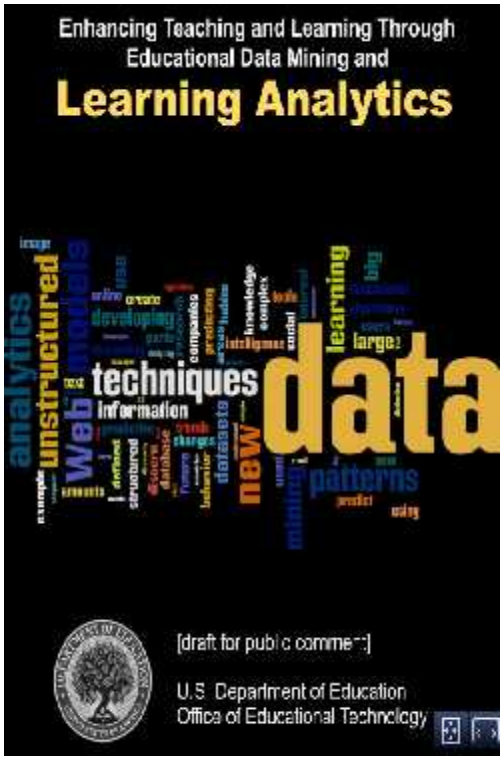
Mining Educational Databases: A Focus on STEM+C Majors for Inclusive Development



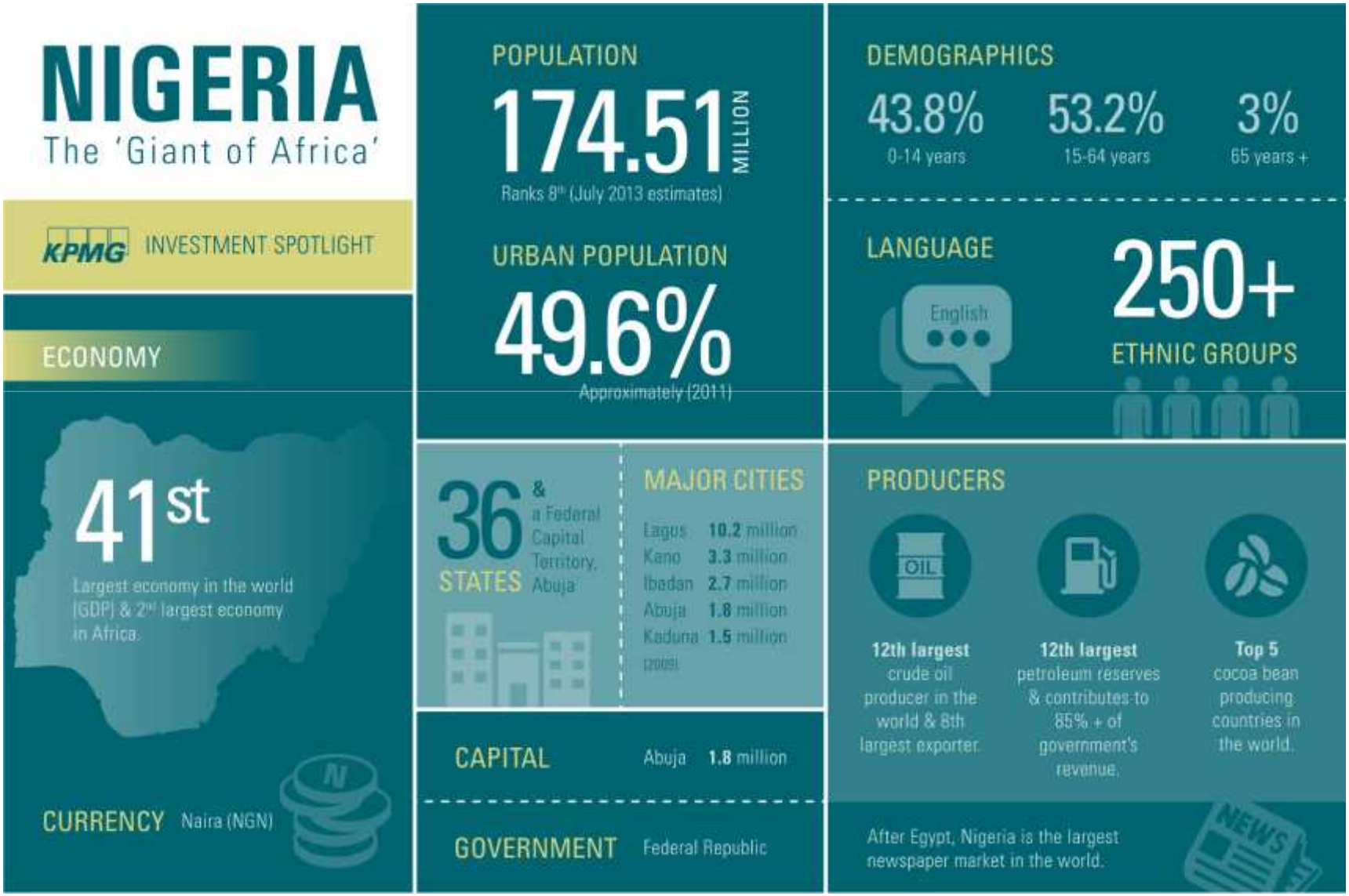
By
Anu A. Gokhale
Professor and Coordinator, Computer Systems Technology
Illinois State University

What is STEM+C?

STEM+C is an acronym coined by the National Science Foundation (NSF) in the U.S. and it stands for:

Science, Technology, Engineering, Mathematics, plus Computing!

Host Country: Interesting Facts



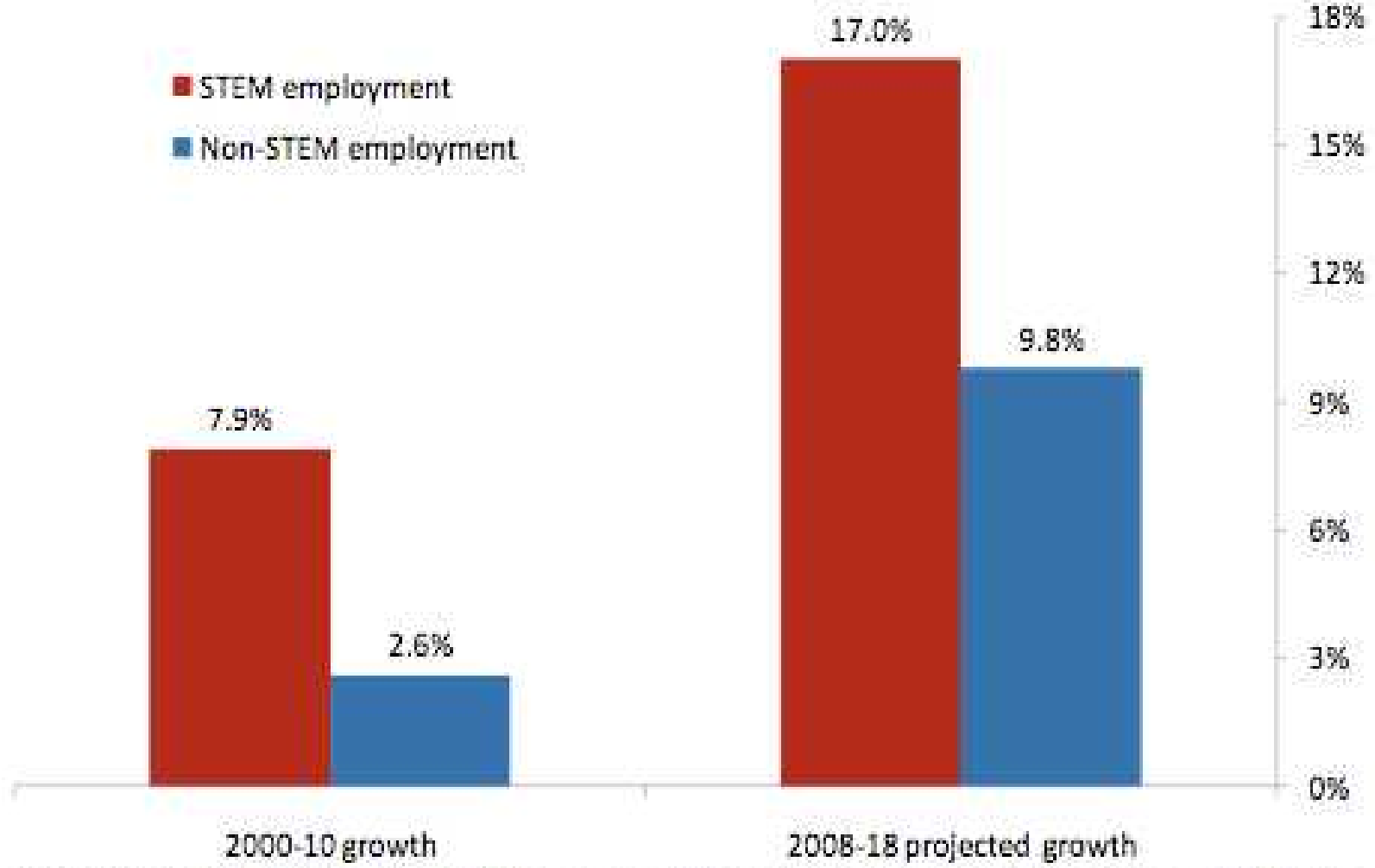
Why Focus on Education?

- The objective is to provide students the cognitive strategies that enable them to think critically, make decisions, and solve problems.

- Business and political leaders are increasingly asking schools to develop skills such as problem solving, critical thinking, communication, collaboration, and self-management — often referred to as "21st century skills."

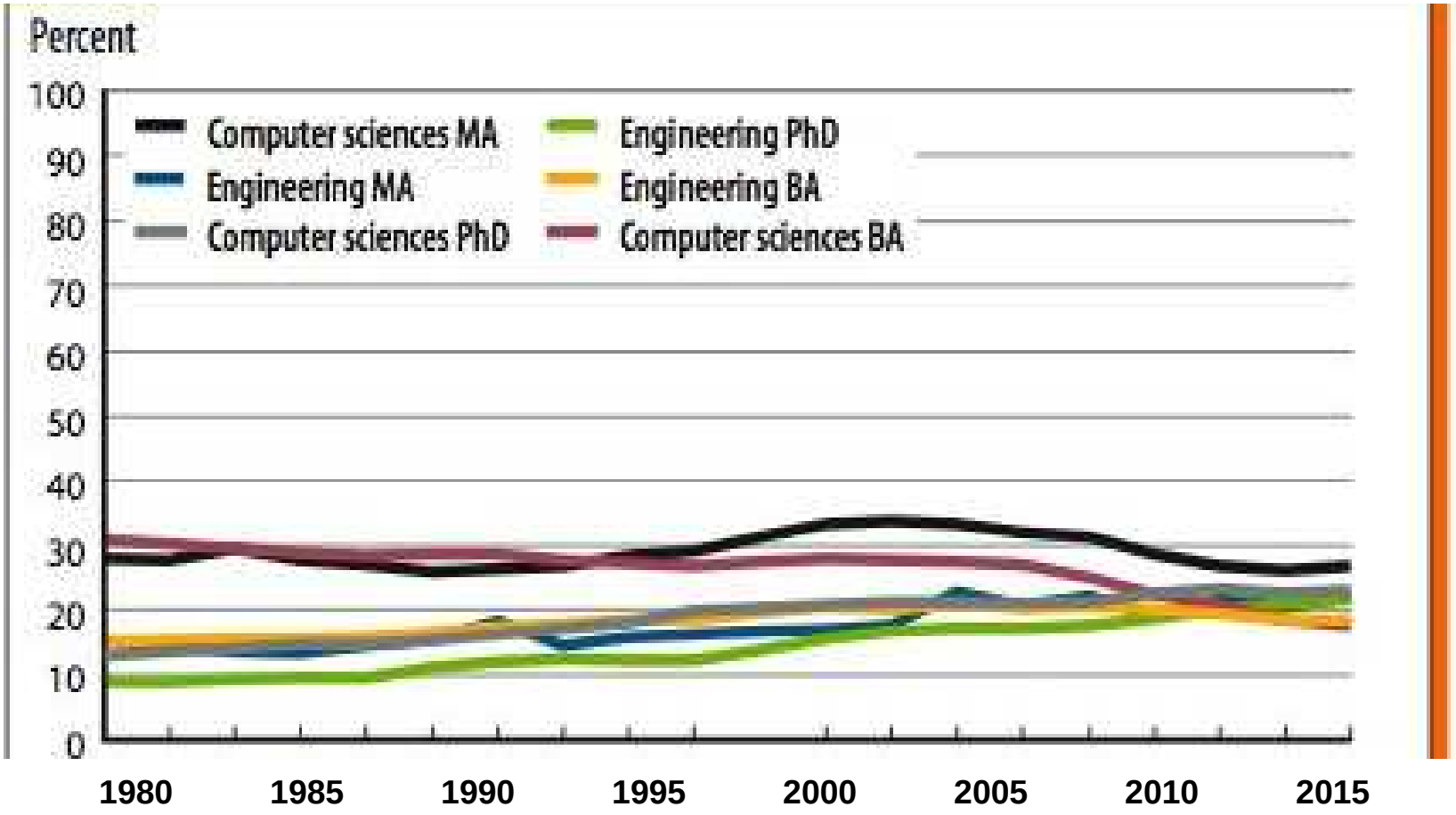
Reference: Pellegrino and Hilton (2013)

Why Should We Care About STEM+C?

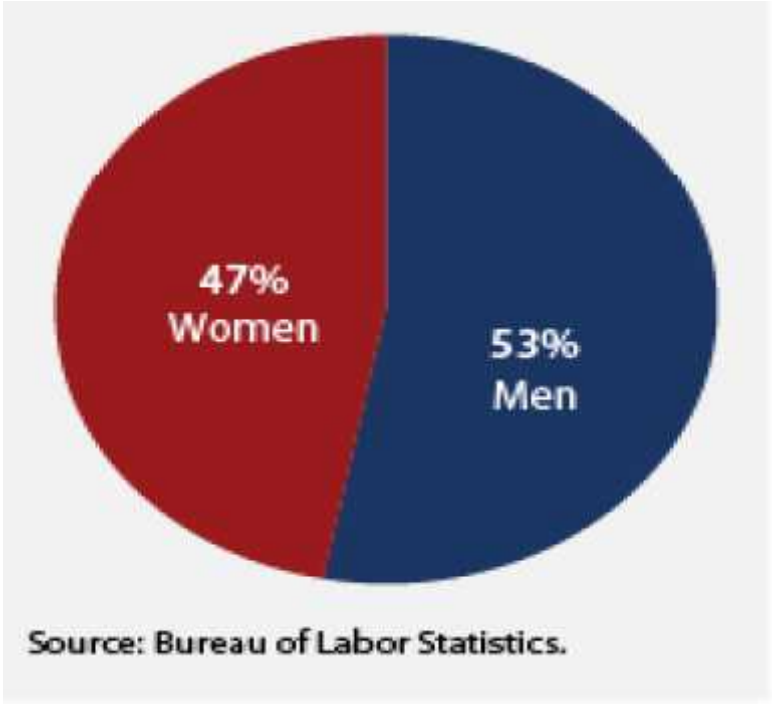


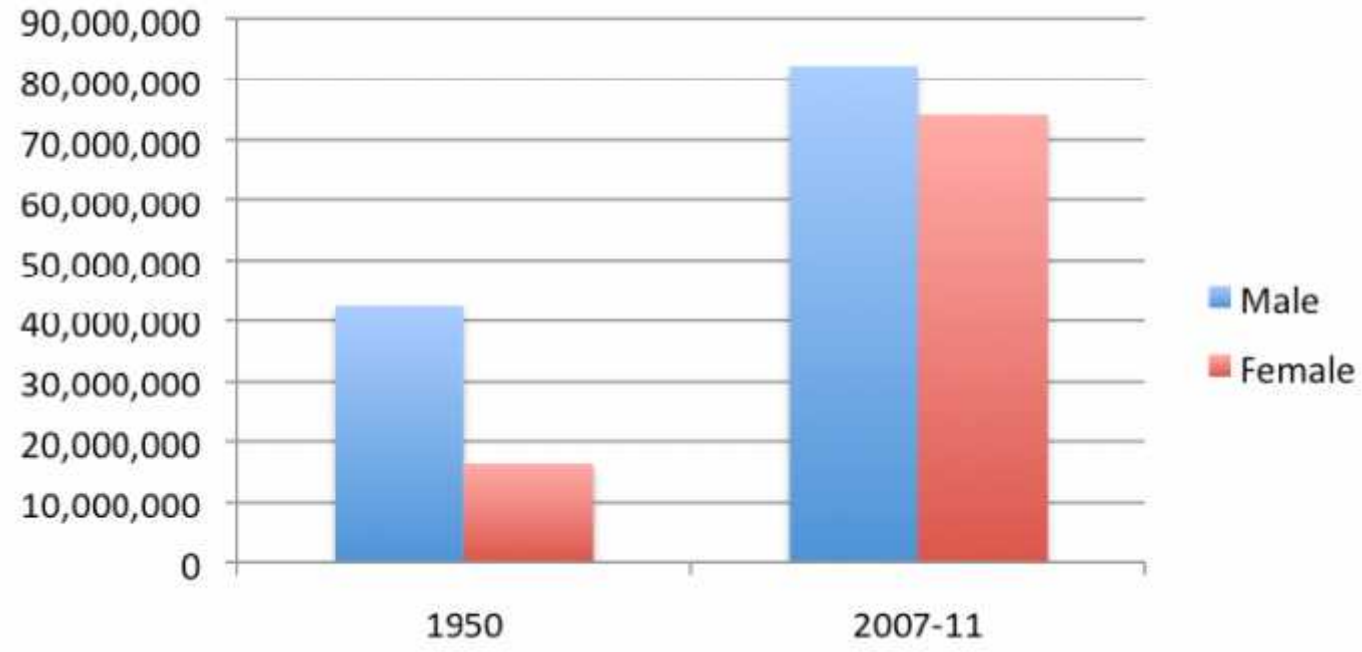
Source: ESA calculations using Current Population Survey public-use microdata and estimates from the Employment Projections Program of the Bureau of Labor Statistics.

STEM+C Degrees: Historical Perspectives



Workforce by Gender

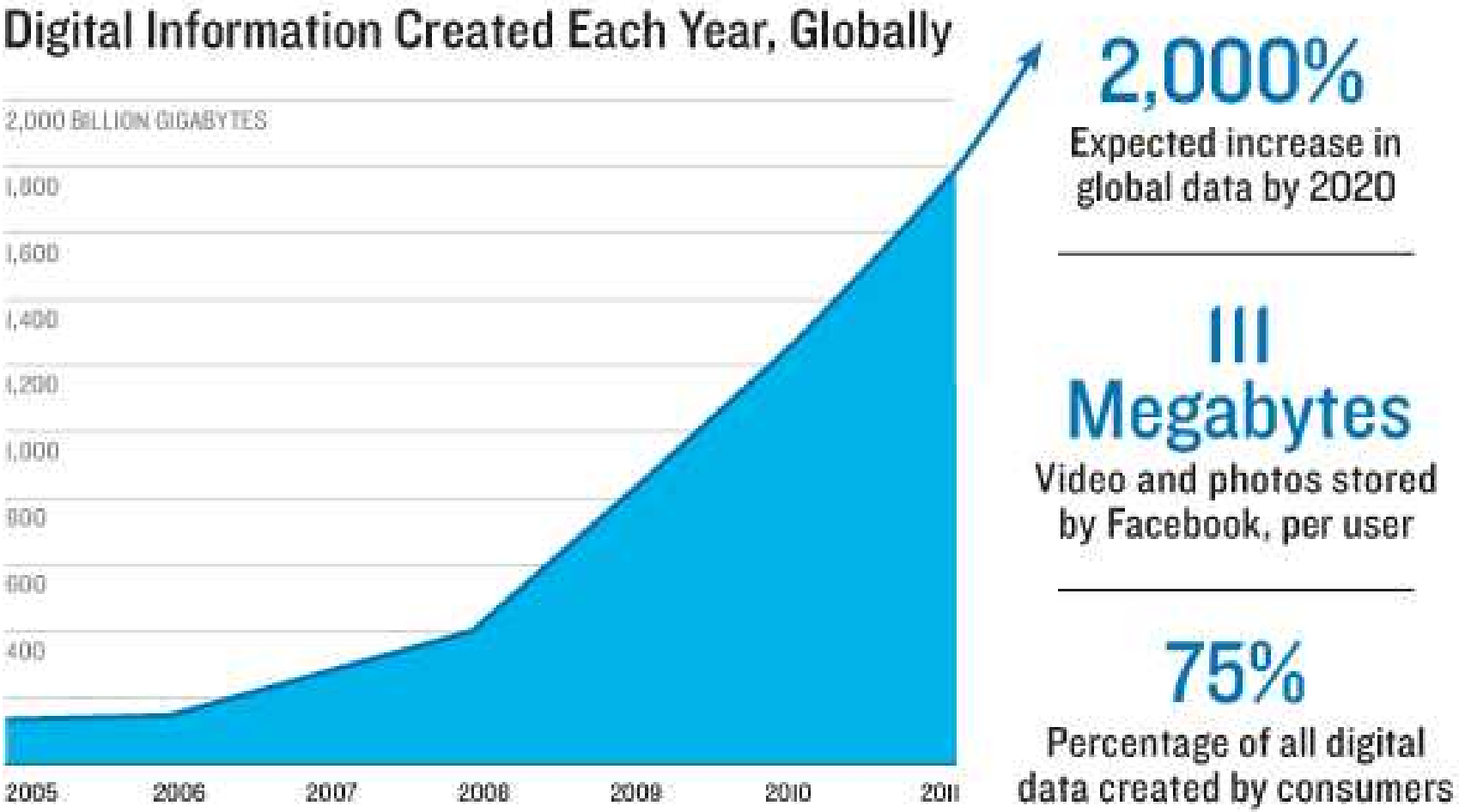




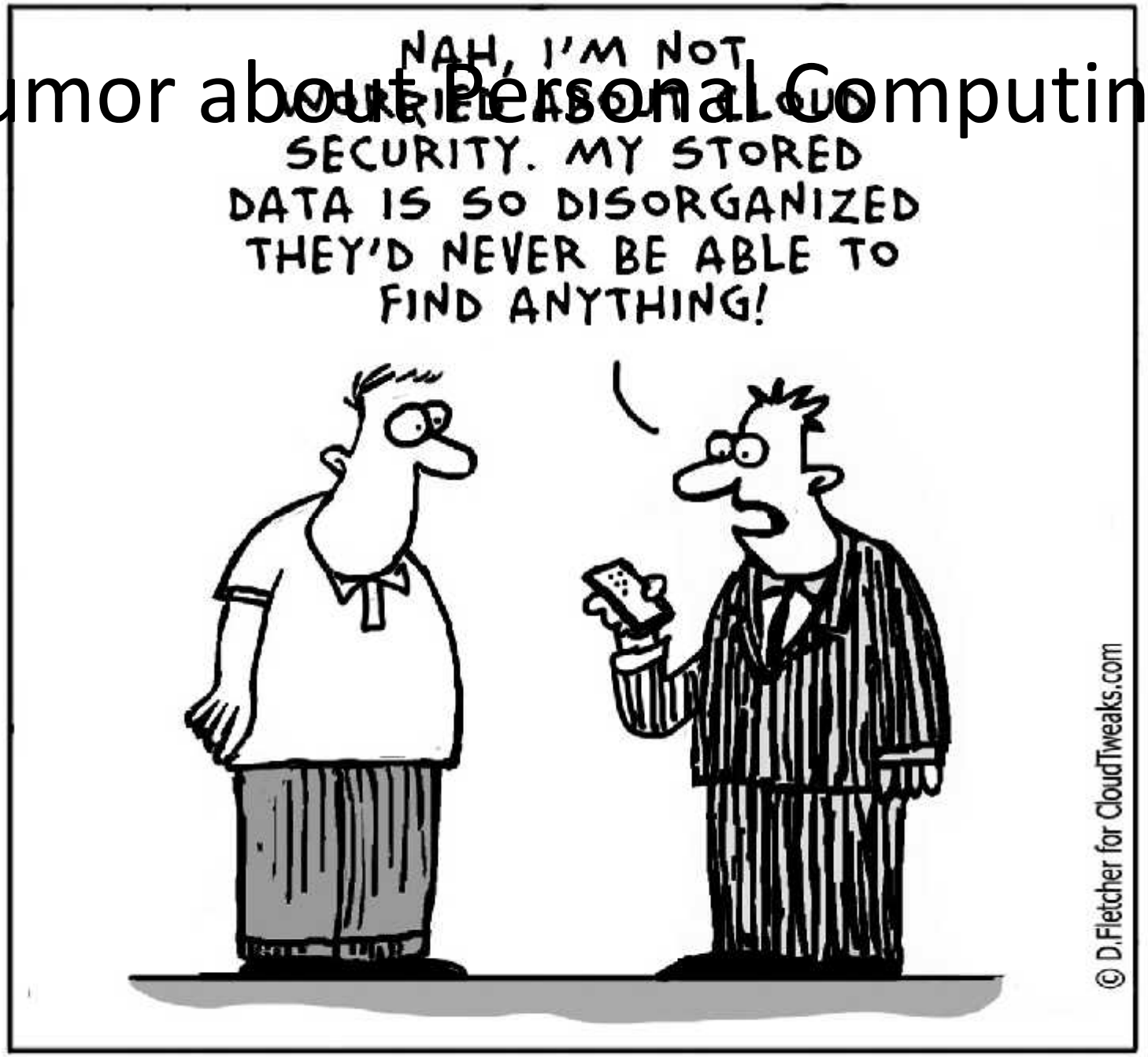
How Do We Recruit and Retain Students in STEM+C Majors and Careers?

- Federal agencies and colleges have **"data inventories" that** could be used for multiple purposes:
  - Adaptive Learning Systems
    - Extensive information about individuals or groups of learners can be analyzed to adapt a learning resource to the learner
  - Assessments Embedded in Learning Systems
    - Non-cognitive affective-domain features, such as interest, persistence, that are recognized as important but that could not be measured reliably and efficiently in the past
  - Linking various types of student data (like family demographics) to academic outcomes

How Much Data is Created Today?



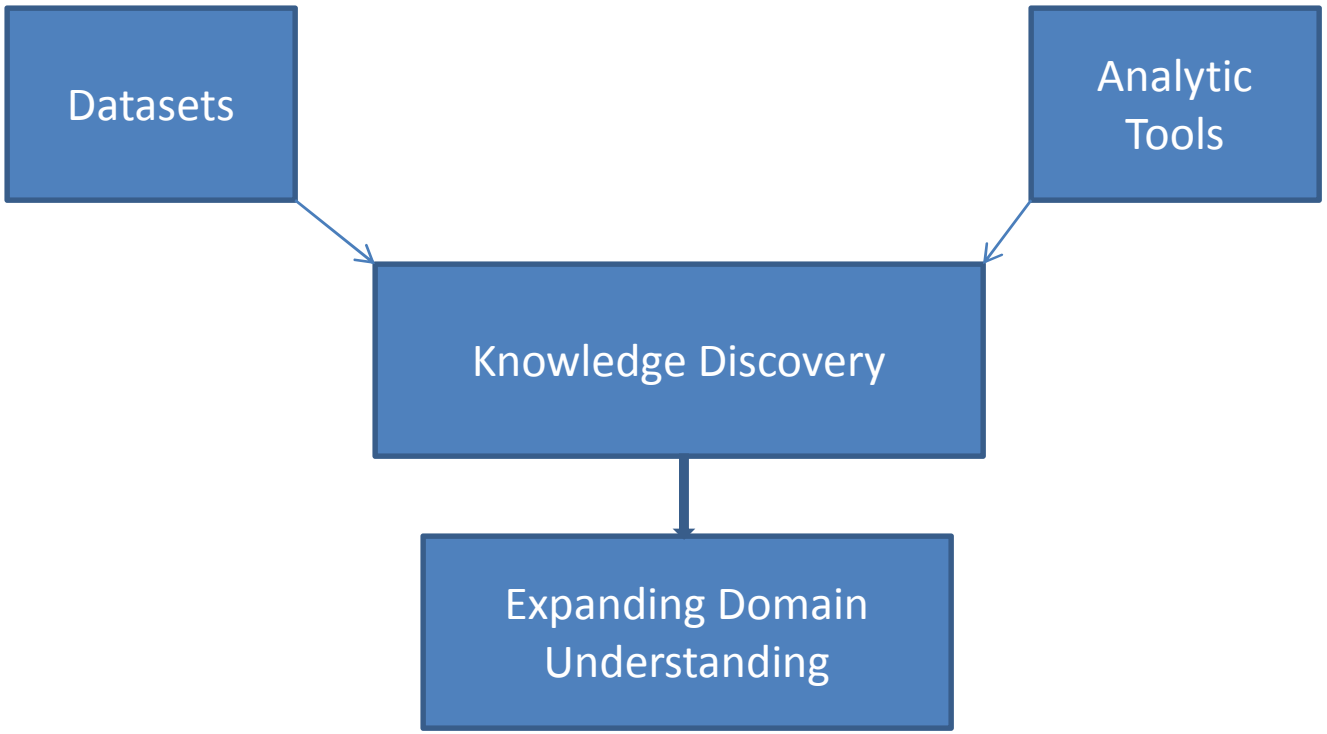
Humor about Personal Computing



What Is Data Mining or KDD? (Knowledge Discovery from Data)

- Data mining, also referred to as KDD, is:
  - The process of finding anomalies, implicit patterns, and correlations within large datasets to predict outcomes, or in other words,
  - The search for relationships and global patterns that exist in large databases but are 'hidden' among the vast amounts of data

Elements of Analyzing Datasets for Knowledge Discovery



Project Goal

- This NSF-funded project is designed to:
  - Use learning communities to recruit and retain a greater number of students in a computing-related major
    - Special emphasis on minorities

    (African-Americans, Hispanics, and Native Americans)
  - Conduct exploratory research, employing KDD techniques
    - The objective is to extract, analyze, and understand information about effectiveness of learning communities from large semi-structured data

Methodology:
Three-Prong Approach

- Online Learning Communities
- Face-to-face Seminars with Professionals
- Faculty Learning Communities

Review of Literature: Why Online Learning Communities?

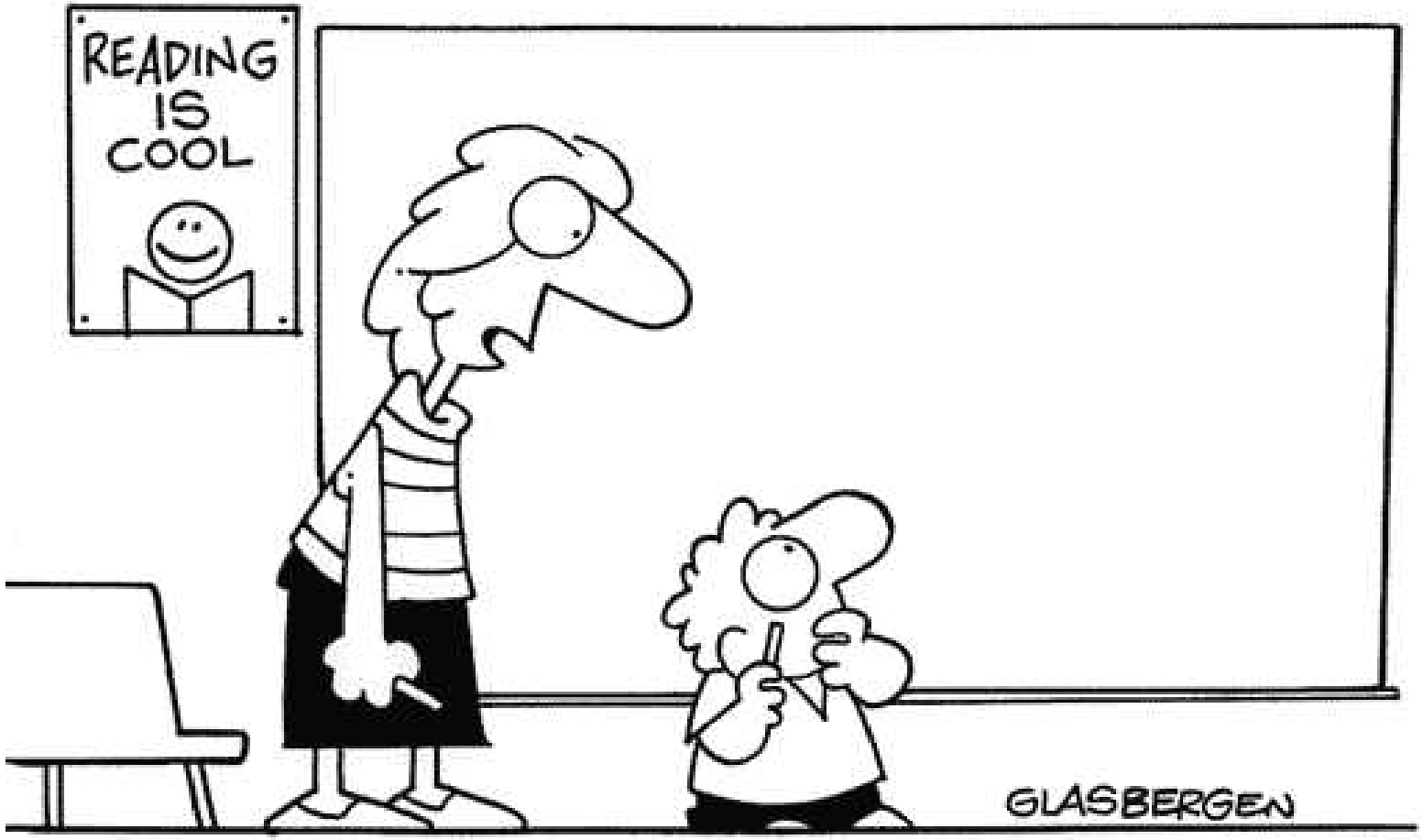
- Principles Underlying Online Learning Communities
  (Ref: Gartner Research; C. Rozwell & D. Morello, Oct. 2011)
  - Are Self-Selecting
  - Are Inclusive
  - Are Efficient; News Travels at Lightning Speed
  - Make an Effective Recruiting and Retention Strategy

- Right Approach to Content:
  (Ref: Gartner Research; J. Lundy, Sept. 2013)
  - Engaging, relevant content for all stakeholders

Who are our Students?

- Most students in colleges or high schools are born between 1980 and 2000
- Members of this age group are often referred to as the Millennials, Net Generation, or Generation Y
- Online is the preferred, or rather the only mode of communication

Some Humor…



"There aren't any icons to click. It's a chalk board."

Benefits of Online Learning Communities: In a Nutshell

- Fits the student culture
- Combines solitary reflection and social interaction
- Can be a powerful promoter of creative and intuitive thinking

Online Learning Communities

- Students (seniors) were selected for blogging
- Four-member blogging/tutor team each semester
  - Blogs posted twice/week during the semester
    - Sunday and Wednesday night
  - Quiz based on the blog earned students extra credit (up to 2 percentage points)
  - Student participants discussed the blog contents and related peripheral issues

Additional Online Activities

- Student Organizations' Chat Rooms
  - IEEE (Institute of Electrical and Electronics Engineers)
  - ACS (Assoc. of Computing Machinery)
  - AITP (Assoc. for Information Technology Professionals)
- Classes-related Online Tutoring
- Networking and Support Structure
  - Mentor Network

Face-to-face Seminars with Professionals

- Professionals in STEM+C fields are invited to visit with students participating in this project
  - Fieldtrips are also arranged to their workplaces
- These after-class informal seminars happen once a week during the semester
- Why?
  - There is no reason to think that students are aware of the work environment and job responsibilities of various STEM+C professionals (FBI Cyber Security, Health Information Systems, Systems Analyst)

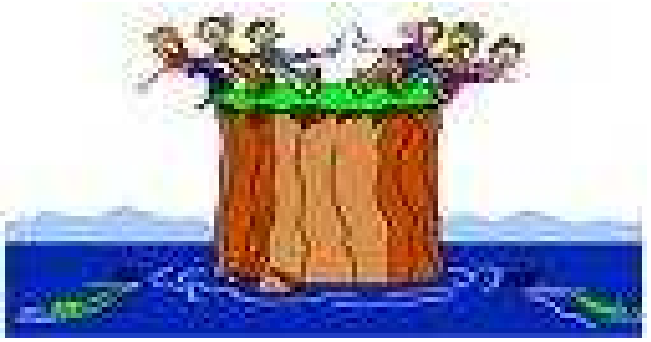
Benefits of Seminars: In a Nutshell

- Goal is to give students a better sense of the range of opportunities in STEM+C careers, and humanize the field

- Professionals were trained by project directors to speak at these meetings, they were asked to:
  - Tell their life story
  - Talk about day-to-day work and the environment
  - Use multimedia to engage students

Faculty Learning Communities

- Teaching style adjustments
  - Focusing on concepts that may not always be familiar to all student groups



- Sensitive to Diversity among Students

Project Activities Over Five Years Generate BIG Data

- Data is of various types
  - Numeric data
    - Number of students who registered in a computing-related course, number of students who declared a computing-related major
  - Non-numeric data
    - Discussion in ‘words’
- Data is unstructured

Big Data Humor…

Benefits & Challenges of Data Generation in Online Learning Systems

- Benefits:
  - One advantage of online learning systems is that they can collect very large amounts of data (big data) from many users quickly
  - As a result, they permit the use of multivariate analytic approaches (analyses of more than one statistical variable at a time) early in the life cycle of an innovation

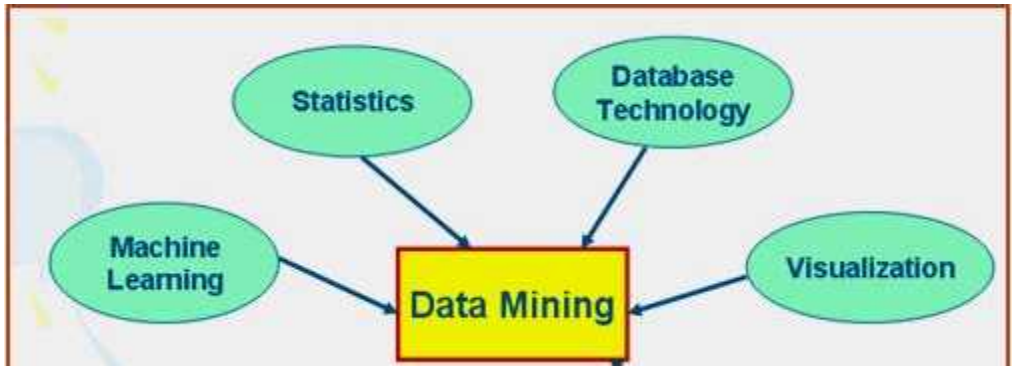
- Challenges:
  - BIG Data

Data Analysis

- KDD framework acknowledges that decisions require different levels of confidence and entail different levels of risk
- Objective is evidence-centered design
- Analytics can uncover patterns of learner behavior that can be used to guide improvement
- A key challenge for the uses of evidence is to identify the relationship between simple user behaviors and complex behavior outcomes

Preparing Data for KDD Analysis

- The process involves:
  - Transferring data into a data warehouse
  - Cleaning the data to remove errors and check for consistency of formats
  - Searching the data using statistical procedures, database queries, visualization, or machine learning methods



KDD Analytic Tools

- The KDD architecture supports a variety of methods.

  Here are some examples:
- Statistical Analysis
  - Used for both summarizing large datasets (i.e., average, min/max, etc.) and in defining models for prediction
  - Most statistical tools prefer to compute over numerical and categorical data organized in columns
- Modeling and Visual Analytics
  - Generates interactive visualizations across many datasets with the hope that users will be able to discover interesting relationships

Some Humor about Data Mining…

Results

- The Eureqa machine learning package was used to uncover trends in data
  - Eureqa is a software tool for automated Machine MAPPing™
    (Modeling, Analyzing, Predicting, Prescribing)

- There is strong agreement between statistical tool (factor analysis) and machine learning tool (Eureqa)

Scoring the Attitude toward IT Scale

- Factor 1 (Interest in Learning about IT)
  - Items: 4, 10, 23, 24, 25, 26, 27, 28
- Factor 2 (IT and the Quality of Life)
  - Items: 6, 9, 15, 16
- Factor 3 (Attitudes Toward Women in IT)
  - Items: 2, 14, 19, 20, 30
- Factor 4 (Opportunities for Women in IT)
  - Items: 1, 7, 12, 13, 21

- Items to be scored negatively:
  - Items: 2, 6, 9, 15, 16, and 20

Reference: Gokhale, A. and Machina, K. "Scale to Measure Attitudes Toward Information Technology." *International Journal of Information and Communication Technology Education*, Volume 9 Issue 3, July 2013, Pages 13-26.

Reliability Coefficients for Factors

- N = 455
- Factor 1: Alpha = 0.8543
- Factor 2: Alpha = 0.6602
- Factor 3: Alpha = 0.7404
- Factor 4: Alpha = 0.7133

Pre-Post ANOVA for Each Factor

- The significant difference occurred for
  - Factor 1: Interest in Learning about STEM
  - Factor 3: Attitudes Toward Women in STEM
  - Factor 4: Opportunities for Women in STEM

Factor	Mean Square	F	Significance
1	1.96	3.834	0.051
2	0.158	0.417	0.519
3	7.947	14.297	0.000
4	1.92	0.319	0.053

Conclusions

- Compared to the Control group, Experimental group students showed:
  - More positive attitudes toward IT
  - More positive attitudes toward women & minorities working in IT
- Online activities are an effective medium to communicate with students, especially the Net-generation
- At the same time, some in-person contact is also important
  - Students liked the seminars where they would ask questions to the professionals who came to the classroom

Humor About Offering _____ as a Service

Further Research…

- The online learning resources continue to collect large amounts of fine-grained data as users interact with them, in real time and over time

- It is critical that we employ sophisticated data mining techniques to decipher information, and verify what's learned


THANK
YOU