teaching and learning. The case method "enables students to discover and develop their own unique framework for approaching, understanding, and dealing with business problems" (Barnes et al., 1994, p. 42). The case method supports experiential, active, and collaborative learning (Heckman and Annabi, 2006). Further, it supports teaching principles, concepts, morals, ethics, strategies, dispositions, and "images of the possible" (Shulman, 1992, p. 3). It helps students learn how to encapsulate a problem, see the inter-relatedness of organizations and processes, and take responsibility in their decision making (Barnes et al., 1994). Kerr and colleagues (2003) indicate that students playing roles in cases report that their learning is enhanced.

Traditional text-based presentation of case materials to individuals, followed by class discussions in the abstract, reduces the effectiveness of the case method for teaching students how to engage complex, organizationally situated project planning, when compared to modern technologies that can support the case method of teaching. The bounded and focused nature of classic written case descriptions eliminates some of the challenges associated with collaborative problem identification. It also reduces some of the ambiguity associated with evidence and argumentation that are common when dealing with planning in organizational settings. Lastly, the linear-bounded nature of reading largely eliminates the interactive, exploratory aspects of organizational decision making, such as those necessary when planning projects. Thus, while the case method is a powerful tool for teaching students how to engage complex problems, traditional case delivery vehicles are subject to significant limitations as a basis for

experiential learning related to organizational problem solving. VWs have features that can be used to augment case-based learning of problem solving and enable more of the active, constructive, collaborative, intentional, complex, contextual, conversational, and reflective activities called for by problem-solving education researchers (Jonassen, 2006), VW learning pioneers (Bronack, 2008), and others (Spiro et al., 1992; Whitehead, 1929/1985). Key VW features include context-situated knowledge spaces, a communicating community, active actions, and facility toolkits (Pan et al., 2006). Knowledge spaces provide information that can help the learner as well as the teacher. These include embedded learning resources such as conceptual definitions, evaluation tools which track how a student or student team arrives at a decision and tutorials or other educational scaffolds that help students with a task the first time they perform it. Communicating communities enable all students to interact, not just those strong and comfortable oral performers during class discussions. These communities include tools such as text or voice chat, email, discussion boards, and support for gesturing. Active action is facilitated by tools that allow learners to act as intensive information providers, problem finders, question answerers, issue analyzers, and solution synthesizers. For more on selecting the appropriate VW for your educational purpose, see Robbins and Butler (2009a) and (2009b).

However, the use of VWs for education in isolation does not naturally lead to learning (Cai et al., 2008; Lakkala et al., 2007; Wells et al., 2008; Windschitl and Sahl, 2002). Therefore, in order to adapt to using new technology within the classroom, careful pedagogical thought about how the technology is to be integrated into the classroom must occur (Badge et al., 2005; Lakkala et al., 2007). One pedagogical approach that can be applied in the context of virtual learning is progressive inquiry learning (Hakkarainen, 2003; Muukkkonen et al., 2005). Progressive inquiry learning focuses on students developing their own questions and creating their own explanations prior to the use of an authoritative source. Progressive inquiry learning can be applied to the learning of solving ill-structured problems, such as the process of developing plans for projects. In order to develop a virtual teaching case that was as exemplary of the ill-structured problem solving that is ubiquitous in industry, but which is scant in our classrooms, and which was apropos for our learning outcomes and educational purposes, we grounded our research, development, teaching, and assessment in the instructional design models suggested by Jonassen (1997) and Choi and Lee (2009).

Research, Development, and Teaching Model

With regards to our desired learning outcomes (Table 1) we seek to help our students use multiple perspectives as they develop project plans. The multiple perspectives that the students should learn to apply include the perspectives of the various managers at the client they are engaged with. Further, the students should apply other perspectives, such as those of their managers or their colleagues within their virtual consulting practice. Finally, if there are other dominant stakeholders, our students should be able to take their perspectives as well – in this teaching case two other dominant stakeholders are the EMEA and the US FDA. We also seek to help our students develop their abilities to justify their identified problems and/or their identified solutions. One problem that can be identified and

Table 1. Desired Learning Outcomes (Adapted from Choi and Lee (2009).

Skill	Literature Support
When Identifying Problems or Generating Solutions, Students Should	
Use Multiple Perspectives	Dewey, 1933; Fleischmann et al. 2009; Jonassen, 1997; Schraw et al., 1995; Shin et al., 2003; Zeichner and Liston, 1996.
Justify Problem	Harrington et al., 1996; Jonassen 1997; Shin et al., 2003; Sinnot, 1989; Voss et al., 1991; Zeichner and Liston, 1996.
Think Critically	Schraw et al., 1995; Zeichner and Liston, 1996.
Use Theory	Bransford, 1993; Chi et al., 1988; Schraw et al., 1995; Shin et al., 2003.

justified in the case is the inability of the client organization (Trilleum) managers to work cooperatively on an ill-defined project – as opposed to their well-structured daily operations. We seek to help our students think critically. For example, students working this case need to come to the realization that the case is about project management, NOT operations management, even though the case is fraught with operations production information. Finally, we seek to help our students apply theory. A core "theory" we ask our students to apply is the Project Management Body of Knowledge (PMBOK). As the designers and developers of the learning environment we first *articulated the problem context* of the virtual case (Figure 1, Activity 1). In order to articulate the problem context (or the setting of

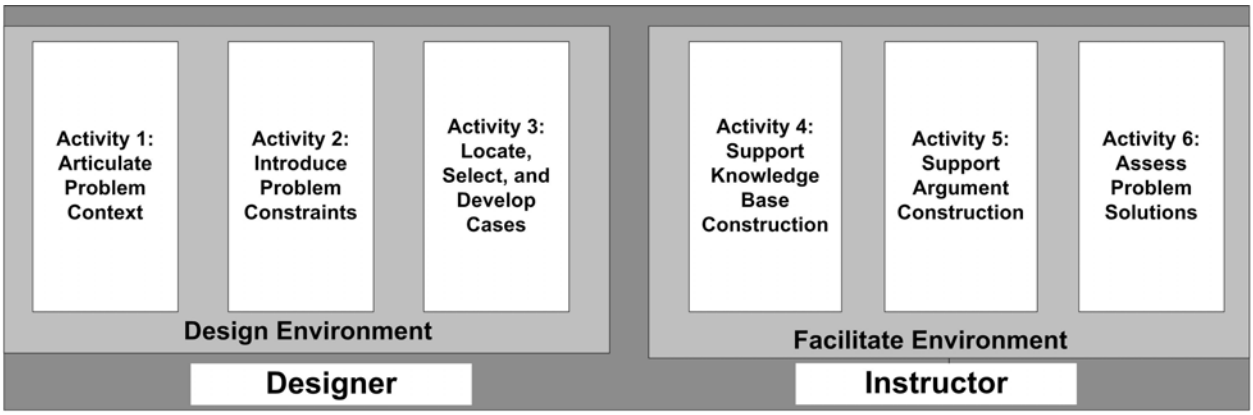


Figure 1: Designing and Instructing Activities (Adapted from Jonassen (1997)).

the problem scenario) we interviewed a subject matter expert (a management consultant who had worked on a similar project 20 years ago) and developed virtual work spaces that are representative of actual real world projects. After we articulated the problem context, we *introduced problem constraints* (Figure 1, Activity 2). Problem constraints in project planning include needs of the various stakeholders who in this case were the seven Trilleum Corporation managers as well as the EMEA and the US FDA. For example, in this case, the Director of Manufacturing is interested in meeting demand, among other things, while the Quality and Safety Manager is interested in assuring that the new integrated processes meet EMEA and US FDA GMP and the VP of Operations is primarily interested in a successful move by the end of the July 4th shutdown.

We further *developed the case* (Figure 1, Activity 3). We did this by determining how the managers at the Trilleum Corporation would be most invested in such a project. Additionally, and based upon feedback from our professional actors in our second pilot, we developed "personas" for each portrayed manager; these personas included detailed job descriptions, "a day in the life of" information for each of the Trilleum managers' roles, information about the managers' genders, ages, hobbies, each manager's last three "positions" accompanied by their affiliated responsibilities, each manager's primary expertise or talent, the biggest crises experienced by each as well as the largest contributions provided by each manager, as well as examples of each of the managers "in action." This contextual information was provided in addition to information about what the professor, alumni/working professional, or professional actor (playing the manager) should indicate when student consultants asked particular questions as well as what the enacted managers' responsibilities were in the geographical transfer and physical integration of three manufacturing operations. Also, to help the students develop their ability to forage information, we built our scripts so that the information that was provided to the students was only provided (by the portrayed managers) if the students requested it, and so that the information would need to be judged, refined, and integrated. For example, some information provided by some managers to student consultants was less than accurate and some information in the case was contradictory (in line with non-virtual reality). We encourage students to develop their abilities to think critically based upon their analyses of their interactions with the manager characters in the case.

Moving from our role as designers to instructors, and to *enable the students' knowledge-base construction*, (Figure 1, Activity 4) we created information artifacts that were embedded within the virtual Trilleum Corporation building. These information artifacts, for example, provided tangible information about the numbers of different types of workstations/fixtures and floor machines within particular current productions lines (dialysis sets, arterial closures, etc.), how maintenance was performed for any particular line, or where production lines would geographically exist before and after their transfer and integration. Supporting the students' *argument construction*, (Figure 1, Activity 5) we provided educational scaffolds within the Consulting Firm's (as opposed to the Trilleum Corporation) virtual building (Linn 1995, Ge and Land 2003, 2004). Note that part of the architecture of the virtual case was supporting problem identification activities in the Trilleum Corporation virtual building while supporting solution generation and consideration abilities in the Consulting Firm's virtual building. Example educational scaffolds (primarily for solution generation and consideration) included sample WBS, schedules, and budgets. Further, other step by step protocols were provided, such as the prescribed steps to develop a proposal. Finally as the students present their solutions to the Trilleum Corporation manager characters towards the end of the case, these characters, represented as avatars played by professors, alumni working professionals, or professional actors, *assess the problem solutions*, (Figure 1, Activity 6) using a rubric provided by the authors.

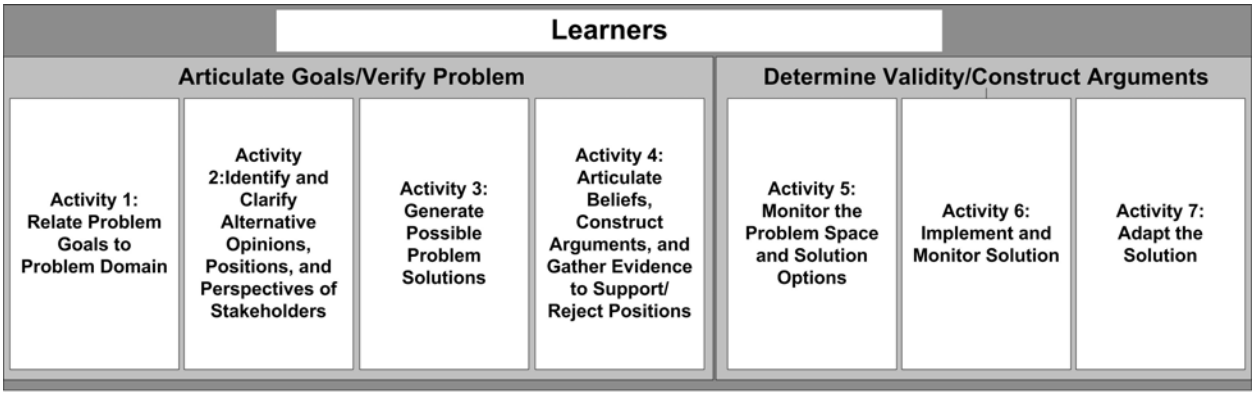


Figure 2: Student Learning Activities (Adapted from Jonassen (1997)).

Moving to Figure 2, Activity 1, we purport that students *relate problem goals* (plan scope, schedule, and budget) *to a problem domain* (the Trilleum Corporation which seeks to geographically transfer and operationally integrate three manufacturing processes by the end of the July 4th shutdown week in order to acquiesce the EMEA and the US FDA. The students also *identify and clarify alternative opinions, positions, and perspectives of stakeholders* (Figure 2, Activity 2) after the student consultants interview up to seven portrayed managers and review GMP as indicated by the EMEA and the US FDA. The students, then, collaboratively, using tools such as private/public text or audio chat, MS Word, MS Excel, Whiteboards, and Shared Desktops provided in our chosen virtual world platform, *generate possible problem solutions* (Figure 2, Activity 3). These include planning activities on weekends, outsourcing the move to professional movers, and focusing on high risk activities first. Then using the information they have gathered from their interviews, interacting with information artifacts, and clarifying their understanding via educational scaffolds, the students *articulate beliefs*, (Figure 2, Activity 4) such as the importance of focusing on project planning as opposed to operational quality, *construct arguments*, (Figure 2, Activity 4) by referring to the [given] fact that their consulting organization that was engaged by the Trilleum Corporation has experience with project management but not with the assurance of quality in medical device manufacturing, and *gather evidence to support or reject positions* (Figure 2, Activity 4) by interviewing Trilleum Corporation manager characters (Andriessen 2006). The case is also designed so that the students can *monitor the problem space and solution options* (Figure 2, Activity 5) by interacting with characters, information artifacts, and educational scaffolds, *implement and monitor a solution*, (Figure 2, Activity 6) by presenting a plan which includes a prospective WBS, schedule, and budget, and *adapt the solution* (Figure 2, Activity 7), when Trilleum manager characters use a rubric provided by the case authors in order to indicate areas in student consultant plans that can be considered or developed further. Finally, at the end of the virtual case, the students learn about one ideal solution to the case, developed by the case authors.

Research Measures for Teaching Case Quality

Our plan for measuring the case's quality is based on Choi and Lee (2009) and Chou and Liu (2005). Choi and Lee (2009) report the design, implementation, and evaluation of an online case-based learning environment for enhancing ill-structured problem solving, and their research model is based on the same instructional design model (Jonassen 1997) that we have used (Figures 1 and 2). Our research questions are in Table 2. The experiment control group will be composed of students learning (by practicing) using written cases as teams. The treatment group will be composed of students learning (by practicing) using the virtual teaching case as teams. Note that Research Questions 1 and 2 focus on understanding whether the virtual teaching case is effective in phases or when comparing the students' learning as a whole to a control, and more traditional, case-based learning environment. Both Research Questions 1 and 2 will be addressed by asking instructors and project managers to apply rubrics (Tables 3 and 4) as expert judges/informants across the treatment/control teams and virtual case stages. We will then compare the degree to which the teams of students in the treatment and control groups applied concepts/skills they learned prior to the experiment, and which are the basis for the rubrics and hence the instructors' and project managers' judgments. Note that Research Questions 3 and 4 are focused on the learning environment and learning satisfaction and therefore these questions are addressed by students completing evaluation surveys and our subsequent analyses of their answers. Research Questions 3 and 4 control/treatment measures are in Tables 5 and 6. Some student teams will be assigned the virtual case treatment condition and others will be assigned to the written

case control. Upon completing the case, all participating students will complete an evaluation survey using a 1-7 Likert scale, including the questions in Tables 5 and 6. The mean answers for students in treatment/control groups will be compared.

Table 2. Research Questions Adapted from Choi and Lee (2009) and Chou and Liu (2005).	
RQ1	Do particular learning activities and the affiliated learning objects in the virtual case improve students' ability to plan scopes, schedules, and budgets? (We will test the gain effects associated with each stage.)
RQ2	Does the overall learning experience (using the VW and the embedded case) improve students' ability to plan projects in a transfer test, when compared to classroom discussions of the written version of the case?
RQ3	Do students who learned in the virtual world report higher levels of satisfaction than their counterparts in classroom discussions of written cases solved by teams of students?
RQ4	What do students who learned in the virtual world report with regards to their learning environment when compared to their counterparts in classroom discussions of written cases solved by teams of students?

Table 3. Rubrics for Research Questions 1 & 2: Learning Efficacy (To Assess Impacts on Students)	
1	To what extent did the students recognize/implement concepts affiliated with the PMBOK Process Groups?
(a)	And concurrently user multiple perspectives? (repeats for #2 through 18 below)
(b)	And concurrently justify their identification of a problem or a solution? (repeats for #2 through 18 below)
(c)	And concurrently think critically? (repeats for #2 through 18 below)
2	To what extent did the students recognize/implement concepts affiliated with PMBOK Knowledge Areas?
3	To what extent did the students recognize/ implement concepts affiliated with the 42 PMBOK processes?
4	To what extent did the students recognize and as appropriate, implement industry specific concepts?
5	To what extent did the students recognize and use WBS Components?
6	To what extent did the students recognize and use WBS Work Packages?
7	To what extent did the students recognize and use WBS Codes?
8	To what extent did the students implement appropriate size activities?
9	To what extent did the students recognize and use Schedule Activities?
10	To what extent did the students recognize and use Schedule Activity Dependencies?
11	To what extent did the students recognize and use Lag and Lead times as appropriate?
12	To what extent did the students recognize and use Schedule Activity Effort and Duration as appropriate?
13	To what extent did the students recognize and use Human Resources?
14	To what extent did the students recognize and use Material Resources?
15	To what extent did the students recognize and attach Human Resources to Schedule Activities?
16	To what extent did the students recognize and attach Material Resources to Schedule Activities?
17	To what extent did the students recognize and set a Baseline?
18	To what extent are the students' WBS, Schedule, and Budget appropriate in terms of scope, time, and cost?

Table 4. Rubrics to measure RQ 1 Gains across VW Case Phases Adapted from Choi and Lee (2009).	
1	To what extent do students justify problems after viewing the initial presentation in the virtual case?
2	To what extent do students use multiple perspectives represented by the characters interviewed in the case?
3	To what extent do students apply theory represented by educational scaffolds that exist in the case?
4	To what extent do students think critically as they collaborate to meld problems, perspectives, approaches?
5	To what extent do students provide a project management plan solution in response to feedback?
6	To what extent are students presented with and discuss an ideal (but non-unique) solution?

Table 5. Measures for Research Question 3: Learning Satisfaction (to be completed by students) Adapted from Chou and Liu (2005).	
1	I am/was satisfied with this learning experience.
2	I am/was satisfied with how I was able to acquire information in this experience.
3	I am/was satisfied with the flexibility in how I could learn in this experience.
4	I am/was satisfied with the level of independence I had in this experience.
5	I am/was satisfied with the instruction provided with this experience.

Table 6. Measures for Research Question 4: Learning Environment (to be completed by students) To what extent did you, as the student…		
1	Find materials that helped you develop the WBS, schedule, and budget?	Williams, 1992
2	Find that the materials helped you develop a strategy to plan the project?	Williams, 1992
3	Receive feedback to assess how well you were learning?	Williams, 1992
4	Provide the teacher with information to assess your learning?	Williams, 1992
5	Find that this case was realistically complex?	Williams, 1992
6	Think that the complexity in the case was manageable?	Williams, 1992
7	Experience "bite-size" pieces of the case when creating your solution?	Williams,1992; Nelson et al. 2008
8	Find the case setting rich and detailed?	Williams, 1992
9	Have the opportunity to actively engage solving problems?	Hackney et al., 2003
10	Find the case authentic?	Hackney et al., 2003
11	Have the opportunity to identify the underlying issues?	Stepich et al., 2001
12	Have the opportunity to clarify the problem(s) in the case?	Stepich et al., 2001
13	Consider multiple factors in tandem?	Stepich et al., 2001
14	See multiple perspectives from various characters?	Stepich et al., 2001
15	Allowed to evolve your solution?	Stepich et al., 2001
16	Consider potential consequences and the implications these might have?	Stepich et al., 2001
17	Develop your ability to reason through a problem to a solution?	Hackney et al., 2003
18	Consider potential impacts upon the client organization?	Hackney et al., 2003
19	Required to make your own decisions?	Keefer, 2005
20	Self-reflective, as you completed this case?	Keefer, 2005
21	Collaborate with your team?	Keefer, 2005
22	Motivated to seek out new knowledge and develop new skills?	Law, 2007
23	Empowered to use alternative means to complete tasks in the case?	Law, 2007
24	Have opportunities to learn from other students' solutions?	Law, 2007
25	To what extent did this questionnaire assess this case?	Williams, 1992

Implications, Limitations, and Future Research

This research, when complete, will represent a prototype of the development of a virtual teaching case and the use and assessment of the initial versions of research instruments whose aim is the assessment of this new form of teaching case, or any type of teaching case, with regards to learning efficacy, gains (across stages of interaction with a case), satisfaction, and environment. While similar cases have been developed, we are unaware of any that will have gone through this intense scrutiny. We contemplate that this paper provides a method for evaluating teaching cases of any form. Perhaps with tools represented in this paper, scholars will continue to evaluate their teaching cases, in order to assure the best possible student learning.

We hope that this virtual case, as has been intimated in three pilots (not reported here), proves to be efficacious. However, this case, as with any teaching case is limited (or not limited) by the abilities of the teachers and students that use that case. As this project closes, we will report this research and provide and disseminate a written teaching case as well that will be used as an experimental control during our data collection and analysis. We also plan to begin developing a second virtual case–a project that was recently funded by our provost. This second virtual case will focus on helping undergraduate students in our introductory management information systems course learn to understand business processes, identify opportunities to use IT to improve/eliminate these business processes, and build arguments that will allow them to obtain financial support for their own (in the future) identified IT opportunities. Finally, we seek to merge the first author's and his colleagues' software agent and online teaching case research with this project (Robbins, 2005; Robbins and Wallace, 2007; Robbins et al., 2009). We would appreciate any suggestions. The materials for this case are freely available. Please contact the first author.

Acknowledgement

We thank Dean John T. Delaney for financial support via the Katz Team Technology Innovation project. We thank the anonymous undergraduate and MBA students at two universities that have participated in three pilots.

References

Andriessen, J. 2006."Arguing to Learn," in *The Cambridge handbook of the learning sciences*, R.K. Sawyer (ed.), Cambridge University Press, Cambridge U.K. pp. 443-459.

Badge, J. L., Cann, A. J., and Scott, J. 2005. " E-learning versus e-teaching: Seeing the pedagogical wood for the technological trees," *Bioscience Education*, (5). Online at http://www.bioscience.heacademy.ac.uk/journal/vol5/beej-5-6.pdf.

Bares, W. H., Zettlemoyer, L. S., and Lester, J. C. 1998. "Habitable 3D learning environments for situated learning," in *Lecture Notes In Computer Science; Vol. 1452, Proceedings of the 4th International Conference on Intelligent Tutoring Systems*, pp. 76-85. Retrieved 01/26/09 from: http://people.csail.mit.edu/lsz/papers/bzl-its-98.pdf.

Barnes, L. B., Christenson, C. R., and Hansen, A. J. 1994. *Teaching and the case method: Text, cases, and readings, third edition*, Boston, MA: Harvard Business School Press.

Bransford, J. D. 1993. "Who ya gonna call? Thoughts about teaching problem solving," in *Cognitive perspectives on educational leadership*, P. Hallinger, K. Leithwood, and J. Murgh (eds.), New York, NY: Teachers College Press, pp. 171–191.

Borrajo, F., Bueno, Y., de Pablo, I., Santos, B., Fernandez, F., Garcia, J., Sagredo, I. 2010. "SIMBA: A simulator for business education and research," *Decision Support Systems* (48:3), pp. 498-506.

Bronack, S., Sanders, R., Cheney, A., Riedl, R., Tashner, J., and Matzen, N. 2008. "Presence pedagogy: Teaching and learning in a 3D virtual immersive world,"*International Journal of Teaching and Learning in Higher Education* (20:1), pp. 59–69.

Cai, H., Sun, B., Farh, P., and Ye, M. 2008. "Virtual Learning Services over 3D Internet: Patterns and Case Studies," in *Proc. 2008 IEEE International Conference on Services Computing* (2), pp. 213-219.

Chi, M. T. H., Glaser, R., and Farr, M. J. 1988. *The nature of expertise.* Hillsdale, NJ: Lawrence Erlbaum.

Choi, I., and Lee, K. 2009. "Designing and implementing a case-based learning environment for enhancing ill-structured problem solving: Classroom management problems for prospective teachers," *Educational Technology Research and Development* (57:1), pp. 99–129.

Chou, S., and Liu, C. 2005. "Learning effectiveness in a Web-based virtual learning environment: A learner control perspective," *Journal of Computer Assisted Learning* (21), pp. 65–76.

Dewey, J. 1933. *How we think: A restatement of the relation of reflective thinking to the educative process.* Lexington, MA: Heath.

Dreher, C., Reiners, T., Dreher, N., and Dreher, H. 2009. "Virtual Worlds as a Context Suited for Information Systems Education: Discussion of Pedagogical Experience and Curriculum Design with Reference to Second Life," *Journal of Information Systems Education* (20:2), pp. 211-224.

European Medicines Agency. (last accessed May 2, 2010). http://www.ema.europa.eu/Inspections/GMPhome.html.

Fleischmann, K. R., Robbins, R. W., and Wallace, W. A. 2009. "Designing educational cases for intercultural information ethics: The importance of diversity, perspectives, values, and pluralism," *Journal of Education for Library and Information Science* (50:1), pp. 4–14.

Ge, X., and Land, S. M. 2003. "Scaffolding students' problem-solving processes in an ill-structured task using question prompts and peer interactions," *Educational Technology Research and Development* (51:1), pp. 21–38.

Ge, X., and Land, S. M. 2004. "A conceptual framework for scaffolding ill-structured problem-solving processes using question prompts and peer interactions," *Educational Technology Research and Development* (52:2), pp. 5–22.

Gupta, S. and Bostrom, R.P. 2009. "Technology-Mediated Learning: A Comprehensive Theoretical Model," *Journal of the Association for Information Systems* (10:9), Article 1.

Hackney, R., McMaster, T., and Harris, A. 2003. "Using cases as a teaching tool in IS education," *Journal of Information Systems Education* (14:3), pp. 229–234.

Hakkarainen, K., Palonen, T., Paavola, S., and Lehtinen, E. *Communities of networked expertise: Professional and educational perspectives.* Amsterdam, Elsevier, 2004.

Harrington, H. L., Quinn-Leering, K., and Hodson, L. 1996. "Written case analyses and critical reflection," *Teaching and Teacher Education* (12:1), pp. 25–37.

Harris, A.L. and Rea, A. 2009. "Web 2.0 and Virtual World Technologies: A Growing Impact on IS Education," *Journal of Information Systems Education* (20:2), pp. 137-144.

Heckman, R., and Annabi, H. 2005. "How the teacher's role changes in on-line case study discussions," *Journal of Information Systems Education*, (17:2), pp. 141–150.

Jonassen, D. H. 1997. "Instructional design models for well-structured and ill-structured problem-solving learning outcomes," *Educational Technology Research and Development* (45:1), pp. 65–94.

Jonassen, D. H. 2006. "Toward a design theory of problem solving," *Educational Technology Research and Development* (50:2), pp. 65–77.

Keefer, M. W. 2005. "Making good use of online case study materials," *Science and Engineering Ethics* (11:3), pp. 413–429.

Keller, C. 2009. "User Acceptance of Virtual Learning Environments: A Case Study from Three Northern European Universities," *Communications of the Association for Information Systems* (25), Article 38. Available at: http://aisel.aisnet.org/cais/vol25/iss1/38

Kerr, D., Troth, A., and Pickering, A. 2003. "The use of role-playing to help students to understand information systems case studies," *Journal of Information Systems Education* (14:2), pp. 167–171.

Lakkala, M., Ilomaki, L., and Palonen, T. 2007. "Implementing virtual collaborative inquiry practices in a middle-school context," *Behavior and Information Technology*, (26:1), pp. 37–53.

Law, W. K. 2007. "Frontiers for learner-centered education," *Journal of Information Systems Education* (18:3), pp. 313–320.

Lin, C., Chou, C. C., and Kuo, M. 2007. "Inhabited virtual learning worlds and impacts on learning behaviors in young school learners," *International Journal of Distance Education* (5:4), pp. 99–112.

Linn, M.C. 1995. "Designing Computer Learning Environments for Engineering and Computer Science: The Scaffolded Knowledge Integration Framework," *Journal of Science Education and Technology* (4:2), pp. 103-126.

Malone, T. W., and Lepper, M. R. 1987. "Making learning fun: A taxonomy of intrinsic motivations for learning," in *Aptitude, learning and instruction III: Conative and affective process analyses*, Snow, R.E. and Farr, M.J. (eds.), Hillsdale, NJ: Lawrence Erlbaum, pp. 223–254.

Messinger, P.R., Stroulia, E., Lyons, K., Bone, M., Niu, R.H., Smirnov, K., and Perelgut, S. 2009. "Virtual Worlds – past, present, and future: New directions in social computing," *Decision Support Systems* (47), pp. 204-228.

Muukkonen, H., Lakkala, M., and Hakkarainen, K. 2005. "Technology-mediation and tutoring: How do they shape progressive inquiry discourse?" *The Journal of the Learning Sciences* (14), pp. 527–565.

Nelson, B., and Erlandson, B. 2008. "Managing cognitive load in educational multi-user virtual environments: Reflection on design practice," *Educational Technology Research and Development* (56:5/6), pp. 619–641. Retrieved January 19, 2009, doi:10.1007/s11423-007-9082-1.

Pan, Z., Cheok, A. D., Yang, H., Zhu, J., and Shi, J. 2005. "Virtual reality and mixed reality for virtual learning environments," *Computers and Graphics* (30), pp. 20–28.

Phang, C.W. and Kankanhalli, A. 2009. "How Do Perceptions of Virtual Worlds Lead to Enhanced Learning? An Empirical Investigation," in *Proceedings of the Thirtieth International Conference on Information Systems*, Association for Information Systems, Phoenix, Arizona, December.

Robbins, R.W. 2005. "Understanding Individual and Group Ethical Problem Solving: A Computational Ethics Approach." Doctoral Dissertation. Rensselaer Polytechnic Institute.

Robbins, R.W. and Butler, B.S. 2009a."Selecting a Virtual World for Learning," *Journal of Information Systems Education*, Special Issue: Impacts of Web 2.0 and Virtual World Technologies on IS Education (20:2), pp. 199-210.

Robbins, R.W. and Butler, B.S. 2009b. "Teaching and Learning Collaboratively and Virtually," in *Proc. 2009 Americas Conference on Information Systems.* Association for Information Systems. San Francisco, CA. Paper No. 655.

Robbins, R.W., Fleischmann, K.R., and Wallace, W.A. 2009. "Computing and Information Ethics Education Research," in *Handbook of Research on Technoethics*, Luppicini, R. and Adell, R. (eds.), New York: Information Science Reference, pp. 391-408.

Robbins, R.W. and Wallace, W.A. 2007."Decision Support for Ethical Problem Solving: A Multi-agent Approach," *Decision Support Systems* (43:4), pp. 1571-1587.

Schraw, G., Dunkle, M. E., and Bendixen, L. D. 1995. "Cognitive processes in well-defined and ill-defined problem solving," *Applied Cognitive Psychology* (9), pp. 1–16.

Shen, J. and Eder, L.B. 2009. "Intentions to Use Virtual Worlds for Education," *Journal of Information Systems Education* (20:2), pp. 225-234.

Shin, N., Jonassen, D. H., and MaGee, S. 2003. "Predictors of well-structured and ill-structured problem solving in an astronomy simulation," *Journal of Research in Science Teaching* (40:1), pp. 7–27.

Shulman, L. S. 1992. "Toward a pedagogy of cases," in *Case methods in teacher education*, Shulman, L.S. (ed.), New York, NY: Teachers College Press, pp. 1-32.

Sinnott, J. D. 1989. "A model of solution of ill-structured problems: Implications for everyday and abstract problem solving," in *Everyday problem solving: Theory and applications*, Sinnott, J. D. (ed.), New York, NY: Praeger, pp. 72–99.

Spiro, R. J., Feltovich, P. J., Jacobson, M. J., and Coulson, R. L. 1992. "Cognitive flexibility, constructivism, and hypertext: Random access instruction for advanced knowledge acquisition in ill-structured domains," in *Constructivism and the technology of instruction, a conversation*, Duffy, T.M. and Jonassen D.H. (eds.), Hillsdale, NJ: Lawrence Erlbaum, pp. 57-75.

Stepich, D. A., Ertmer, P. A., and Lane, M. M. 2001. "Problem solving in a case-based course: Strategies for facilitating coached expertise," *Educational Technology Research and Development* (49:3), pp. 1042–1629.

Topi, H.; Valacich, J.S.; Wright, R.T.; Kaiser, K; Nunamaker, Jr., J.F.; Sipior, J.C.; and de Vreede, G-J. 2010. "IS 2010: Curriculum Guidelines for Undergraduate Degree Programs in Information Systems," *Communications of the Association for Information Systems* (26), Article 18.

United States Food and Drug Administration. 2010. http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/PostmarketRequirements/HumanFactors/ucm119213.htm, Accessed May 2, 2010.

Virtual World News, 2009. http://www.virtualworldsnews.com/2007/10/protonmedia-par.html, Accessed January 24, 2009.

Voss, J. F., Wolfe, C. R., Lawrence, J. A., and Engle, R. A. 1991. "From representation to decision: An analysis of problem solving in international relations," in *Complex problem solving: Principles and mechanisms*, Sternberg, R. J. and Frensch, P.A. (eds.), Hillsdale, NJ: Lawrence Erlbaum, pp. 119–157.

Wagner, C. and Ip, R.K.F. 2009. "Action Learning with Second Life – A Pilot Study," *Journal of Information Systems Education* (20:2), pp. 249-258.

Wang, Y. and Braman, J. 2009. "Extending the Classroom through Second Life," *Journal of Information Systems Education* (20:2), pp. 235-248.

Wells, P., de Lange, P., and Fieger, P. 2008. "Integrating a virtual learning environment into a second-year accounting course: Determinants of overall student perception," *Accounting and Finance* (48), pp. 503–518.

Whitehead, A.N. 1929/1985. *The Aims of Education and Other Essays*, New York, NY: Free Press.

Whitworth, A. 2005. "The politics of virtual learning environments: Environmental change, conflict, and e-learning," *British Journal of Educational Technology* (36:4), pp. 685–691.

Williams, S. M. 1992. "Putting case-based instruction into context: Examples from legal and medical education," *The Journal of the Learning Sciences* (2:4), pp. 367–427.

Windschitl, M. and Sahl, K. 2002. **"**Tracing Teachers' Use of Technology in a Laptop Computer School: The Interplay of Teacher Beliefs, Social Dynamics, and Institutional Culture," *American Educational Research Journal* (39), pp. 165-205.

Wu, D., Hiltz, S.R., and Bieber, M. 2010. "Acceptance of Educational Technology: Field Studies of Asynchronous Participatory Examinations," *Communications of the Association for Information Systems* (26), Article 21.

Zeichner, K. M., and Liston, D. P. 1996. *Reflective teaching: An introduction*, Mahweh, NJ: Lawrence Erlbaum.

Section J - Data: Collaborating with individuals in industries

This section describes how I (and my students) collaborated with individuals in industries.

(Please turn the page.)

In my teaching, I sought to provide meaningful experiences to students in ways that I could help them. One of the ways of learning that I value immensely is via projects. Projects have a wonderful way of helping people see things that they did not recognize or value before. (This said, projects have a way of not providing the necessary structure that is sometimes necessary to learn complex skills. These are easily learned via carefully crafted problems.)

I developed relationships with organizations in industry. Four of these were MapInfo, NXP Semiconductors (previously a division of Philips Electronics), GE Specialty Materials, and IBM. My most intensive relationships were with MapInfo and NXP Semiconductors. At both MapInfo and NXP Semiconductors, the individuals I interacted with were stimulated by interacting with students, were able to find problems that could be addressed by the students, and were in organizations that allowed them to experiment carefully with how students could help their organizations. Unfortunately, my experience with GE Specialty Materials and IBM indicated that the project sponsors had a large number of constraints and goals, and my students who were basically unskilled, were not able to help the project sponsors as much as the sponsors or their organizations would have liked. However, in the case of IBM, IBM did recruit one of my students after a project concluded.

Sample projects included the design and/or development of a decision support system/knowledgebase to help management troubleshoot fabrication plant tools and analyze root causes, a software review system, a problem tracking system, and an educational certification system. Other projects developed by student teams have included a projects binder, a proposal development aide, and a site locator, a registration system, a calendar, and a training evaluation system. Finally, one student team, unrelated to a corporation, designed and developed a wireless remote control device enabling interaction with an interactive teddy bear for children with limited motor skills, (e.g., cerebral palsy). These students formed an entrepreneurial venture and joined the Rensselaer Incubator Center.

Special Note:
To protect these individuals privacy, this testimonial has been anonymized.

PHILIPS

Philips Semiconductors

Date removed.

To whom it may concern:

We are writing to express our impressions and gratitude to Marist College, for the opportunity to participate in a mutually beneficial program to both Philips Semiconductors and the Marist College students.

In the spring of [year removed], the graduate course in Information Analysis, took on a project with Philips. The objective being to give the students some practical experience, solving a problem facing an operating manufacturing organization. The expected result was for the students to gain some experience in defining a problem, its project scope, architecting a solution and provide a system specification with a working prototype. Philips hoped to benefit by gaining some insight into how to better organize many sources of data into a comprehensive and intuitive knowledge based system.

The students led by Professor Russ Robbins, took on the task with enthusiasm and showed a high degree of commitment to their learning and the success of the project. The weekly reviews and information exchanges were both concise and informative. The project results exceed our initial expectations and will provide Philips with an excellent foundation to build upon.

We found working with Mr. Robbins to be both a professional and stimulating experience from our initial engagement discussions, through the progress reviews and final conclusions and results. He is truly engaged with his students providing the right amount of guidance and mentorship. We are impressed with his commitment to their success and to the development of an on-going relationship with area industries.

Once again we appreciate the effort and look forward to the continued partnership in future endeavors.

Closing, two signatures, and names removed.

Address removed.

Section K - Examples: Thinking outside the box

This section indicates how I have “thought outside the box” in the past.

(Please turn the page.)

Note that these dichotomies are meant to show how, in the past, I approached things in a different way. The standard foci" below are not listed to make irrefutable claims. Instead the information in the "standard foci" columns are simply what I saw (or what I think I saw) when I studied or worked in small portions of these disciplines.

Standard Foci in B-Schools	**My Past Foci**
One discipline "like horse blinders"	Many disciplines
Describe	Build
Teach concepts	Teach skills

Table 1: Thinking out of the box in B-Schools

Standard Foci in NLP	**My Past Foci**
Statistical	Integrating Statistical and Functional
See a document as a set of symbols.	See the document as a person's output.
Study at the document level.	Study at the clause level.
Prepositions considered "stop words."	Prepositional phrases are contextual.
Auxiliary verbs considered "stop words."	Auxiliary verbs indicate desire, duty, promises, possibility, probability, surety, ability, existence, properties, possession, execution…much of what makes us human.

Table 2: Thinking out the box in NLP

Standard foci in Cognitive Psychology	**My Past Foci**
Confirmatory research	Engineering research
Statistical models	Processes
Divisive	Integrative
Focus on explaining the brain	Focus on mimicking the brain
Theory focused	Theory execution focused
Focus on looking into the head	Focus on looking from in head outward.

Table 3: Thinking out of the box in Cognitive Psychology

Section L - Example: Cleaning and manipulating data

This section provides an example of cleaning and manipulating data.

(Please turn the page.)

This example uses the R programming language to tidy data so that each output file contains only one variable. This is valuable when using software such as R which does not scale well.

```
###This file run_analysis.R merges the following files

## features.txt
## activityunderscorelabels
## train/subjectunderscoretrain.txt
## train/yunderscoretrain.txt
## test/subjectunderscoretest.text
## test/yunderscoretest.txt

###and outputs 561 seperate files into the folder tidyMerged. The new files each represent one
measurement type, and have all measurements in the data linked directly to participant ID and
activity name.

###Each file has three columns: one for indicating subject, one for indicating activity, and
one for indicating the variable that is reported. Each file has 10,299 observations.


#install.packages("doBy")
#require("doBy")

#Define a function that renames activities

setActivityNamesByObs<-function(actsByObs){
  for(i in seq_along(actsByObs[,1])){
    actsByObs[i,1]<-gsub("1", "Walking", actsByObs[i,1])
    actsByObs[i,1]<-gsub("2", "Walking Up Stairs", actsByObs[i,1])
    actsByObs[i,1]<-gsub("3", "Walking Down Stairs", actsByObs[i,1])
    actsByObs[i,1]<-gsub("4", "Standing", actsByObs[i,1])
    actsByObs[i,1]<-gsub("5", "Sitting", actsByObs[i,1])
    actsByObs[i,1]<-gsub("6", "Lying", actsByObs[i,1])
  }
  assign("actsByObs",actsByObs,.GlobalEnv)

}

###

#Define a function that reads data about observations, subjects attached to observations,
activities attached to observations,
#activities performed by subjects, and variables (also known as features or measurements)
captured or already created.
getData <- function(observationsFile, variablesFile, SubjectsByObjsFile, ActivitiesByObjsFile,
activitiesFile){

  tempObservations<-read.table(observationsFile,sep="",header=FALSE)
  assign("tempObservations",tempObservations,.GlobalEnv)

  variables<-read.table(variablesFile, sep="", header=FALSE)

  variableNames<-character(0)

  variableNames<-as.character(variables[,2])
  assign("variableNames",variableNames,.GlobalEnv)
```

```
  tempSubjsByObs<-read.table(SubjectsByObjsFile,sep="", header=FALSE)
  assign("tempSubjsByObs",tempSubjsByObs,.GlobalEnv)

  tempActsByObs<-read.table(ActivitiesByObjsFile,sep="", header=FALSE)
  assign("tempActsByObs",tempActsByObs,.GlobalEnv)

  activityNames<-read.table(activitiesFile, sep="", header=FALSE)
  assign("activityNames",activityNames,.GlobalEnv)

  colnames(activityNames)<-c("Number","Name")

}

###

#Define a function that creates variable (i.e., measurement) names that are more readable by a
analyst.
clarifyVariableNames <- function (variableNames){
  tempVariableNames<-variableNames

  for(i in 1:length(tempVariableNames)){

    if(grepl("[b][a][n][d][s][E][n][e][r][g][y]",variableNames[i])){

      x<-variableNames[i]
      assign("x",x,.GlobalEnv)

      variableNames[i]<-paste(variableNames[i],"A", sep="")
      assign("variableNames",variableNames,.GlobalEnv)

      for(j in 1:length(tempVariableNames)){

        if(variableNames[j]==x){

          y<-variableNames[j]

          assign("y",y,.GlobalEnv)
          variableNames[j]<-paste(variableNames[j],"B", sep="")

          if(variableNames[j+14]==x){
            variableNames[j+14]<-paste(variableNames[j+14],"C", sep="")

            assign("variableNames",variableNames,.GlobalEnv)

            for(k in 1:length(tempVariableNames)){

              if(variableNames[k]==variableNames[j]){

                z<-variableNames[k]
                assign("z",z,.GlobalEnv)
                variableNames[k]<-paste(variableNames[k],"C", sep="")
                assign("variableNames",variableNames,.GlobalEnv)
```

```
            }

          }

        }

      }

    }

  }

}

for(l in 1:length(tempVariableNames)){

  if(grepl("[C][B][A]",variableNames[l])==TRUE){
    variableNames[l]<-gsub("CBA","C", variableNames[l])
    assign("variableNames",variableNames,.GlobalEnv)

  }

}

for(m in 1:length(tempVariableNames)){

  if(grepl("[B][A]",variableNames[m])==TRUE){
    variableNames[m]<-gsub("BA","B", variableNames[m])
    assign("variableNames",variableNames,.GlobalEnv)

  }

}
for(n in 1:length(tempVariableNames)){
  if(grepl("[C][A]",variableNames[n])==TRUE){
    variableNames[n]<-gsub("CA","C", variableNames[m])
    assign("variableNames",variableNames,.GlobalEnv)
  }
}

for(i in 1:length(tempVariableNames)){

  if(!grepl("[b][a][n][d][s][E][n][e][r][g][y]",variableNames[i])){
    if(grepl("[A][c][c]",variableNames[i])){

      x<-variableNames[i]
      assign("x",x,.GlobalEnv)

      variableNames[i]<-paste(variableNames[i],"A", sep="")
      assign("variableNames",variableNames,.GlobalEnv)

      for(j in 1:length(tempVariableNames)){
```

```
        if(identical(variableNames[j],x)){

          y<-variableNames[j]

          assign("y",y,.GlobalEnv)
          variableNames[j]<-paste(variableNames[j],"B", sep="")
          assign("variableNames",variableNames,.GlobalEnv)

          for(k in 1:length(tempVariableNames)){

            if(isTRUE(all.equal(variableNames[k],variableNames[j]))){

              z<-variableNames[k]

              assign("z",z,.GlobalEnv)
              variableNames[k]<-paste(variableNames[k],"C", sep="")
              assign("variableNames",variableNames,.GlobalEnv)
            }

          }

        }

      }

    }

  }

}

for(l in 1:length(tempVariableNames)){

  if(grepl("[C][B][A]",variableNames[l])==TRUE){
    variableNames[l]<-gsub("CBA","C", variableNames[l])
    assign("variableNames",variableNames,.GlobalEnv)

  }

}

for(m in 1:length(tempVariableNames)){

  if(grepl("[B][A]",variableNames[m])==TRUE){
    variableNames[m]<-gsub("BA","B", variableNames[m])
    assign("variableNames",variableNames,.GlobalEnv)

  }

}

for(n in 1:length(tempVariableNames)){
```

```
    if(grepl("[C][A]",variableNames[n])==TRUE){
      variableNames[n]<-gsub("CA","C", variableNames[m])
      assign("variableNames",variableNames,.GlobalEnv)

    }

  }

  for(n in 1:length(tempVariableNames)){

    if(grepl("[A][A]",variableNames[n])==TRUE){
      variableNames[n]<-gsub("AA","A", variableNames[m])
      assign("variableNames",variableNames,.GlobalEnv)

    }

  }

  for(i in 1:length(tempVariableNames)){

    if(!grepl("[b][a][n][d][s][E][n][e][r][g][y]",variableNames[i])){

      if(grepl("[G][y][r][o]",variableNames[i])){

      x<-variableNames[i]
      assign("x",x,.GlobalEnv)

      variableNames[i]<-paste(variableNames[i],"A", sep="")
      assign("variableNames",variableNames,.GlobalEnv)


      for(j in 1:length(tempVariableNames)){

        if(identical(variableNames[j],x)){

          y<-variableNames[j]

          assign("y",y,.GlobalEnv)
          variableNames[j]<-paste(variableNames[j],"B", sep="")
          assign("variableNames",variableNames,.GlobalEnv)

          for(k in 1:length(tempVariableNames)){

            if(isTRUE(all.equal(variableNames[k],variableNames[j]))){

              z<-variableNames[k]
              assign("z",z,.GlobalEnv)
              variableNames[k]<-paste(variableNames[k],"C", sep="")
              assign("variableNames",variableNames,.GlobalEnv)

            }

          }
```

```
      }

    }

  }

}

  for(l in 1:length(tempVariableNames)){

    if(grepl("[C][B][A]",variableNames[l])==TRUE){
      variableNames[l]<-gsub("CBA","C", variableNames[l])
      assign("variableNames",variableNames,.GlobalEnv)

    }

  }

  for(m in 1:length(tempVariableNames)){

    if(grepl("[B][A]",variableNames[m])==TRUE){
      variableNames[m]<-gsub("BA","B", variableNames[m])
      assign("variableNames",variableNames,.GlobalEnv)

    }

  }

  for(n in 1:length(tempVariableNames)){

    if(grepl("[C][A]",variableNames[n])==TRUE){
      variableNames[n]<-gsub("CA","C", variableNames[m])
      assign("variableNames",variableNames,.GlobalEnv)

    }

  }

  for(n in 1:length(variableNames)){

    if(grepl("[A][A]",variableNames[n])==TRUE){
      variableNames[n]<-gsub("AA","A", variableNames[m])
      assign("variableNames",variableNames,.GlobalEnv)
    }
  }

  for(i in 1:length(variableNames)){

    variableNames[i]<-gsub("entropy","Decline", variableNames[i])
    variableNames[i]<-gsub("arCoeff","Coeff", variableNames[i])
```

```
    variableNames[i]<-gsub("^f","", variableNames[i])
    variableNames[i]<-gsub("^t","", variableNames[i])
    variableNames[i]<-gsub("Acc","Acceleration", variableNames[i])
    variableNames[i]<-gsub("GravityAcceleration","GravitationalPull",variableNames[i])
    variableNames[i]<-gsub("GravityAcceleration","GravitationalPull",variableNames[i])
    variableNames[i]<-gsub("GyroThrust","RotationalVelocity",variableNames[i])
    variableNames[i]<-gsub("Gyro","Rotational",variableNames[i])
    variableNames[i]<-gsub("mean","Mean", variableNames[i])
    variableNames[i]<-gsub("correlation","Corr", variableNames[i])
    variableNames[i]<-gsub("Mag","Magnitude", variableNames[i])
    variableNames[i]<-gsub("mad","Median", variableNames[i])
    variableNames[i]<-gsub("bandsEnergy","EnergyInRange", variableNames[i])
    variableNames[i]<-gsub("angle","AngleBetweenVectors", variableNames[i])
    variableNames[i]<-gsub("sma","Magnitude", variableNames[i])
    variableNames[i]<-gsub("maxInds","MaxMagnitude", variableNames[i])
    variableNames[i]<-gsub("iqr","InterQuartileRange", variableNames[i])
    variableNames[i]<-gsub("AngleBetweenVectors","DirectionalChange", variableNames[i])
    variableNames[i]<-gsub("AccelerationThrust","Acceleration", variableNames[i])
    variableNames[i]<-gsub("std","StdDev", variableNames[i])
    variableNames[i]<-gsub("energy","Energy", variableNames[i])
    variableNames[i]<-gsub("max","Max", variableNames[i])
    variableNames[i]<-gsub("min","Min", variableNames[i])
    variableNames[i]<-gsub("-","",variableNames[i])
    variableNames[i]<-gsub("[(]","",variableNames[i])
    variableNames[i]<-gsub("[)]","",variableNames[i])
    variableNames[i]<-gsub("WalkingUpWalkingUp","WalkingUp",variableNames[i])

    variableNames[i]<-gsub("DirectionalChangetAccelerationMean,gravityMean","DirectionalChangeInT
    hrustRelativeToGravityMean",variableNames[i])

    variableNames[i]<-gsub("DirectionalChangetRotationalVelocityMean,gravityMean","DirectionalCha
    ngeThrustRotationalVelocityMeanRelativeToGravity",variableNames[i])
    variableNames[i]<-gsub(",gravity","RelativeToGravity",variableNames[i])

    variableNames[i]<-gsub("DirectionalChangeRotationalMeanRelativeToGravityMean","DirectionalCha
    ngetRotationalRelativeToGravityMean",variableNames[i])
    assign("variableNames",variableNames,.GlobalEnv)

  }

}


#Execute the getData function and pass it the appropriate files related to training.
getData("train/X_train.txt","features.txt","train/subject_train.txt", "train/y_train.txt",
"activity_labels.txt")

setActivityNamesByObs(tempActsByObs)

assign("tempActsByObs",tempActsByObs,.GlobalEnv)

#Pass the observations, subjects by observation, and acts by observation data from temp to the
appropriate data frames.
```

```
observations<-tempObservations
subjsByObs<-tempSubjsByObs
actsByObs<-tempActsByObs

getData("test/X_test.txt","features.txt","test/subject_test.txt", "test/y_test.txt",
"activity_labels.txt")
observations<-rbind(observations, tempObservations)

subjsByObs<-rbind(subjsByObs, tempSubjsByObs)

actsByObs<-rbind(actsByObs, tempActsByObs)

setActivityNamesByObs(actsByObs)
actsByObs<-data.frame(actsByObs)

#Execute the function which alters the variable names and pass it the current names.

clarifyVariableNames(variableNames)

combineObsVarsSubjsActs<-function(observations, variableNames, subjsByObs, actsByObs){

  colnames(observations)<-variableNames

  assign("observations",observations,.GlobalEnv)
  colnames(subjsByObs)[1]<-"subjsByObs"

  subjsByObsIds<-as.character(subjsByObs$subjsByObs)
  assign("subjsByObsIds",subjsByObsIds,.GlobalEnv)

  colnames(actsByObs)[1]<-"actsByObs"

  actsByObs<-as.character(actsByObs$actsByObs)
  assign("actsByObs",actsByObs,.GlobalEnv)

}

combineObsVarsSubjsActs(observations, variableNames, subjsByObs, actsByObs)

actsByObs<-data.frame(actsByObs)
subjsByObsIds<-data.frame(subjsByObsIds)

observations<-cbind(actsByObs,observations)
observations<-cbind(subjsByObsIds,observations)
colnames(observations)<-c("subjsIds", "activities", variableNames)

summaryFunction <- function(x) c(means = mean(x))

for (i in 1:length(variableNames)){

  index<-i+2

  if(!file.exists("./tidyMerged")){dir.create("./tidyMerged")}
```

```
  measureToAggregate<-data.frame(observations[,c(1,2,index)])
  colnames(measureToAggregate)<-c("subjsByObsIds", "actsByObs", "measure")
  measureToAggregate$subjsByObsIds<-as.numeric(as.character(measureToAggregate$subjsByObsIds))
  measureToAggregate <- measureToAggregate[order(measureToAggregate$subjsByObsIds,
  measureToAggregate$actsByObs) , ]
  write.table(measureToAggregate,paste("./tidyMerged/",
  variableNames[i],".txt",sep=""),row.names=FALSE)

  if(grepl("([M][e][a][n]|[S][t][d][D][e][v])",colnames(observations)[index])==TRUE){
    if(!file.exists("./tidyExtracted")){dir.create("./tidyExtracted")}
    extractionMeanStdDev<-data.frame(observations[,c(1,2,index)])
    colnames(extractionMeanStdDev)<-c("subjsByObsIds", "actsByObs", "measure")

    extractionMeanStdDev$subjsByObsIds<-as.numeric(as.character(extractionMeanStdDev$subjsByObsId
    s))
    extractionMeanStdDev <- extractionMeanStdDev[order(extractionMeanStdDev$subjsByObsIds,
    extractionMeanStdDev$actsByObs) , ]
    write.table(extractionMeanStdDev,paste("./tidyExtracted/",
    variableNames[i],".txt",sep=""),row.names=FALSE)

  }

  if(!file.exists("./tidyAveraged")){dir.create("./tidyAveraged")}

  aggregatedMeasure<-summaryBy(measure ~ subjsByObsIds + actsByObs, data=measureToAggregate,
  FUN=summaryFunction)

  colnames(aggregatedMeasure)<-c("subjIds", "activities", "means")
  names(aggregatedMeasure)
  aggregatedMeasure$subjIds<-as.numeric(as.character(aggregatedMeasure$subjIds))
  aggregatedMeasure <- aggregatedMeasure[order(aggregatedMeasure$subjIds,
  aggregatedMeasure$activities ) , ]

  write.table(aggregatedMeasure,paste("./tidyAveraged/",
  variableNames[i],".txt",sep=""),row.names=FALSE)

}

write.table(variableNames,"tidyVariableNames.txt", sep=" ", quote=FALSE)
```

Section M - Example: Building predictive models using machine learning algorithms

This section provides an example of building a statistical model using machine learning algorithms.

(Please turn the page.)

I built a model that predicted the release year of a song given a set of audio features. This entailed using a method referred to as gradient descent to identify the most appropriate linear regression model.

Sometimes, closed form solutions to identify best fit models are computationally prohibitively expensive. Machine learning researchers and practitioners frequently apply a version of an algorithm known as gradient descent. Gradient descent guides changing parameter values in a estimated model by seeking to minimize a quality metric (e.g. RSS, RMSE, etc.) by seeking the minimum on a convex function (which in 2d appears similar to a parabola).

I then tuned models using a technique referred to as grid search and added quadratic features to improve predictions' accuracies.

#Building a song recommender

#Fire up GraphLab Create

In [1]:
```
import graphlab
import matplotlib
```

#Load music data

In [2]:
```
song_data = graphlab.SFrame('song_data.gl/')
```

#Explore data

Music data shows how many times a user listened to a song, as well as the details of the song.

In [3]:
```
song_data.head()
```

##Showing the most popular songs in the dataset

In [4]:
```
graphlab.canvas.set_target('ipynb')
```

In [ ]:
```
song_data['song'].show()
```

In [6]:
```
len(song_data)
```

Out[6]: 1116609

Question 3

In [80]:
```
train_data_q3,test_data_q3 = song_data.random_split(.8,seed=0)
```

In [82]:
```
personalized_model_q3 = graphlab.item_similarity_recommender.create(train_data_q3,
                                                                    user_id='user_id',
                                                                    item_id='song')
```

```
PROGRESS: Recsys training: model = item_similarity
PROGRESS: Warning: Ignoring columns song_id, listen_count, title, artist;
PROGRESS:     To use one of these as a target column, set target = <column_name>
PROGRESS:     and use a method that allows the use of a target.
PROGRESS: Preparing data set.
PROGRESS:     Data has 893580 observations with 66085 users and 9952 items.
PROGRESS:     Data prepared in: 0.720572s
PROGRESS: Computing item similarity statistics:
PROGRESS: Computing most similar items for 9952 items:
PROGRESS: +-----------------+-----------------+
PROGRESS: | Number of items | Elapsed Time    |
PROGRESS: +-----------------+-----------------+
PROGRESS: | 1000            | 0.450162        |
PROGRESS: | 2000            | 0.478016        |
PROGRESS: | 3000            | 0.507115        |
PROGRESS: | 4000            | 0.532584        |
PROGRESS: | 5000            | 0.557696        |
PROGRESS: | 6000            | 0.584154        |
PROGRESS: | 7000            | 0.611167        |
PROGRESS: | 8000            | 0.639596        |
PROGRESS: | 9000            | 0.673179        |
PROGRESS: +-----------------+-----------------+
PROGRESS: Finished training in 0.847551s
```

In [84]:
```
subset_test_users = test_data_q3['user_id'].unique()[0:10000]
```

In [84]:
```
subset_test_users = test_data_q3['user_id'].unique()[0:10000]
```

In [86]:
```
q3_recommendations = personalized_model_q3.recommend(subset_test_users,k=1)
```

```
PROGRESS: recommendations finished on 1000/10000 queries. users per second: 2556.88
PROGRESS: recommendations finished on 2000/10000 queries. users per second: 2647.65
PROGRESS: recommendations finished on 3000/10000 queries. users per second: 2676.48
PROGRESS: recommendations finished on 4000/10000 queries. users per second: 2696.35
PROGRESS: recommendations finished on 5000/10000 queries. users per second: 2686.07
PROGRESS: recommendations finished on 6000/10000 queries. users per second: 2676.49
PROGRESS: recommendations finished on 7000/10000 queries. users per second: 2681.58
PROGRESS: recommendations finished on 8000/10000 queries. users per second: 2698.41
PROGRESS: recommendations finished on 9000/10000 queries. users per second: 2690.2
PROGRESS: recommendations finished on 10000/10000 queries. users per second: 2682
```

In [88]:
```
len(q3_recommendations)
```

Out[88]:
```
10000
```

In [98]:
```
song_listen_counts_q3 = q3_recommendations.groupby(key_columns='song',operations={'count'
```

In [99]:
```
song_listen_counts_q3.head()
```

Out[99]:

song	count
Arco Arena - Cake	1
Too Deep - Girl Talk	2
Guys Like Me - Eric Church ...	2
Freedom - Akon	2
Wish You Were Here - Incubus ...	1
Change - Blind Melon	1

Section N - Example: Using statistical analysis tools such as R and RapidMiner

This section provides an example using RapidMiner.

(Please turn the page.)

My use of R (and Python) are becoming broad. My use of RapidMiner is narrow. However, one rewarding experience with RapidMiner occurred when I built a step-by-step learning workshop to help people learn how to perform K-means clustering.

Susquehanna University
Sigmund Weis School of Business
Robbins

"Hands-on" Demonstration
K-Means Clustering
via RapidMiner

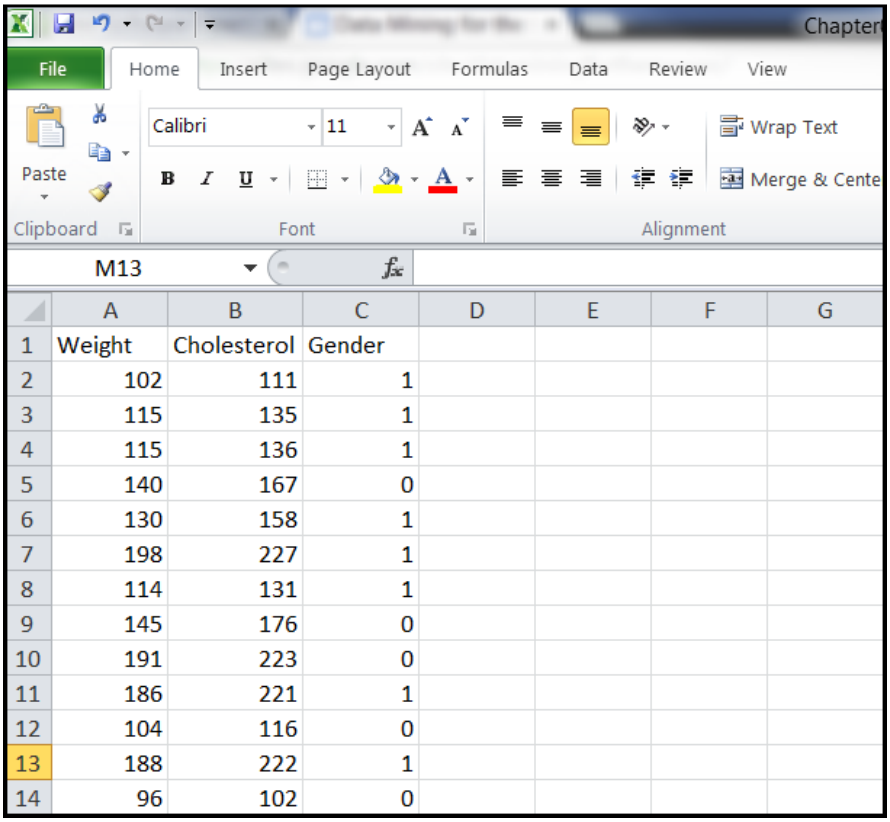
Special Note: Learning can be a difficult endeavor. At any point in this hands on demonstration, if you have any trouble, PLEASE reach out to me at 570-372-4465 or robbinsr@susqu.edu. I am very serious about your learning.

Imagine that you are a healthy lifestyle program director for a healthcare insurance company[1] known as the Prosperential Insurance Company of America. Your responsibilities at Prosperential include offering programs and incentives to your customers to help them transition to healthier lifestyles. If your programs are successful, some customers will need less health care services than would be the case otherwise. As part of this responsibility, you want to identify clusters of individuals that your company insures who are at risk of developing coronary heart disease as a result of being overweight and/or having high cholesterol and who may be benefit from participating in a "heart-healthy" program.

1. If you have not installed RapidMiner Studio (Starter Version), please do so prior to working through this demonstration. A free "Starter" version of RapidMiner is available at http://rapidminer.com.

2. Please download the file named **Chapter06DataSet.csv** from **http://bit.ly/1h0URh4** into your Downloads folder.

3. Open the **Chapter06DataSet.csv** file in Excel if you would like to review the data. The data has 547 records. Each record represents a person and has three values: a person's gender, a person's cholesterol level, and a person's weight, in pounds. A screenshot of the first 14 records is shown below so that you can become familiar with how the data is provided to us, prior to our use of it in RapidMiner. Note that in this data, a "1" represents a male and a "0" represents a female.



	A	B	C	D	E	F	G
1	Weight	Cholesterol	Gender				
2	102	111	1				
3	115	135	1				
4	115	136	1				
5	140	167	0				
6	130	158	1				
7	198	227	1				
8	114	131	1				
9	145	176	0				
10	191	223	0				
11	186	221	1				
12	104	116	0				
13	188	222	1				
14	96	102	0				

Please proceed to next page.

[1] This sample demonstration uses a data set and extends an example presented in the text Data Mining for the Masses, authored by Matt North and available for purchase at http://amzn.to/1mKkibB. The data set is available at http://bit.ly/1h0URh4.

4. Open RapidMiner through your **Start** button.

5. Import the **Chapter06DataSet.csv** file into RapidMiner by…

a. Selecting File (File)

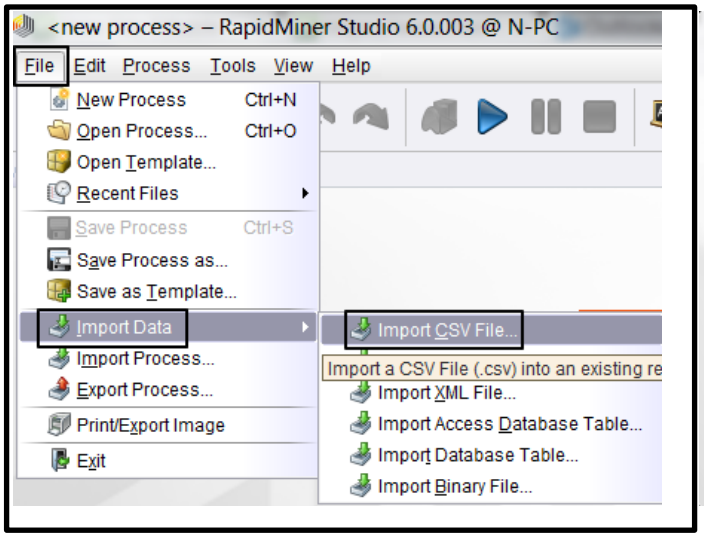
b. Selecting Import Data (Import Data)

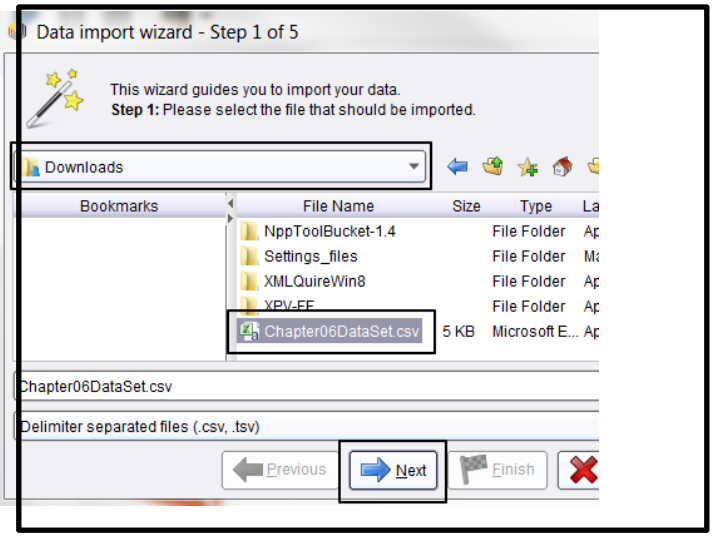
c. Selecting Import CSV File... (Import CSV File)

d. Selecting your Downloads (Downloads) folder.

e. Selecting the Chapter06DataSet.csv (Chapter06DataSet.csv) file.

f. Selecting Next (Next).





6. Continue importing the **Chapter06DataSet.csv** file into RapidMiner by…

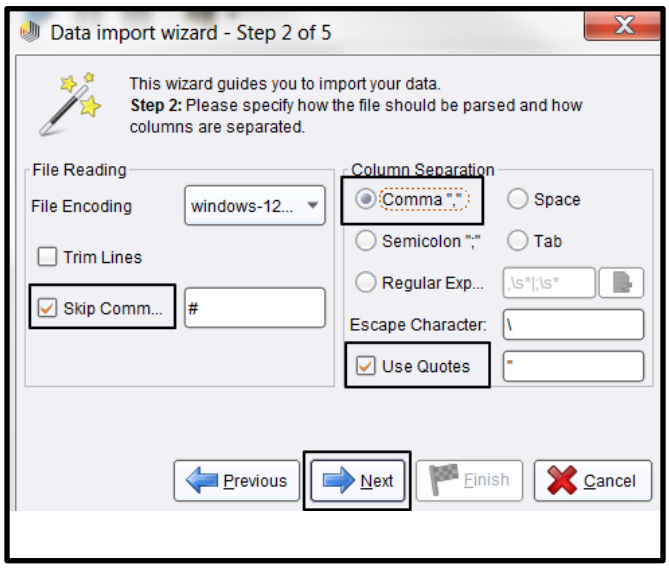
a. Assuring that Skip Comm... (Skip Comments) is checked.

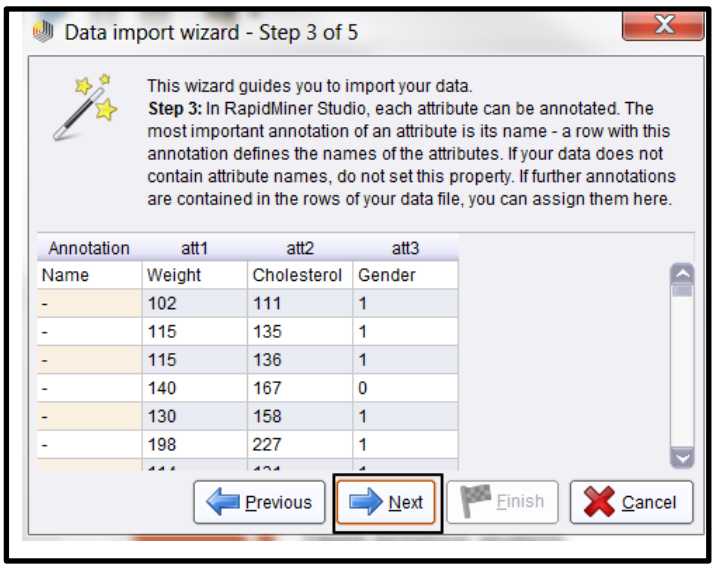
b. Assuring that the radio button for Comma "," (Comma) is selected.

c. Assuring that Use Quotes (Use Quotes) is checked.

d. Selecting Next (Next) on the Step 2 of 5 screen.

e. Selecting Next (Next) on the Step 3 of 5 screen. **Please proceed to next page.**





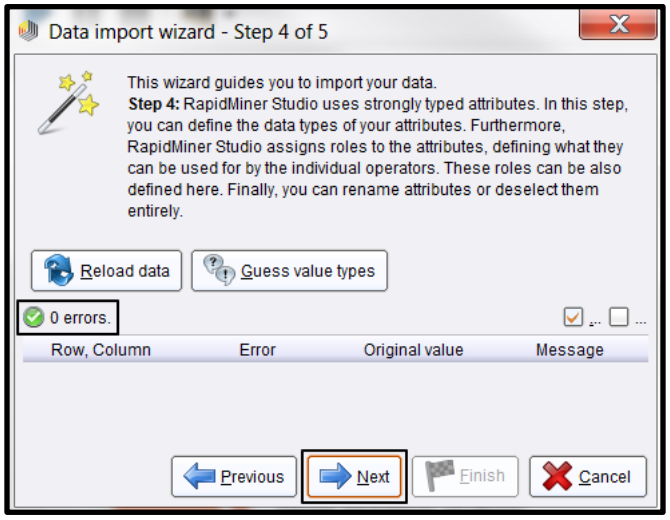
f. Assuring that 0 errors. (0 errors) have occurred.

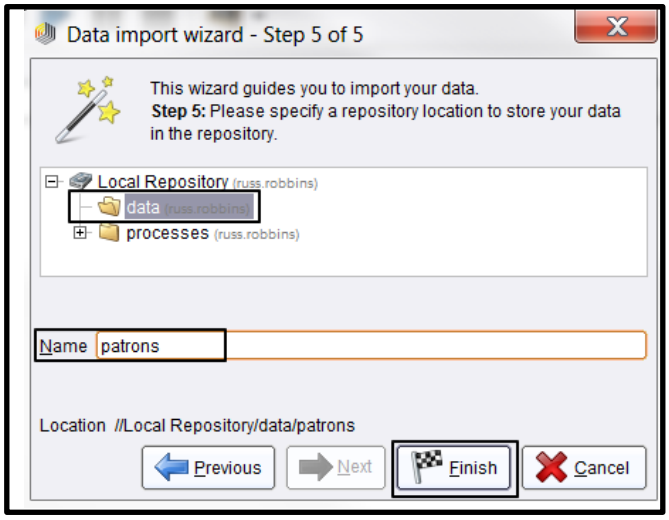
g. Selecting Next (Next) on the Step 3 of 5 screen.

h. Selecting the data (russ.robbins) (data) folder.

i. Typing **patrons** in the Name (name) field.

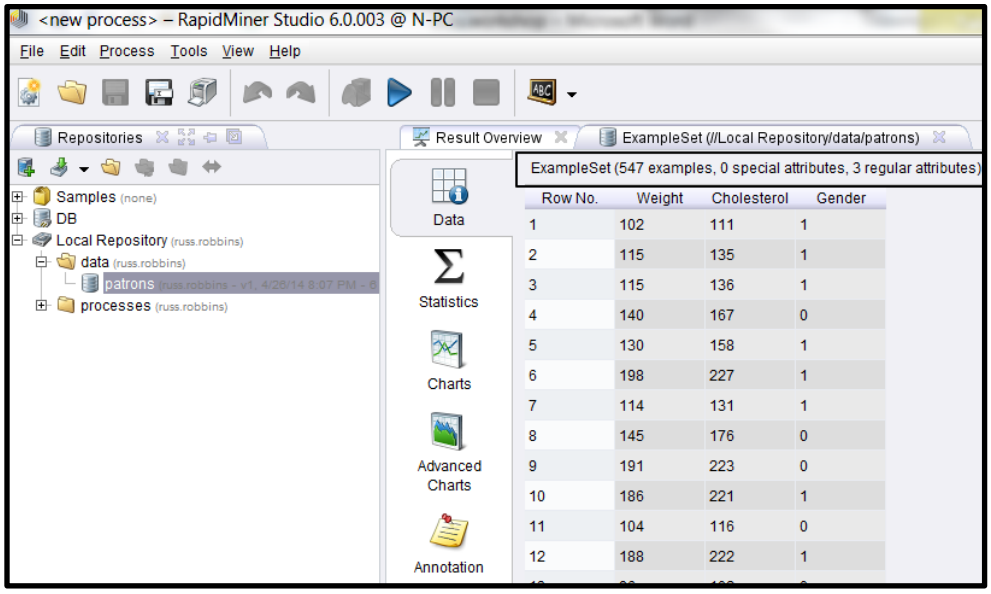
j. Selecting Finish (Finish).



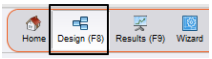


7. You should see a screen similar to the screen below. If you do not, please start at step 1 again. (No worries.)

a. Assure you have imported ExampleSet (547 examples, 0 special attributes, 3 regular attributes) (547 examples, 0 special attributes, 3 regular attributes).

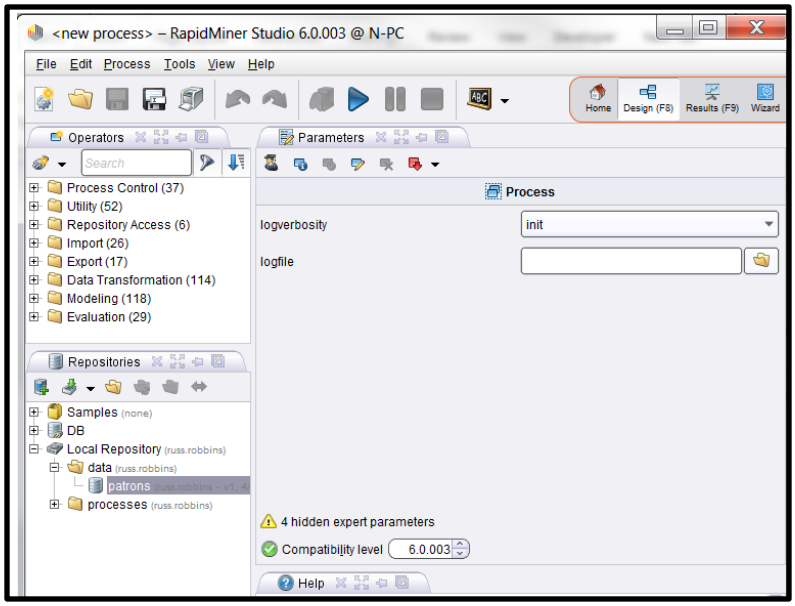


Please proceed to next page.

8. At this point we can to use RapidMiner to do "k-means" clustering on our data. Please choose the 

(Design View) from the top right of your screen so that we can design a process that will create clusters from this data.

a. You should see a screen similar to the one below.



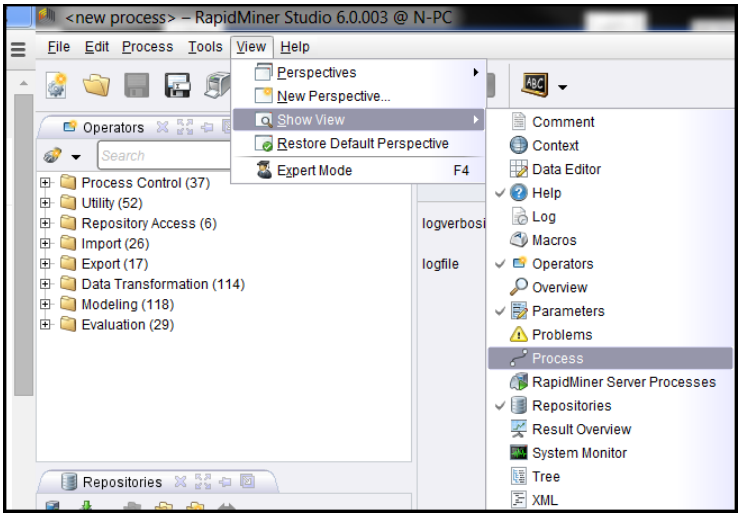
9. K-means clustering is simply a mathematical algorithm that compares records of data by comparing how "close" each pair of records are and then creates groups of records. In our case, you are trying to identify a subset of individuals you can focus on in terms of offering each person in the group incentives and the opportunity to participate in a "heart healthy" program. So for example, a person with a cholesterol level of 215 who is male and weighs 190 and is male may be placed in the same group (or cluster) as another person who has a cholesterol level of 210, weighs 198, and is male. If you are interested, a quick video that explains k-means clustering is here:

10. At this point you want to add a new "window" to your screen so that you can begin designing the process that will perform the k-means clustering algorithm.

a. Select View (View) from the top menu.

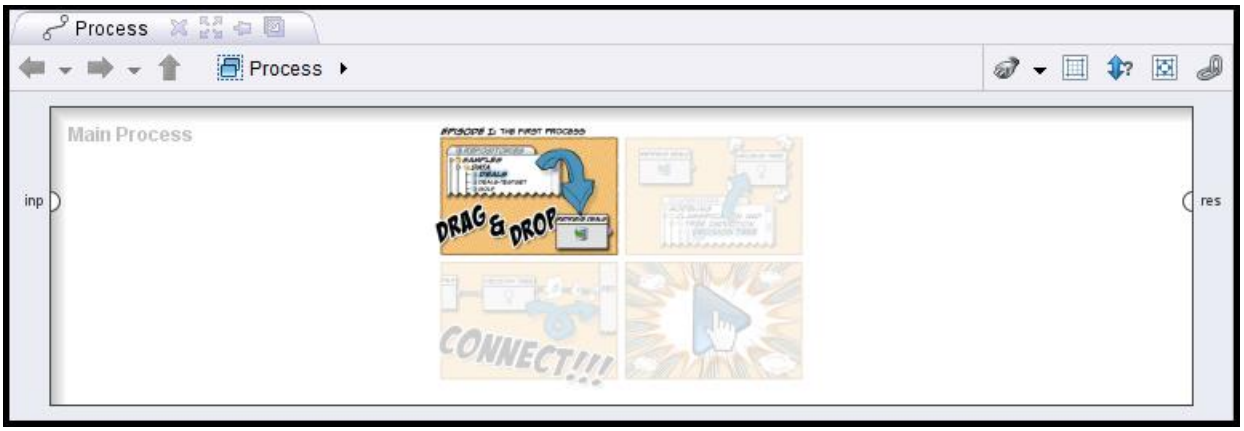
b. On the submenu, select Show View (Show View).

c. On the next submenu, select Process (Process).



Please proceed to next page.

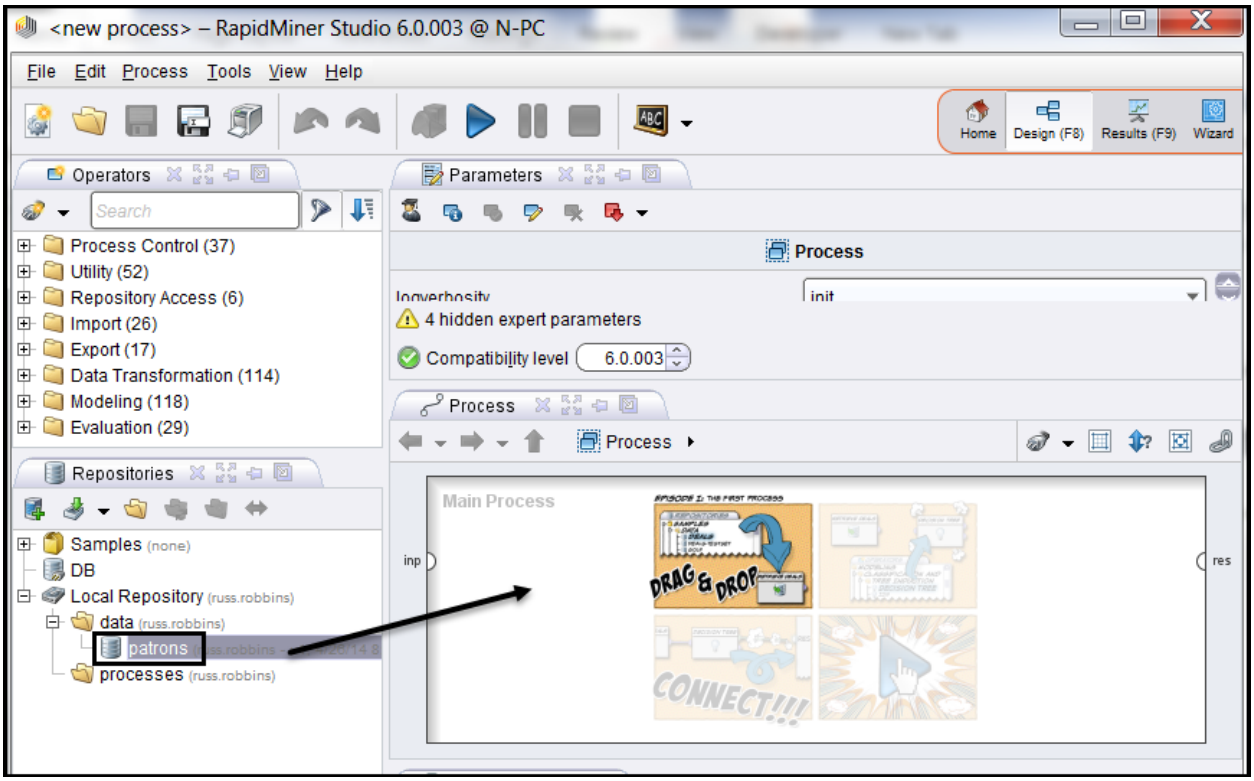
11. Your screen should now have the following Process window within it.
a. This window is known as the Process Perspective.



12. Select the data that you have imported 

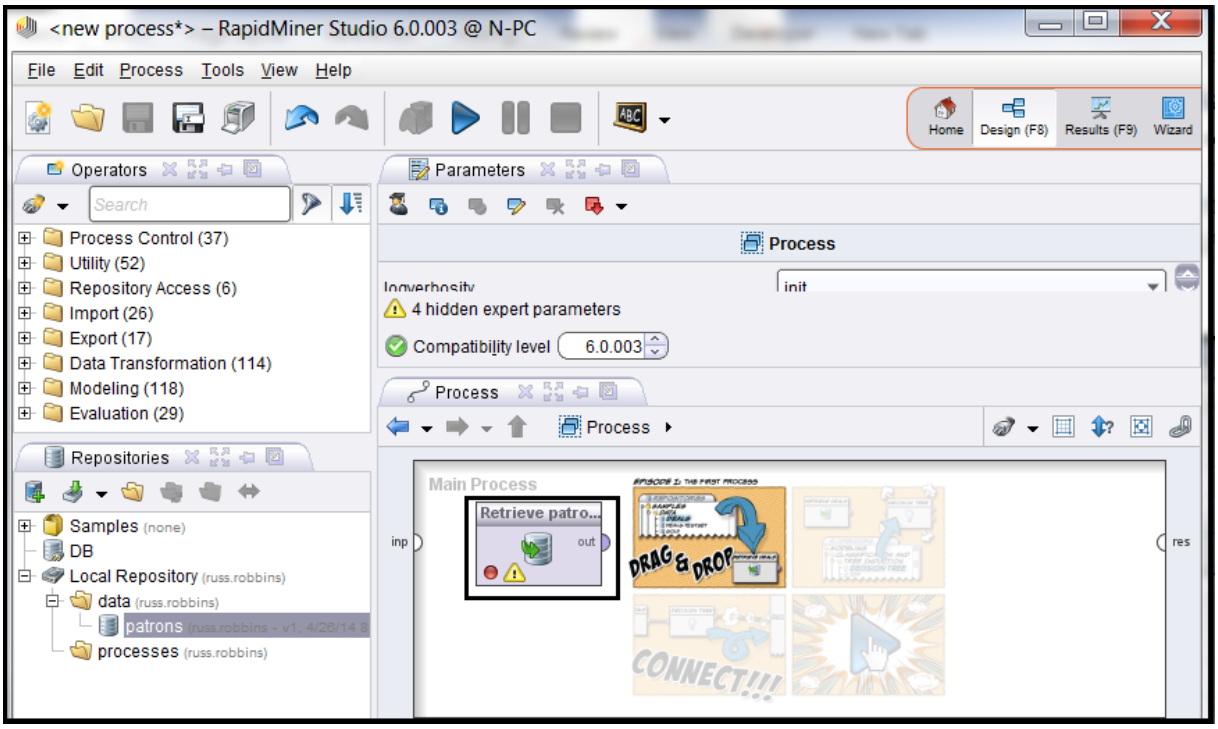
(patrons) and "pull" this data into the 

(Process) perspective. By placing the imported data into the process you are indicating that the software should retrieve and use this data.



Please proceed to next page.

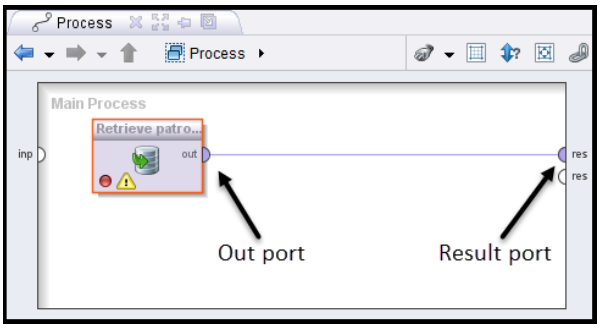
13. The result should be as shown below.
b. Note that the square box that now in the Process perspective has the name "Retrieve patrons".
c. Note that this square box is referred to as an "operator."



14. On the right upper side of the "Retrieve patrons" operator, you will see a half-circle that is purple. This is the "out" "port" of the operator.

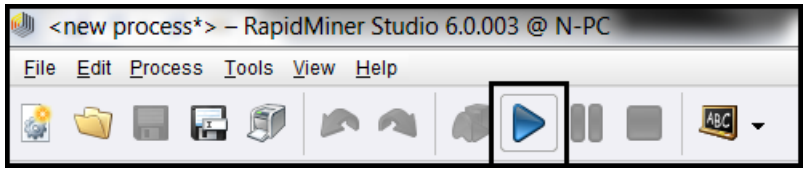
15. On the right side of the Process perspective, you will see another half circle that is purple, and which is named "res". This is a "result" port. (A "port" is just a place where data goes out or in, just like an ocean port.)

16. Select the out port and drag your mouse over to the res port and then release your mouse. The result should be what you see below.

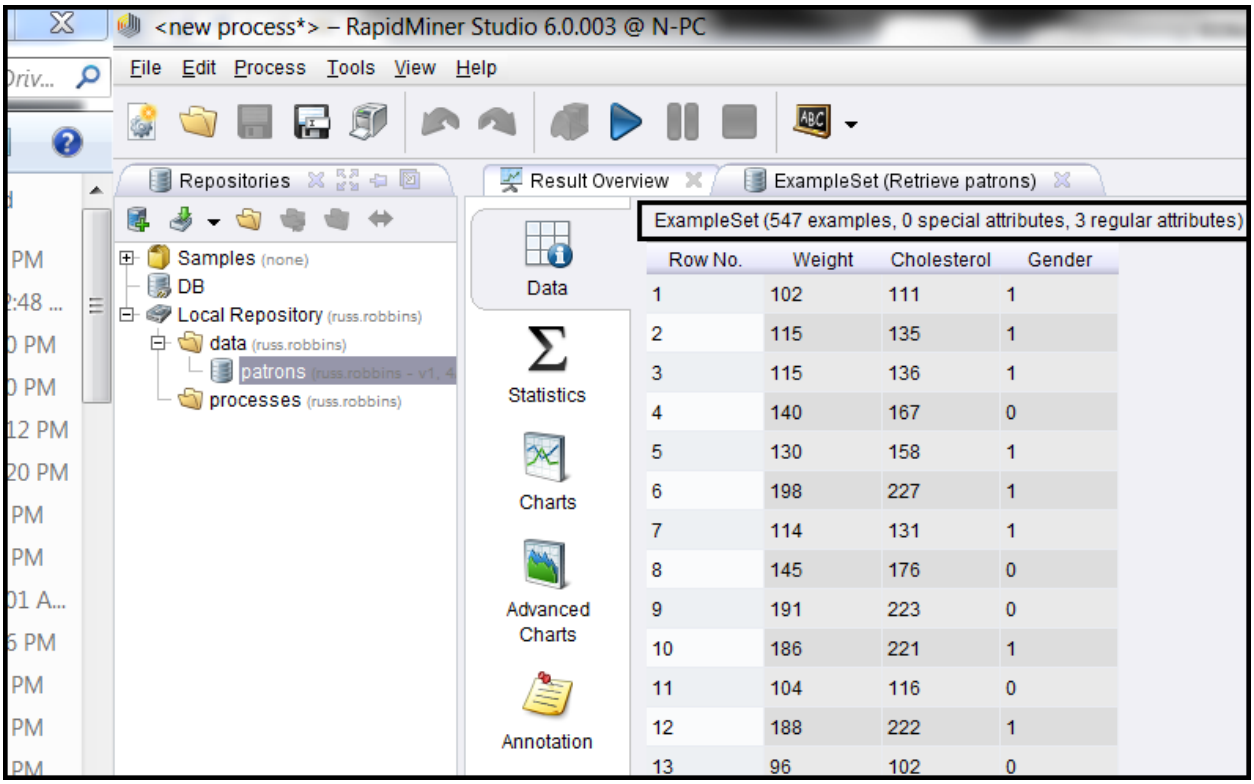


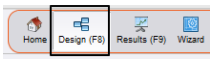
Please proceed to next page.

17. Click the large blue play button at the top of your screen to run the model.



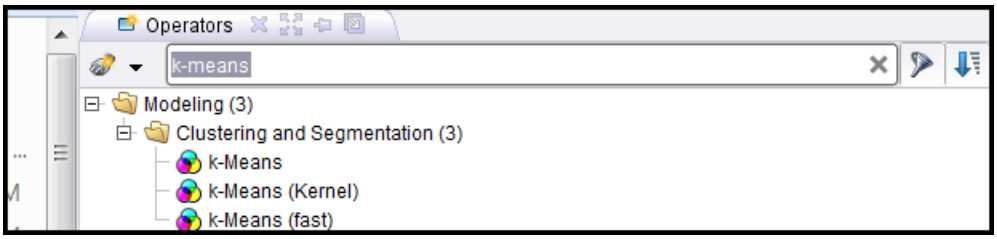
18. Your results should look like the following screenshot. As before, confirm 547 examples, 0 special attributes, and 3 regular attributes.



19. Select Design View. 

20. In the Operators Search Box in the top left of your screen, type k-means (k-means).
(Be sure to include the hyphen.)

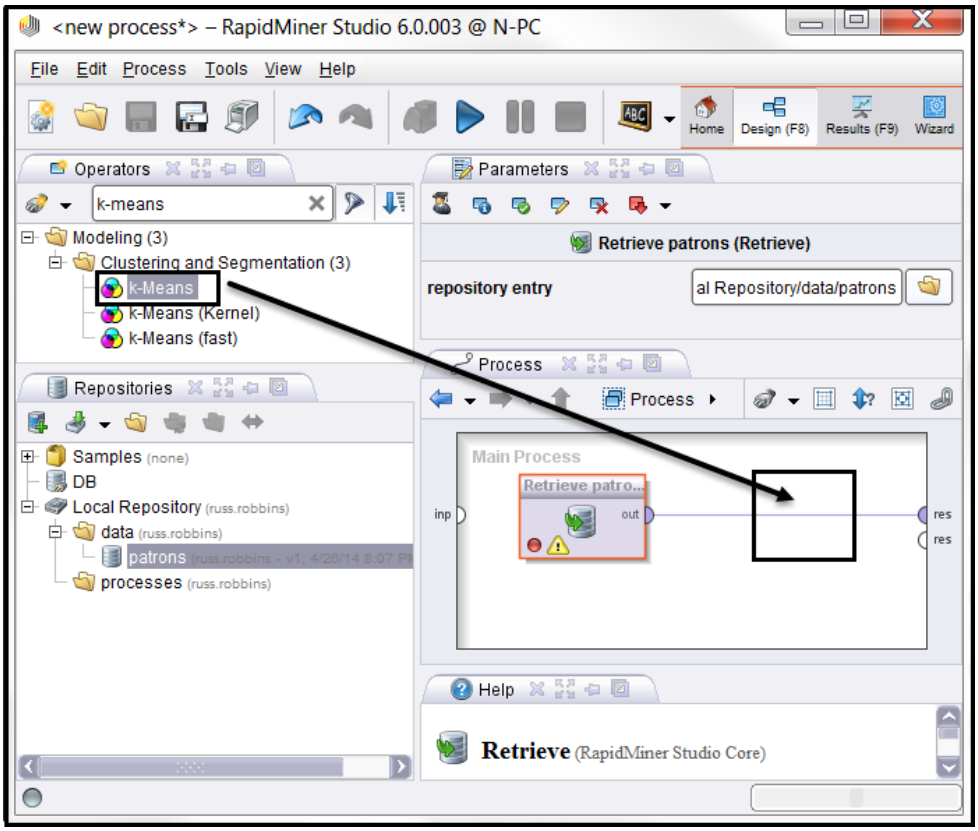
21. After typing "k-means" you should see the following. **Please proceed to next page.**



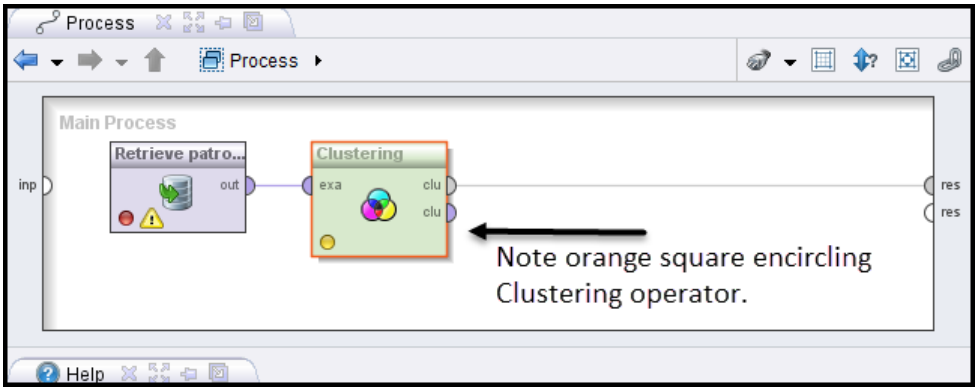
22. Select the k-Means (k-means) algorithm operator and "pull" it onto the purple line connecting the Retrieve patrons out operator and the result operator on the right side of the Process perspective.

a. Do not release your mouse button until the purple line between the out port and the res port becomes bold purple.

b. If your k-means operator is surrounded with an orange-lined square, then you know that you have placed the operator correctly.

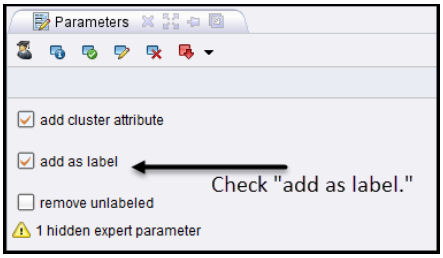


23. Your resulting Process perspective should look like the screenshot below.



Please proceed to next page.

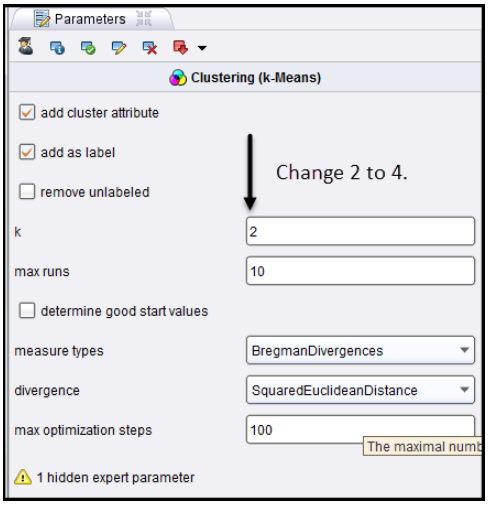
24. If "add as label" is not checked, please do this.



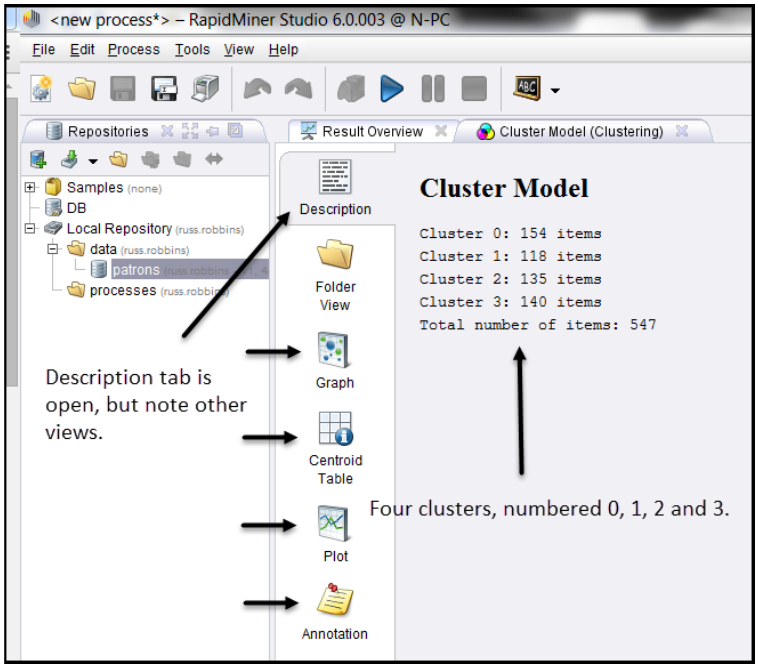
25. At this point you could run the "k-means" algorithm that you have "programmed" into RapidMiner. However, as the program director, you have decided you want four "clusters" or groupings of patrons, instead of what RapidMiner does as a default, which is two. So, if necessary, maximize your 

(Parameters) perspective, by clicking the 

location noted that is next to the X. You should see a perspective (a window) that looks like what is below.



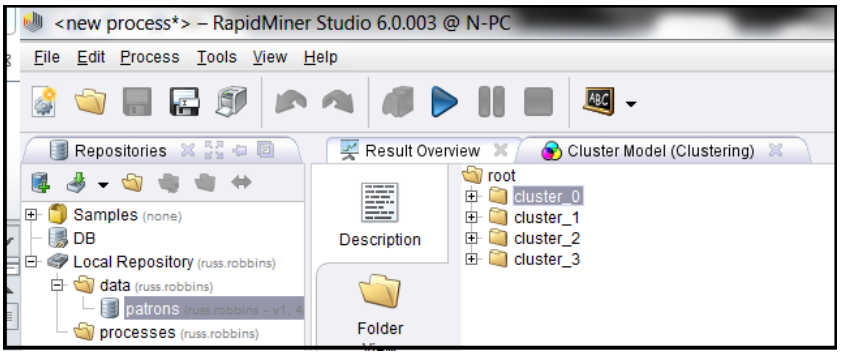
26. Run the k-means clustering algorithm on the data by now pressing the  (Play) button. Note that four clusters have been created. Note that there are different "views" of the results. **Please proceed to next page.**



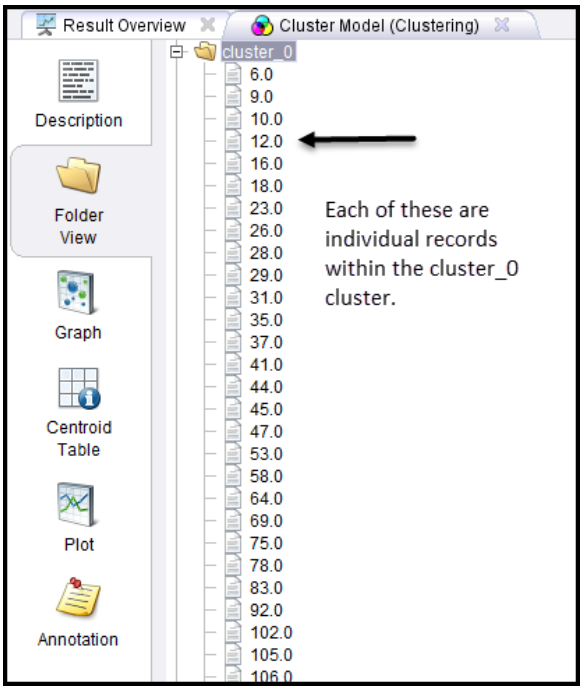
27. Select the folder view.

a. Note there are four clusters (cluster_0, cluster_1, cluster_2, and cluster_3).

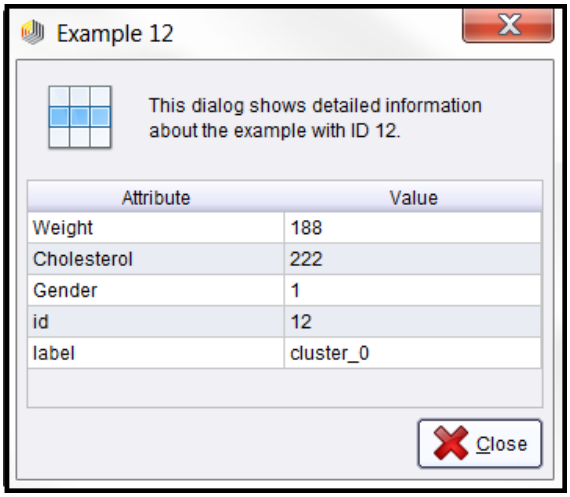
b. Note the names of the clusters include the underscore (_) character.



28. Double -click on cluster_0. You should see a view similar to below.

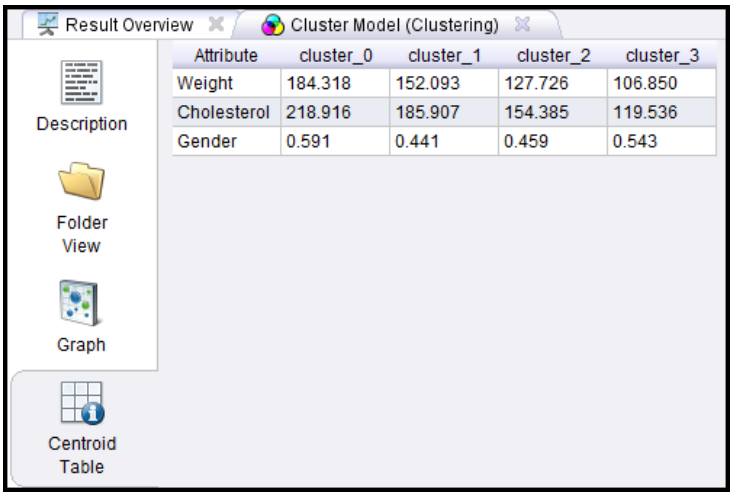


29. Now use your mouse to click on record 12. This screen represents one record that has been classified into cluster_0, using the k-means algorithm. Note that 12 is the "id" of the record and not data from the CSV file.

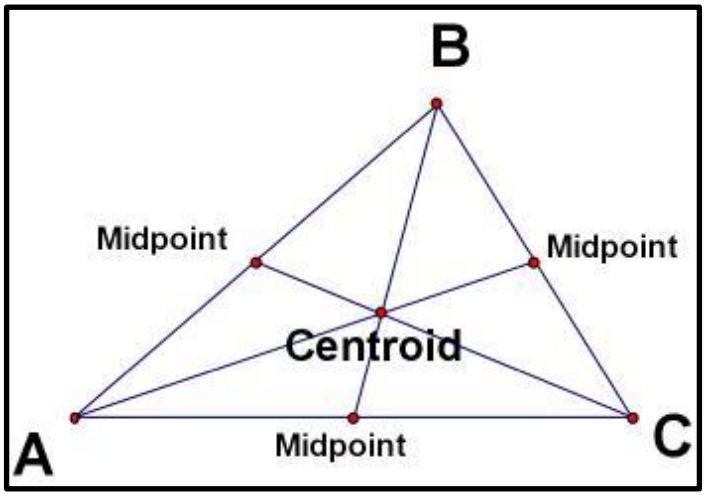


Please proceed to next page.

30. Select the Centroid View.
a. Note the centroid weights, cholesterol, and gender.
b. A centroid is not identical to the average but in this example, they are identical.
c. I have placed a graphic of a centroid to the right of the centroid table to help you see that a centroid is about geographical space while an average is a computation that can be made without any consideration of space.



Attribute	cluster_0	cluster_1	cluster_2	cluster_3
Weight	184.318	152.093	127.726	106.850
Cholesterol	218.916	185.907	154.385	119.536
Gender	0.591	0.441	0.459	0.543



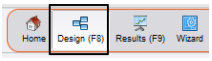
31. Which cluster do you think you will want to target with your "heart healthy" programs? Why?

32. Which cluster would you target second?

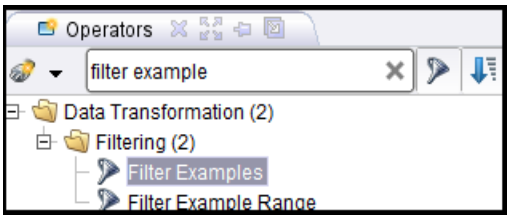
33. What does the gender centroid of .591 for cluster_0 mean?

34. After answering these questions, feel free to review the Graph and Plot views.

35. To get more statistical information about the clusters, we can now add an operator to our process. The operator we are going to add is "Filter Examples."

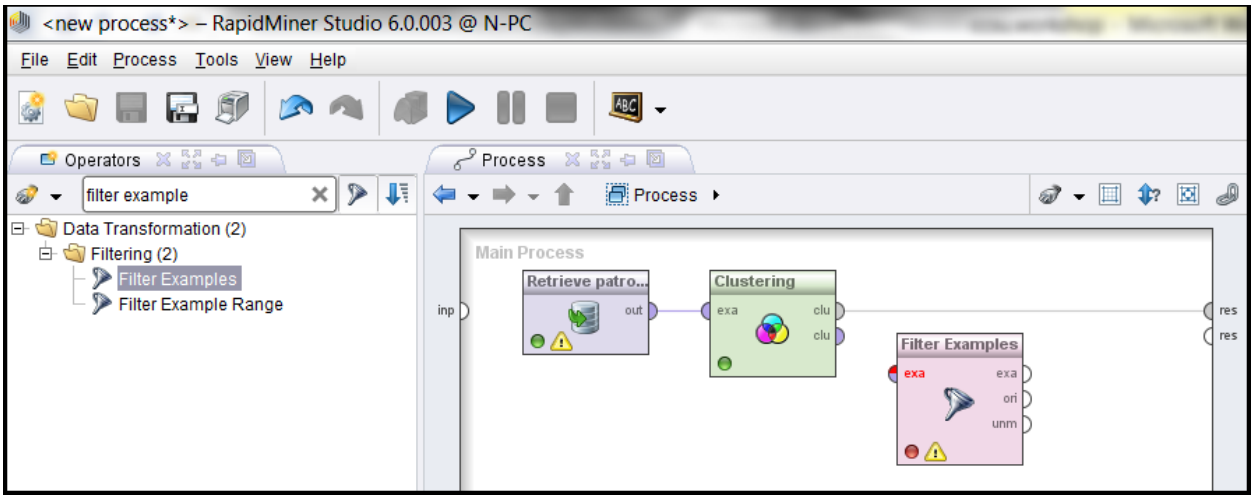
36. Select Design View.  Then type "Filter Examples" into the Operators search field.

37. The result should be as shown below.

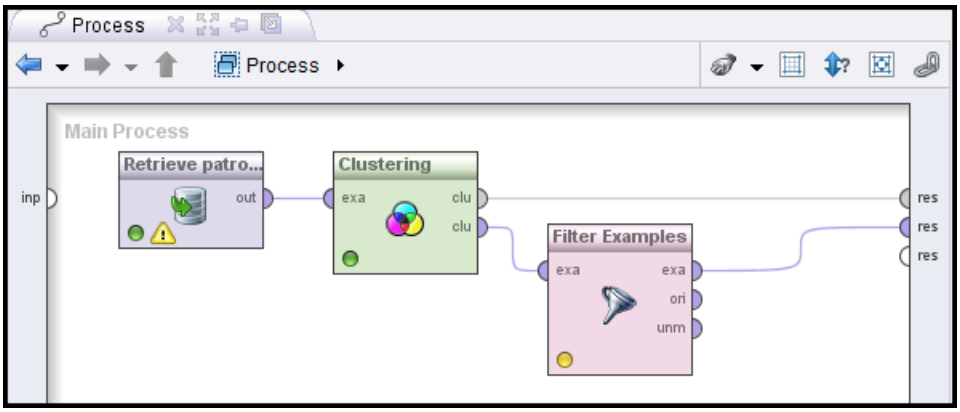


Please proceed to next page.

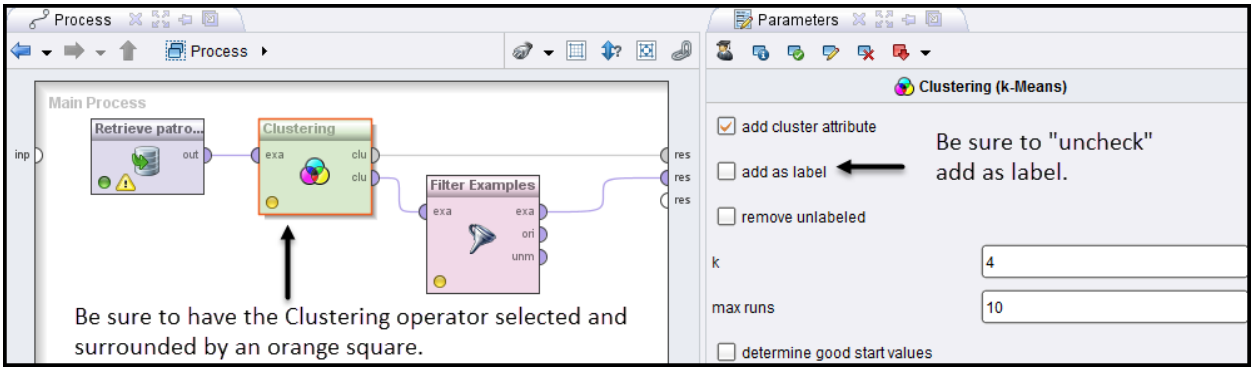
38. Then "pull" the Filter Examples operator into the Process perspective, so that your Filter Examples operator is as below.



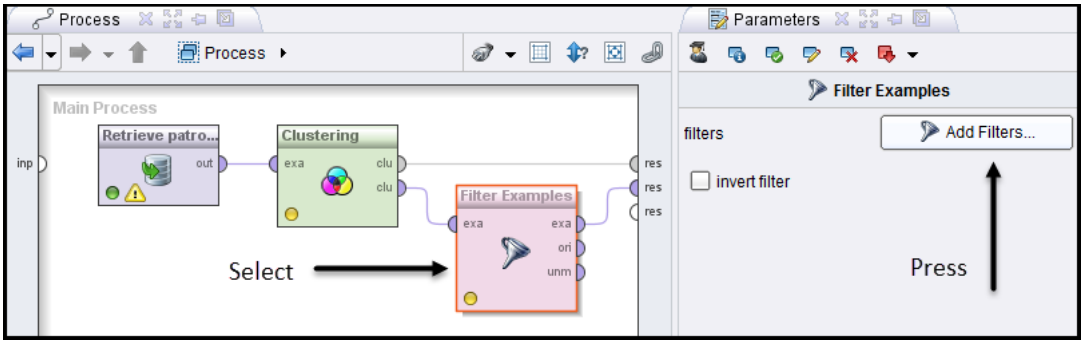
39. Now connect the clu (clustering) port on the right side of the Clustering operator with the exa (examples) port on the left of the Filter Examples operator and the exa (examples) port on the right of the Filter Examples port to the res (results) port on the right of the Process perspective. The result should be as shown below.



40. Before we use the new process, we have to set parameters on the Clustering and the Filter Examples operators.
a. Select the Clustering operator in the Process perspective.
b. In the Parameters perspective, be sure to uncheck the "add as label" checkbox. **Please proceed to next page.**

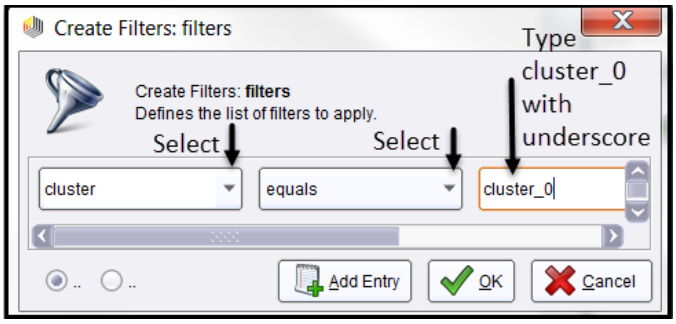


41. Now let's set parameters on Filter Examples operator.
a. Select the Filter Examples operator in the Process perspective.
b. In the Parameters perspective, press the "Add Filters..." button.

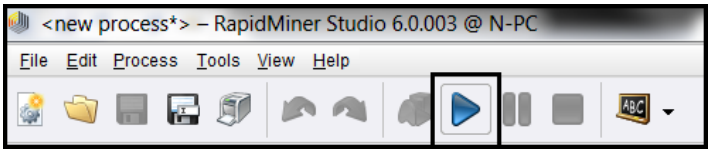


42. Fill in the Create Filters dialog box as below. Be sure to type "cluster_0" exactly as below.

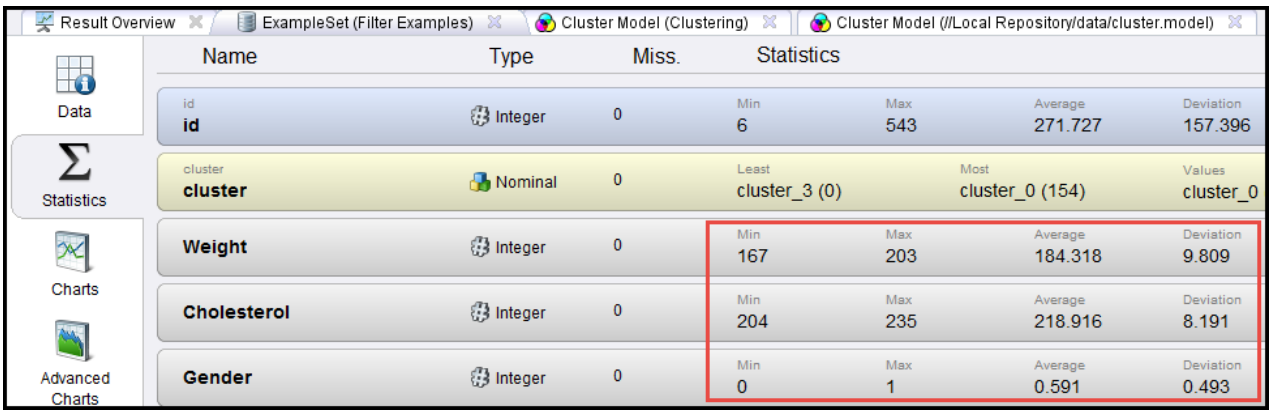
Press OK (OK) when you are finished.



43. Click the large blue play button at the top of your screen to run the model.

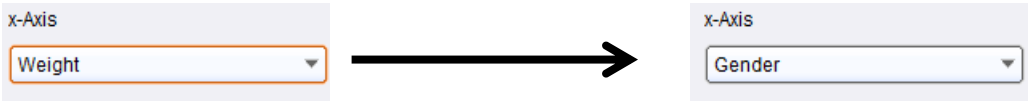


44. Switch to the Statistics view. Σ Statistics Note this view which summarizes weight, cholesterol, and gender for cluster_0. Note that most of the individuals in cluster_0 are men, since the "average" gender is 0.591. The average is 0.591 because there are more records with 1 in the gender field than there are 0. If you remember, 1 refers to males.



Please proceed to next page.

46. Remain in the Charts View and change the x-axis from Weight to Gender.



47. The chart you should now see should be as below.

a. Note that it appears that both genders (male, 1) and (female, 0) in the cluster_0 seem to have cholesterol in the above 200 range, and, at least by looking at this chart, it doesn't seem like a person's gender can "explain" why a person might have high (over 200) cholesterol.

Conclusion:

By using the k-means clustering algorithm, and knowing that those with the highest levels of cholesterol and weight are the most susceptible to heart disease, you have identified some patrons with the highest risk. With this information, you will be able to return to your company's database and identify the names and contact information for these individuals insured by your company. Using their contact information, you can begin to reach out to these individuals so that they can consider whether they want to be involved in your upcoming "heart healthy" program.

Congratulations! on completing this "hands on" demonstration of the use of the k-Means clustering algorithm.

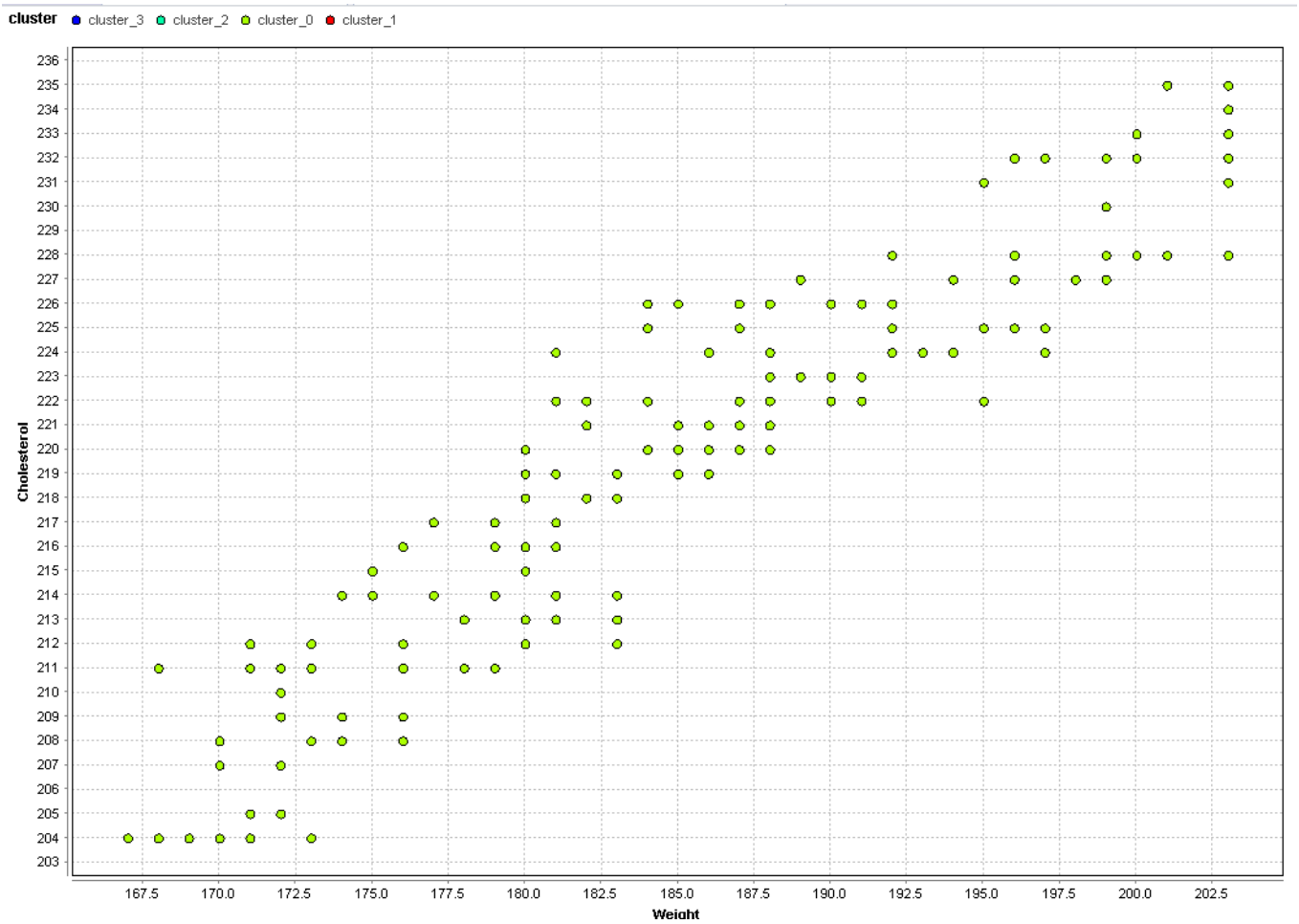
45. Switch to the 

(Charts) view of the cluster_0 data.

a. Note the only data on the chart is for records summarizing individuals in cluster_0.

b. Note the linear relationship between weight and cholesterol. (As weight goes up, so does cholesterol.)

(You have an upcoming meeting with clinicians to begin developing the "heart healthy" programs. Perhaps you could ask them about this possible relationship between weight and cholesterol, and how this may figure into the development of the heart healthy programs.)



Please proceed to next page.

Section O- Example: Using machine learning tools

This section provides an example of using machine learning (ML) tools.

(Please turn the page.)

I have used all major R and Python integrated development environments, a long list of R libraries (including caret, a dominant ML library), a long list of Python packages (including scikit-learn (a dominant ML package), as well as Apache Spark, an open source distributed processing ML tool, which when used for some applications, can by 100 times as fast as Hadoop methods.

In [2]:
```
import graphlab
```

In [3]:
```
products = graphlab.SFrame('amazon_baby.gl/')
```

```
[INFO] This commercial license of GraphLab Create is assigned to engr@dato.com.



[INFO] Start server at: ipc:///tmp/graphlab_server-121368 - Server binary: /home/ubuntu/anacon
rver_1440696851.log

[INFO] GraphLab Server Version: 1.5.2
```

In [3]:
```
products.head()
```

Out[3]:

name	review	rating
Planetwise Flannel Wipes	These flannel wipes are OK, but in my opinion ...	3.0
Planetwise Wipe Pouch	it came early and was not disappointed. i love ...	5.0
Annas Dream Full Quilt with 2 Shams ...	Very soft and comfortable and warmer than it ...	5.0
Stop Pacifier Sucking without tears with ...	This is a product well worth the purchase. I ...	5.0
Stop Pacifier Sucking without tears with ...	All of my kids have cried non-stop when I tried to ...	5.0
Stop Pacifier Sucking without tears with ...	When the Binky Fairy came to our house, we didn't ...	5.0
A Tale of Baby's Days with Peter Rabbit ...	Lovely book, it's bound tightly so you may no ...	4.0
Baby Tracker® - Daily Childcare Journal, ...	Perfect for new parents. We were able to keep ...	5.0
Baby Tracker® - Daily Childcare Journal, ...	A friend of mine pinned this product on Pinte ...	5.0
Baby Tracker® - Daily Childcare Journal, ...	This has been an easy way for my nanny to record ...	4.0

[10 rows x 3 columns]

In [8]:
```
giraffe_reviews = products[products['name'] == 'Vulli Sophie the Giraffe Teether']
```

In [9]:
```
len(giraffe_reviews)
```

Out[9]: 785

In [10]:
```
giraffe_reviews['rating'].show(view='Categorical')
```

#Build a sentiment classifier

In [11]:
```
products['rating'].show(view='Categorical')
```

In [3]:
```
products.head()
```

Out[3]:

name	review	rating
Planetwise Flannel Wipes	These flannel wipes are OK, but in my opinion ...	3.0
Planetwise Wipe Pouch	it came early and was not disappointed. i love ...	5.0
Annas Dream Full Quilt with 2 Shams ...	Very soft and comfortable and warmer than it ...	5.0
Stop Pacifier Sucking without tears with ...	This is a product well worth the purchase. I ...	5.0
Stop Pacifier Sucking without tears with ...	All of my kids have cried non-stop when I tried to ...	5.0
Stop Pacifier Sucking without tears with ...	When the Binky Fairy came to our house, we didn't ...	5.0
A Tale of Baby's Days with Peter Rabbit ...	Lovely book, it's bound tightly so you may no ...	4.0
Baby Tracker® - Daily Childcare Journal, ...	Perfect for new parents. We were able to keep ...	5.0
Baby Tracker® - Daily Childcare Journal, ...	A friend of mine pinned this product on Pinte ...	5.0
Baby Tracker® - Daily Childcare Journal, ...	This has been an easy way for my nanny to record ...	4.0

[10 rows x 3 columns]

In [14]:
```
#ignore all 3* reviews
products = products[products['rating'] != 3]
```

In [15]:
```
#positive sentiment = 4* or 5* reviews
products['sentiment'] = products['rating'] >=4
```

In [16]:
```
products.head()
```

Out[16]:

name	review	rating	word_count	sentiment
Planetwise Wipe Pouch	it came early and was not disappointed. i love ...	5.0	{'and': 3, 'love': 1, 'it': 2, 'highly': 1, ...	1
Annas Dream Full Quilt with 2 Shams ...	Very soft and comfortable and warmer than it ...	5.0	{'and': 2, 'quilt': 1, 'it': 1, 'comfortable': ...	1
Stop Pacifier Sucking without tears with ...	This is a product well worth the purchase. I ...	5.0	{'ingenious': 1, 'and': 3, 'love': 2, ...	1
Stop Pacifier Sucking without tears with ...	All of my kids have cried non-stop when I tried to ...	5.0	{'and': 2, 'parents!!': 1, 'all': 2, 'puppet.': ...	1
Stop Pacifier Sucking without tears with ...	When the Binky Fairy came to our house, we didn't ...	5.0	{'and': 2, 'cute': 1, 'help': 2, 'doll': 1, ...	1
A Tale of Baby's Days with Peter Rabbit ...	Lovely book, it's bound tightly so you may no ...	4.0	{'shop': 1, 'be': 1, 'is': 1, 'it': 1, 'as': ...	1
Baby Tracker® - Daily Childcare Journal, ...	Perfect for new parents. We were able to keep ...	5.0	{'feeding,': 1, 'and': 2, 'all': 1, 'right': 1, ...	1

In [6]:
```
graphlab.canvas.set_target('ipynb')
```

In [7]:
```
products['name'].show()
```

In [8]:
```
giraffe_reviews = products[products['name'] == 'Vulli Sophie the Giraffe Teether']
```

In [9]:
```
len(giraffe_reviews)
```

Out[9]: 785

In [10]:
```
giraffe_reviews['rating'].show(view='Categorical')
```

In [17]:
```
train_data,test_data = products.random_split(.8, seed=0)
```

In [18]:
```
sentiment_model = graphlab.logistic_classifier.create(train_data,
                                                      target='sentiment',
                                                      features=['word_count'],
                                                      validation_set=test_data)
```

```
PROGRESS: Logistic regression:
PROGRESS: --------------------------------------------------------
PROGRESS: Number of examples          : 133448
PROGRESS: Number of classes           : 2
PROGRESS: Number of feature columns   : 1
PROGRESS: Number of unpacked features : 219217
PROGRESS: Number of coefficients    : 219218
PROGRESS: Starting L-BFGS
PROGRESS: --------------------------------------------------------
PROGRESS: +-----------+----------+-----------+-------------+-------------------+---------------------+
PROGRESS: | Iteration | Passes   | Step size | Elapsed Time | Training-accuracy | Validation-accuracy |
PROGRESS: +-----------+----------+-----------+-------------+-------------------+---------------------+
PROGRESS: | 1         | 5        | 0.000002  | 1.737795     | 0.841481          | 0.839989            |
PROGRESS: | 2         | 9        | 3.000000  | 2.388519     | 0.947425          | 0.894877            |
PROGRESS: | 3         | 10       | 3.000000  | 2.630476     | 0.923768          | 0.866232            |
PROGRESS: | 4         | 11       | 3.000000  | 2.868154     | 0.971779          | 0.912743            |
PROGRESS: | 5         | 12       | 3.000000  | 3.106237     | 0.975511          | 0.908900            |
PROGRESS: | 6         | 13       | 3.000000  | 3.348601     | 0.899991          | 0.825967            |
PROGRESS: | 10        | 18       | 1.000000  | 4.446451     | 0.988715          | 0.916256            |
PROGRESS: +-----------+----------+-----------+-------------+-------------------+---------------------+
```

#Evaluate the sentiment model

In [19]:
```
sentiment_model.evaluate(test_data, metric='roc_curve')
```

Out[19]:
```
{'roc_curve': Columns:
        threshold       float
        fpr     float
        tpr     float
        p       int
        n       int

 Rows: 1001
```

In [20]:
```
giraffe_reviews['predicted_sentiment'] = sentiment_model.predict(giraffe_reviews, output
```

In [21]:
```
giraffe_reviews.head()
```

Out[21]:

name	review	rating	word_count	predicted_sentiment
Vulli Sophie the Giraffe Teether ...	He likes chewing on all the parts especially the ...	5.0	{'and': 1, 'all': 1, 'because': 1, 'it': 1, ...	0.999513023521
Vulli Sophie the Giraffe Teether ...	My son loves this toy and fits great in the diaper ...	5.0	{'and': 1, 'right': 1, 'help': 1, 'just': 1, ...	0.999320678306
Vulli Sophie the Giraffe Teether ...	There really should be a large warning on the ...	1.0	{'and': 2, 'all': 1, 'latex.': 1, 'being': 1, ...	0.013558811687
Vulli Sophie the Giraffe Teether ...	All the moms in my moms' group got Sophie for ...	5.0	{'and': 2, 'one!': 1, 'all': 1, 'love': 1, ...	0.995769474148
Vulli Sophie the Giraffe Teether ...	I was a little skeptical on whether Sophie was ...	5.0	{'and': 3, 'all': 1, 'old': 1, 'her.': 1, ...	0.662374415673
Vulli Sophie the Giraffe Teether ...	I have been reading about Sophie and was going ...	5.0	{'and': 6, 'seven': 1, 'already': 1, 'love': 1, ...	0.999997148186
Vulli Sophie the Giraffe Teether ...	My neice loves her sophie and has spent hours ...	5.0	{'and': 4, 'drooling.': 1, 'love': 1, 'her.': 1, ...	0.989190989536
Vulli Sophie the Giraffe Teether ...	What a friendly face! And those mesmerizing ...	5.0	{'and': 3, 'chew': 1, 'don't': 1, 'is': 1, ...	0.999563518413
Vulli Sophie the Giraffe Teether ...	We got this just for my son to chew on instea ...	5.0	{'chew': 2, 'because': 1, 'just': 2, 'what': 1, ...	0.970160542725
Vulli Sophie the Giraffe Teether ...	My baby seems to like this toy, but I could ...	3.0	{'and': 2, 'already': 1, 'in': 1, 'some': 1, ' ...	0.195367644588

[10 rows x 5 columns]

In [20]:
```
giraffe_reviews['predicted_sentiment'] = sentiment_model.predict(giraffe_reviews, output
```

In [21]:
```
giraffe_reviews.head()
```

Out[21]:

	name	review	rating	word_count	predicted_sentiment
	Vulli Sophie the Giraffe Teether ...	He likes chewing on all the parts especially the ...	5.0	{'and': 1, 'all': 1, 'because': 1, 'it': 1, ...	0.999513023521
	Vulli Sophie the Giraffe Teether ...	My son loves this toy and fits great in the diaper ...	5.0	{'and': 1, 'right': 1, 'help': 1, 'just': 1, ...	0.999320678306
	Vulli Sophie the Giraffe Teether ...	There really should be a large warning on the ...	1.0	{'and': 2, 'all': 1, 'latex.': 1, 'being': 1, ...	0.013558811687
	Vulli Sophie the Giraffe Teether ...	All the moms in my moms' group got Sophie for ...	5.0	{'and': 2, 'one!': 1, 'all': 1, 'love': 1, ...	0.995769474148
	Vulli Sophie the Giraffe Teether ...	I was a little skeptical on whether Sophie was ...	5.0	{'and': 3, 'all': 1, 'old': 1, 'her.': 1, ...	0.662374415673
	Vulli Sophie the Giraffe Teether ...	I have been reading about Sophie and was going ...	5.0	{'and': 6, 'seven': 1, 'already': 1, 'love': 1, ...	0.999997148186
	Vulli Sophie the Giraffe Teether ...	My neice loves her sophie and has spent hours ...	5.0	{'and': 4, 'drooling,': 1, 'love': 1, 'her.': 1, ...	0.989190989536
	Vulli Sophie the Giraffe Teether ...	What a friendly face! And those mesmerizing ...	5.0	{'and': 3, 'chew': 1, 'don't': 1, 'is': 1, ...	0.999563518413
	Vulli Sophie the Giraffe Teether ...	We got this just for my son to chew on instea ...	5.0	{'chew': 2, 'because': 1, 'just': 2, 'what': 1, ...	0.970160542725
	Vulli Sophie the Giraffe Teether ...	My baby seems to like this toy, but I could ...	3.0	{'and': 2, 'already': 1, 'in': 1, 'some': 1, ' ...	0.195367644588

[10 rows x 5 columns]

Data:

threshold	fpr	tpr	p	n
0.0	0.216672925272	0.00536278722875	28157	5326
0.0010000000475	0.783327074728	0.994637212771	28157	5326
0.00200000009499	0.744461134059	0.993323152324	28157	5326
0.00300000002608	0.722305670297	0.992577334233	28157	5326
0.00400000018999	0.707097258731	0.992044607025	28157	5326
0.00499999988824	0.695456252347	0.991760485847	28157	5326
0.00600000005215	0.686068343973	0.991156728345	28157	5326
0.00700000021607	0.675553886594	0.990766061725	28157	5326
0.00800000037998	0.665790461885	0.990339879959	28157	5326
0.00899999961257	0.658467893353	0.989913698192	28157	5326

[1001 rows x 5 columns]

Note: Only the head of the SFrame is printed.

You can use print_rows(num_rows=m, num_columns=n) to print more rows and columns.}

[20]:
```
sentiment_model.show(view='Evaluation')
```

In [22]:
```
giraffe_reviews = giraffe_reviews.sort('predicted_sentiment', ascending=False)
```

In [24]:
```
giraffe_reviews.head()
```

Out[24]:

name	review	rating	word_count	predicted_sentiment
Vulli Sophie the Giraffe Teether ...	Sophie, oh Sophie, your time has come. My ...	5.0	{'giggles': 1, 'all': 1, 'violet's': 2, 'food' ...	1.0
Vulli Sophie the Giraffe Teether ...	I'm not sure why Sophie is such a hit with the ...	4.0	{'peace': 1, 'month': 1, 'bright': 1, 'softer' ...	0.999999999703
Vulli Sophie the Giraffe Teether ...	I'll be honest...I bought this toy because all the ...	4.0	{'all': 2, 'pops': 1, 'existence.': 1, ...	0.999999999392
Vulli Sophie the Giraffe Teether ...	We got this little giraffe as a gift from a ...	5.0	{'all': 2, 'don't': 1, '(literally).so': 1, ...	0.99999999919
Vulli Sophie the Giraffe Teether ...	As a mother of 16month old twins; I bought ...	5.0	{'cute': 1, 'all': 1, 'reviews.': 2, 'just' ...	0.999999998657
Vulli Sophie the Giraffe Teether ...	Sophie the Giraffe is the perfect teething toy. ...	5.0	{'just': 2, 'both': 1, 'month': 1, 'ears,': 1, ...	0.999999997108
Vulli Sophie the Giraffe Teether ...	Sophie la giraffe is absolutely the best toy ...	5.0	{'and': 5, 'the': 1, 'all': 1, 'that': 2, ...	0.999999995589
Vulli Sophie the Giraffe Teether ...	My 5-mos old son took to this immediately. The ...	5.0	{'just': 1, 'shape': 2, 'mutt': 1, ''dog': 1, ...	0.999999995573
Vulli Sophie the Giraffe Teether ...	My nephews and my four kids all had Sophie in ...	5.0	{'and': 4, 'chew': 1, 'all': 1, 'perfect,': 1, ...	0.999999989527
Vulli Sophie the Giraffe Teether ...	Never thought I'd see my son French kissing a ...	5.0	{'giggles': 1, 'all': 1, 'out,': 1, 'over': 1, ...	0.999999985069

[10 rows x 5 columns]

In [25]:
```
giraffe_reviews[0]['review']
```

Out[25]:
"Sophie, oh Sophie, your time has come. My granddaughter, Violet is 5 months old and starting to teeth. What joy little Sophie brings to Violet. Sophie is made of a very pliable rubber that is sturdy but not tough. It is quite easy for Violet to twist Sophie into unheard of positions to get Sophie into her mouth. The little nose and hooves fit perfectly into small mouths, and the drooling has purpose. The paint on Sophie is food quality.Sophie was born in 1961 in France. The maker had wondered why there was nothing available for babies and made Sophie from the finest rubber, phthalate-free on St Sophie's Day, thus the name was born. Since that time millions of Sophie's populate the world. She is soft and for babies little hands easy to grasp. Violet especially loves the bumpy head and horns of Sophie. Sophie has a long neck that easy to grasp and twist. She has lovely, sizable spots that attract Violet's attention. Sophie has happy little squeaks that bring squeals of delight from Violet. She is able to make Sophie squeak and that brings much joy. Sophie's smooth skin is soothing to Violet's little gums. Sophie is 7 inches tall and is the exact correct size for babies to hold and love.As you well know the first thing babies grasp, goes into their mouths- how wonderful to have a toy that stimulates all of the senses and helps with the issue of teething. Sophie is small enough to fit into any size pocket or bag. Sophie is the perfect find for babies from a few months to a year old. How wonderful to hear the giggles and laughs that emanate from babies who find Sophie irresistible. Viva La Sophie!Highly Recommended. prisrob 12-11-09"

In [26]:
```
giraffe_reviews[1]['review']
```

Out[26]:
"I'm not sure why Sophie is such a hit with the little ones, but my 7 month old baby girl is one of her adoring fans. The rubber is softer and more pleasant to handle, and my daughter has enjoyed chewing on her legs and the nubs on her head even before she started teething. She also loves the squeak that Sophie makes when you squeeze her. Not sure what it is but if Sophie is amongst a pile of her other toys, my daughter will more often than not reach for Sophie. And I have the peace of mind of knowing that only edible and safe paints and materials have been used to make Sophie, as opposed to Bright Starts and other baby toys made in China. Now that the research is out on phthalates and other toxic substances in baby toys, I think it's more important than ever to find good quality toys that are also safe for our babies to handle and put in their mouths. Sophie is a must-have for every new mom in my opinion. Even if your kid is one of the few that can take or leave her, it's worth a try. Vulli, the makers of Sophie, also make natural rubber teething rings that my daughter loves as well."

##Show most negative reviews for giraffe

In [27]:
```
giraffe_reviews[-1]['review']
```

Out[27]:
"My son (now 2.5) LOVED his Sophie, and I bought one for every baby shower I've gone to. Now, my daughter (6 months) just today nearly choked on it and I will never give it to her again. Had I not been within hearing range it could have been fatal. The strange sound she was making caught my attention and when I went to her and found the front curved leg shoved well down her throat and her face a purply/blue I panicked. I pulled it out and she vomited all over the carpet before screaming her head off. I can't believe how my opinion of this toy has changed from a must-have to a must-not-use. Please don't disregard any of the choking hazard comments, they are not over exaggerated!"

In [28]:
```
giraffe_reviews[-2]['review']
```

Out[28]:
"This children's toy is nostalgic and very cute. However, there is a distinct rubber smell and a very odd taste, yes I tried it, that my baby did not enjoy. Also, if it is soiled it is extremely difficult to clean as the rubber is a kind of porus material and does not clean well. The final thing is the squeaking device inside which stopped working after the first couple of days. I returned this item feeling I had overpaid for a toy that was defective and did not meet my expectations. Please do not be swayed by the cute packaging and hype surounding it as I was. One more thing, I was given a full refund from Amazon without any problem."

In [ ]:
```
```

Section P - Example: Using machine learning techniques

This section provides an example of a machine learning (ML) technique.

(Please turn the page.)

ML techniques I use include those that help:

- predict an object/event's category using other information (classification)
- predict an object/event's measurement value using other data (regression)
- identify previously unknown structure in groups of objects/events (clustering)
- use data subsets to measure effectiveness of an algorithm (resampling)
- choose predictive models that are likely to be more accurate (model selection)
- create additional information to use in predictive models (regularization)

```
{
 "cells": [
  {
   "cell_type": "markdown",
   "metadata": {},
   "source": [
    "#Using deep features to build an image classifier\n",
    "\n",
    "#Fire up GraphLab Create"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 26,
   "metadata": {},
   "outputs": [],
   "source": [
    "import graphlab"
   ]
  },
  {
   "cell_type": "markdown",
   "metadata": {},
   "source": [
    "#Load a common image analysis dataset\n",
    "\n",
    "We will use a popular benchmark dataset in computer vision called CIFAR-10.  \n",
    "\n",
    "(We've reduced the data to just 4 categories = {'cat','bird','automobile','dog'}.)\n",
    "\n",
    "This dataset is already split into a training set and test set.  "
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 27,
   "metadata": {},
   "outputs": [],
   "source": [
    "image_train = graphlab.SFrame('image_train_data/')\n",
    "image_test = graphlab.SFrame('image_test_data/')"
   ]
  },
  {
   "cell_type": "markdown",
   "metadata": {},
   "source": [
    "#Exploring the image data"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 28,
```

```
  "metadata": {},
  "outputs": [],
  "source": [
   "graphlab.canvas.set_target('ipynb')"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 4,
  "metadata": {},
  "outputs": [],
  "source": [
   "image_train['image'].show()"
  ]
 },
 {
  "cell_type": "markdown",
  "metadata": {},
  "source": [
   "#Train a classifier on the raw image pixels\n",
   "\n",
   "We first start by training a classifier on just the raw pixels of the image."
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 30,
  "metadata": {},
  "outputs": [],
  "source": [
   "raw_pixel_model = graphlab.logistic_classifier.create(image_train,target='label',\n",
   "                                                      features=['image_array'])"
  ]
 },
 {
  "cell_type": "markdown",
  "metadata": {},
  "source": [
   "#Make a prediction with the simple model based on raw pixels"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 31,
  "metadata": {},
  "outputs": [],
  "source": [
   "image_test[0:3]['image'].show()"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 32,
```

```
  "metadata": {},
  "outputs": [],
  "source": [
   "image_test[0:3]['label']"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 33,
  "metadata": {},
  "outputs": [],
  "source": [
   "raw_pixel_model.predict(image_test[0:3])"
  ]
 },
 {
  "cell_type": "markdown",
  "metadata": {},
  "source": [
   "The model makes wrong predictions for all three images."
  ]
 },
 {
  "cell_type": "markdown",
  "metadata": {},
  "source": [
   "#Evaluating raw pixel model on test data"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 34,
  "metadata": {},
  "outputs": [],
  "source": [
   "raw_pixel_model.evaluate(image_test)"
  ]
 },
 {
  "cell_type": "markdown",
  "metadata": {},
  "source": [
   "The accuracy of this model is poor, getting only about 46% accuracy."
  ]
 },
 {
  "cell_type": "markdown",
  "metadata": {},
  "source": [
   "#Can we improve the model using deep features\n",
   "\n",
   "We only have 2005 data points, so it is not possible to train a deep neural network
   effectively with so little data.  Instead, we will use transfer learning: using deep
```

```
   features trained on the full ImageNet dataset, we will train a simple model on this small
   dataset."
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 35,
  "metadata": {},
  "outputs": [],
  "source": [
   "len(image_train)"
  ]
 },
 {
  "cell_type": "markdown",
  "metadata": {},
  "source": [
   "##Computing deep features for our images\n",
   "\n",
   "The two lines below allow us to compute deep features.  This computation takes a little
   while, so we have already computed them and saved the results as a column in the data you
   loaded. \n",
   "\n",
   "(Note that if you would like to compute such deep features and have a GPU on your machine,
   you should use the GPU enabled GraphLab Create, which will be significantly faster for this
   task.)"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 36,
  "metadata": {},
  "outputs": [],
  "source": [
   "#deep_learning_model =
   graphlab.load_model('http://s3.amazonaws.com/GraphLab-Datasets/deeplearning/imagenet_model_it
   er45')\n",
   "#image_train['deep_features'] = deep_learning_model.extract_features(image_train)"
  ]
 },
 {
  "cell_type": "markdown",
  "metadata": {},
  "source": [
   "As we can see, the column deep_features already contains the pre-computed deep features
   for this data. "
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 37,
  "metadata": {},
  "outputs": [],
```

```
  "source": [
   "image_train.head()"
  ]
 },
 {
  "cell_type": "markdown",
  "metadata": {},
  "source": [
   "#Given the deep features, let's train a classifier"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 13,
  "metadata": {},
  "outputs": [],
  "source": [
   "deep_features_model = graphlab.logistic_classifier.create(image_train,\n",
   "                                                          features=['deep_features'],\n",
   "                                                          target='label')"
  ]
 },
 {
  "cell_type": "markdown",
  "metadata": {},
  "source": [
   "#Apply the deep features model to first few images of test set"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 14,
  "metadata": {},
  "outputs": [],
  "source": [
   "image_test[0:3]['image'].show()"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 15,
  "metadata": {},
  "outputs": [],
  "source": [
   "deep_features_model.predict(image_test[0:3])"
  ]
 },
 {
  "cell_type": "markdown",
  "metadata": {},
  "source": [
   "The classifier with deep features gets all of these images right!"
  ]
```

```
},
{
 "cell_type": "markdown",
 "metadata": {},
 "source": [
  "#Compute test_data accuracy of deep_features_model\n",
  "\n",
  "As we can see, deep features provide us with significantly better accuracy (about 78%)"
 ]
},
{
 "cell_type": "code",
 "execution_count": 16,
 "metadata": {},
 "outputs": [],
 "source": [
  "deep_features_model.evaluate(image_test)"
 ]
},
{
 "cell_type": "code",
 "execution_count": null,
 "metadata": {},
 "outputs": [],
 "source": [
  "knn_model.evaluate(image_test)"
 ]
},
{
 "cell_type": "code",
 "execution_count": null,
 "metadata": {},
 "outputs": [],
 "source": [
  ""
 ]
},
{
 "cell_type": "code",
 "execution_count": null,
 "metadata": {},
 "outputs": [],
 "source": [
  ""
 ]
},
{
 "cell_type": "code",
 "execution_count": null,
 "metadata": {},
 "outputs": [],
 "source": [
  ""
```

```
   ]
  },
  {
   "cell_type": "code",
   "execution_count": null,
   "metadata": {},
   "outputs": [],
   "source": [
    ""
   ]
  }
 ],
 "metadata": {
  "kernelspec": {
   "display_name": "Python 2",
   "language": "python",
   "name": "python2"
  },
  "language_info": {
   "codemirror_mode": {
    "name": "ipython",
    "version": 2.0
   },
   "file_extension": ".py",
   "mimetype": "text/x-python",
   "name": "python",
   "nbconvert_exporter": "python",
   "pygments_lexer": "ipython2",
   "version": "2.7.10"
  }
 },
 "nbformat": 4,
 "nbformat_minor": 0
}
```

Section Q - Example: Using natural language processing (NLP) tools and techniques

This section provides an example of using natural language processing (NLP) tools and techniques.

(Please turn the page.)

I have analyzed the sentiment of authors of product reviews. I trained a sentiment analysis model using identified key polarizing words, verified the weights learned for each of these words, and compared the results of this simpler classifier with classifiers that use all words that are present in a product review. I also use the following NLP techniques in my declarative model of written, grammatically correct English:

- combining symbols to describe sentence parts (regular expressions)
- choosing whether a word is a noun, etc., using data (part-of-speech tagging)
- describing the ways that words can be combined (grammars)
- representing a sentence using relationships between words (syntactic parsing)
- deciding among word meanings based on data (computational semantics)

```
{
 "cells": [
  {
   "cell_type": "markdown",
   "metadata": {},
   "source": [
    "#Document retrieval from wikipedia data\n",
    "\n",
    "#Fire up GraphLab Create"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 85,
   "metadata": {},
   "outputs": [],
   "source": [
    "import graphlab"
   ]
  },
  {
   "cell_type": "markdown",
   "metadata": {},
   "source": [
    "#Load some text data - from wikipedia, pages on people"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 86,
   "metadata": {},
   "outputs": [],
   "source": [
    "people = graphlab.SFrame('people_wiki.gl/')"
   ]
  },
  {
   "cell_type": "markdown",
   "metadata": {},
   "source": [
    "Data contains:  link to wikipedia article, name of person, text of article."
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 87,
   "metadata": {},
   "outputs": [
    {
     "data": {
      "text/html": [
       "<div style=\"max-height:1000px;max-width:1500px;overflow:auto;\"><table frame=\"box\"
       rules=\"cols\">\n",
```

```
"    <tr>\n",
"        <th style=\"padding-left: 1em; padding-right: 1em; text-align:
center\">URI</th>\n",
"        <th style=\"padding-left: 1em; padding-right: 1em; text-align:
center\">name</th>\n",
"        <th style=\"padding-left: 1em; padding-right: 1em; text-align:
center\">text</th>\n",
"    </tr>\n",
"    <tr>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">&lt;http://dbpedia.org/resou<br>rce/Digby_Morrell&gt; ...</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">Digby Morrell</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">digby morrell born 10<br>october 1979 is a former ...</td>\n",
"    </tr>\n",
"    <tr>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">&lt;http://dbpedia.org/resou<br>rce/Alfred_J._Lewy&gt; ...</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">Alfred J. Lewy</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">alfred j lewy aka sandy<br>lewy graduated from ...</td>\n",
"    </tr>\n",
"    <tr>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">&lt;http://dbpedia.org/resou<br>rce/Harpdog_Brown&gt; ...</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">Harpdog Brown</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">harpdog brown is a singer<br>and harmonica player who ...</td>\n",
"    </tr>\n",
"    <tr>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">&lt;http://dbpedia.org/resou<br>rce/Franz_Rottensteiner&gt;
...</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">Franz Rottensteiner</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">franz rottensteiner born<br>in waidmannsfeld lower ...</td>\n",
"    </tr>\n",
"    <tr>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">&lt;http://dbpedia.org/resou<br>rce/G-Enka&gt; ...</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">G-Enka</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">henry krvits born 30<br>december 1974 in tallinn ...</td>\n",
"    </tr>\n",
"    <tr>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">&lt;http://dbpedia.org/resou<br>rce/Sam_Henderson&gt; ...</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
```

```
      vertical-align: top\">Sam Henderson</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">sam henderson born<br>october 18 1969 is an ...</td>\n",
      "    </tr>\n",
      "    <tr>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">&lt;http://dbpedia.org/resou<br>rce/Aaron_LaCrate&gt; ...</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">Aaron LaCrate</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">aaron lacrate is an<br>american music producer ...</td>\n",
      "    </tr>\n",
      "    <tr>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">&lt;http://dbpedia.org/resou<br>rce/Trevor_Ferguson&gt; ...</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">Trevor Ferguson</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">trevor ferguson aka john<br>farrow born 11 november ...</td>\n",
      "    </tr>\n",
      "    <tr>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">&lt;http://dbpedia.org/resou<br>rce/Grant_Nelson&gt; ...</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">Grant Nelson</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">grant nelson born 27<br>april 1971 in london  ...</td>\n",
      "    </tr>\n",
      "    <tr>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">&lt;http://dbpedia.org/resou<br>rce/Cathy_Caruth&gt; ...</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">Cathy Caruth</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">cathy caruth born 1955 is<br>frank h t rhodes ...</td>\n",
      "    </tr>\n",
      "</table>\n",
      "[10 rows x 3 columns]<br/>\n",
      "</div>"
     ]
    },
    "output_type": "execute_result",
    "metadata": {}
   }
  ],
  "source": [
   "people.head()"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 88,
  "metadata": {},
```

```
  "outputs": [
   {
    "data": {
     "text/plain": [
      "59071"
     ]
    },
    "execution_count": 88,
    "output_type": "execute_result",
    "metadata": {}
   }
  ],
  "source": [
   "len(people)"
  ]
 },
 {
  "cell_type": "markdown",
  "metadata": {},
  "source": [
   "#Explore the dataset and checkout the text it contains\n",
   "\n",
   "##Exploring the entry for president Obama"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 89,
  "metadata": {},
  "outputs": [],
  "source": [
   "obama = people[people['name'] == 'Barack Obama']"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 90,
  "metadata": {},
  "outputs": [
   {
    "data": {
     "text/html": [
      "<div style=\"max-height:1000px;max-width:1500px;overflow:auto;\"><table frame=\"box\"
      rules=\"cols\">\n",
      "    <tr>\n",
      "        <th style=\"padding-left: 1em; padding-right: 1em; text-align:
      center\">URI</th>\n",
      "        <th style=\"padding-left: 1em; padding-right: 1em; text-align:
      center\">name</th>\n",
      "        <th style=\"padding-left: 1em; padding-right: 1em; text-align:
      center\">text</th>\n",
      "    </tr>\n",
      "    <tr>\n",
```

```
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">&lt;http://dbpedia.org/resou<br>rce/Barack_Obama&gt; ...</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">Barack Obama</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">barack hussein obama ii<br>brk husen bm born august ...</td>\n",
      "    </tr>\n",
      "</table>\n",
      "[? rows x 3 columns]<br/>Note: Only the head of the SFrame is printed. This SFrame is
      lazily evaluated.<br/>You can use len(sf) to force materialization.\n",
      "</div>"
     ]
    },
    "output_type": "execute_result",
    "metadata": {}
   }
  ],
  "source": [
   "obama"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 91,
  "metadata": {},
  "outputs": [
   {
    "data": {
     "text/plain": [
      "dtype: str\nRows: ?\n['barack hussein obama ii brk husen bm born august 4 1961 is the
      44th and current president of the united states and the first african american to hold
      the office born in honolulu hawaii obama is a graduate of columbia university and
      harvard law school where he served as president of the harvard law review he was a
      community organizer in chicago before earning his law degree he worked as a civil rights
      attorney and taught constitutional law at the university of chicago law school from 1992
      to 2004 he served three terms representing the 13th district in the illinois senate from
      1997 to 2004 running unsuccessfully for the united states house of representatives in
      2000in 2004 obama received national attention during his campaign to represent illinois
      in the united states senate with his victory in the march democratic party primary his
      keynote address at the democratic national convention in july and his election to the
      senate in november he began his presidential campaign in 2007 and after a close primary
      campaign against hillary rodham clinton in 2008 he won sufficient delegates in the
      democratic party primaries to receive the presidential nomination he then defeated
      republican nominee john mccain in the general election and was inaugurated as president
      on january 20 2009 nine months after his election obama was named the 2009 nobel peace
      prize laureateduring his first two years in office obama signed into law economic
      stimulus legislation in response to the great recession in the form of the american
      recovery and reinvestment act of 2009 and the tax relief unemployment insurance
      reauthorization and job creation act of 2010 other major domestic initiatives in his
      first term included the patient protection and affordable care act often referred to as
      obamacare the doddfrank wall street reform and consumer protection act and the dont ask
      dont tell repeal act of 2010 in foreign policy obama ended us military involvement in
      the iraq war increased us troop levels in afghanistan signed the new start arms control
```

```
      treaty with russia ordered us military involvement in libya and ordered the military
      operation that resulted in the death of osama bin laden in january 2011 the republicans
      regained control of the house of representatives as the democratic party lost a total of
      63 seats and after a lengthy debate over federal spending and whether or not to raise
      the nations debt limit obama signed the budget control act of 2011 and the american
      taxpayer relief act of 2012obama was reelected president in november 2012 defeating
      republican nominee mitt romney and was sworn in for a second term on january 20 2013
      during his second term obama has promoted domestic policies related to gun control in
      response to the sandy hook elementary school shooting and has called for full equality
      for lgbt americans while his administration has filed briefs which urged the supreme
      court to strike down the defense of marriage act of 1996 and californias proposition 8
      as unconstitutional in foreign policy obama ordered us military involvement in iraq in
      response to gains made by the islamic state in iraq after the 2011 withdrawal from iraq
      continued the process of ending us combat operations in afghanistan and has sought to
      normalize us relations with cuba', ... ]"
     ]
    },
    "execution_count": 91,
    "output_type": "execute_result",
    "metadata": {}
   }
  ],
  "source": [
   "obama['text']"
  ]
 },
 {
  "cell_type": "markdown",
  "metadata": {},
  "source": [
   "##Exploring the entry for actor George Clooney"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 92,
  "metadata": {},
  "outputs": [
   {
    "data": {
     "text/plain": [
      "dtype: str\nRows: ?\n['george timothy clooney born may 6 1961 is an american actor
      writer producer director and activist he has received three golden globe awards for his
      work as an actor and two academy awards one for acting and the other for
      producingclooney made his acting debut on television in 1978 and later gained wide
      recognition in his role as dr doug ross on the longrunning medical drama er from 1994 to
      1999 for which he received two emmy award nominations while working on er he began
      attracting a variety of leading roles in films including the superhero film batman robin
      1997 and the crime comedy out of sight 1998 in which he first worked with a director who
      would become a longtime collaborator steven soderbergh in 1999 clooney took the lead
      role in three kings a wellreceived war satire set during the gulf warin 2001 clooneys
      fame widened with the release of his biggest commercial success the heist comedy oceans
      eleven the first of the film trilogy a remake of the 1960 film with frank sinatra as
```

```
      danny ocean he made his directorial debut a year later with the biographical thriller
      confessions of a dangerous mind and has since directed the drama good night and good
      luck 2005 the sports comedy leatherheads 2008 the political drama the ides of march 2011
      and the comedydrama war film the monuments men 2014he won an academy award for best
      supporting actor for the middle east thriller syriana 2005 and subsequently earned best
      actor nominations for the legal thriller michael clayton 2007 the comedydrama up in the
      air 2009 and the drama the descendants 2011 in 2013 he received the academy award for
      best picture for producing the political thriller argo alongside ben affleck and grant
      heslov he is the only person ever to be nominated for academy awards in six
      categoriesclooney is sometimes described as one of the most handsome men in the world in
      2005 tv guide ranked clooney no 1 on its 50 sexiest stars of all time list in 2009 he
      was included in times annual time 100 as one of the most influential people in the world
      clooney is also noted for his political activism and has served as one of the united
      nations messengers of peace since january 31 2008 his humanitarian work includes his
      advocacy of finding a resolution for the darfur conflict raising funds for the 2010
      haiti earthquake 2004 tsunami and 911 victims and creating documentaries such as sand
      and sorrow to raise awareness about international crises he is also a member of the
      council on foreign relations', ... ]"
     ]
    },
    "execution_count": 92,
    "output_type": "execute_result",
    "metadata": {}
   }
  ],
  "source": [
   "clooney = people[people['name'] == 'George Clooney']\n",
   "clooney['text']"
  ]
 },
 {
  "cell_type": "markdown",
  "metadata": {},
  "source": [
   "#Get the word counts for Obama article"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 93,
  "metadata": {},
  "outputs": [],
  "source": [
   "obama['word_count'] = graphlab.text_analytics.count_words(obama['text'])"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 94,
  "metadata": {},
  "outputs": [
   {
    "name": "stdout",
```

```
    "output_type": "stream",
    "text": [
     "[{'operations': 1, 'represent': 1, 'office': 2, 'unemployment': 1, 'doddfrank': 1,
     'over': 1, 'unconstitutional': 1, 'domestic': 2, 'major': 1, 'years': 1, 'against': 1,
     'proposition': 1, 'seats': 1, 'graduate': 1, 'debate': 1, 'before': 1, 'death': 1, '20':
     2, 'taxpayer': 1, 'representing': 1, 'obamacare': 1, 'barack': 1, 'to': 14, '4': 1,
     'policy': 2, '8': 1, 'he': 7, '2011': 3, '2010': 2, '2013': 1, '2012': 1, 'bin': 1,
     'then': 1, 'his': 11, 'march': 1, 'gains': 1, 'cuba': 1, 'school': 3, '1992': 1, 'new':
     1, 'not': 1, 'during': 2, 'ending': 1, 'continued': 1, 'presidential': 2, 'states': 3,
     'husen': 1, 'osama': 1, 'californias': 1, 'equality': 1, 'prize': 1, 'lost': 1, 'made':
     1, 'inaugurated': 1, 'january': 3, 'university': 2, 'rights': 1, 'july': 1, 'gun': 1,
     'stimulus': 1, 'rodham': 1, 'troop': 1, 'withdrawal': 1, 'brk': 1, 'nine': 1, 'where': 1,
     'referred': 1, 'affordable': 1, 'attorney': 1, 'on': 2, 'often': 1, 'senate': 3,
     'regained': 1, 'national': 2, 'creation': 1, 'related': 1, 'hawaii': 1, 'born': 2,
     'second': 2, 'defense': 1, 'election': 3, 'close': 1, 'operation': 1, 'insurance': 1,
     'sandy': 1, 'afghanistan': 2, 'initiatives': 1, 'for': 4, 'reform': 1, 'house': 2,
     'review': 1, 'representatives': 2, 'ended': 1, 'current': 1, 'state': 1, 'won': 1,
     'limit': 1, 'victory': 1, 'unsuccessfully': 1, 'reauthorization': 1, 'keynote': 1,
     'full': 1, 'patient': 1, 'august': 1, 'degree': 1, '44th': 1, 'bm': 1, 'mitt': 1,
     'attention': 1, 'delegates': 1, 'lgbt': 1, 'job': 1, 'harvard': 2, 'term': 3, 'served':
     2, 'ask': 1, 'november': 2, 'debt': 1, 'by': 1, 'wall': 1, 'care': 1, 'received': 1,
     'great': 1, 'signed': 3, 'libya': 1, 'receive': 1, 'of': 18, 'months': 1, 'urged': 1,
     'foreign': 2, 'american': 3, 'protection': 2, 'economic': 1, 'act': 8, 'military': 4,
     'hussein': 1, 'or': 1, 'first': 3, 'control': 4, 'named': 1, 'clinton': 1, 'dont': 2,
     'campaign': 3, 'russia': 1, 'civil': 1, 'reinvestment': 1, 'into': 1, 'address': 1,
     'primary': 2, 'community': 1, 'mccain': 1, 'down': 1, 'hook': 1, '63': 1, 'americans': 1,
     'elementary': 1, 'total': 1, 'earning': 1, 'repeal': 1, 'from': 3, 'raise': 1,
     'district': 1, 'spending': 1, 'republican': 2, 'legislation': 1, 'three': 1, 'relations':
     1, 'nobel': 1, 'start': 1, 'tell': 1, 'iraq': 4, 'convention': 1, 'resulted': 1, 'john':
     1, 'was': 5, '2012obama': 1, 'form': 1, 'that': 1, 'tax': 1, 'sufficient': 1,
     'republicans': 1, 'strike': 1, 'hillary': 1, 'street': 1, 'arms': 1, 'honolulu': 1,
     'filed': 1, 'worked': 1, 'hold': 1, 'with': 3, 'obama': 9, 'ii': 1, 'has': 4, '1997': 1,
     '1996': 1, 'whether': 1, 'reelected': 1, 'budget': 1, 'us': 6, 'nations': 1, 'recession':
     1, 'while': 1, 'taught': 1, 'marriage': 1, 'policies': 1, 'promoted': 1, 'called': 1,
     'and': 21, 'supreme': 1, 'ordered': 3, 'nominee': 2, 'process': 1, '2000in': 1, 'is': 2,
     'romney': 1, 'briefs': 1, 'defeated': 1, 'general': 1, '13th': 1, 'as': 6, 'at': 2, 'in':
     30, 'sought': 1, 'organizer': 1, 'shooting': 1, 'increased': 1, 'normalize': 1,
     'lengthy': 1, 'united': 3, 'court': 1, 'recovery': 1, 'laden': 1, 'laureateduring': 1,
     'peace': 1, 'administration': 1, '1961': 1, 'illinois': 2, 'other': 1, 'which': 1,
     'party': 3, 'primaries': 1, 'sworn': 1, 'relief': 2, 'war': 1, 'columbia': 1, 'combat':
     1, 'after': 4, 'islamic': 1, 'running': 1, 'levels': 1, 'two': 1, 'involvement': 3,
     'response': 3, 'included': 1, 'president': 4, 'law': 6, 'nomination': 1, '2008': 1, 'a':
     7, '2009': 3, 'chicago': 2, 'constitutional': 1, 'defeating': 1, 'treaty': 1, 'federal':
     1, '2007': 1, '2004': 3, 'african': 1, 'the': 40, 'democratic': 4, 'consumer': 1,
     'began': 1, 'terms': 1}]\n"
    ]
   }
  ],
  "source": [
   "print obama['word_count']"
  ]
 },
 {
```

```
  "cell_type": "markdown",
  "metadata": {},
  "source": [
   "##Sort the word counts for the Obama article"
  ]
 },
 {
  "cell_type": "markdown",
  "metadata": {},
  "source": [
   "###Turning dictonary of word counts into a table"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 95,
  "metadata": {},
  "outputs": [],
  "source": [
   "obama_word_count_table = obama[['word_count']].stack('word_count', new_column_name =
   ['word','count'])"
  ]
 },
 {
  "cell_type": "markdown",
  "metadata": {},
  "source": [
   "###Sorting the word counts to show most common words at the top"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 96,
  "metadata": {},
  "outputs": [
   {
    "data": {
     "text/html": [
      "<div style=\"max-height:1000px;max-width:1500px;overflow:auto;\"><table frame=\"box\"
      rules=\"cols\">\n",
      "    <tr>\n",
      "        <th style=\"padding-left: 1em; padding-right: 1em; text-align:
      center\">word</th>\n",
      "        <th style=\"padding-left: 1em; padding-right: 1em; text-align:
      center\">count</th>\n",
      "    </tr>\n",
      "    <tr>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">normalize</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">1</td>\n",
      "    </tr>\n",
      "    <tr>\n",
```

```
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">sought</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">1</td>\n",
"    </tr>\n",
"    <tr>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">combat</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">1</td>\n",
"    </tr>\n",
"    <tr>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">continued</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">1</td>\n",
"    </tr>\n",
"    <tr>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">unconstitutional</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">1</td>\n",
"    </tr>\n",
"    <tr>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">8</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">1</td>\n",
"    </tr>\n",
"    <tr>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">californias</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">1</td>\n",
"    </tr>\n",
"    <tr>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">1996</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">1</td>\n",
"    </tr>\n",
"    <tr>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">marriage</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">1</td>\n",
"    </tr>\n",
"    <tr>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">defense</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">1</td>\n",
"    </tr>\n",
```

```
      "</table>\n",
      "[10 rows x 2 columns]<br/>\n",
      "</div>"
     ]
    },
    "output_type": "execute_result",
    "metadata": {}
   }
  ],
  "source": [
   "obama_word_count_table.head()"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 97,
  "metadata": {},
  "outputs": [
   {
    "data": {
     "text/html": [
      "<div style=\"max-height:1000px;max-width:1500px;overflow:auto;\"><table frame=\"box\"
      rules=\"cols\">\n",
      "    <tr>\n",
      "        <th style=\"padding-left: 1em; padding-right: 1em; text-align:
      center\">word</th>\n",
      "        <th style=\"padding-left: 1em; padding-right: 1em; text-align:
      center\">count</th>\n",
      "    </tr>\n",
      "    <tr>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">the</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">40</td>\n",
      "    </tr>\n",
      "    <tr>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">in</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">30</td>\n",
      "    </tr>\n",
      "    <tr>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">and</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">21</td>\n",
      "    </tr>\n",
      "    <tr>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">of</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">18</td>\n",
      "    </tr>\n",
```

```
      "    <tr>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">to</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">14</td>\n",
      "    </tr>\n",
      "    <tr>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">his</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">11</td>\n",
      "    </tr>\n",
      "    <tr>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">obama</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">9</td>\n",
      "    </tr>\n",
      "    <tr>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">act</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">8</td>\n",
      "    </tr>\n",
      "    <tr>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">a</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">7</td>\n",
      "    </tr>\n",
      "    <tr>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">he</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">7</td>\n",
      "    </tr>\n",
      "</table>\n",
      "[273 rows x 2 columns]<br/>Note: Only the head of the SFrame is printed.<br/>You can
      use print_rows(num_rows=m, num_columns=n) to print more rows and columns.\n",
      "</div>"
     ]
    },
    "output_type": "execute_result",
    "metadata": {}
   }
  ],
  "source": [
   "obama_word_count_table.sort('count',ascending=False)"
  ]
 },
 {
  "cell_type": "markdown",
  "metadata": {},
```

```
   "source": [
    "Most common words include uninformative words like \"the\", \"in\", \"and\",..."
   ]
  },
  {
   "cell_type": "markdown",
   "metadata": {},
   "source": [
    "#Compute TF-IDF for the corpus \n",
    "\n",
    "To give more weight to informative words, we weigh them by their TF-IDF scores."
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 98,
   "metadata": {},
   "outputs": [
    {
     "data": {
      "text/html": [
       "<div style=\"max-height:1000px;max-width:1500px;overflow:auto;\"><table frame=\"box\"
       rules=\"cols\">\n",
       "    <tr>\n",
       "        <th style=\"padding-left: 1em; padding-right: 1em; text-align:
       center\">URI</th>\n",
       "        <th style=\"padding-left: 1em; padding-right: 1em; text-align:
       center\">name</th>\n",
       "        <th style=\"padding-left: 1em; padding-right: 1em; text-align:
       center\">text</th>\n",
       "        <th style=\"padding-left: 1em; padding-right: 1em; text-align:
       center\">word_count</th>\n",
       "    </tr>\n",
       "    <tr>\n",
       "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
       vertical-align: top\">&lt;http://dbpedia.org/resou<br>rce/Digby_Morrell&gt; ...</td>\n",
       "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
       vertical-align: top\">Digby Morrell</td>\n",
       "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
       vertical-align: top\">digby morrell born 10<br>october 1979 is a former ...</td>\n",
       "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
       vertical-align: top\">{'since': 1, 'carltons':<br>1, 'being': 1, '2005' ...</td>\n",
       "    </tr>\n",
       "    <tr>\n",
       "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
       vertical-align: top\">&lt;http://dbpedia.org/resou<br>rce/Alfred_J._Lewy&gt; ...</td>\n",
       "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
       vertical-align: top\">Alfred J. Lewy</td>\n",
       "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
       vertical-align: top\">alfred j lewy aka sandy<br>lewy graduated from ...</td>\n",
       "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
       vertical-align: top\">{'precise': 1, 'thomas':<br>1, 'closely': 1, ...</td>\n",
       "    </tr>\n",
```

```
"    <tr>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">&lt;http://dbpedia.org/resou<br>rce/Harpdog_Brown&gt; ...</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">Harpdog Brown</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">harpdog brown is a singer<br>and harmonica player who ...</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">{'just': 1, 'issued': 1,<br>'mainly': 1, 'nominat ...</td>\n",
"    </tr>\n",
"    <tr>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">&lt;http://dbpedia.org/resou<br>rce/Franz_Rottensteiner&gt;
...</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">Franz Rottensteiner</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">franz rottensteiner born<br>in waidmannsfeld lower ...</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">{'all': 1,<br>'bauforschung': 1, ...</td>\n",
"    </tr>\n",
"    <tr>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">&lt;http://dbpedia.org/resou<br>rce/G-Enka&gt; ...</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">G-Enka</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">henry krvits born 30<br>december 1974 in tallinn ...</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">{'legendary': 1,<br>'gangstergenka': 1, ...</td>\n",
"    </tr>\n",
"    <tr>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">&lt;http://dbpedia.org/resou<br>rce/Sam_Henderson&gt; ...</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">Sam Henderson</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">sam henderson born<br>october 18 1969 is an ...</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">{'now': 1, 'currently':<br>1, 'less': 1, 'being' ...</td>\n",
"    </tr>\n",
"    <tr>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">&lt;http://dbpedia.org/resou<br>rce/Aaron_LaCrate&gt; ...</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">Aaron LaCrate</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">aaron lacrate is an<br>american music producer ...</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">{'exclusive': 2,<br>'producer': 1, 'tribe': ...</td>\n",
"    </tr>\n",
"    <tr>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
```

```
      vertical-align: top\">&lt;http://dbpedia.org/resou<br>rce/Trevor_Ferguson&gt; ...</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">Trevor Ferguson</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">trevor ferguson aka john<br>farrow born 11 november ...</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">{'taxi': 1, 'salon': 1,<br>'gangs': 1, 'being': 1, ...</td>\n",
      "    </tr>\n",
      "    <tr>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">&lt;http://dbpedia.org/resou<br>rce/Grant_Nelson&gt; ...</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">Grant Nelson</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">grant nelson born 27<br>april 1971 in london  ...</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">{'houston': 1, 'frankie':<br>1, 'labels': 1, ...</td>\n",
      "    </tr>\n",
      "    <tr>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">&lt;http://dbpedia.org/resou<br>rce/Cathy_Caruth&gt; ...</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">Cathy Caruth</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">cathy caruth born 1955 is<br>frank h t rhodes ...</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">{'phenomenon': 1,<br>'deborash': 1, ...</td>\n",
      "    </tr>\n",
      "</table>\n",
      "[10 rows x 4 columns]<br/>\n",
      "</div>"
     ]
    },
    "output_type": "execute_result",
    "metadata": {}
   }
  ],
  "source": [
   "people['word_count'] = graphlab.text_analytics.count_words(people['text'])\n",
   "people.head()"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 99,
  "metadata": {},
  "outputs": [
   {
    "data": {
     "text/html": [
      "<div style=\"max-height:1000px;max-width:1500px;overflow:auto;\"><table frame=\"box\"
      rules=\"cols\">\n",
      "    <tr>\n",
```

```
      "        <th style=\"padding-left: 1em; padding-right: 1em; text-align:
      center\">docs</th>\n",
      "    </tr>\n",
      "    <tr>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">{'since':<br>1.455376717308041, ...</td>\n",
      "    </tr>\n",
      "    <tr>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">{'precise':<br>6.44320060695519, ...</td>\n",
      "    </tr>\n",
      "    <tr>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">{'just':<br>2.7007299687108643, ...</td>\n",
      "    </tr>\n",
      "    <tr>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">{'all':<br>1.6431112434912472, ...</td>\n",
      "    </tr>\n",
      "    <tr>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">{'legendary':<br>4.280856294365192, ...</td>\n",
      "    </tr>\n",
      "    <tr>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">{'now': 1.96695239252401,<br>'currently': ...</td>\n",
      "    </tr>\n",
      "    <tr>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">{'exclusive':<br>10.455187230695827, ...</td>\n",
      "    </tr>\n",
      "    <tr>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">{'taxi':<br>6.0520214560945025, ...</td>\n",
      "    </tr>\n",
      "    <tr>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">{'houston':<br>3.935505942157149, ...</td>\n",
      "    </tr>\n",
      "    <tr>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">{'phenomenon':<br>5.750053426395245, ...</td>\n",
      "    </tr>\n",
      "    "</table>\n",
      "    "[59071 rows x 1 columns]<br/>Note: Only the head of the SFrame is printed.<br/>You can
      use print_rows(num_rows=m, num_columns=n) to print more rows and columns.\n",
      "</div>"
     ]
    },
    "output_type": "execute_result",
    "metadata": {}
   }
  ],
```

```
  "source": [
   "tfidf = graphlab.text_analytics.tf_idf(people['word_count'])\n",
   "tfidf"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 100,
  "metadata": {},
  "outputs": [],
  "source": [
   "people['tfidf'] = tfidf['docs']"
  ]
 },
 {
  "cell_type": "markdown",
  "metadata": {},
  "source": [
   "##Examine the TF-IDF for the Obama article"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 101,
  "metadata": {},
  "outputs": [],
  "source": [
   "obama = people[people['name'] == 'Barack Obama']"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 102,
  "metadata": {},
  "outputs": [
   {
    "data": {
     "text/html": [
      "<div style=\"max-height:1000px;max-width:1500px;overflow:auto;\"><table frame=\"box\"
      rules=\"cols\">\n",
      "    <tr>\n",
      "        <th style=\"padding-left: 1em; padding-right: 1em; text-align:
      center\">word</th>\n",
      "        <th style=\"padding-left: 1em; padding-right: 1em; text-align:
      center\">tfidf</th>\n",
      "    </tr>\n",
      "    <tr>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">obama</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">43.2956530721</td>\n",
      "    </tr>\n",
      "    <tr>\n",
```

```
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">act</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">27.678222623</td>\n",
"    </tr>\n",
"    <tr>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">iraq</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">17.747378588</td>\n",
"    </tr>\n",
"    <tr>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">control</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">14.8870608452</td>\n",
"    </tr>\n",
"    <tr>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">law</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">14.7229357618</td>\n",
"    </tr>\n",
"    <tr>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">ordered</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">14.5333739509</td>\n",
"    </tr>\n",
"    <tr>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">military</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">13.1159327785</td>\n",
"    </tr>\n",
"    <tr>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">involvement</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">12.7843852412</td>\n",
"    </tr>\n",
"    <tr>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">response</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">12.7843852412</td>\n",
"    </tr>\n",
"    <tr>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">democratic</td>\n",
"        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
vertical-align: top\">12.4106886973</td>\n",
"    </tr>\n",
```

```
      "</table>\n",
      "[273 rows x 2 columns]<br/>Note: Only the head of the SFrame is printed.<br/>You can
      use print_rows(num_rows=m, num_columns=n) to print more rows and columns.\n",
      "</div>"
     ]
    },
    "output_type": "execute_result",
    "metadata": {}
   }
  ],
  "source": [

   "obama[['tfidf']].stack('tfidf',new_column_name=['word','tfidf']).sort('tfidf',ascending=Fals
   e)"
  ]
 },
 {
  "cell_type": "markdown",
  "metadata": {},
  "source": [
   "Words with highest TF-IDF are much more informative."
  ]
 },
 {
  "cell_type": "markdown",
  "metadata": {},
  "source": [
   "#Manually compute distances between a few people\n",
   "\n",
   "Let's manually compare the distances between the articles for a few famous people.  "
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 103,
  "metadata": {},
  "outputs": [],
  "source": [
   "clinton = people[people['name'] == 'Bill Clinton']"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 104,
  "metadata": {},
  "outputs": [],
  "source": [
   "beckham = people[people['name'] == 'David Beckham']"
  ]
 },
 {
  "cell_type": "markdown",
  "metadata": {},
```

```
   "source": [
    "##Is Obama closer to Clinton than to Beckham?\n",
    "\n",
    "We will use cosine distance, which is given by\n",
    "\n",
    "(1-cosine_similarity) \n",
    "\n",
    "and find that the article about president Obama is closer to the one about former
    president Clinton than that of footballer David Beckham."
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 105,
   "metadata": {},
   "outputs": [
    {
     "data": {
      "text/plain": [
       "0.8339854936884276"
      ]
     },
     "execution_count": 105,
     "output_type": "execute_result",
     "metadata": {}
    }
   ],
   "source": [
    "graphlab.distances.cosine(obama['tfidf'][0],clinton['tfidf'][0])"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 106,
   "metadata": {},
   "outputs": [
    {
     "data": {
      "text/plain": [
       "0.9791305844747478"
      ]
     },
     "execution_count": 106,
     "output_type": "execute_result",
     "metadata": {}
    }
   ],
   "source": [
    "graphlab.distances.cosine(obama['tfidf'][0],beckham['tfidf'][0])"
   ]
  },
  {
   "cell_type": "markdown",
```

```
  "metadata": {},
  "source": [
   "\n",
   "#Build a nearest neighbor model for document retrieval\n",
   "\n",
   "We now create a nearest-neighbors model and apply it to document retrieval.  "
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 107,
  "metadata": {},
  "outputs": [
   {
    "name": "stdout",
    "output_type": "stream",
    "text": [
     "PROGRESS: Starting brute force nearest neighbors model training.\n"
    ]
   }
  ],
  "source": [
   "knn_model = graphlab.nearest_neighbors.create(people,features=['tfidf'],label='name')"
  ]
 },
 {
  "cell_type": "markdown",
  "metadata": {},
  "source": [
   "#Applying the nearest-neighbors model for retrieval"
  ]
 },
 {
  "cell_type": "markdown",
  "metadata": {},
  "source": [
   "##Who is closest to Obama?"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 108,
  "metadata": {},
  "outputs": [
   {
    "name": "stdout",
    "output_type": "stream",
    "text": [
     "PROGRESS: Starting pairwise querying.\nPROGRESS:
     +--------------+---------+-------------+--------------+\nPROGRESS: | Query points | #
     Pairs | % Complete. | Elapsed Time |\nPROGRESS:
     +--------------+---------+-------------+--------------+\nPROGRESS: | 0            |
     1       | 0.00169288  | 15.062ms     |\nPROGRESS: | Done         |         | 100
```

```
     | 171.082ms    |\nPROGRESS: +--------------+---------+------------+--------------+\n"
    ]
   },
   {
    "data": {
     "text/html": [
      "<div style=\"max-height:1000px;max-width:1500px;overflow:auto;\"><table frame=\"box\"
      rules=\"cols\">\n",
      "    <tr>\n",
      "        <th style=\"padding-left: 1em; padding-right: 1em; text-align:
      center\">query_label</th>\n",
      "        <th style=\"padding-left: 1em; padding-right: 1em; text-align:
      center\">reference_label</th>\n",
      "        <th style=\"padding-left: 1em; padding-right: 1em; text-align:
      center\">distance</th>\n",
      "        <th style=\"padding-left: 1em; padding-right: 1em; text-align:
      center\">rank</th>\n",
      "    </tr>\n",
      "    <tr>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">0</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">Barack Obama</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">0.0</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">1</td>\n",
      "    </tr>\n",
      "    <tr>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">0</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">Joe Biden</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">0.794117647059</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">2</td>\n",
      "    </tr>\n",
      "    <tr>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">0</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">Joe Lieberman</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">0.794685990338</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">3</td>\n",
      "    </tr>\n",
      "    <tr>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">0</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">Kelly Ayotte</td>\n",
```

```
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">0.811989100817</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">4</td>\n",
      "    </tr>\n",
      "    <tr>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">0</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">Bill Clinton</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">0.813852813853</td>\n",
      "        <td style=\"padding-left: 1em; padding-right: 1em; text-align: center;
      vertical-align: top\">5</td>\n",
      "    </tr>\n",
      "</table>\n",
      "[5 rows x 4 columns]<br/>\n",
      "</div>"
     ]
    },
    "output_type": "execute_result",
    "metadata": {}
   }
  ],
  "source": [
   "knn_model.query(obama)"
  ]
 },
 {
  "cell_type": "markdown",
  "metadata": {},
  "source": [
   "As we can see, president Obama's article is closest to the one about his vice-president
   Biden, and those of other politicians.  "
  ]
 },
 {
  "cell_type": "markdown",
  "metadata": {},
  "source": [
   "##Other examples of document retrieval"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 109,
  "metadata": {},
  "outputs": [],
  "source": [
   "swift = people[people['name'] == 'Taylor Swift']"
  ]
 },
 {
```

```
  "cell_type": "code",
  "execution_count": 29,
  "metadata": {},
  "outputs": [],
  "source": [
   "knn_model.query(swift)"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 30,
  "metadata": {},
  "outputs": [],
  "source": [
   "jolie = people[people['name'] == 'Angelina Jolie']"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 31,
  "metadata": {},
  "outputs": [],
  "source": [
   "knn_model.query(jolie)"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 32,
  "metadata": {},
  "outputs": [],
  "source": [
   "arnold = people[people['name'] == 'Arnold Schwarzenegger']"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 33,
  "metadata": {},
  "outputs": [],
  "source": [
   "knn_model.query(arnold)"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 122,
  "metadata": {},
  "outputs": [],
  "source": [
   "victoria = people[people['name'] == 'Victoria Beckham']"
  ]
 },
```

```
{
 "cell_type": "code",
 "execution_count": 123,
 "metadata": {},
 "outputs": [],
 "source": [
  "stephen = people[people['name'] == 'Stephen Dow Beckham']"
 ]
},
{
 "cell_type": "code",
 "execution_count": 124,
 "metadata": {},
 "outputs": [],
 "source": [
  "louis = people[people['name'] == 'Louis Molloy']"
 ]
},
{
 "cell_type": "code",
 "execution_count": 125,
 "metadata": {},
 "outputs": [],
 "source": [
  "mary = people[people['name'] == 'Mary Fitzgerald (artist)']"
 ]
},
{
 "cell_type": "code",
 "execution_count": 126,
 "metadata": {},
 "outputs": [
  {
   "data": {
    "text/plain": [
     "0.31219127606005204"
    ]
   },
   "execution_count": 126,
   "output_type": "execute_result",
   "metadata": {}
  }
 ],
 "source": [
  "graphlab.distances.cosine(victoria['word_count'][0],stephen['word_count'][0])"
 ]
},
{
 "cell_type": "code",
 "execution_count": 127,
 "metadata": {},
 "outputs": [
  {
```

```
    "data": {
     "text/plain": [
      "0.31050751295927514"
     ]
    },
    "execution_count": 127,
    "output_type": "execute_result",
    "metadata": {}
   }
  ],
  "source": [
   "graphlab.distances.cosine(victoria['word_count'][0],louis['word_count'][0])"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 128,
  "metadata": {},
  "outputs": [
   {
    "data": {
     "text/plain": [
      "0.20730703611504997"
     ]
    },
    "execution_count": 128,
    "output_type": "execute_result",
    "metadata": {}
   }
  ],
  "source": [
   "graphlab.distances.cosine(victoria['word_count'][0],mary['word_count'][0])"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 129,
  "metadata": {},
  "outputs": [],
  "source": [
   "cliff = people[people['name'] == 'Cliff Richard']"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 130,
  "metadata": {},
  "outputs": [],
  "source": [
   "george = people[people['name'] == 'George Bush']"
  ]
 },
 {
```

```
   "cell_type": "code",
   "execution_count": 131,
   "metadata": {},
   "outputs": [
    {
     "data": {
      "text/plain": [
       "0.16142415258967036"
      ]
     },
     "execution_count": 131,
     "output_type": "execute_result",
     "metadata": {}
    }
   ],
   "source": [
    "graphlab.distances.cosine(elton['word_count'][0],cliff['word_count'][0])"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": null,
   "metadata": {},
   "outputs": [],
   "source": [
    "graphlab.distances.cosine(elton['word_count'][0],george['word_count'][0])"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": null,
   "metadata": {},
   "outputs": [],
   "source": [
    ""
   ]
  }
 ],
"metadata": {
 "kernelspec": {
  "display_name": "Python 2",
  "language": "python",
  "name": "python2"
 },
 "language_info": {
  "codemirror_mode": {
   "name": "ipython",
   "version": 2.0
  },
  "file_extension": ".py",
  "mimetype": "text/x-python",
  "name": "python",
  "nbconvert_exporter": "python",
```

```
   "pygments_lexer": "ipython2",
   "version": "2.7.10"
  }
 },
 "nbformat": 4,
 "nbformat_minor": 0
}
```

Section R - Example: Using data mining to analyze unstructured data

This section provides an example of analyzing large volumes of unstructured data using distributed processing tools.

(Please turn the page.)

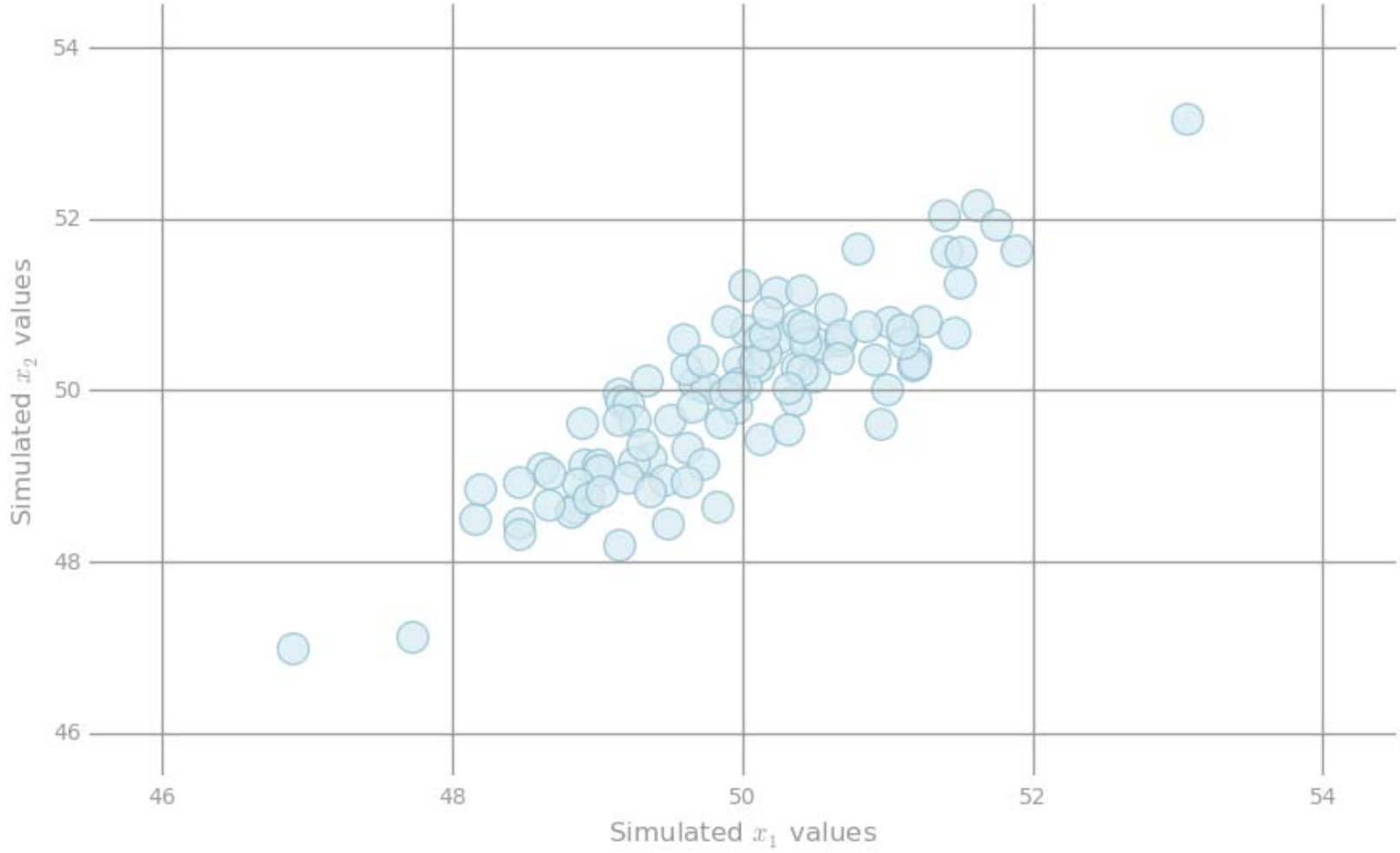
In [ ]:
```
'This is a lab in the Scalable Machine Learning course at Berkeley / EdX.'
'Disclaimer: Much of this code was provided as scaffolding for the lab.'
'Code cells validated by test assertion cells were written by me.'
```

In [63]:
```
import timeit
start_time = timeit.default_timer()
labVersion = 'cs190_week5_v_1_2'
```

In [64]:
```
import matplotlib.pyplot as plt
import numpy as np

def preparePlot(xticks, yticks, figsize=(10.5, 6), hideLabels=False, gridColor='#999999',
                gridWidth=1.0):
    """Template for generating the plot layout."""
    plt.close()
    fig, ax = plt.subplots(figsize=figsize, facecolor='white', edgecolor='white')
    ax.axes.tick_params(labelcolor='#999999', labelsize='10')
    for axis, ticks in [(ax.get_xaxis(), xticks), (ax.get_yaxis(), yticks)]:
        axis.set_ticks_position('none')
        axis.set_ticks(ticks)
        axis.label.set_color('#999999')
        if hideLabels: axis.set_ticklabels([])
    plt.grid(color=gridColor, linewidth=gridWidth, linestyle='-')
    map(lambda position: ax.spines[position].set_visible(False), ['bottom', 'top', 'left', 'right']
)
    return fig, ax

def create2DGaussian(mn, sigma, cov, n):
    """Randomly sample points from a two-dimensional Gaussian distribution"""
    np.random.seed(142)
    return np.random.multivariate_normal(np.array([mn, mn]), np.array([[sigma, cov], [cov, sigma]])
, n)
```

In [65]:
```
dataRandom = create2DGaussian(mn=50, sigma=1, cov=0, n=100)
# generate layout and plot data
fig, ax = preparePlot(np.arange(46, 55, 2), np.arange(46, 55, 2))
ax.set_xlabel(r'Simulated $x_1$ values'), ax.set_ylabel(r'Simulated $x_2$ values')
ax.set_xlim(45, 54.5), ax.set_ylim(45, 54.5)
plt.scatter(dataRandom[:,0], dataRandom[:,1], s=14**2, c='#d6ebf2', edgecolors='#8cbfd0', alpha=0.7
5)
pass
```

In [66]:
```
dataCorrelated = create2DGaussian(mn=50, sigma=1, cov=.9, n=100)

# generate layout and plot data
fig, ax = preparePlot(np.arange(46, 55, 2), np.arange(46, 55, 2))
ax.set_xlabel(r'Simulated $x_1$ values'), ax.set_ylabel(r'Simulated $x_2$ values')
ax.set_xlim(45.5, 54.5), ax.set_ylim(45.5, 54.5)
plt.scatter(dataCorrelated[:,0], dataCorrelated[:,1], s=14**2, c='#d6ebf2',
            edgecolors='#8cbfd0', alpha=0.75)
pass
```



In [67]:
```
# TODO: Replace <FILL IN> with appropriate code
correlatedData = sc.parallelize(dataCorrelated)
meanCorrelated = correlatedData.mean()

correlatedDataZeroMean = correlatedData.map(lambda x: ([(row - [col for col in meanCorrelated]) for
 row in x][0][0],
                                                        [(row - [col for col in meanCorrelated]) for
 row in x][1][1]))
```

In [68]:
```
# TEST Interpreting PCA (1a)
from test_helper import Test
Test.assertTrue(np.allclose(meanCorrelated, [49.95739037, 49.97180477]),
                'incorrect value for meanCorrelated')
Test.assertTrue(np.allclose(correlatedDataZeroMean.take(1)[0], [-0.28561917, 0.10351492]),
                'incorrect value for correlatedDataZeroMean')
```

```
1 test passed.
1 test passed.
```

In [69]:
```
# TODO: Replace <FILL IN> with appropriate code
# Compute the covariance matrix using outer products and correlatedDataZeroMean

correlatedCov = (correlatedDataZeroMean.map(lambda x: np.outer(x, x))
                 .sum() / correlatedDataZeroMean.count())
```

In [70]:
```
# TEST Sample covariance matrix (1b)
covResult = [[ 0.99558386,  0.90148989], [0.90148989, 1.08607497]]
Test.assertTrue(np.allclose(covResult, correlatedCov), 'incorrect value for correlatedCov')
```

```
1 test passed.
```

In [71]:
```
# TODO: Replace <FILL IN> with appropriate code
def estimateCovariance(data):
    """Compute the covariance matrix for a given rdd.

    Note:
    The multi-dimensional covariance array should be calculated
    using outer products.  Don't
    forget to normalize the data by first subtracting the mean.

    Args:
        data (RDD of np.ndarray):  An `RDD` consisting of NumPy arrays.

    Returns:
        np.ndarray: A multi-dimensional array where the number of rows and columns both equal the
            length of the arrays in the input `RDD`.
    """

    meanData = data.mean()

    return (data.map(lambda x: np.outer(x - meanData, x - meanData))
                    .sum() / data.count())

correlatedCovAuto = estimateCovariance(correlatedData)
```

In [72]:
```
# TEST Covariance function (1c)
correctCov = [[ 0.99558386,  0.90148989], [0.90148989, 1.08607497]]
Test.assertTrue(np.allclose(correctCov, correlatedCovAuto),
                'incorrect value for correlatedCovAuto')
```

```
1 test passed.
```

In [73]:
```
# TODO: Replace <FILL IN> with appropriate code
from numpy.linalg import eigh

# Calculate the eigenvalues and eigenvectors from correlatedCovAuto
eigVals, eigVecs = eigh(correlatedCovAuto)
#print 'eigenvalues: {0}'.format(eigVals)
#print '\neigenvectors: \n{0}'.format(eigVecs)

# Use np.argsort to find the top eigenvector based on the largest eigenvalue
inds = np.argsort(eigVals, axis=0)[::-1]

topComponent = eigVecs[:,inds[0]]
```

In [74]:
```
# TEST Eigendecomposition (1d)
def checkBasis(vectors, correct):
    return np.allclose(vectors, correct) or np.allclose(np.negative(vectors), correct)
Test.assertTrue(checkBasis(topComponent, [0.68915649, 0.72461254]),
                'incorrect value for topComponent')
```

```
1 test passed.
```

(1e) PCA scores

We just computed the top principal component for a 2-dimensional non-spherical dataset. Now let's use this principal component to derive a one-dimensional representation for the original data. To compute these compact representations, which are sometimes called PCA "scores", calculate the dot product between each data point in the raw data and the top principal component.

In [75]:
```
# TODO: Replace <FILL IN> with appropriate code
# Use the topComponent and the data from correlatedData to generate PCA scores
correlatedDataScores = correlatedData.map(lambda x: (x.dot(topComponent)))
```

In [76]:
```
# TEST PCA Scores (1e)
firstThree = [70.51682806, 69.30622356, 71.13588168]
Test.assertTrue(checkBasis(correlatedDataScores.take(3), firstThree),
                'incorrect value for correlatedDataScores')
```

```
1 test passed.
```

In [77]:
```
# TODO: Replace <FILL IN> with appropriate code
def pca(data, k=2):
    """Computes the top `k` principal components, corresponding scores, and all eigenvalues.

    Note:
        All eigenvalues should be returned in sorted order (largest to smallest). `eigh` returns
        each eigenvectors as a column.  This function should also return eigenvectors as columns.

    Args:
        data (RDD of np.ndarray): An `RDD` consisting of NumPy arrays.
        k (int): The number of principal components to return.

    Returns:
        tuple of (np.ndarray, RDD of np.ndarray, np.ndarray): A tuple of (eigenvectors, `RDD` of
            scores, eigenvalues).  Eigenvectors is a multi-dimensional array where the number of
            rows equals the length of the arrays in the input `RDD` and the number of columns equal
s
            `k`.  The `RDD` of scores has the same number of rows as `data` and consists of arrays
            of length `k`.  Eigenvalues is an array of length d (the number of features).
    """
    eigVals, eigVecs = eigh(estimateCovariance(data))

    inds = np.argsort(eigVals, axis=0)[::-1]

    topComponents = eigVecs[:,inds[0:k]]
    dataScores = data.map(lambda x: (x.dot(topComponents)))
    eigVals = eigVals[::-1]

    return topComponents, dataScores, eigVals

print(pca(correlatedData, 2))

# Run pca on correlatedData with k = 2
topComponentsCorrelated, correlatedDataScoresAuto, eigenvaluesCorrelated = pca(correlatedData, 2)

# # Create a higher dimensional test set
pcaTestData = sc.parallelize([np.arange(x, x + 4) for x in np.arange(0, 20, 4)])
componentsTest, testScores, eigenvaluesTest = pca(pcaTestData, 3)
```

```
(array([[ 0.68915649, -0.72461254],
       [ 0.72461254,  0.68915649]]), PythonRDD[129] at RDD at PythonRDD.scala:43, array([ 1.94345
403,  0.13820481]))
```

In [78]:
```
# TEST PCA Function (2a)
Test.assertTrue(checkBasis(topComponentsCorrelated.T,
                           [[0.68915649,  0.72461254], [-0.72461254, 0.68915649]]),
                'incorrect value for topComponentsCorrelated')
firstThreeCorrelated = [[70.51682806, 69.30622356, 71.13588168], [1.48305648, 1.5888655, 1.86710679
]]
Test.assertTrue(np.allclose(firstThreeCorrelated,
                            np.vstack(np.abs(correlatedDataScoresAuto.take(3))).T),
                'incorrect value for firstThreeCorrelated')
Test.assertTrue(np.allclose(eigenvaluesCorrelated, [1.94345403, 0.13820481]),
                           'incorrect values for eigenvaluesCorrelated')
topComponentsCorrelatedK1, correlatedDataScoresK1, eigenvaluesCorrelatedK1 = pca(correlatedData, 1)
Test.assertTrue(checkBasis(topComponentsCorrelatedK1.T, [0.68915649,  0.72461254]),
                'incorrect value for components when k=1')
Test.assertTrue(np.allclose([70.51682806, 69.30622356, 71.13588168],
                            np.vstack(np.abs(correlatedDataScoresK1.take(3))).T),
                'incorrect value for scores when k=1')
Test.assertTrue(np.allclose(eigenvaluesCorrelatedK1, [1.94345403, 0.13820481]),
                           'incorrect values for eigenvalues when k=1')
Test.assertTrue(checkBasis(componentsTest.T[0], [ .5, .5, .5, .5]),
                 'incorrect value for componentsTest')
Test.assertTrue(np.allclose(np.abs(testScores.first()[0]), 3.),
                'incorrect value for testScores')
Test.assertTrue(np.allclose(eigenvaluesTest, [ 128, 0, 0, 0 ]), 'incorrect value for eigenvaluesTes
t')
```

```
1 test passed.
1 test passed.
1 test passed.
1 test passed.
1 test passed.
1 test passed.
1 test passed.
1 test passed.
1 test passed.
```

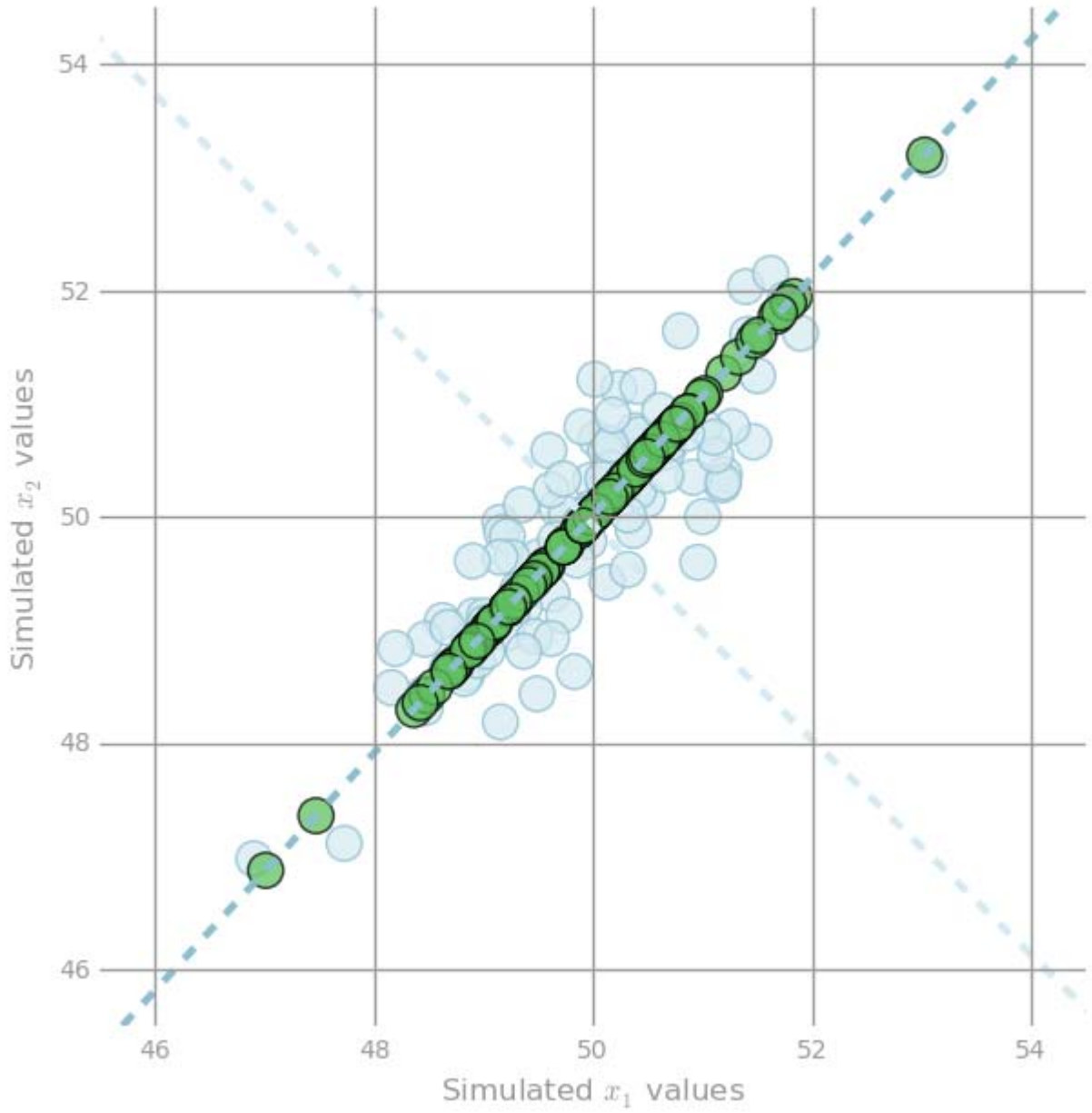
In [79]:
```
# TODO: Replace <FILL IN> with appropriate code
randomData = sc.parallelize(dataRandom)

# Use pca on randomData
topComponentsRandom, randomDataScoresAuto, eigenvaluesRandom = pca(randomData)
```

In [80]:
```
# TEST PCA on `dataRandom` (2b)
Test.assertTrue(checkBasis(topComponentsRandom.T,
                           [[-0.2522559 ,  0.96766056], [-0.96766056,  -0.2522559]]),
                'incorrect value for topComponentsRandom')
firstThreeRandom = [[36.61068572,  35.97314295,  35.59836628],
                    [61.3489929 ,  62.08813671,  60.61390415]]
Test.assertTrue(np.allclose(firstThreeRandom, np.vstack(np.abs(randomDataScoresAuto.take(3))).T),
                'incorrect value for randomDataScoresAuto')
Test.assertTrue(np.allclose(eigenvaluesRandom, [1.4204546, 0.99521397]),
                            'incorrect value for eigenvaluesRandom')
```

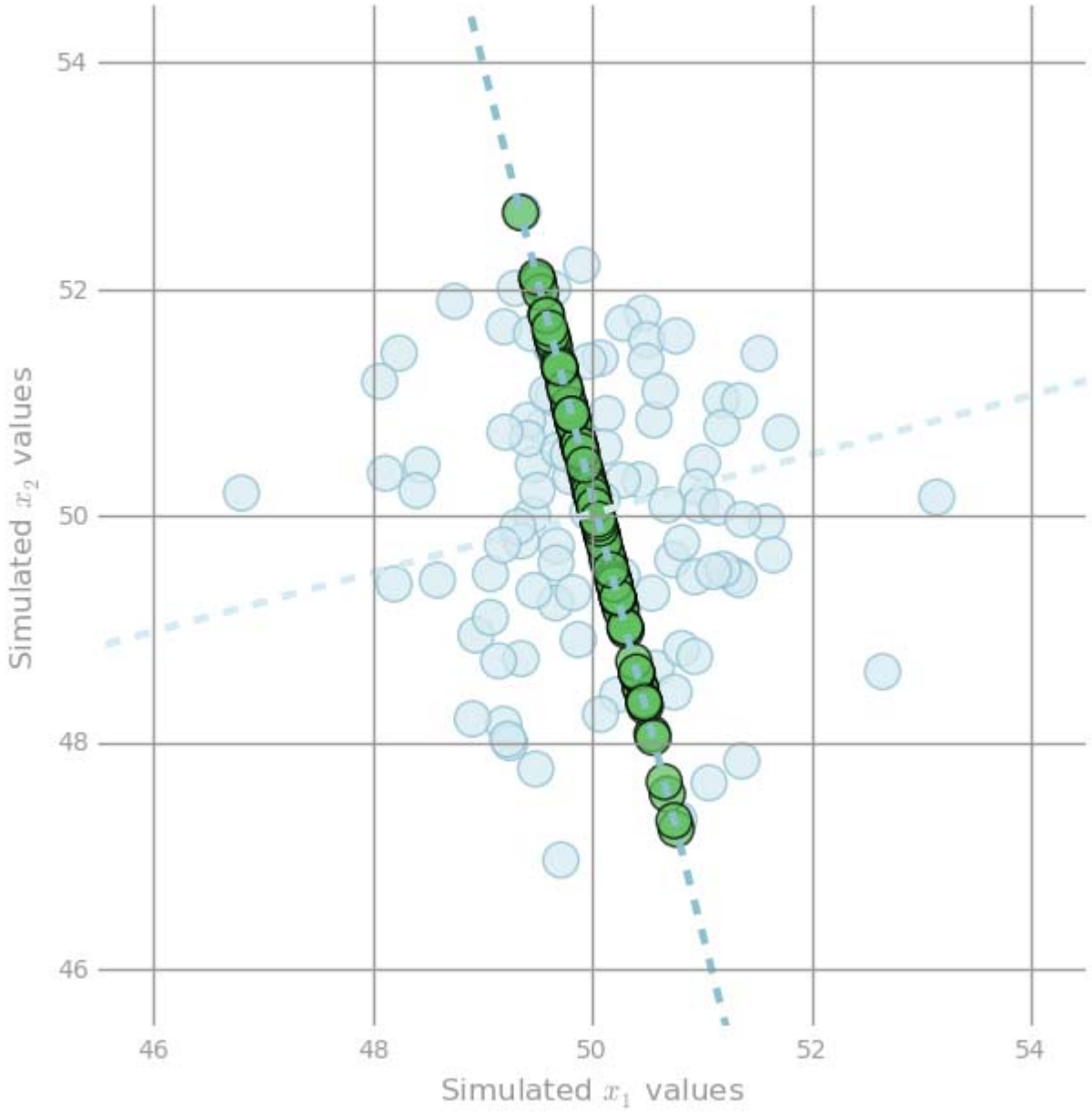
```
1 test passed.
1 test passed.
1 test passed.
```

In [81]:
```
def projectPointsAndGetLines(data, components, xRange):
    """Project original data onto first component and get line details for top two components."""
    topComponent= components[:, 0]
    slope1, slope2 = components[1, :2] / components[0, :2]

    means = data.mean()[:2]
    demeaned = data.map(lambda v: v - means)
    projected = demeaned.map(lambda v: (v.dot(topComponent) /
                                        topComponent.dot(topComponent)) * topComponent)
    remeaned = projected.map(lambda v: v + means)
    x1,x2 = zip(*remeaned.collect())

    lineStartP1X1, lineStartP1X2 = means - np.asarray([xRange, xRange * slope1])
    lineEndP1X1, lineEndP1X2 = means + np.asarray([xRange, xRange * slope1])
    lineStartP2X1, lineStartP2X2 = means - np.asarray([xRange, xRange * slope2])
    lineEndP2X1, lineEndP2X2 = means + np.asarray([xRange, xRange * slope2])

    return ((x1, x2), ([lineStartP1X1, lineEndP1X1], [lineStartP1X2, lineEndP1X2]),
            ([lineStartP2X1, lineEndP2X1], [lineStartP2X2, lineEndP2X2]))
```

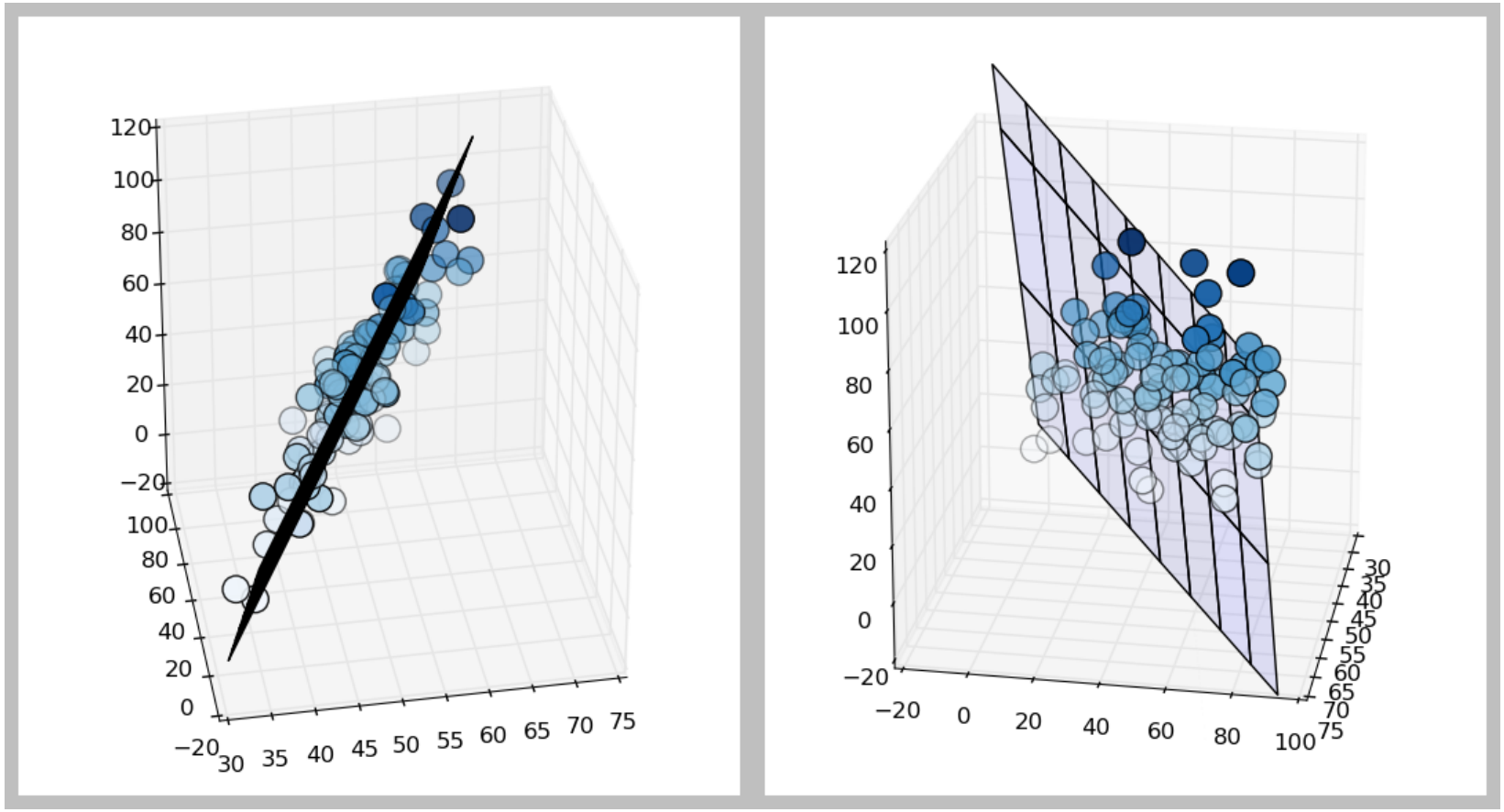
In [82]:
```
((x1, x2), (line1X1, line1X2), (line2X1, line2X2)) = \
    projectPointsAndGetLines(correlatedData, topComponentsCorrelated, 5)

# generate layout and plot data
fig, ax = preparePlot(np.arange(46, 55, 2), np.arange(46, 55, 2), figsize=(7, 7))
ax.set_xlabel(r'Simulated $x_1$ values'), ax.set_ylabel(r'Simulated $x_2$ values')
ax.set_xlim(45.5, 54.5), ax.set_ylim(45.5, 54.5)
plt.plot(line1X1, line1X2, linewidth=3.0, c='#8cbfd0', linestyle='--')
plt.plot(line2X1, line2X2, linewidth=3.0, c='#d6ebf2', linestyle='--')
plt.scatter(dataCorrelated[:,0], dataCorrelated[:,1], s=14**2, c='#d6ebf2',
            edgecolors='#8cbfd0', alpha=0.75)
plt.scatter(x1, x2, s=14**2, c='#62c162', alpha=.75)
pass
```


54
52
50
48
46
Simulated x_2 values
46
48
50
52
54
Simulated x_1 values

In [83]:
```
((x1, x2), (line1X1, line1X2), (line2X1, line2X2)) = \
    projectPointsAndGetLines(randomData, topComponentsRandom, 5)

# generate layout and plot data
fig, ax = preparePlot(np.arange(46, 55, 2), np.arange(46, 55, 2), figsize=(7, 7))
ax.set_xlabel(r'Simulated $x_1$ values'), ax.set_ylabel(r'Simulated $x_2$ values')
ax.set_xlim(45.5, 54.5), ax.set_ylim(45.5, 54.5)
plt.plot(line1X1, line1X2, linewidth=3.0, c='#8cbfd0', linestyle='--')
plt.plot(line2X1, line2X2, linewidth=3.0, c='#d6ebf2', linestyle='--')
plt.scatter(dataRandom[:,0], dataRandom[:,1], s=14**2, c='#d6ebf2',
            edgecolors='#8cbfd0', alpha=0.75)
plt.scatter(x1, x2, s=14**2, c='#62c162', alpha=.75)
pass
```


54
52
50
48
46
Simulated x_2 values
46
48
50
52
54
Simulated x_1 values

In [84]:
```
from mpl_toolkits.mplot3d import Axes3D

m = 100
mu = np.array([50, 50, 50])
r1_2 = 0.9
r1_3 = 0.7
r2_3 = 0.1
sigma1 = 5
sigma2 = 20
sigma3 = 20
c = np.array([[sigma1 ** 2, r1_2 * sigma1 * sigma2, r1_3 * sigma1 * sigma3],
              [r1_2 * sigma1 * sigma2, sigma2 ** 2, r2_3 * sigma2 * sigma3],
              [r1_3 * sigma1 * sigma3, r2_3 * sigma2 * sigma3, sigma3 ** 2]])
np.random.seed(142)
dataThreeD = np.random.multivariate_normal(mu, c, m)

from matplotlib.colors import ListedColormap, Normalize
from matplotlib.cm import get_cmap
norm = Normalize()
cmap = get_cmap("Blues")
clrs = cmap(np.array(norm(dataThreeD[:,2])))[:,0:3]

fig = plt.figure(figsize=(11, 6))
ax = fig.add_subplot(121, projection='3d')
ax.azim=-100
ax.scatter(dataThreeD[:,0], dataThreeD[:,1], dataThreeD[:,2], c=clrs, s=14**2)

xx, yy = np.meshgrid(np.arange(-15, 10, 1), np.arange(-50, 30, 1))
normal = np.array([0.96981815, -0.188338, -0.15485978])
z = (-normal[0] * xx - normal[1] * yy) * 1. / normal[2]
xx = xx + 50
yy = yy + 50
z = z + 50

ax.set_zlim((-20, 120)), ax.set_ylim((-20, 100)), ax.set_xlim((30, 75))
ax.plot_surface(xx, yy, z, alpha=.10)

ax = fig.add_subplot(122, projection='3d')
ax.azim=10
ax.elev=20
#ax.dist=8
ax.scatter(dataThreeD[:,0], dataThreeD[:,1], dataThreeD[:,2], c=clrs, s=14**2)
```

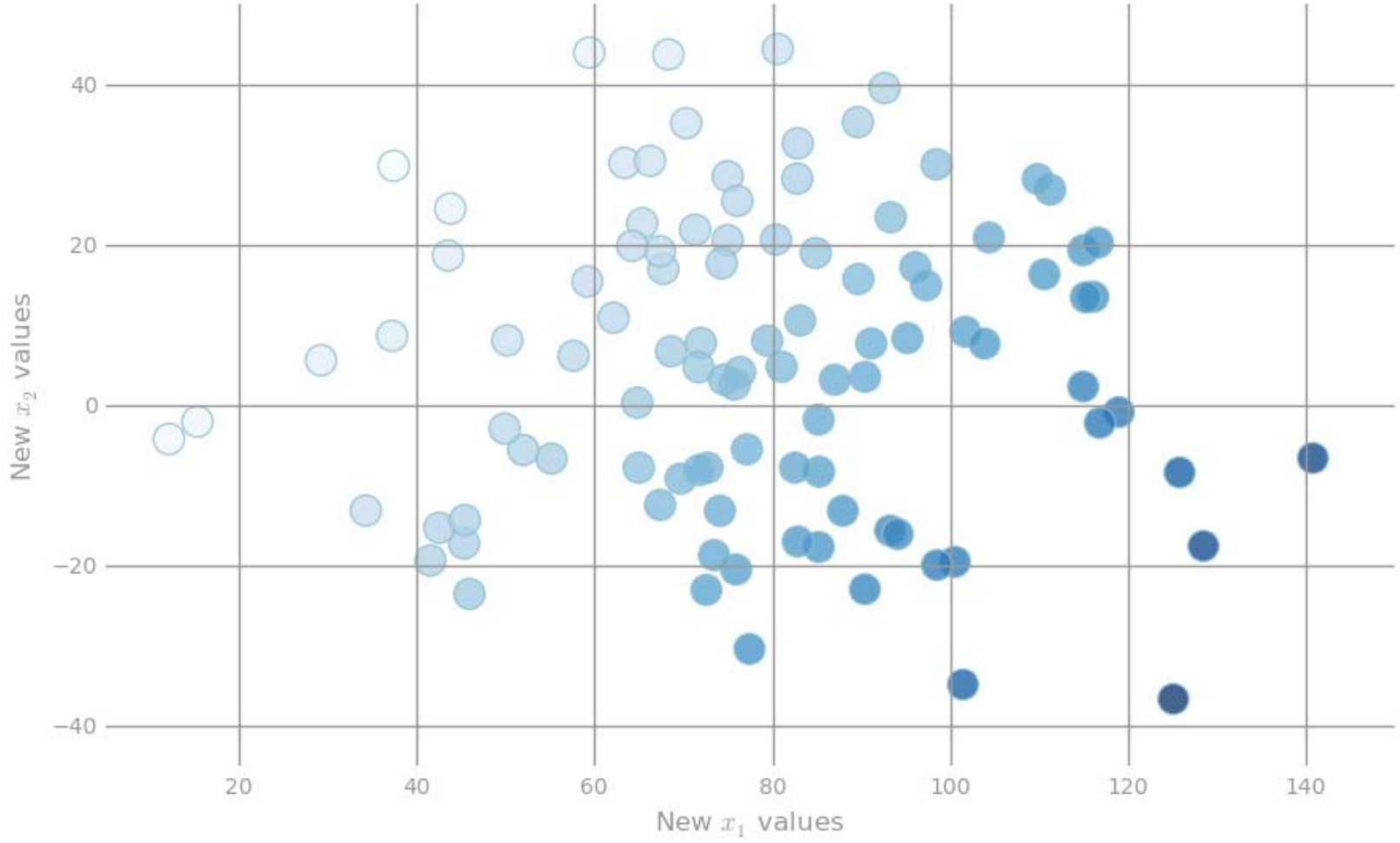


In [85]:
```
# TODO: Replace <FILL IN> with appropriate code
threeDData = sc.parallelize(dataThreeD)
componentsThreeD, threeDScores, eigenvaluesThreeD = pca(threeDData, k=2)
```

In [86]:
```
# TEST 3D to 2D (2c)
Test.assertEquals(componentsThreeD.shape, (3, 2), 'incorrect shape for componentsThreeD')
Test.assertTrue(np.allclose(np.sum(eigenvaluesThreeD), 969.796443367),
                'incorrect value for eigenvaluesThreeD')
Test.assertTrue(np.allclose(np.abs(np.sum(componentsThreeD)), 1.77238943258),
                'incorrect value for componentsThreeD')
Test.assertTrue(np.allclose(np.abs(np.sum(threeDScores.take(3))), 237.782834092),
                'incorrect value for threeDScores')
```

```
1 test passed.
1 test passed.
1 test passed.
1 test passed.
```

In [87]:
```
scoresThreeD = np.asarray(threeDScores.collect())

# generate layout and plot data
fig, ax = preparePlot(np.arange(20, 150, 20), np.arange(-40, 110, 20))
ax.set_xlabel(r'New $x_1$ values'), ax.set_ylabel(r'New $x_2$ values')
ax.set_xlim(5, 150), ax.set_ylim(-45, 50)
plt.scatter(scoresThreeD[:,0], scoresThreeD[:,1], s=14**2, c=clrs, edgecolors='#8cbfd0', alpha=0.75
)
pass
```



In [88]:
```
# TODO: Replace <FILL IN> with appropriate code
def varianceExplained(data, k=1):
    """Calculate the fraction of variance explained by the top `k` eigenvectors.

    Args:
        data (RDD of np.ndarray): An RDD that contains NumPy arrays which store the
            features for an observation.
        k: The number of principal components to consider.

    Returns:
        float: A number between 0 and 1 representing the percentage of variance explained
            by the top `k` eigenvectors.
    """
    components, scores, eigenvalues = pca(data,k)
    numerator = eigenvalues[0:k].sum(dtype=np.float64)
    denominator = eigenvalues.sum(dtype=np.float64)
    ratio = numerator/denominator
    return ratio

start_time = timeit.default_timer()
varianceRandom1 = varianceExplained(randomData, 1)
varianceCorrelated1 = varianceExplained(correlatedData, 1)
varianceRandom2 = varianceExplained(randomData, 2)
varianceCorrelated2 = varianceExplained(correlatedData, 2)
varianceThreeD2 = varianceExplained(threeDData, 2)
```

In [89]:
```
# TEST Variance explained (2d)
Test.assertTrue(np.allclose(varianceRandom1, 0.588017172066), 'incorrect value for varianceRandom1'
)
Test.assertTrue(np.allclose(varianceCorrelated1, 0.933608329586),
                'incorrect value for varianceCorrelated1')
Test.assertTrue(np.allclose(varianceRandom2, 1.0), 'incorrect value for varianceRandom2')
Test.assertTrue(np.allclose(varianceCorrelated2, 1.0), 'incorrect value for varianceCorrelated2')
Test.assertTrue(np.allclose(varianceThreeD2, 0.993967356912), 'incorrect value for varianceThreeD2'
)
```

```
1 test passed.
1 test passed.
1 test passed.
1 test passed.
1 test passed.
```

In [90]:
```
import os
baseDir = os.path.join('data')
inputPath = os.path.join('cs190', 'neuro.txt')

inputFile = os.path.join(baseDir, inputPath)
lines = sc.textFile(inputFile)

# Check that everything loaded properly
assert len(lines.first()) == 1397
assert lines.count() == 46460
```

In [91]:
```
# TODO: Replace <FILL IN> with appropriate code
def parse(line):
    """Parse the raw data into a (`tuple`, `np.ndarray`) pair.

    Note:
        You should store the pixel coordinates as a tuple of two ints and the elements of the pixel
 intensity
        time series as an np.ndarray of floats.

    Args:
        line (str): A string representing an observation.  Elements are separated by spaces.  The
            first two elements represent the coordinates of the pixel, and the rest of the elements
            represent the pixel intensity over time.

    Returns:
        tuple of tuple, np.ndarray: A (coordinate, pixel intensity array) `tuple` where coordinate
is
            a `tuple` containing two values and the pixel intensity is stored in an NumPy array
            which contains 240 values.
    """
    raw_split = line.split(" ")
    coordinate = (int(raw_split[0]), int(raw_split[1]))
    pia =  np.asarray(raw_split[2:], dtype=float)
    return coordinate, pia

rawData = lines.map(parse)
rawData.cache()
entry = rawData.first()
print 'Length of movie is {0} seconds'.format(len(entry[1]))
print 'Number of pixels in movie is {0:,}'.format(rawData.count())
print ('\nFirst entry of rawData (with only the first five values of the NumPy array):\n({0}, {1})'
       .format(entry[0], entry[1][:5]))
```

```
Length of movie is 240 seconds
Number of pixels in movie is 46,460

First entry of rawData (with only the first five values of the NumPy array):
((0, 0), [ 103.   103.7  103.2  102.7  103.8])
```

In [92]:
```
# TEST Parse the data (3b)
Test.assertTrue(isinstance(entry[0], tuple), "entry's key should be a tuple")
Test.assertEquals(len(entry), 2, 'entry should have a key and a value')
Test.assertTrue(isinstance(entry[0][1], int), 'coordinate tuple should contain ints')
Test.assertEquals(len(entry[0]), 2, "entry's key should have two values")
Test.assertTrue(isinstance(entry[1], np.ndarray), "entry's value should be an np.ndarray")
Test.assertTrue(isinstance(entry[1][0], np.float), 'the np.ndarray should consist of np.float value
s')
Test.assertEquals(entry[0], (0, 0), 'incorrect key for entry')
Test.assertEquals(entry[1].size, 240, 'incorrect length of entry array')
Test.assertTrue(np.allclose(np.sum(entry[1]), 24683.5), 'incorrect values in entry array')
```

```
1 test passed.
1 test passed.
1 test passed.
1 test passed.
1 test passed.
1 test passed.
1 test passed.
1 test passed.
1 test passed.
```

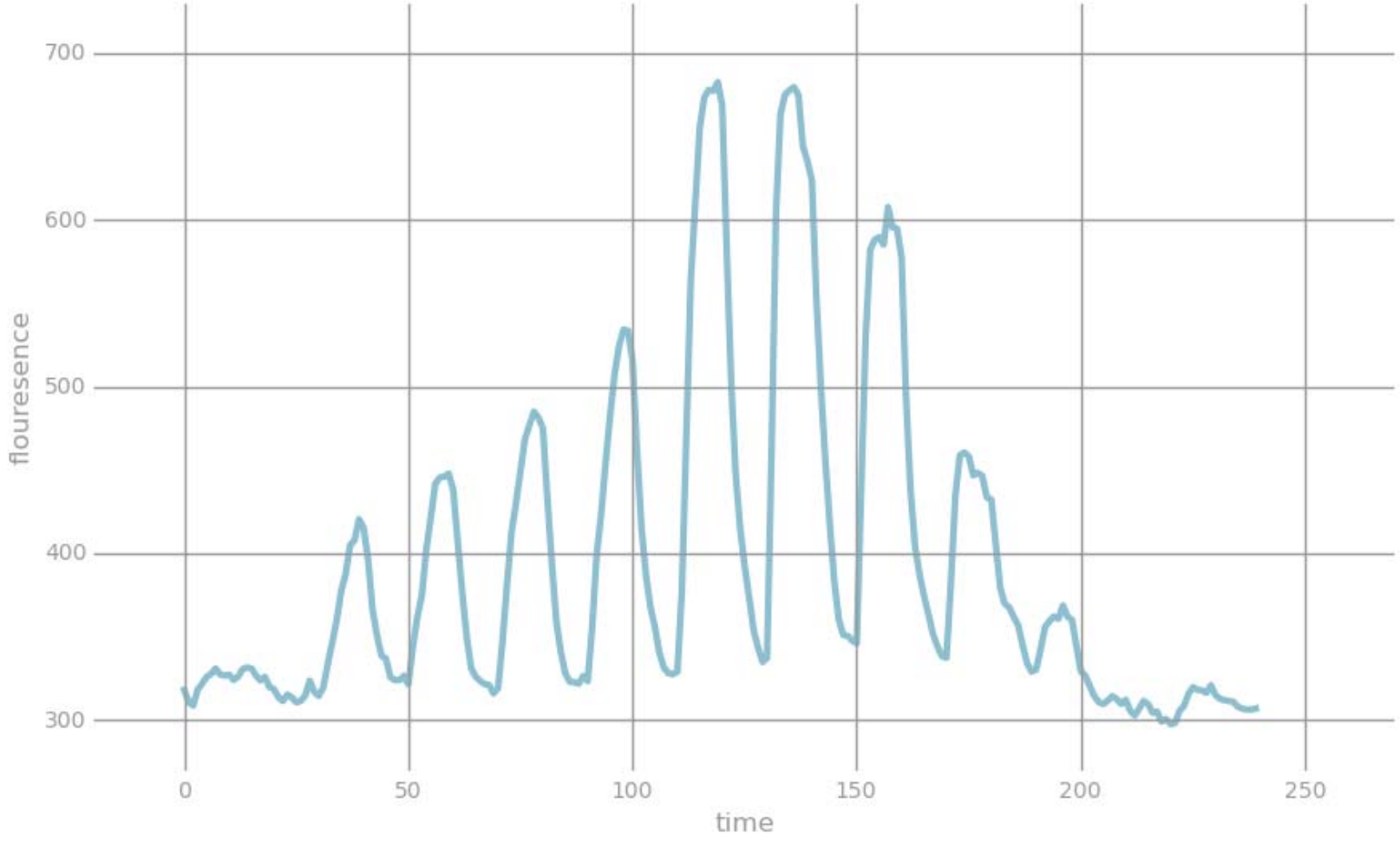
In [93]:
```
# TODO: Replace <FILL IN> with appropriate code
fp = rawData.flatMap(lambda x: x[1])
mn = fp.min()
mx = fp.max()

print mn, mx
```

```
100.6 940.8
```

In [94]:
```
# TEST Min and max flouresence (3c)
Test.assertTrue(np.allclose(mn, 100.6), 'incorrect value for mn')
Test.assertTrue(np.allclose(mx, 940.8), 'incorrect value for mx')
```

```
1 test passed.
1 test passed.
```

In [95]:
```
example = rawData.filter(lambda (k, v): np.std(v) > 100).values().first()

# generate layout and plot data
fig, ax = preparePlot(np.arange(0, 300, 50), np.arange(300, 800, 100))
ax.set_xlabel(r'time'), ax.set_ylabel(r'flouresence')
ax.set_xlim(-20, 270), ax.set_ylim(270, 730)
plt.plot(range(len(example)), example, c='#8cbfd0', linewidth='3.0')
pass
```

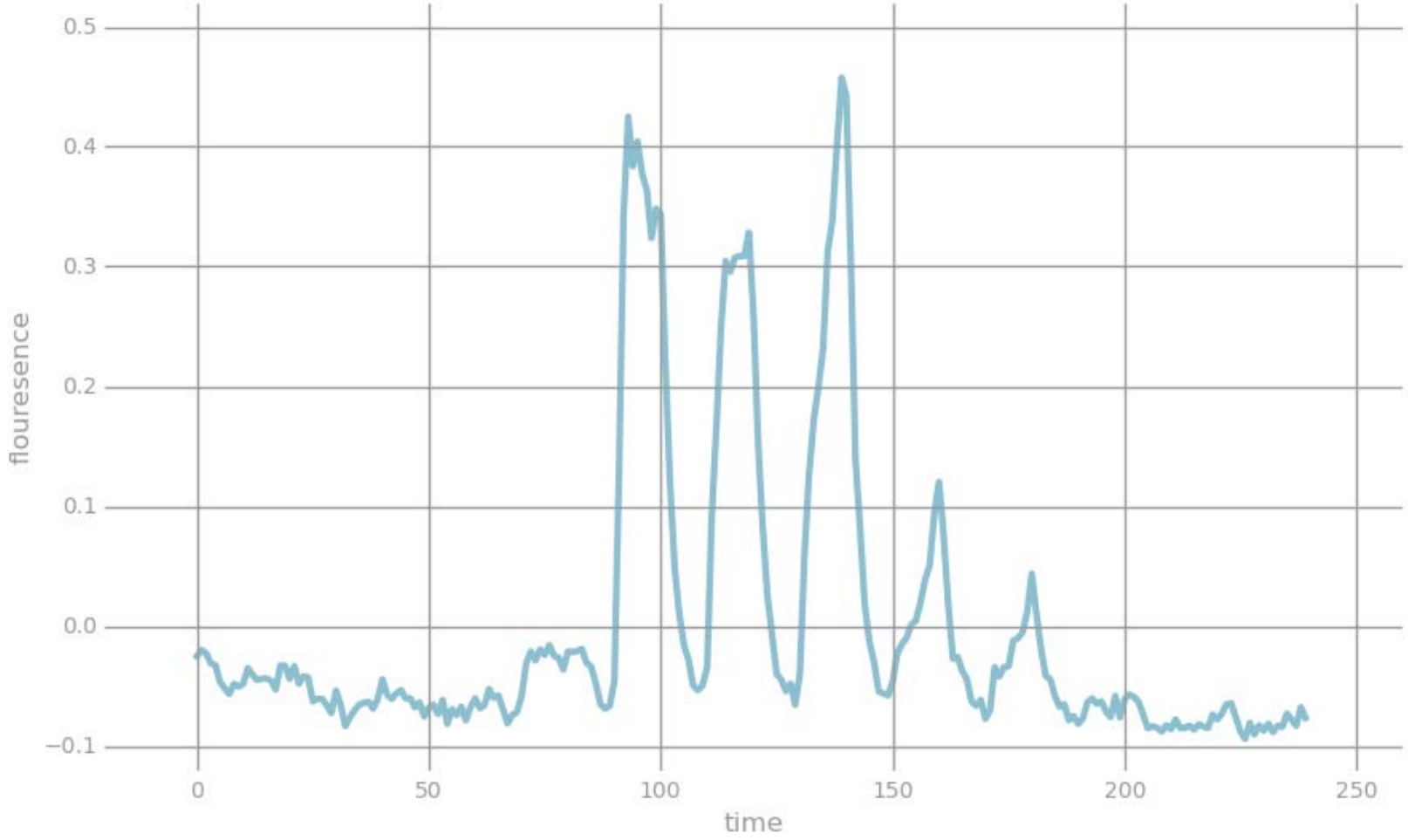


In [96]:
```
# TODO: Replace <FILL IN> with appropriate code

def rescale(ts):
    """Take a np.ndarray and return the standardized array by
    subtracting and dividing by the mean.

    Note:
        You should first subtract the mean and then divide by the mean.

    Args:
        ts (np.ndarray): Time series data (`np.float`) representing pixel intensity.

    Returns:
        np.ndarray: The times series adjusted by subtracting the mean and dividing by the mean.
    """
    a = ts
    b = ts.mean()
    c = a - b
    d = c / b
    return d

scaledData = rawData.mapValues(lambda v: rescale(v))
mnScaled = scaledData.map(lambda (k, v): v).map(lambda v: min(v)).min()
mxScaled = scaledData.map(lambda (k, v): v).map(lambda v: max(v)).max()
```

In [97]:
```
# TEST Fractional signal change (3d)
Test.assertTrue(isinstance(scaledData.first()[1], np.ndarray), 'incorrect type returned by rescale'
)
Test.assertTrue(np.allclose(mnScaled, -0.27151288), 'incorrect value for mnScaled')
Test.assertTrue(np.allclose(mxScaled, 0.90544876), 'incorrect value for mxScaled')
```

```
1 test passed.
1 test passed.
1 test passed.
```

In [98]:
```
example = scaledData.filter(lambda (k, v): np.std(v) > 0.1).values().first()

# generate layout and plot data
fig, ax = preparePlot(np.arange(0, 300, 50), np.arange(-.1, .6, .1))
ax.set_xlabel(r'time'), ax.set_ylabel(r'flouresence')
ax.set_xlim(-20, 260), ax.set_ylim(-.12, .52)
plt.plot(range(len(example)), example, c='#8cbfd0', linewidth='3.0')
pass
```

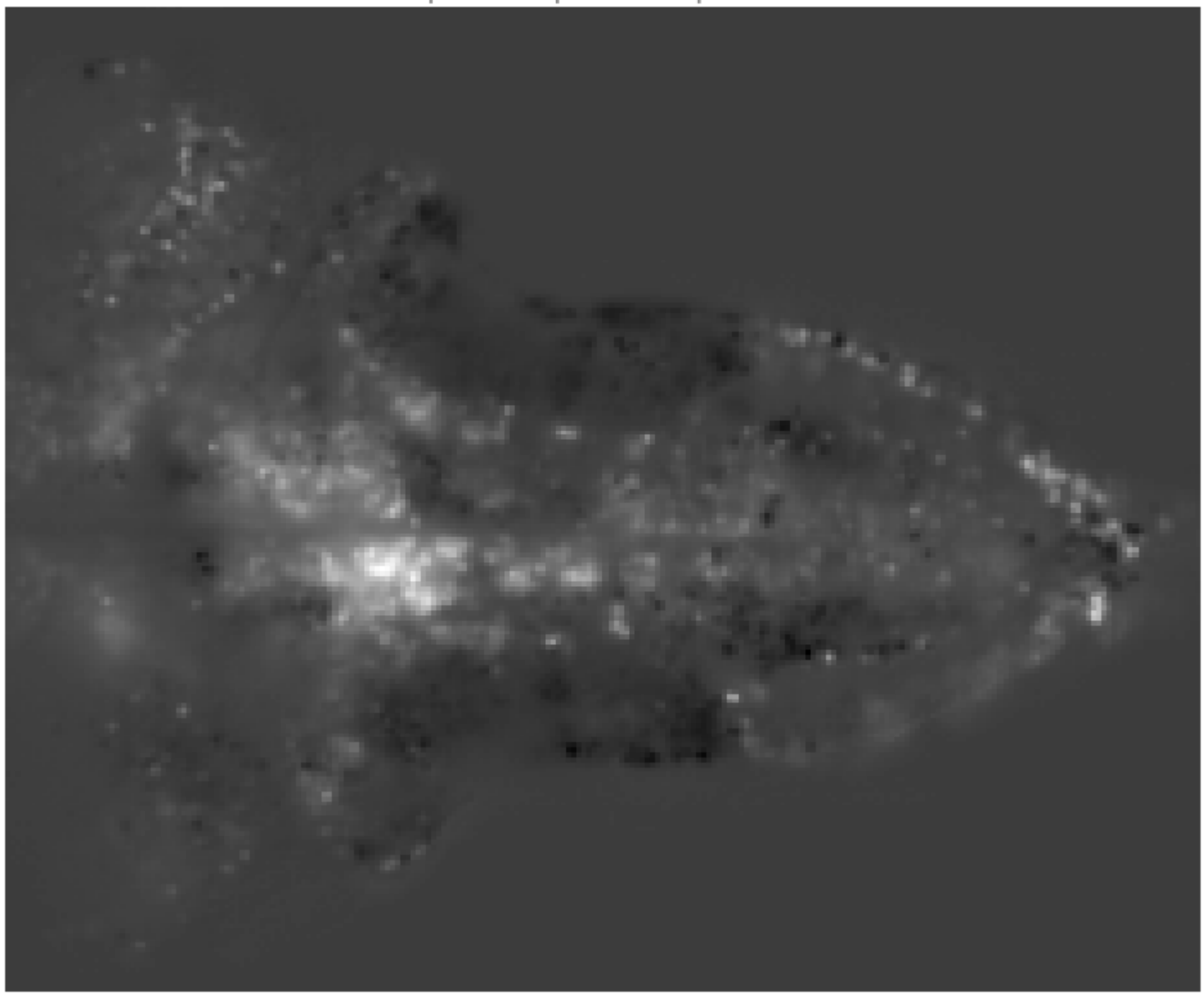


In [99]:
```
# TODO: Replace <FILL IN> with appropriate code
# Run pca using scaledData

just_pia = scaledData.map(lambda x: x[1])
componentsScaled, scaledScores, eigenvaluesScaled = pca(just_pia, 3)
```

In [100]:
```
# TEST PCA on the scaled data (3e)
Test.assertEquals(componentsScaled.shape, (240, 3), 'incorrect shape for componentsScaled')
Test.assertTrue(np.allclose(np.abs(np.sum(componentsScaled[:5, :])), 0.283150995232),
                'incorrect value for componentsScaled')
Test.assertTrue(np.allclose(np.abs(np.sum(scaledScores.take(3))), 0.0285507449251),
                'incorrect value for scaledScores')
Test.assertTrue(np.allclose(np.sum(eigenvaluesScaled[:5]), 0.206987501564),
                'incorrect value for eigenvaluesScaled')
```

```
1 test passed.
1 test passed.
1 test passed.
1 test passed.
```

In [101]:
```
import matplotlib.cm as cm

scoresScaled = np.vstack(scaledScores.collect())
imageOneScaled = scoresScaled[:,0].reshape(230, 202).T

# generate layout and plot data
fig, ax = preparePlot(np.arange(0, 10, 1), np.arange(0, 10, 1), figsize=(9.0, 7.2), hideLabels=Tru
e)
ax.grid(False)
ax.set_title('Top Principal Component', color='#888888')
image = plt.imshow(imageOneScaled,interpolation='nearest', aspect='auto', cmap=cm.gray)
pass
```


Top Principal Component

In [84]:
```
subset_test_users = test_data_q3['user_id'].unique()[0:10000]
```

In [86]:
```
q3_recommendations = personalized_model_q3.recommend(subset_test_users,k=1)
```

```
PROGRESS: recommendations finished on 1000/10000 queries. users per second: 2556.88
PROGRESS: recommendations finished on 2000/10000 queries. users per second: 2647.65
PROGRESS: recommendations finished on 3000/10000 queries. users per second: 2676.48
PROGRESS: recommendations finished on 4000/10000 queries. users per second: 2696.35
PROGRESS: recommendations finished on 5000/10000 queries. users per second: 2686.07
PROGRESS: recommendations finished on 6000/10000 queries. users per second: 2676.49
PROGRESS: recommendations finished on 7000/10000 queries. users per second: 2681.58
PROGRESS: recommendations finished on 8000/10000 queries. users per second: 2698.41
PROGRESS: recommendations finished on 9000/10000 queries. users per second: 2690.2
PROGRESS: recommendations finished on 10000/10000 queries. users per second: 2682
```

In [88]:
```
len(q3_recommendations)
```

Out[88]:
```
10000
```

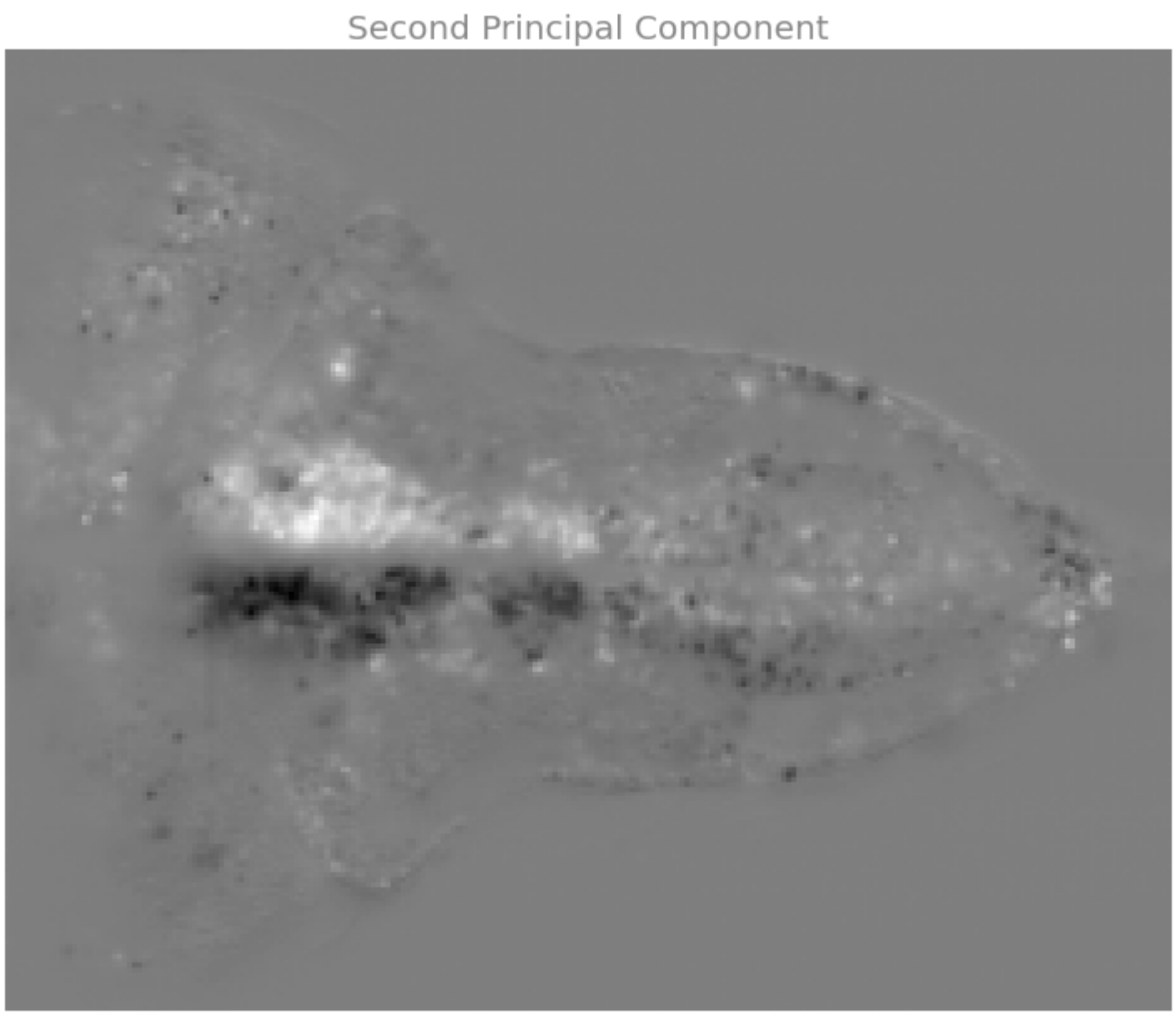
In [98]:
```
song_listen_counts_q3 = q3_recommendations.groupby(key_columns='song',operations={'count'
```

In [99]:
```
song_listen_counts_q3.head()
```

Out[99]:

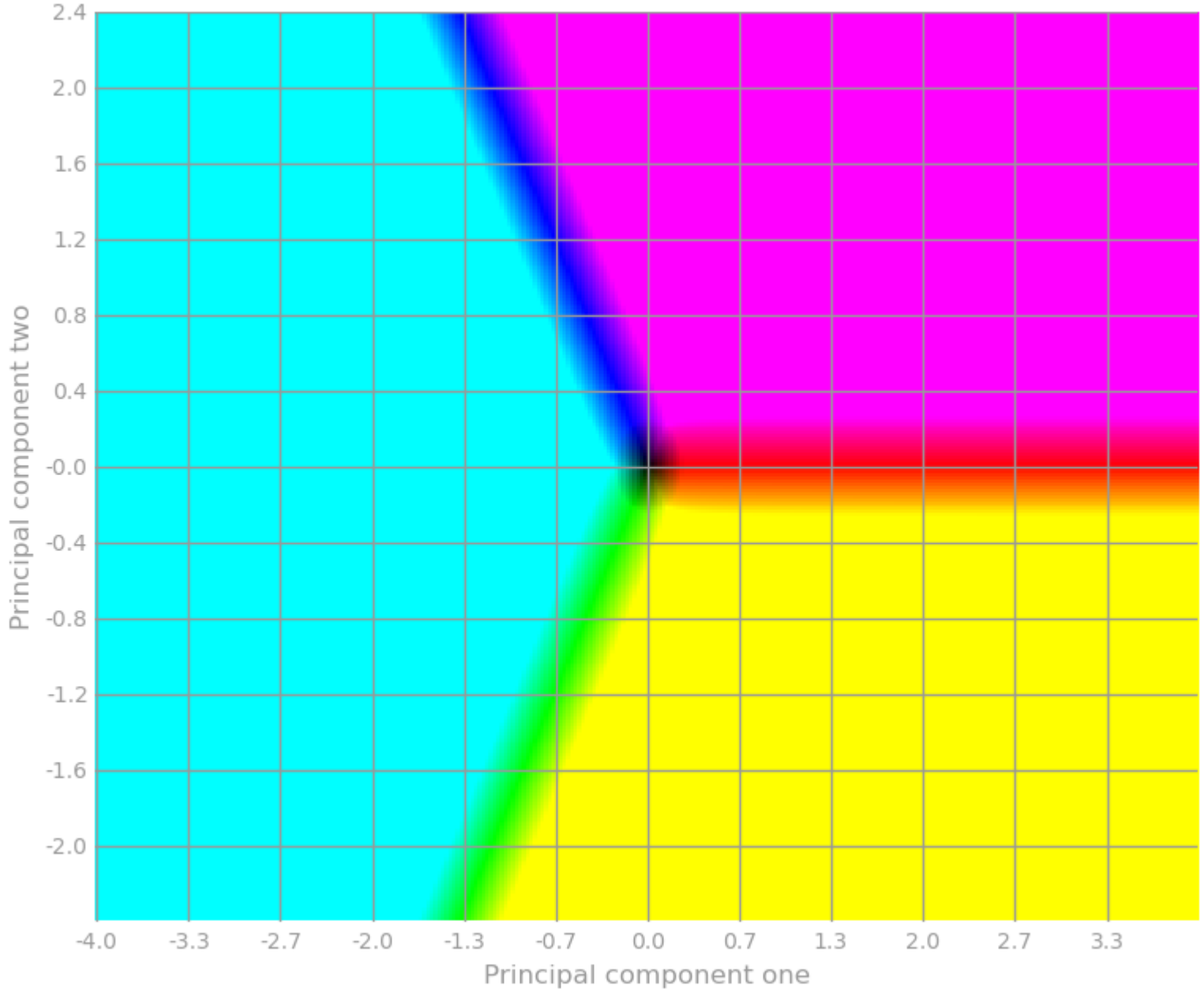
song	count
Arco Arena - Cake	1
Too Deep - Girl Talk	2
Guys Like Me - Eric Church ...	2
Freedom - Akon	2
Wish You Were Here - Incubus ...	1
Change - Blind Melon	1

In [102]:
```
imageTwoScaled = scoresScaled[:,1].reshape(230, 202).T

# generate layout and plot data
fig, ax = preparePlot(np.arange(0, 10, 1), np.arange(0, 10, 1), figsize=(9.0, 7.2), hideLabels=Tru
e)
ax.grid(False)
ax.set_title('Second Principal Component', color='#888888')
image = plt.imshow(imageTwoScaled,interpolation='nearest', aspect='auto', cmap=cm.gray)
pass
```

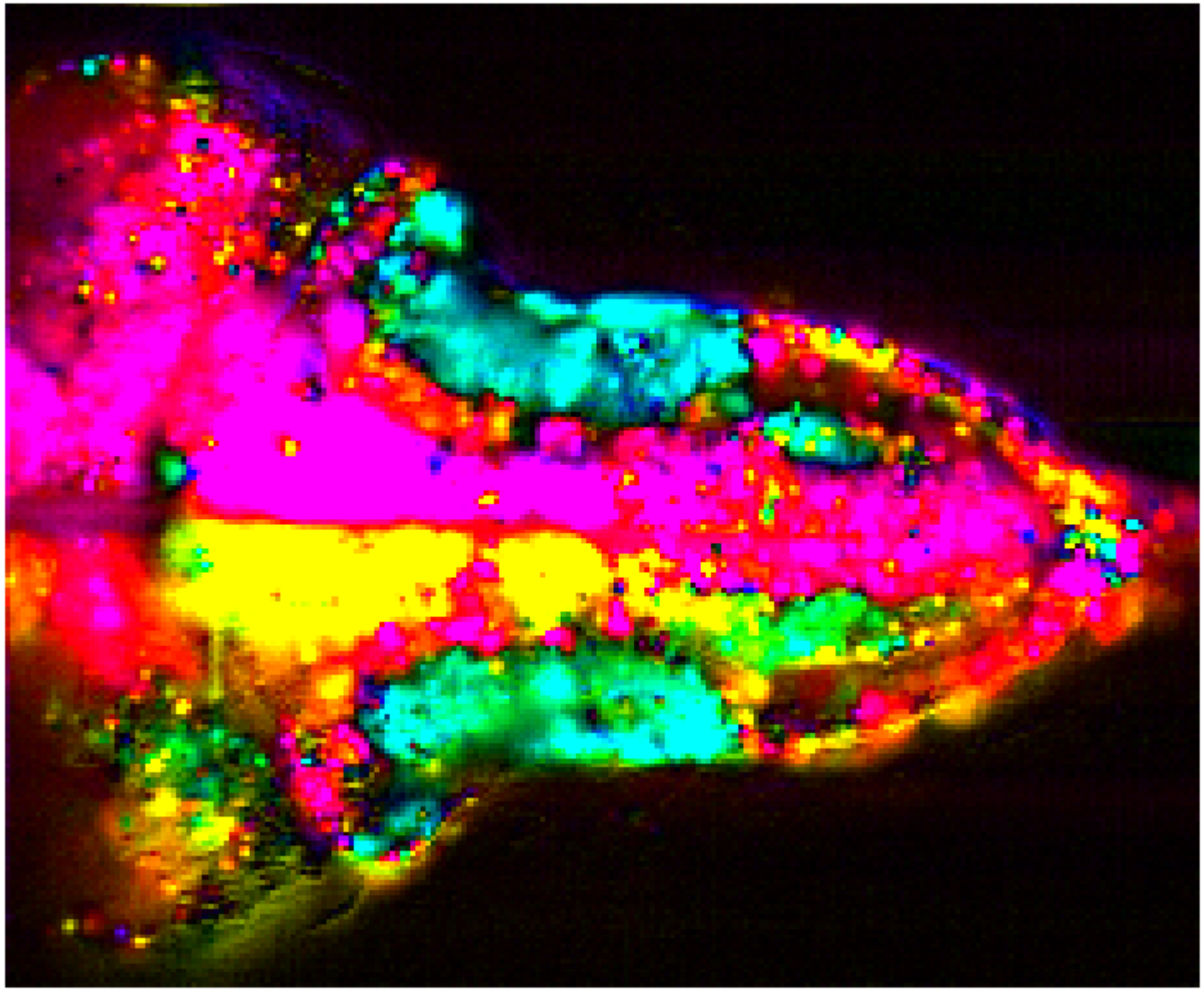

Second Principal Component

In [103]:
```
# Adapted from python-thunder's Colorize.transform where cmap='polar'.
# Checkout the library at: https://github.com/thunder-project/thunder and
# http://thunder-project.org/

def polarTransform(scale, img):
    """Convert points from cartesian to polar coordinates and map to colors."""
    from matplotlib.colors import hsv_to_rgb

    img = np.asarray(img)
    dims = img.shape

    phi = ((np.arctan2(-img[0], -img[1]) + np.pi/2) % (np.pi*2)) / (2 * np.pi)
    rho = np.sqrt(img[0]**2 + img[1]**2)
    saturation = np.ones((dims[1], dims[2]))

    out = hsv_to_rgb(np.dstack((phi, saturation, scale * rho)))

    return np.clip(out * scale, 0, 1)
```

In [104]:
```
# Show the polar mapping from principal component coordinates to colors.
x1AbsMax = np.max(np.abs(imageOneScaled))
x2AbsMax = np.max(np.abs(imageTwoScaled))

numOfPixels = 300
x1Vals = np.arange(-x1AbsMax, x1AbsMax, (2 * x1AbsMax) / numOfPixels)
x2Vals = np.arange(x2AbsMax, -x2AbsMax, -(2 * x2AbsMax) / numOfPixels)
x2Vals.shape = (numOfPixels, 1)

x1Data = np.tile(x1Vals, (numOfPixels, 1))
x2Data = np.tile(x2Vals, (1, numOfPixels))

# Try changing the first parameter to lower values
polarMap = polarTransform(2.0, [x1Data, x2Data])

gridRange = np.arange(0, numOfPixels + 25, 25)
fig, ax = preparePlot(gridRange, gridRange, figsize=(9.0, 7.2), hideLabels=True)
image = plt.imshow(polarMap, interpolation='nearest', aspect='auto')
ax.set_xlabel('Principal component one'), ax.set_ylabel('Principal component two')
gridMarks = (2 * gridRange / float(numOfPixels) - 1.0)
x1Marks = x1AbsMax * gridMarks
x2Marks = -x2AbsMax * gridMarks
ax.get_xaxis().set_ticklabels(map(lambda x: '{0:.1f}'.format(x), x1Marks))
ax.get_yaxis().set_ticklabels(map(lambda x: '{0:.1f}'.format(x), x2Marks))
pass
```


2.4
2.0
1.6
1.2
0.8
0.4
-0.0
-0.4
-0.8
-1.2
-1.6
-2.0
Principal component two
-4.0
-3.3
-2.7
-2.0
-1.3
-0.7
0.0
0.7
1.3
2.0
2.7
3.3
Principal component one

In [105]:
```
# Use the same transformation on the image data
# Try changing the first parameter to lower values
brainmap = polarTransform(2.0, [imageOneScaled, imageTwoScaled])

# generate layout and plot data
fig, ax = preparePlot(np.arange(0, 10, 1), np.arange(0, 10, 1), figsize=(9.0, 7.2), hideLabels=Tru
e)
ax.grid(False)
image = plt.imshow(brainmap,interpolation='nearest', aspect='auto')
pass
```



Part 4: Feature-based aggregation and PCA

In [106]:
```
# TODO: Replace <FILL IN> with appropriate code
vector = np.array([0., 1., 2., 3., 4., 5.])

# Create a multi-dimensional array that when multiplied (using .dot)
# against vector, results in a two element array where the first element
# is the sum of the 0, 2, and 4 indexed elements of
# vector and the second element is the sum of the 1, 3, and 5
# indexed elements of vector.
# This should be a 2 row by 6 column array

sumEveryOther = np.array([[1,0,1,0,1,0],[0,1,0,1,0,1]])

# Create a multi-dimensional array that when multiplied (using .dot)
# against vector, results in a
# three element array where the first element is the sum of the 0 and 3
# indexed elements of vector,
# the second element is the sum of the 1 and 4 indexed elements of vector,
# and the third element is
# the sum of the 2 and 5 indexed elements of vector.
# This should be a 3 row by 6 column array

sumEveryThird = np.array([[1,0,0,1,0,0],[0,1,0,0,1,0],[0,0,1,0,0,1]])

# Create a multi-dimensional array that can be used to sum the
# first three elements of vector and
# the last three elements of vector, which returns a two element
# array with those values when dotted
# with vector.
# sumByThree = np.array([[0,0,0,1,0,0],[0,0,0,0,3,0]])
sumByThree = np.array([[ 1.,  1.,  1.,  0.,  0.,  0.], [ 0.,  0.,  0.,  1.,  1.,  1.]])

# Create a multi-dimensional array that sums the first two elements, second two elements, and
# last two elements of vector, which returns a three element array with those values when dotted
# with vector.
# This should be a 3 row by 6 column array
# sumByTwo = np.array([[0,1,0,0,0,0],[0,0,0,0,0,1],[0,0,0,3,0,0]])
sumByTwo = (np.array([[ 1.,  1.,  0.,  0.,  0.,  0.],
 [ 0.,  0.,  1.,  1.,  0.,  0.],
 [ 0.,  0.,  0.,  0.,  1.,  1.]]))
print 'sumEveryOther.dot(vector):\t{0}'.format(sumEveryOther.dot(vector))
print 'sumEveryThird.dot(vector):\t{0}'.format(sumEveryThird.dot(vector))
```

```
sumEveryOther.dot(vector):      [ 6.  9.]
sumEveryThird.dot(vector):      [ 3.  5.  7.]

sumByThree.dot(vector): [  3.  12.]
sumByTwo.dot(vector):   [ 1.  5.  9.]
```

In [107]:
```
# TEST Aggregation using arrays (4a)
Test.assertEquals(sumEveryOther.shape, (2, 6), 'incorrect shape for sumEveryOther')
Test.assertEquals(sumEveryThird.shape, (3, 6), 'incorrect shape for sumEveryThird')
Test.assertTrue(np.allclose(sumEveryOther.dot(vector), [6, 9]), 'incorrect value for sumEveryOther
')
Test.assertTrue(np.allclose(sumEveryThird.dot(vector), [3, 5, 7]),
                'incorrect value for sumEveryThird')
Test.assertEquals(sumByThree.shape, (2, 6), 'incorrect shape for sumByThree')
Test.assertEquals(sumByTwo.shape, (3, 6), 'incorrect shape for sumByTwo')
Test.assertTrue(np.allclose(sumByThree.dot(vector),  [3, 12]), 'incorrect value for sumByThree')
Test.assertTrue(np.allclose(sumByTwo.dot(vector), [1, 5, 9]), 'incorrect value for sumByTwo')
```

```
1 test passed.
1 test passed.
1 test passed.
1 test passed.
1 test passed.
1 test passed.
1 test passed.
1 test passed.
```

In [108]:
```
# Reference for what to recreate
print 'sumEveryOther: \n{0}'.format(sumEveryOther)
print '\nsumEveryThird: \n{0}'.format(sumEveryThird)
```

```
sumEveryOther:
[[1 0 1 0 1 0]
 [0 1 0 1 0 1]]

sumEveryThird:
[[1 0 0 1 0 0]
 [0 1 0 0 1 0]
 [0 0 1 0 0 1]]
```

In [109]:
```
# TODO: Replace <FILL IN> with appropriate code
# Use np.tile and np.eye to recreate the arrays
sumEveryOtherTile = np.tile(np.eye(2), 3)
sumEveryThirdTile = np.tile(np.eye(3), 2)

print sumEveryOtherTile
print 'sumEveryOtherTile.dot(vector): {0}'.format(sumEveryOtherTile.dot(vector))
print '\n', sumEveryThirdTile
print 'sumEveryThirdTile.dot(vector): {0}'.format(sumEveryThirdTile.dot(vector))
```

```
[[ 1.  0.  1.  0.  1.  0.]
 [ 0.  1.  0.  1.  0.  1.]]
sumEveryOtherTile.dot(vector): [ 6.  9.]

[[ 1.  0.  0.  1.  0.  0.]
 [ 0.  1.  0.  0.  1.  0.]
 [ 0.  0.  1.  0.  0.  1.]]
sumEveryThirdTile.dot(vector): [ 3.  5.  7.]
```

In [110]:
```
# TEST Recreate with `np.tile` and `np.eye` (4b)
Test.assertEquals(sumEveryOtherTile.shape, (2, 6), 'incorrect shape for sumEveryOtherTile')
Test.assertEquals(sumEveryThirdTile.shape, (3, 6), 'incorrect shape for sumEveryThirdTile')
Test.assertTrue(np.allclose(sumEveryOtherTile.dot(vector), [6, 9]),
                'incorrect value for sumEveryOtherTile')
Test.assertTrue(np.allclose(sumEveryThirdTile.dot(vector), [3, 5, 7]),
                'incorrect value for sumEveryThirdTile')
```

```
1 test passed.
1 test passed.
1 test passed.
1 test passed.
```

In [111]:
```
# Reference for what to recreate
print 'sumByThree: \n{0}'.format(sumByThree)
print '\nsumByTwo: \n{0}'.format(sumByTwo)
```

```
sumByThree:
[[ 1.  1.  1.  0.  0.  0.]
 [ 0.  0.  0.  1.  1.  1.]]

sumByTwo:
[[ 1.  1.  0.  0.  0.  0.]
 [ 0.  0.  1.  1.  0.  0.]
 [ 0.  0.  0.  0.  1.  1.]]
```

In [112]:
```
# TODO: Replace <FILL IN> with appropriate code
# Use np.kron, np.eye, and np.ones to recreate the arrays
sumByThreeKron = np.kron(np.eye(2),np.ones(3))
sumByTwoKron = np.kron(np.eye(3),np.ones(2))

print sumByThreeKron
print 'sumByThreeKron.dot(vector): {0}'.format(sumByThreeKron.dot(vector))
print '\n', sumByTwoKron
print 'sumByTwoKron.dot(vector): {0}'.format(sumByTwoKron.dot(vector))
```

```
[[ 1.  1.  1.  0.  0.  0.]
 [ 0.  0.  0.  1.  1.  1.]]
sumByThreeKron.dot(vector): [  3.  12.]

[[ 1.  1.  0.  0.  0.  0.]
 [ 0.  0.  1.  1.  0.  0.]
 [ 0.  0.  0.  0.  1.  1.]]
sumByTwoKron.dot(vector): [ 1.  5.  9.]
```

In [113]:
```
# TEST Recreate with `np.kron` (4c)
Test.assertEquals(sumByThreeKron.shape, (2, 6), 'incorrect shape for sumByThreeKron')
Test.assertEquals(sumByTwoKron.shape, (3, 6), 'incorrect shape for sumByTwoKron')
Test.assertTrue(np.allclose(sumByThreeKron.dot(vector),  [3, 12]),
                'incorrect value for sumByThreeKron')
Test.assertTrue(np.allclose(sumByTwoKron.dot(vector), [1, 5, 9]),
                'incorrect value for sumByTwoKron')
```

```
1 test passed.
1 test passed.
1 test passed.
1 test passed.
```

In [114]:
```
# TODO: Replace <FILL IN> with appropriate code
# Create a multi-dimensional array to perform the aggregation
#T = np.tile(?, 12)
T = np.tile(np.eye(20), 12)
print(T.shape)
# Transform scaledData using T.  Make sure to retain the keys.
timeData = scaledData.map(lambda (k, v): (k, T.dot(v)))

print(timeData.values().first())
```

```
(20, 240)
[ 0.00802155  0.00607693 -0.0075354   0.00121539  0.02163388  0.00121539
 -0.03087082  0.00510462  0.01191079  0.02455081 -0.0182308   0.00802155
 -0.00948002 -0.00948002  0.02163388 -0.02212004  0.00704924  0.00121539
 -0.01142464 -0.00850771]
```

In [115]:
```
# TEST Aggregate by time (4d)
Test.assertEquals(T.shape, (20, 240), 'incorrect shape for T')
timeDataFirst = timeData.values().first()
timeDataFifth = timeData.values().take(5)[4]
Test.assertEquals(timeData.count(), 46460, 'incorrect length of timeData')
Test.assertEquals(timeDataFirst.size, 20, 'incorrect value length of timeData')
Test.assertEquals(timeData.keys().first(), (0, 0), 'incorrect keys in timeData')
Test.assertTrue(np.allclose(timeDataFirst[:2], [0.00802155, 0.00607693]),
                'incorrect values in timeData')
Test.assertTrue(np.allclose(timeDataFifth[-2:],[-0.00636676, -0.0179427]),
                'incorrect values in timeData')
```

```
1 test passed.
1 test passed.
1 test passed.
1 test passed.
1 test passed.
1 test passed.
```

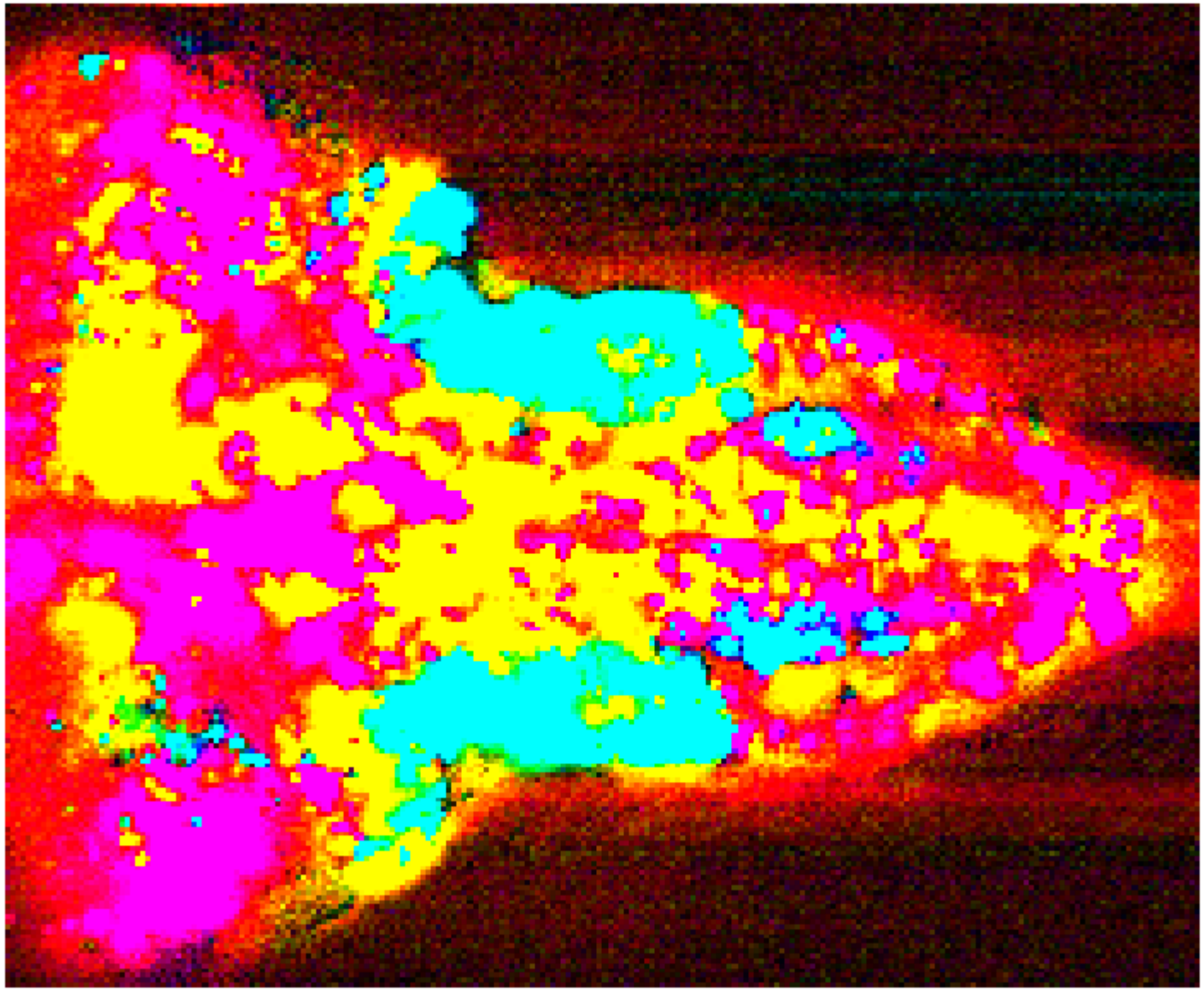
In [116]:
```
# TODO: Replace <FILL IN> with appropriate code

just_tpia = timeData.map(lambda x: x[1])
componentsTime, timeScores, eigenvaluesTime = pca(just_tpia, 3)

print 'componentsTime: (first five) \n{0}'.format(componentsTime[:5,:])
print ('\ntimeScores (first three): \n{0}'
       .format('\n'.join(map(str, timeScores.take(3)))))
print '\neigenvaluesTime: (first five) \n{0}'.format(eigenvaluesTime[:5])
```

```
componentsTime: (first five)
[[ 0.27392702 -0.16152431  0.01388556]
 [ 0.09941893 -0.31968127 -0.34738824]
 [-0.03376505 -0.32933108 -0.35606954]
 [-0.12092744 -0.2845482  -0.27232364]
 [-0.18219248 -0.22998061 -0.12248985]]

timeScores (first three):
[-0.00720617 -0.00292979 -0.00223645]
[ 0.02353076 -0.00197457  0.00362094]
[ 0.01310623  0.00123069 -0.00582974]

eigenvaluesTime: (first five)
[ 0.77528991  0.05038881  0.01173423  0.0059711   0.00138073]
```

In [117]:
```
# TEST Obtain a compact representation (4e)
Test.assertEquals(componentsTime.shape, (20, 3), 'incorrect shape for componentsTime')
Test.assertTrue(np.allclose(np.abs(np.sum(componentsTime[:5, :])), 2.37299020),
                'incorrect value for componentsTime')
Test.assertTrue(np.allclose(np.abs(np.sum(timeScores.take(3))), 0.0213119114),
                'incorrect value for timeScores')
Test.assertTrue(np.allclose(np.sum(eigenvaluesTime[:5]), 0.844764792),
                'incorrect value for eigenvaluesTime')
```

```
1 test passed.
1 test passed.
1 test passed.
1 test passed.
```

In [118]:
```
scoresTime = np.vstack(timeScores.collect())
imageOneTime = scoresTime[:,0].reshape(230, 202).T
imageTwoTime = scoresTime[:,1].reshape(230, 202).T
brainmap = polarTransform(3, [imageOneTime, imageTwoTime])

# generate layout and plot data
fig, ax = preparePlot(np.arange(0, 10, 1), np.arange(0, 10, 1), figsize=(9.0, 7.2), hideLabels=Tru
e)
ax.grid(False)
image = plt.imshow(brainmap,interpolation='nearest', aspect='auto')
pass
```



In [119]:
```
# TODO: Replace <FILL IN> with appropriate code
# Create a multi-dimensional array to perform the aggregation


D = np.kron(np.eye(12),np.ones(20))
# Transform scaledData using D.  Make sure to retain the keys.
directionData = scaledData.map(lambda (k, v): (k, D.dot(v)))

directionData.cache()
print directionData.count()
print directionData.first()
```

```
46460
((0, 0), array([ 0.03346365,  0.03638058, -0.02195799, -0.02487492,  0.00721129,
        0.00332206, -0.02098568,  0.00915591, -0.00542873, -0.01029027,
        0.0081836 , -0.01417951]))
```

In [120]:
```
# TEST Aggregate by direction (4f)
Test.assertEquals(D.shape, (12, 240), 'incorrect shape for D')
directionDataFirst = directionData.values().first()
directionDataFifth = directionData.values().take(5)[4]
Test.assertEquals(directionData.count(), 46460, 'incorrect length of directionData')
Test.assertEquals(directionDataFirst.size, 12, 'incorrect value length of directionData')
Test.assertEquals(directionData.keys().first(), (0, 0), 'incorrect keys in directionData')
Test.assertTrue(np.allclose(directionDataFirst[:2], [ 0.03346365,  0.03638058]),
                'incorrect values in directionData')
Test.assertTrue(np.allclose(directionDataFifth[:2], [ 0.01479147, -0.02090099]),
                'incorrect values in directionData')
```

```
1 test passed.
1 test passed.
1 test passed.
1 test passed.
1 test passed.
1 test passed.
```

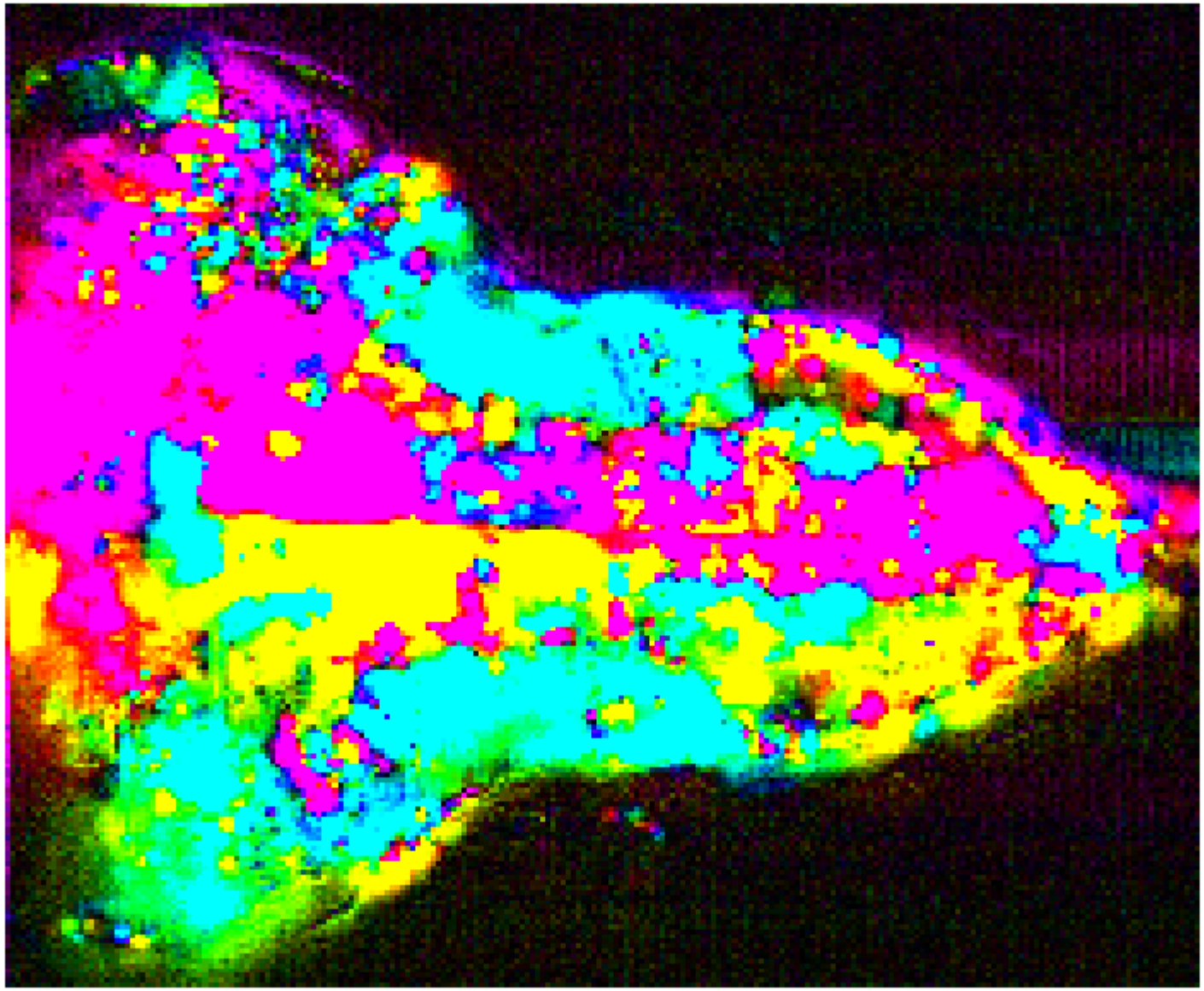
In [121]:
```
# TODO: Replace <FILL IN> with appropriate code
just_dpia = directionData.map(lambda x: x[1])

componentsDirection, directionScores, eigenvaluesDirection = pca(just_dpia, 3)

print 'componentsDirection: (first five) \n{0}'.format(componentsDirection[:5,:])
print ('\ndirectionScores (first three): \n{0}'
       .format('\n'.join(map(str, directionScores.take(3)))))
print '\neigenvaluesDirection: (first five) \n{0}'.format(eigenvaluesDirection[:5])
```

```
componentsDirection: (first five)
[[-0.25952179  0.16201941  0.24947433]
 [-0.31369506 -0.09185175  0.29464223]
 [-0.21716693 -0.35944645  0.35296454]
 [-0.11517273 -0.37356905  0.07169062]
 [ 0.02996577 -0.36272623 -0.14783897]]

directionScores (first three):
[-0.01622513  0.01322998  0.01322204]
[ 0.00999482  0.0652367  -0.04524758]
[ 0.004646    0.05751097  0.00756383]

eigenvaluesDirection: (first five)
[ 0.96411048  0.77613553  0.12762987  0.09775924  0.04333691]
```

In [122]:
```
# TEST Compact representation of direction data (4g)
Test.assertEquals(componentsDirection.shape, (12, 3), 'incorrect shape for componentsDirection')
Test.assertTrue(np.allclose(np.abs(np.sum(componentsDirection[:5, :])), 1.080232069),
                'incorrect value for componentsDirection')
Test.assertTrue(np.allclose(np.abs(np.sum(directionScores.take(3))), 0.10993162084),
                'incorrect value for directionScores')
Test.assertTrue(np.allclose(np.sum(eigenvaluesDirection[:5]), 2.0089720377),
                'incorrect value for eigenvaluesDirection')
```

```
1 test passed.
1 test passed.
1 test passed.
1 test passed.
```

In [61]:
```
scoresDirection = np.vstack(directionScores.collect())
imageOneDirection = scoresDirection[:,0].reshape(230, 202).T
imageTwoDirection = scoresDirection[:,1].reshape(230, 202).T
brainmap = polarTransform(2, [imageOneDirection, imageTwoDirection])
# with thunder: Colorize(cmap='polar', scale=2).transform([imageOneDirection, imageTwoDirection])

# generate layout and plot data
fig, ax = preparePlot(np.arange(0, 10, 1), np.arange(0, 10, 1), figsize=(9.0, 7.2), hideLabels=True
)
ax.grid(False)
image = plt.imshow(brainmap, interpolation='nearest', aspect='auto')
pass
```



In [ ]:
```
'This is a lab in the Scalable Machine Learning course at Berkeley / EdX.'
```

Section S - Example: Using distributed processing to analyze structured data

This section provides an example of using distributed processing to analyze structured data.

(Please turn the page.)

Validating that a program does what you expect

Author: Russ Robbins

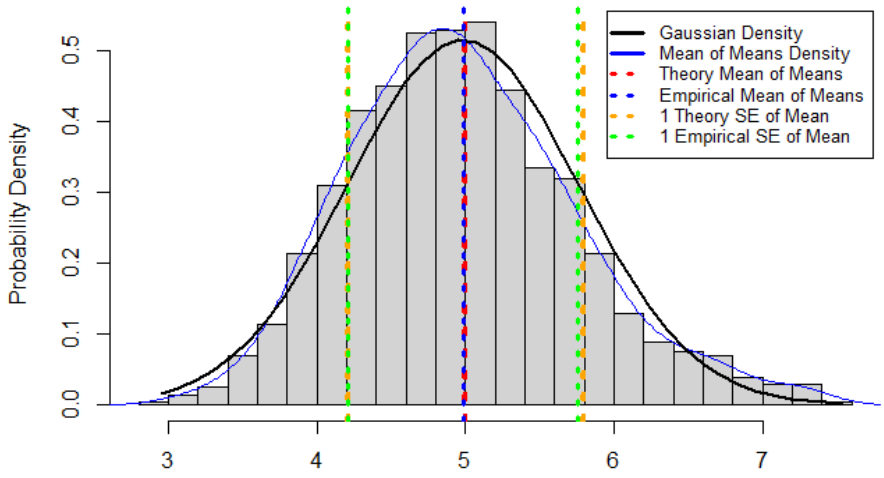
Affiliated Code Repository (right click, and open new window or tab)

Executive Summary

Statistical theory indicates that sample statistics are unbiased estimators of the population parameters they seek to represent. The R programming language provides a built-in function named rexp(). The purpose of rexp() is to generate values for the theoretical exponential distribution.

This example shows that the data generated by the R function rexp() indeed, over 40 draws performed 1000 times, does provide data that maps to statistical theory. Therefore, at least in the context shown here, the function rexp() acts as expected. Note that if rexp() acts as expected, rexp() can then be used for simulation experiments that use rexp's outputs.

Fig. 1: Comparing a function's output to its expected output



Means of 1000 Samples of 40 Numbers drawn randomly from the Exponential Distribution

Technical Discussion

The average mean of the distributions generated by rexp() is centered at 5.0365828. This very similar to what theory purports. Theory, for this lambda, indicated it would be centered at 5.

The standard error of the mean of the distributions generated by rexp() is 0.7883693. Theory, for this lambda, indicated it would be 0.7905694.

Note that the Mean of Means Density distribution (the solid blue line forming a bell shape) that was generated approximates the Gaussian distribution (the solid black line).

This is what is expected when sample statistics are indeed unbiased estimator. This is the behavior the rexp() function seeks to provide.

Steps in Analysis

Note that the results (as shown in the figures above) differ from those below since a different set of simulations were run. Also note that the R code that underlays the executive summary is hidden (i.e., echo=FALSE).

1. First I set the parameters to be used with rexp().

```
a_lambda <- 0.2
a_n=40
a_num_sims=1:1000
a_sample_mean<-1/lambda
a_SD<-1/a_lambda
a_exp_mean_of_means <- a_sample_mean
a_exp_SE_of_means <- SD / sqrt(n)
```

2. Then I checked to see that the function rexp() actually creates different distributions each time.

```
par(mfrow = c(1,3))

set.seed(1020)
a_rexp_sim1<-rexp(a_n,a_lambda)
a_rexp_mean1<-round(mean(a_rexp_sim1),3)
a_rexp_sd1<-round(sd(a_rexp_sim1),3)
a_xname<-paste("Mean: ", a_rexp_mean1," SD: ", a_rexp_sd1)
hist(x=a_rexp_sim1,xlab=a_xname)

set.seed(3888)
a_rexp_sim2<-rexp(a_n,a_lambda)
a_rexp_mean2<-round(mean(a_rexp_sim2),3)
a_rexp_sd2<-round(sd(a_rexp_sim2),3)
a_xname<-paste("Mean: ", a_rexp_mean2," SD: ", a_rexp_sd2)
hist(x=a_rexp_sim2,xlab=a_xname)

set.seed(4242)
a_rexp_sim3<-rexp(a_n,a_lambda)
a_rexp_mean3<-round(mean(a_rexp_sim3),3)
a_rexp_sd3<-round(sd(a_rexp_sim3),3)
a_xname<-paste("Mean: ", a_rexp_mean3," SD: ", a_rexp_sd3)
hist(x=a_rexp_sim3,xlab=a_xname)
```

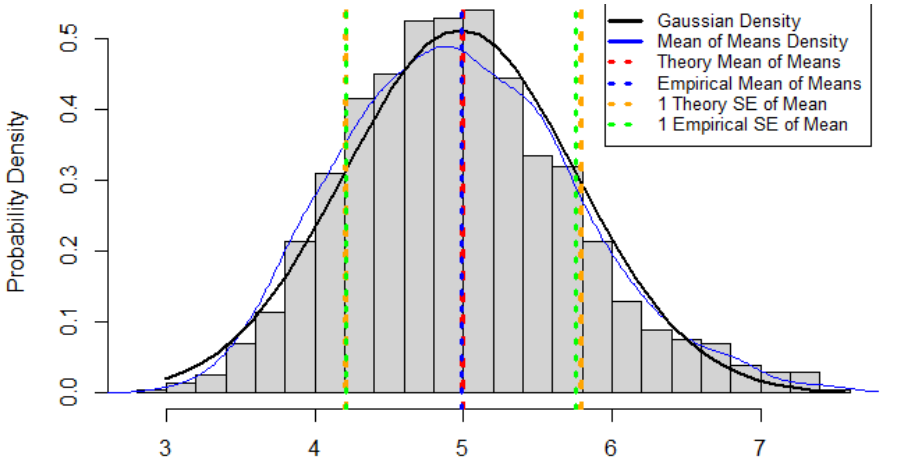
3. Then I ran 1000 simulations of samples of 40 randomly drawn exponentials and computed the empirical mean as well as the empircal standard error.

```
a_rexp_sims<-data.frame(lapply(a_num_sims, function (x) {rexp(a_n,a_lambda)}))
colnames(a_rexp_sims)<-a_num_sims
a_rexp_sims_sample_means<-data.frame(sapply(a_rexp_sims,mean))
colnames(a_rexp_sims_sample_means)<-c("means")
a_emp_mean_of_means<-mean(a_rexp_sims_sample_means[,1])
a_emp_SE_of_means<-sd(a_rexp_sims_sample_means[,1])
a_h<-unlist(a_rexp_sims_sample_means[,1])
```

4. Then I created a table and a histogram to show that the theory and the empirical results actually agreed. Note that this figure is slightly different to that generated above because the analysis was run again in as this **steps in the analysis** section was built. In other words, the rexp behaves randomly each time. Again, this is what it claims to do. More importantly, this kind of behavior would be very helpful when you do simulation studies. You wouldn't want every set of simulations to do the same thing. Instead you would want your simulations to be driven by randomly generated values, so that you could mimic a prospective future, that you do not know yet.

```
statistics<-c(a_exp_mean_of_means,
              a_emp_mean_of_means,
              a_exp_SE_of_means,
              a_emp_SE_of_means)

a_t<-matrix(data=c(a_exp_mean_of_means,
                   a_emp_mean_of_means,a_exp_SE_of_means,a_emp_SE_of_means),
            nrow=2, ncol=2, byrow=TRUE)
a_t<-data.frame(a_t)
colnames(a_t)<-c("Theoretical","Empirical")
rownames(a_t)<-c("Mean","SE")

hist(h,breaks=20,col="light gray",prob=TRUE,main="Fig. 2: Comparing a function's output to its expected output"
,ylab="Probability Density",xlab="Means of 1000 Samples of 40 Numbers drawn randomly from the Exponential Distr
ibution")
lines(density(a_h),col="blue")
legend("topright",legend=c("Gaussian Density", "Mean of Means Density", "Theory Mean of Means", "Empirical Mean
 of Means","1 Theory SE of Mean", "1 Empirical SE of Mean"),lty=c(1,1,3,3,3,3),lwd=c(3,3,3,3,3,3), col=c("black
","blue","red", "blue","orange","green"),merge=TRUE,inset=.01,cex=.8,adj=0)
a_xfit<-seq(min(a_h),max(a_h),length=40)
    a_yfit<-dnorm(a_xfit,mean=mean(a_h),sd=sd(a_h))
lines(a_xfit, a_yfit, col="black", lwd=2)
abline(v=exp_mean_of_means,col="red", lwd=4, lty=3)
abline(v=emp_mean_of_means,col="blue", lwd=3, lty=3)
abline(v=exp_mean_of_means-exp_SE_of_means,col="orange", lwd=4, lty=3)
abline(v=exp_mean_of_means+exp_SE_of_means,col="orange", lwd=4, lty=3)
abline(v=emp_mean_of_means-emp_SE_of_means,col="green", lwd=3, lty=3)
abline(v=emp_mean_of_means+emp_SE_of_means,col="green", lwd=3, lty=3)
```

Fig. 2: Comparing a function's output to its expected output



Means of 1000 Samples of 40 Numbers drawn randomly from the Exponential Distribution

Section T - Data: Skills List

Note that the skills list in this section (and its affiliated links) can be reached at:

http://robbinsr.github.io/assets/skills/skills.pdf

(Please turn the page.)

Russ Robbins
Skills – Detailed List
http://robbinsr.squarespace.com
Contact Me

Analytics Fundamentals

- Relational Algebra (basic)
- Structured Query Language (SQL) (proficient)
- Multidimensional Data Modeling (proficient)
- Online Analytical Processing (proficient)
- Extraction, Transformation, Load (proficient)
- Linear Algebra-Matrix/Vector Operations (basic)
- Extensible Markup Language (basic)
- JavaScript Object Notation (basic)
- Comma Separated Values Files (proficient)
- Graph Theory (basic)

Databases

- Cassandra (formal training)
- Ontotext GraphDB (used)
- MongoDB (formal training)
- MySQL (used)
- Neo4j (evaluated)
- Oracle (used)
- SQL Server (used)
- Stardog (used)
- Teradata (evaluated very lightly)
- Virtuoso (evaluated)

Development Environments

- Anaconda (evaluated)
- Databricks (used)
- Eclipse (used)
- Enthought Canopy (evaluated)
- IDLE (evaluated)
- iPython interpreter (evaluated)
- iPython notebook(used)
- Komodo (evaluated)
- NetBeans (used)
- Oracle SQL Developer (used)
- Oracle Applications (used)
- Pycharm (used)
- Revolution R (used)
- R Studio (used)

Development Environments (continued)

- Spyder (used)
- Stanford Protege (used)
- Teradata Studio Express (evaluated very lightly)
- TopBraid Composer (used)
- Visual Studio (evaluated)
- Web Storm (used)
- Wing (evaluated)
- WinPython (used)

Development Processes

- Quality (e.g., security, reliability, usability) Assurance (basic)
- User Interface Design Principles (proficient)
- Functional Requirements (expert)
- Listening (proficient, but one can always work on this...)
- System Requirements (basic)
- Nonfunctional Requirements (proficient)
- Tracing Requirements (proficient)
- Software Design Patterns (rudimentary)
- UML Graphical Modeling (expert)
- Software Testing (basic)
- IEEE Software Engineering Standards (proficient)

Drag and Drop Toolboxes

- Cognos (evaluated)
- Informatica (used)
- Oracle Business Intelligence 11g (evaluated)
- Oracle Fusion/Essbase (evaluated)
- RapidMiner (used/taught)
- SAP Business ByDesign (used/taught)
- SAP Business Objects (formal training/used)
- SAP Crystal Reports (used/taught)
- SPSS (used)
- Tableau (evaluated)
- Brio (absorbed by Hyperion then by Oracle) (used)
- Sequitur (defunct or absorbed, not sure)

Russ Robbins
Skills – Detailed List
http://robbinsr.squarespace.com
Contact Me

Languages

- C (very, very long time ago)
- C++ (very long time ago)
- Java (basic)
- JavaScript/CSS/HTML (basic)
- JSON (basic)
- Markdown (basic)
- Pandoc (basic)
- Python (familiar)
- OWL (basic)
- R (proficient)
- RDF (basic)
- Regular Expressions (between basic and proficient)
- Spark (familiar)
- SPARQL (basic)
- SQL (proficient)
- UML (proficient)
- XML (basic)

Machine Learning Fundamentals

- Classification (basic)
- Regression (basic)
- Resampling (rudimentary)
- Model Selection (rudimentary)
- Regularization (rudimentary)
- Non-linear Models (rudimentary)
- Tree-based Methods (rudimentary)
- Support Vector Machines (rudimentary)
- Clustering (rudimentary)

Miscellaneous

- Adobe Illustrator
- Photoshop (used)
- Altova XML Spy (used)
- Apache Jena (Java API>(used)
- Apache Tomcat (Web Server)> (used)
- Bootstrap (Web Framework)> (used)
- Foundation (Web Framework)> (used)

Miscellaneous (continued)

- Google Apps (used)
- IBM Rational Software (used)
- Jack Intelligent Agents (used)
- Jadex Active Components (Java API) (used)
- Microsoft Project (used)
- Microsoft Visio (used)
- SharePoint (used)
- Access (used)
- OpenRDF Sesame (Java API)(used)
- Pellet (lightly evaluated)
- Pure (Web Framework)> (used)
- Research Cyc (lightly evaluated)
- WordNet (used)
- Verbnet (used)
- FrameNe (used)
- PropBank (used)
- LemonUBY (Vocabularies) (used)
- LSEG4 (used)
- Lexical Markup Framework
- isoCAT (Grammars) (used)
- CCAE
- Penn TreeBank
- British National Corpuses (used)
- Windows Movie Maker (used)
- Open Source or Free Alternatives (proficient)

Programming Fundamentals

- Input/Output (proficient)
- Control Flow (proficient)
- Classes (basic)
- Objects (basic)
- Functions (proficient)
- Methods (proficient)
- Dictionaries (between basic and proficient)
- Lists (between basic and proficient)
- Sets (between basic and proficient)
- Graphs (between basic and proficient)
- Iteration (proficient)
- Vectorization (rudimentary)

Programming Fundamentals (continued)

- Recursion (basic)
- Exception Handing (rudimentary)
- Lambda Expressions (basic)
- Map (basic)
- Reduce (basic)
- Closures (rudimentary)
- Search Algorithms (rudimentary)

Programming (Statistics) Toolboxes

- MATLAB (evaluated)
- Octave (used)
- Minitab (used)
- Python (used)
- R (used)
- Rattle (learned)
- Revolution R (used)
- SAS (evaluated)
- Stata (used)

Programming Utilities

- AWS Compute / Storage / Database (using)
- Heroku / Python / NodeJS (using)
- Cygwin (used)
- GitHub (using)
- JSFiddle
- CodePen (used)
- Maven (evaluated)
- Oracle Virtual Box (used)
- Putty (used)
- Regex101, RegexPal, Regexr, Pythex, (using)
- RPubs (using)
- ShinyApps (using)
- Notepad++, Sublime, EmEditor, Nano, Gedit, etc. (using)
- Ubuntu Linux (using)
- Unix (used)
- VMWare (evaluated)

Russ Robbins
Skills – Detailed List
http://robbinsr.squarespace.com
Contact Me

Project Management

- Reference for all items in this section
- Risk Management (basic)
- Quality Management (basic)
- Scope (Requirements) Management (proficient)
- Time (Schedule) Management (basic)
- Cost (Budget) Management (basic)
- Employee Management (basic)
- Vendor Management (basic)
- Customer Management (basic)
- Interface with Software Engineering (expert)

Statistics Fundamentals

- Reference for all items in this section
- Descriptive Statistics (proficient)
- Distributions (basic)
- Probability Theory (proficient)
- Bayes Theorem (basic)
- Hypothesis Testing (between basic and proficient)
- Simple & Multiple Linear Regression (proficient)
- Oneway & Multifactor ANOVA (basic)
- Logistic & Ordinal Regression (proficient)
- Binomial Test (basic)
- Chi-square Contingency Tables (basic)
- Non-parametric Alternatives (proficient)

Python Packages

- Reference for all items in this section
- Beautiful Soup (used)
- Bottle (used)
- Core NLP (evaluated)
- iPython (used)
- MatPlotLib (evaluated)
- NumPy (used)
- Pandas (used)
- PyMongo (used)
- pyR (evaluated)
- PySpark (used)
- Re (used)

Python Packages (continued)

- SciPy (used)
- NLTK (used)
- PyQt (used)

R Libraries

- Reference for all items in this section
- caret (used)
- ggplot2 (used)
- data.table (used)
- doBy (used)
- Hmisc (used)
- knitr (used)
- MASS (used)
- lattice (used)
- leaps (used)
- plyr (used)
- rCharts (used)
- regex (used)
- reshape2 (used)
- rPython (evaluated)

Training Approaches & Tools

- Project-based Learning (expert)
- Problem-based Learning (expert)
- Case Based Learning (expert)
- Collaboration and Learning (expert)
- Technology and Learning (expert)
- Learning Outcomes (proficient)
- Learning Outcomes Assessment (proficient)
- Camtasia
- Audacity
- Adobe Creative Cloud (used)
- WebEx
- GoToMeeting
- Google Hangouts (used)
- PollEverywhere
- Socrative (used)
- Screencast

Training Approaches & Tools (continued)

- YouTube (used)
- Articulate
- Captivate
- Lectora (evaluated)
- WebCT (now part of Blackboard)
- Sakai
- Blackboard
- Moodle (used)
- Simulations
- Virtual Worlds
- Augmented Reality(used)