MINUTES OF THE STATISTICS SUBCOMMITTEE 83RD MEETING, MAY 26 – 28, 2005

THURSDAY MAY 26, 2005

Present: Bruce Kadonoff (Coquitlam College), Costa Karavas (Vancouver Community College), Kevin Keen (University of Northern British Columbia), Colin Lawrence (British Columbia Institute of Technology), Alex Liu (Kwantlen University College), Yuriko Riesen (Northwest Community College), Lang Wu (University of British Columbia)

Acting Chair: Kevin Keen

Acting Secretary: Alex Liu

1. **Approval of Agenda**

   Motion: That the agenda as proposed be adopted. Moved by Colin Lawrence and seconded by Costa Karavas. **Carried unanimously**.

2. **Approval of Notes and Minutes of the Statistics Subcommittee from the 82nd Articulation Meeting**

   Motion: That the notes and minutes of the Statistics Subcommittee from the 82nd Articulation Meeting as circulated be adopted. Moved by Colin Lawrence and seconded by Costa Karavas. **Carried unanimously**.

3. **Articulation and Transfer Issues for Courses in Statistics**

   As a point of information, the Acting Chair noted three streams in introductory statistics courses: (i) two calculus-based courses (6–8 credit hours) in probability and statistics at the 200 or 300 level; (ii) a single calculus-based course (3–4 credit hours) in probability and statistics at the 200 or 300 level, and; (iii) a single non-calculus course (3–4 credit hours) in statistics at the 100, 200, or 300 level. Non-calculus introductory statistics courses are sometimes also taught by economics, business management, and psychology departments at postsecondary institutions.

4. Articulation between Lower and Upper Division Courses

Because introductory statistics courses that are taught without calculus content can be assigned 100, 200, or 300 level course numbers and those introductory calculus-based statistics courses can be assigned 200 or 300 level course numbers, students with lower level credit can request and receive transfer credit from a lower division course to an upper division course. Discussion revealed no opposition to this practice.

5. Statistics Service Courses Taught by Non-Statisticians

The Acting Chair reported the results of a survey in 2003 of 199 liberal arts colleges in the United States authored by Tom Moore (Grinnell College, Iowa) and Julie Legler (St. Olaf's College, Minnesota). This survey, with a response rate of 67%, showed that only 40% of colleges responding have at least one PhD statistician on faculty with 50% of introductory statistics courses taught outside mathematical science departments. Only 25% of these colleges offered a second course in statistics. Statistical software was reported used in 75% of non-calculus introductory courses, 74% of calculus-based introductory statistics courses, and 55% of probability and statistics two-course sequences. Student projects were reported in 69% of non-calculus introductory courses, 67% of calculus-based introductory statistics courses, and 48% of probability and statistics two-course sequences.

The subcommittee discussed whether these results would also be representative of BC colleges, universities, and university-colleges without reaching any conclusions or making a recommendation for a similar survey in BC. It was noted that limited course offerings in statistics form a hiring barrier for mathematical sciences departments in the colleges and it was generally agreed that quality of instruction is a stronger consideration than whether an instructor's doctorate is in statistics.

Moore, T. & Legler, J. (2005). Survey on statistics within the liberal arts colleges. *Amstat News*, **335**, 14–17.

6. **Content in Statistics Service Courses for Engineering Students**

   Motion: That the Chair write a letter to the Canadian Engineering Accreditation Board to enquire about content required in statistics courses for engineering students. Moved by Colin Lawrence and seconded by Costa Karavas. **Carried unanimously**.

7. **Information Items**

   Colin Lawrence reported that a new offering for the British Columbia Institute of Technology will be the Degree Transfer Program in Science and Technology to begin in September 2005. This new program will include as an option an introductory probability and statistics course worth 5 credit hours.

8. **Recommendations for the Statistical Subcommittee at Future Articulation Meetings**

   Recommendations that were considered to encourage greater attendance by statisticians at future Articulation Meetings included adding scientific sessions and joint, or contiguous, meetings with the Pacific Northwest Statistics Group.

9. **New Business**

   There was no new business.

10. **Motion to Adjourn**

    Colin Lawrence moved to adjourn. **Carried unanimously**.

SATURDAY MAY 27, 2005

9:00 – 9:50

Seminar: Accreditation of Statisticians and Undergraduate Statistics Programs by the Statistical Society of Canada

Presenter Kevin J. Keen, Ph.D., P.Stat.

Kevin changed hats from Acting Chair of the Statistics Subcommittee and UNBC Representative to BCCUPMS to that of a Representative of the Statistical Society of Canada (SSC) to advise BCCUPMS on professional accreditation programs of the SSC. Attendance at the Annual Meeting of BCCUMPS by a Representative from the SSC to advise on professional matters will become a permanent feature and will serve as a further distinction between the Mathematics and Statistics Subcommittees of BCCUPMS. Kevin presented information concerning educational and professional requirements for individuals to be accredited as either an Associate Statistician (A.Stat.) or a Professional Statistician (P.Stat.). Kevin also presented details on the draft requirements and procedures for Universities to apply for professional accreditation of undergraduate programs. (See Annex to these minutes for contents of the 2nd draft document.) Successful graduates of accredited undergraduate programs will be able to follow a simplified procedure when applying for the A.Stat. designation. The P.Stat. designation can be attained after the A.Stat. with 6 more years of education and professional practice supervised by a P.Stat.

SATURDAY MAY 27, 2005

10:00 – 10:50

Roundtable Discussion: How much methodology is enough in an introductory statistics course?

Present: Hongbin Cui (Northern Lights College), Costa Karavas (Vancouver Community College), Kevin Keen (University of Northern British Columbia), David Leeming (University of Victoria), Alex Liu (Kwantlen University College), Lang Wu (University of British Columbia)

Acting Chair: Kevin Keen

Acting Secretary: Alex Liu

1. The temptation in a non-calculus introductory statistics course is to overload the course with topics and methodology to the detriment of the learning experience.

2. The roundtable participants agreed on the following list of topics and methodology for a non-calculus introductory statistics course:

   a) Visualizing data (bar graphs, boxplots, histograms, normal probability plots, scattergrams, residual plots)
   b) Summary statistics (sample mean, sample standard deviation, five-number summary, correlation, and regression)
   c) Probability theory (events, union of events, intersection of events, disjoint events, independence and the multiplication rule, means and variance of a finite random variable, binomial distribution, normal distribution)
   d) Design of sample surveys and experiments (simple random sampling, stratified sampling, completely randomized design, randomization, bias, voluntary response limitations, difference between observational studies and experiments)
   e) Inference and estimation (sampling distribution of a statistic, bias and variation of a statistic, confidence intervals, *P*-values [lower tail, upper tail, two-tailed], level of significance)
   f) Means in one and two populations (to included matched pairs)
   g) Proportions in one and two populations
   h) Simple linear regression
   i) Multiple linear regression

3. It was generally agreed that students could be exposed to but not examined on the probabilities of Type I and Type II errors if time allows.

4. Some instructors noted that they would include the topics of conditional probability and Bayes’s Rule but the majority of participants indicated that these additional topics would not be covered.

5. Some instructors noted that one-way and two-way ANOVA and the randomized block design might be substituted for multiple linear regression.

SATURDAY MAY 27, 2005

11:00 – 11:50

Seminar: Statistical Learning Machines: A New Paradigm

Presenter Kevin J. Keen, Ph.D., P.Stat.

Handouts: See Annex 2 for handouts of Powerpoint slides.

Guidelines for Accrediting Program
Draft 2 –31 July 2004

This draft reflects suggestions made by the Chairs/Directors of Statistical Program who attended the SSC Annual Meeting in Montreal in June 2004.

1. Introduction

There are two levels of qualifications for the profession practice of statistics in Canada – P.Stat. (Professional Statistician) and A.Stat. (Associate Statistician).

The A.Stat. designation is intended to indicate that the holder has completed a course of study equivalent to a major or honours degree in statistics, or in exceptional instances, has otherwise demonstrated an advanced understanding of statistical theory and its application. It is expected that most students who have completed a Masters degree and better undergraduates would be suitably qualified for the A.Stat. designation. An A.Stat. would be regarded as the entry level requirements for a Statistician practicing in Canada under the direction of a P.Stat. or other suitably qualified individual. It is expected that most A.Stat. would work towards obtaining their P.Stat. designation.

The qualification of P.Stat. is intended to indicate that the holder has the necessary academic qualifications, and a minimum of six years of professional experience in the application of statistics.

The educational qualifications for an A.Stat. are outlined in Appendices A and F of the SSC accreditation document and are reproduced in this document for convenience. According to the SSC Accreditation documents, it is planned that an A.Stat. designation will be automatically awarded upon the successful completion of a an accredited program in Statistics in Canada. As of the time of writing this document, there are no accredited programs yet established in Canada.

This document will outline the requirements for a program to be awarded an accredited status, and how accreditation is renewed. While the guidelines below are couched in terms of an academic department at a University or College, usually in the area of Statistics or Mathematics and Statistics, other structures are possible.

2. Applying for Accredited Status for a Program

The application should be submitted to the Accreditation Committee of the SSC electronically (e.g. as a PDF file).

2.1 Demonstration of internal support

An application for accreditation of a course of study is usually sponsored by a Department. The application should have the demonstrated support of at least three faculty within the Department (e.g. cosigning the application), and the support of the more senior administration (e.g. the Dean should also cosign the application). As there is little or no financial costs to accreditation of programs, lack of support by members internal to the Department, or more senior administration, would not be desirable.

2.2 Documentation of program

While the intent is that the A.Stat. educational requirements are equivalent to a major degree in statistics, it is not necessary that only institutions offering such degrees be eligible for accreditation, nor must the accredited program be exactly the same as a Department's degree program. This allows flexibility for Departments that are too small to offer a specialized program in Statistics, or for students who have different educational paths.

As noted below, the education requirements for an A.Stat. have been grouped into modules. A module may or may not be equivalent to a particular course offered by a Department. The Department will ensure that their documentation makes it clear which module is covered by which course. For each module (particularly for the statistical and probability modules), a copy of the official detailed course outline showing topics covered in the course and the textbook used, some sample assignments, some sample term tests, and a sample final exam should be submitted.

2.3 Length of accreditation

A successfully accredited program shall maintain its status for five years from the date that accreditation was awarded by the SSC.

At the end of the five year period, a new application should be submitted in full. The resubmission will help Departments avoid “drift” in the program and its standards and to update their program as courses change over time.

2.4 Revoking of accredited status

The Board of the SSC (upon recommendation from the Accreditation Appeals Committee) may revoke accredited status at any time. Normally, the Department involved would be invited to make a submission to the Board before such a decision is made.

3. Standards for an accredited program

The standards for an accredited program have been broken into a number of areas corresponding to mathematical prerequisites; statistical methodology; computer skills; communication skills; and substantive knowledge in an application area.

Within each area, a number of modules have been identified. It is envisioned that approximately 18-20 courses would be necessary to fulfill the requirements for an A.Stat. of which approximately half would be in Statistics with the remainder providing the mathematical, computation, communication, and breadth requirements. A course is defined as approximately 30 hours of instruction, e.g. a standard thirteen week course that meets three times a week for 50 minutes.

Each university should specify a minimum standard (both a minimum grade in each course, and a minimum GPA to be computed over at least 10 courses) that will be sufficient for an A.Stat designation. Individual departments are free to set higher standards (e.g. more courses, minimum grades in courses, or a GPA requirement that exceeds their university graduation requirements). Each Departments standards will be reviewed by the Accreditation Committee who may recommend changes to the Department for consideration.

The textbooks listed are exemplary only to indicate the expected level of instruction, and do not constitute an endorsement by the SSC nor are Departments obligated to use these textbooks.

3.1 Mathematical modules – approximately 4 courses

(a) Calculus I
(b) Calculus II
(c) Calculus III
These modules should cover the standard topics in differentiation, single variable integration , and multivariable integration. These are to a great extent standard topics offered in introductory calculus courses and so little detail is provided here.

(d) Linear Algebra
This modules should cover matrix manipulations, vector spaces, singular values, eigenvalues and eigenvectors. These topics are to a great extent standard topics in an introductory linear algebra course.

3.2 Statistical and probability modules

Note that some of these course may require additional introductory courses in statistics and probability before completion which are usually not counted towards completion of the A.Stat. requirements.

The following modules can usually be covered in approximately 8 courses. The first five modules are core, i.e. all A.Stat. applicants should have completed these topics, while the last set of modules is elective.

3.2.1 Mathematical statistics modules – approximately 2 courses

(a) Distributional theory (moments, transformations, moment generating functions)
(b) Basic distributions (normal, t, chi-square, F, exponential, weibull, uniform, etc)
(c) Relationships among basic distributions.
(d) Basic theory of estimation; sufficiency; method of moments; maximum likelihood estimation; basic Bayes estimation; confidence intervals; credible intervals; prediction intervals
(e) Basic theory of hypothesis testing; likelihood ratio tests; chi-square tests;
(f) Basic probability theory; convergence types;

These modules should cover the majority of the topics in books such a Hogg and Craig (Introduction to Mathematical Statistics) or Mood, Bose, and Graybill (Introduction to the Theory of Statistics)

3.2.2 Linear Regression module – approximately 1 course

(a) Single variable regression;
(b) Multiple regression using matrix notation; diagnostics;
(c) Model selection; forwards, backwards, stepwise, Cp, AIC, etc.

This is a standard course in regression methods as covered in Netter and Wasserman (Applied Linear Models).

3.2.3 Design and analysis of experiments module – approximately 1 course

(a) Completely randomized designs;
(b) Complete block designs;
(c) Split-plot designs;
(d) Fractional factorial designs;
(e) Response surface designs;

These topics should also discuss sample size determination and power determination. There should be practice in both DESIGN and ANALYSIS of experiments. These are standard topics covered in books such as Montgomery (Design and Analysis of Experiments)

3.2.4. Survey sampling module – approximately 1 course

(a) Simple random samples;
(b) Systematic samples;
(c) Cluster samples;
(d) Two stage samples;

These topics should cover stratification; ratio and regression estimation; domain estimation; estimates of means, total, proportions, and ratios. These are standard topics in books such as Lohr (Survey Sampling)

3.2.5. Other modules – approximately 4 courses

The applicant should complete an additional 4 course that can incorporate a wide variety of topics. Some of the potential topics are listed below – this list is exemplary rather than exhaustive – Departments can use other topics with approval from the Accreditation Committee.

(a). Generalized linear models

Logistic regression; log-linear models; contingency tables

(b). Modern computational methods
Bootstrapping; jackknifing, and other resampling methods

(c) Computational Bayesian methods

(d) Generalized estimating equations

(e) Survival analysis

(f) Data mining

(g) Statistical consulting

(h) Time series

(i) Multivariate methods

(j) Non-parametric methods

(k) Quality control

(l) Data analysis/capstone course
In this course students should take an integrative approach to data analysis using such topics as visualization, model building, model validation, etc.

(m) Econometrics

(n) Actuarial Science courses

(o) Categorical data analysis

3.3 Computer skills - (approximately 2 courses)

Students should be able to use the standard productivity tools, be able to use common statistical packages, and should also be able to program non-standard analyses. Many programs integrate these topics through out the undergraduate experience without formal courses in productivity tools or statistical packages.

(a).Productivity tools – word processors; spreadsheets; drawing programs; web usage

(b). Statistical packages
Students should have experience in at least one of S-Plus, R, SAS, SPSS, etc.

(c). Formal computer language

Students should have a basic understanding of programming at the base level using a language such as FORTRAN, C, Basic, Matlab, S-Plus, R, or similar languages.

3.4 Communication Skills – approximately 1 course

(a) Written and oral communication

In some programs, student may take specialized courses in these areas. In other courses, these skills may be integrated into the program over a broad array of courses. For example, some courses in a program may be designated as writing intensive courses. The student should receive substantial feedback to help develop their communication skills.

3.5 Substantive Area – approximately 4 courses

The student should develop expertise in a substantive area other than statistics. In many programs, this would be obtained by a minor in another area consisting of four courses after the first year. A “minor” in mathematics is acceptable.

Appendix A

Educational Guidelines for Accrediting Statisticians

These Educational Guidelines will serve as the non-binding basis for awarding the A.Stat. (Associate Statistician) designation. They are also part of the requirements for receiving the P.Stat. (Professional Statistician) designation.

An A.Stat. should have the equivalent of at least a major or honours degree in Statistics, or in exceptional instances, have otherwise demonstrated an advanced understanding of statistical theory and its application (see Appendix B). Substantial work in developing curriculum guidelines for such programs is underway in the American Statistical Association. Their general guidelines for an undergraduate program in statistics are available on their web site and presented in Appendix F.

In particular: “Effective statisticians at any level display a combination of skills that are not exclusively mathematical. Programs should provide some background in these areas:

> * Statistical: Graduates should have training and experience in statistical reasoning, in designing studies (including practical aspects), in exploratory analysis of data by graphical and other means, and in a variety of formal inference procedures.
> * Mathematical: Undergraduate major programs should include study of probability and statistical theory along with the prerequisite mathematics, especially calculus and linear algebra....
> * Computational: Working with data requires more than basic
> computing skills. Programs should require familiarity with a standard statistical software package and should encourage study of data management and algorithmic problem solving.
> * Nonmathematical: Graduates should be expected to write clearly, to speak fluently, and to have developed skills in collaboration and teamwork and in organizing and managing projects.....
> * Substantive area: Because statistics is a methodological discipline, statistics programs should include some depth in an area of application.”

The Accreditation Committee recommends that applicants who are not from accredited programs (Appendix E) review the list of core topics below. In creating this list, the Committee is mindful of the observation by Moore (2001, P.5) that with “diminished expectations: we cannot teach a wide audience what we might like to ‘ cover’....Niss warned against the ‘dreaded disease syllabitis’ that assesses a course or programme by the length of list of topics”. Bryce et al. (2001) and Ritter et al. (2001) also discuss the undergraduate curriculum for a degree in Statistics.

Some of the topics appear to be graduate level material (e.g., survival analysis, data mining, or neural nets). The decision to include them required careful thought. The Committee agrees with Ritter et al. (2001) “that no student could have studied all the topics....nor could realistic undergraduate programs be constructed to cover every topic....what most employers want are bright individuals who have a good core knowledge of statistics, good computing capability, and good people skills.” At the same time, the Committee is mindful of another comment by Moore (2001), who states that “no undergraduate programme is intended to train professional statisticians. For better or worse, statisticians are defined as having at least a master’s degree or equivalent experience. Holders of a bachelor’s degree may eventually reach this status via on the job training and practical experience, but their degree does not equip them for professional practice.” Too many employers think that an honours degree will do as long as the person can run a statistical package without supervision by higher level personnel.

1. Mathematical Background
* single and multivariable calculus (integration and differentiation)
* linear algebra
* matrix algebra
* linear systems of equations
* eigenvalues/eigenvectors, singular value decomposition

2. Statistical Background
* probability theory and stochastic processes
* distributional theory (e.g., relationships among the standard distributions)
* estimation and hypothesis testing theory
* foundations (sufficiency, etc...)
* methods of moments
* maximum likelihood
* general estimating equations
* Bayesian methods
* core methodology
* data visualization and exploration
* single/multiple/logistic regression
* chi square and generalized linear models
* design and analysis of experiments
* single and multifactor designs
* crd, rcb, split plot, repeated measures, fractionation
* design and analysis of surveys
* srs, cluster, multistage sampling designs
* variance reduction: stratification, ratio, regression
* bootstrapping and jackknifing

3. Computational skills
* basic programming skills with procedural languages
* using statistical packages effectively
* databases and data management
* simulation and modelling
* data transfers between different formats (e.g., Excel > SAS > ACCESS)

4. Communication skills
* effective technical writing and presentations
* teamwork and collaboration

5. Specialization (depending upon area of expertise). Some examples are:
Industry/Manufacturing/Engineering
* quality/process control, time series, reliability
* neural nets

Medical
* survival analysis, categorical data analysis
* generalized estimating equations

Business and Management
* multivariate analysis, time series, quality/process control
* data mining

Government
* multivariate analysis, privacy issues, advanced survey sampling

Biology/Ecology
* capture/recapture, Taylor's power law
* principal components, multivariate analysis methods
* randomization tests

Social Sciences
* factor analysis, principal components, survey instrument design

Bryce, G.R, Gould, R., Notz, W.L., and Peck, R.L. (2001). Curriculum Guidelines for Bachelor of Science Degrees in Statistical Science, American Statistician, 55, 7-13.

Moore, D.S. (2001). Undergraduate Programs and the Future of Academic Statistics, American Statistician, 55, 1-6.
Ritter, M.A., Starbuck, R.R. and Hogg, R.V. (2001). Advice from Prospective Employers on Training BS Statisticians, American Statistician, 7, 14- 18.

Appendix E

Accrediting Educational Programs

Institutions (universities, colleges, and others) will submit “programs” for consideration by the Accreditation Committee, and approval by the Board of the SSC. Students who successfully complete accredited programs with a specified level of performance would automatically receive the A.Stat. designation.

The package brought for approval will include detailed course outlines, sample assignments, sample examinations, and a written statement on how the program meets the educational guidelines. For an initial submission, there should be an indication of the length of time that the program has been operative. Accredited programs will be reviewed every five years.

Programs proposed for accreditation should follow the American Statistical Association guidelines on undergraduate programs in statistical science, as given in Appendix F.

Appendix F

American Statistical Association

Curriculum Guidelines for Undergraduate Programs in Statistical Science (quotes are used below inthe skills needed area to indicate minor modifications from: http://www.amstat.org/education/Curriculum_Guidelines.html).

The American Statistical Association endorses the value of undergraduate programs in statistical science, both for statistical science majors and for students in other majors seeking a minor or concentration. This document provides guidelines for development of curricula for such programs.

Principles

Undergraduate programs in statistics are intended to equip students with quantitative skills that they can employ and build on in flexible ways. Some students will plan graduate work in statistics or other fields, while others will seek employment after their first degree. Programs should be sufficiently flexible to accommodate varying goals. Undergraduate programs are not intended to train professional statisticians, though some graduates may reach this level through work experience and/or further study.

Institutions vary greatly in the type and intensity of programs they are able to offer. The ASA believes that almost all institutions can provide a level of statistical education that is useful to both students and employers. We encourage flexibility in adapting these guidelines to institutional constraints. In many cases, statistics min ors or concentrations for quantitatively oriented students in fields such as biology, business, and behavioral and social science may be more feasible than a full statistics major.

Undergraduate statistics programs should emphasize concepts and tools for working with data and provide experience in designing data collection and in analyzing real data that go beyond the content of a first course in statistical methods. The detailed statistical content may vary, and may be accompanied by varying levels of study in computing, mathematics, and a field of application.

Though statistics requires mathematics for the development of its underlying theory, statistics is distinct from mathematics and uses many nonmathematical skills; thus, the curriculum must be more than a sequence of mathematics courses. It is essential that faculty trained in statistics and experienced in working with data be involved in developing statistics programs and in teaching or supervising courses required by the programs.

Skills Needed

Effective statisticians at any level display a combination of skills that are not exclusively mathematical. Programs should provide some background in these areas:
* Statistical Graduates should have training and experience in statistical reasoning, in designing studies (including practical aspects), in exploratory analysis of data by graphical and other means, and in a variety of formal inference procedures “at both univariate and multivariate levels”.
* Mathematical Undergraduate major programs should include study of probability and statistical theory along with the prerequisite mathematics, especially calculus and linear algebra. Programs for non-majors may require less study of mathematics. Programs preparing for graduate work may require additional mathematics.
* Computational Working with data requires more than basic computing skills. Programs should require familiarity with a standard statistical software package and should encourage study of data management and algorithmic problem-solving.
* Nonmathematical Graduates should be expected to write clearly, to speak fluently, and to have developed skills in collaboration and teamwork and in organizing and managing projects. Academic programs often fail to offer adequate preparation in these are as.
* Substantive area Because statistics is a methodological

discipline, statistics programs should include some depth in an area of application “and integration of statistical principles in an applied context”.

Curriculum Topics for Undergraduate Degrees in Statistical Science
The approach to teaching the following topics should:
* Emphasize real data and authentic applications.
* Present data in a context that is both meaningful to students and indicative of the science behind the data.
* Include experience with statistical computing.
* Encourage synthesis of theory, methods, and applications.
* Offer frequent opportunities to develop communication skills.

Statistical Topics:
* Statistical theory (e.g., distributions of random variables, point and interval estimation, hypothesis testing, Bayesian methods).
* Graphical data analysis methods.
* Statistical modelling (e.g., simple, multiple, and logistic regression; categorical data; diagnostics; data mining).
* Design of studies (e.g., random assignment, replication, blocking, analysis of variance, fixed and random effects, diagnostics in experiments; random sampling, stratification in sample surveys; data exploration in observational studies).

Mathematical Topics:
* Calculus (integration and differentiation) through multivariable calculus.
* Applied linear algebra (emphasis on matrix manipulations, linear transformations, projections in Euclidean space, eigenvalue/eigenvector decomposition and singular value decomposition).

Probability:
* Emphasis on connections between concepts and their applications in statistics.

Computational Topics:
* Programming concepts; database concepts and technology.
* Professional statistical software appropriate to a variety of tasks.

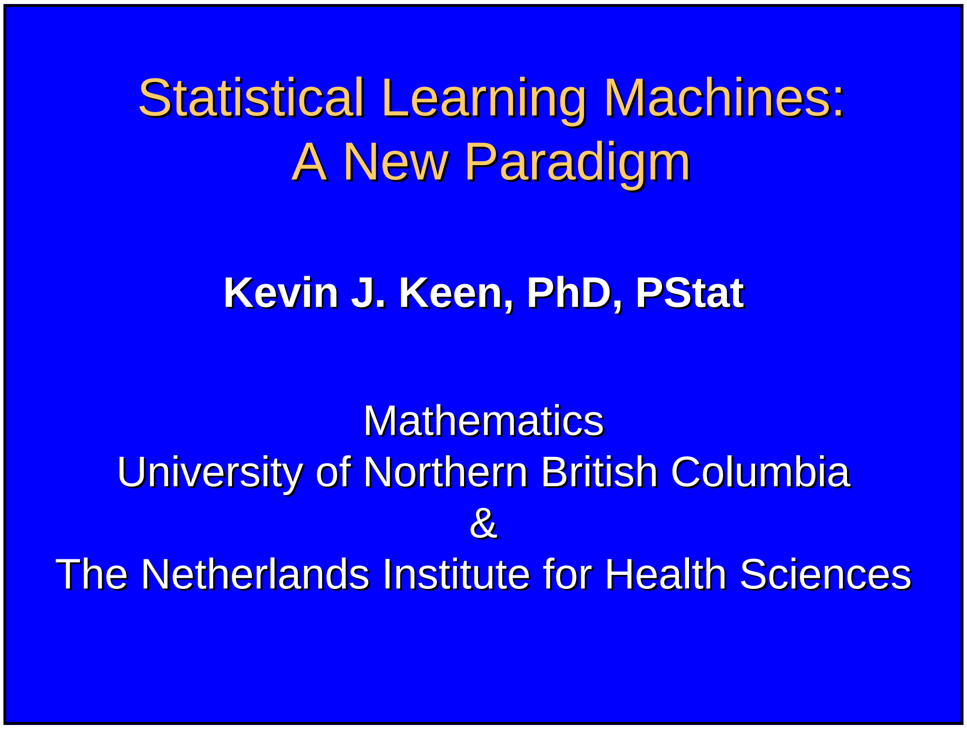
Non-mathematical Topics:
* Effective technical writing and presentations.
* Teamwork and collaboration.
* Planning for data collection.
* Data management.

Electives:
There are many electives that might be included in a statistics major. Since resources will vary among institutions, the identification of what will be offered is left to the discretion of individual units.
Practice:

When possible, the undergraduate experience should include an internship, a senior-level “capstone” course, a consulting experience of some kind, or a combination of these. These and other opportunities to practice statistics should be included in a variety of venues in an undergraduate program.

Handouts of Powerpoint slides for seminar on

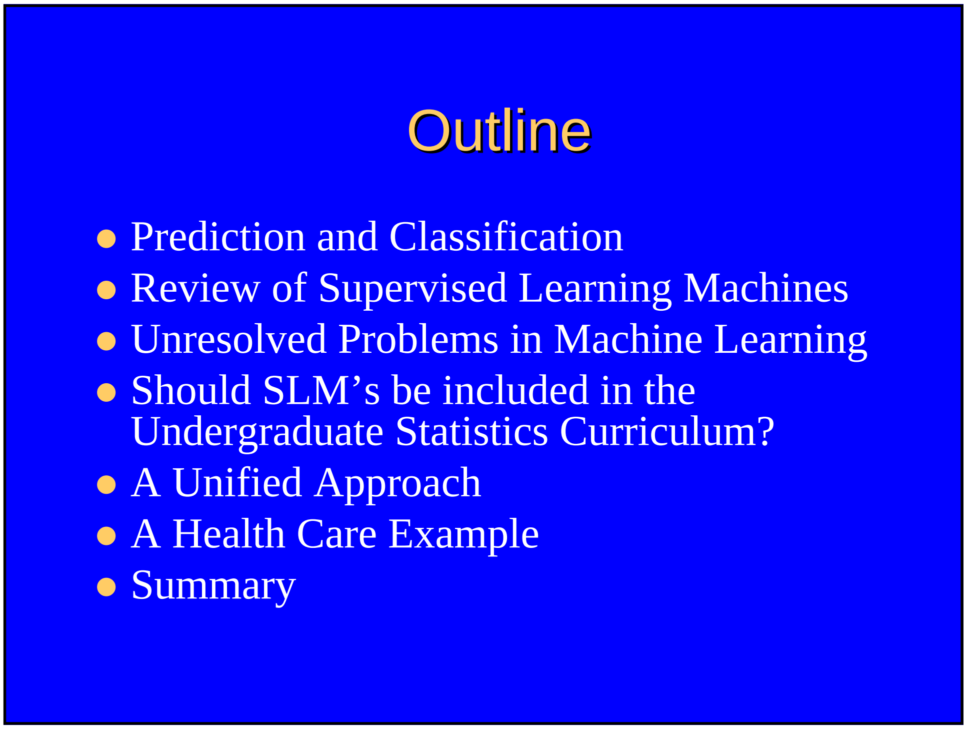
Statistical Learning Machines: A New Paradigm



Statistical Learning Machines: A New Paradigm

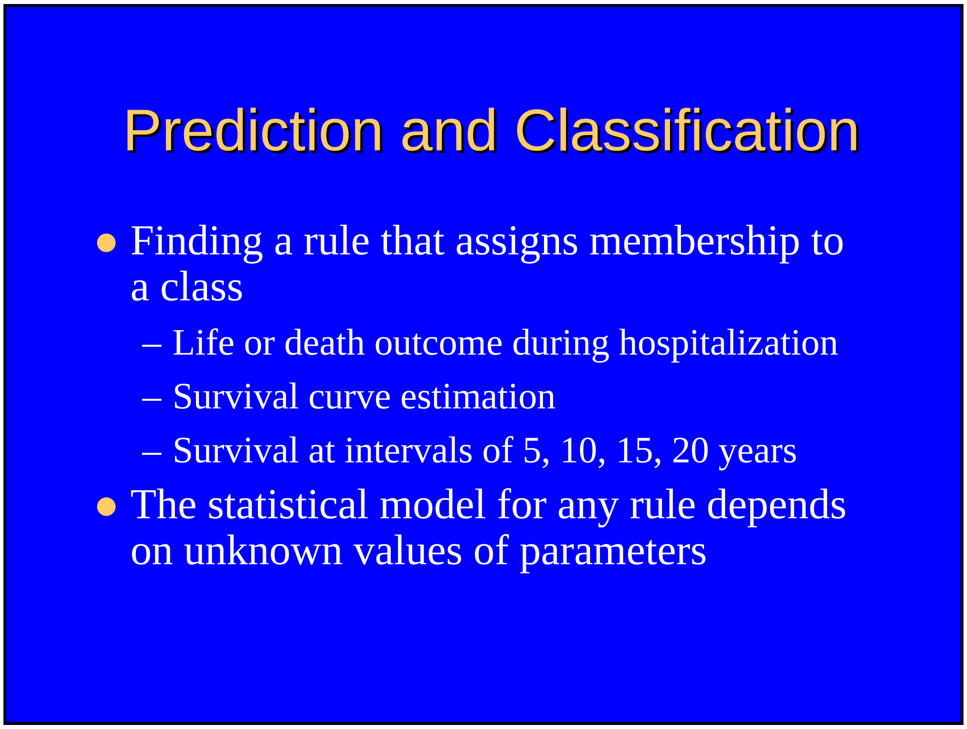
Kevin J. Keen, PhD, PStat

Mathematics
University of Northern British Columbia
&
The Netherlands Institute for Health Sciences



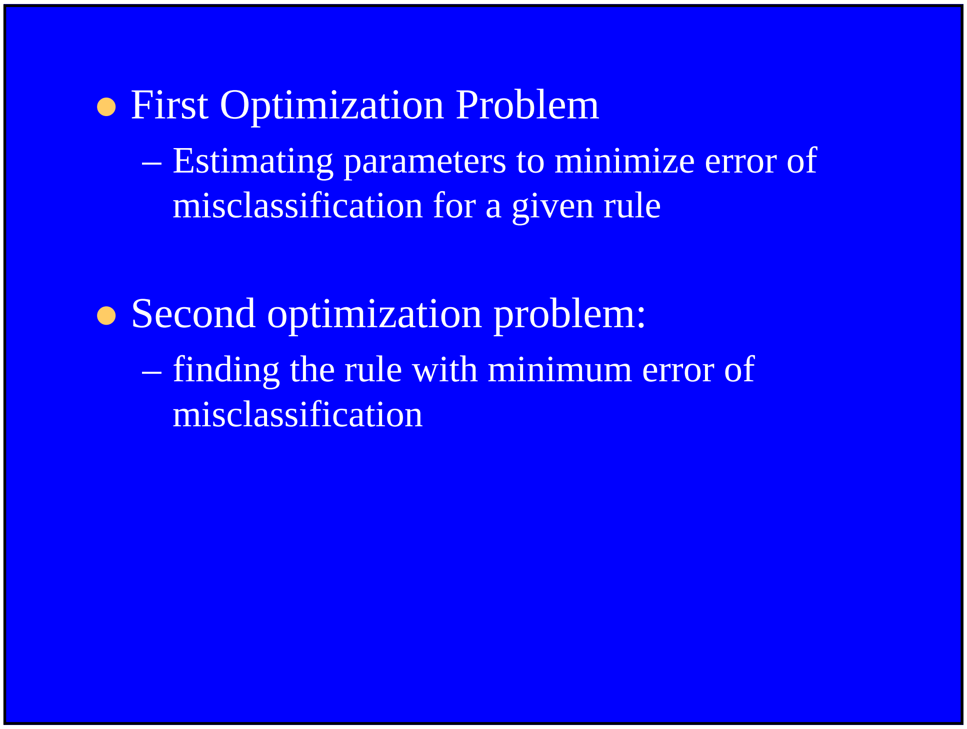
Outline

- Prediction and Classification
- Review of Supervised Learning Machines
- Unresolved Problems in Machine Learning
- Should SLM’s be included in the Undergraduate Statistics Curriculum?
- A Unified Approach
- A Health Care Example
- Summary

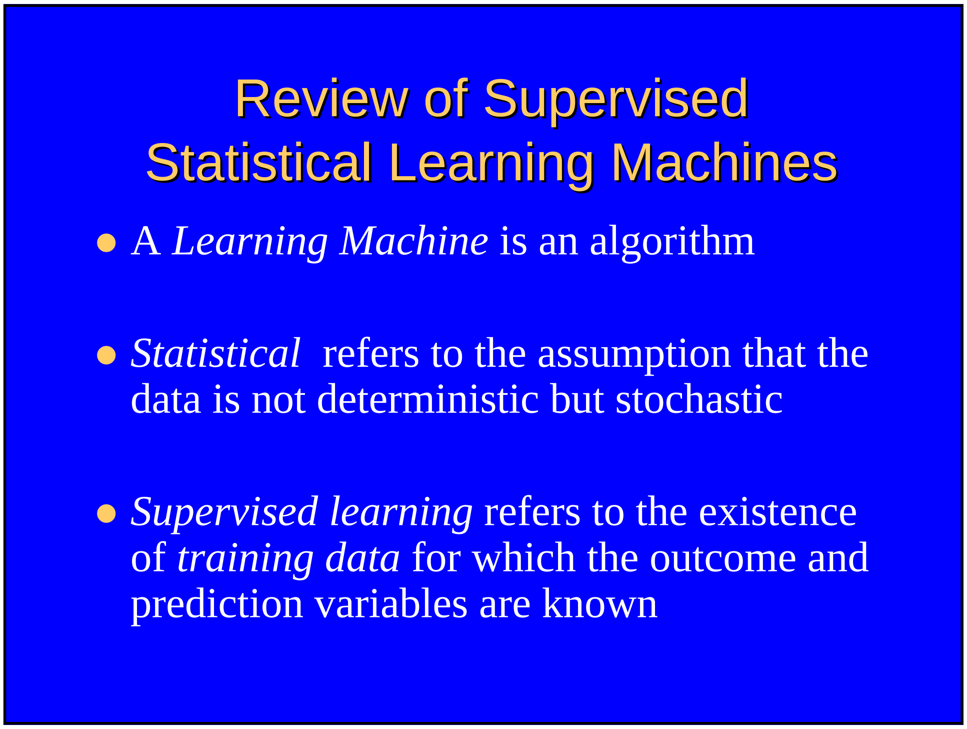


Prediction and Classification

- Finding a rule that assigns membership to a class
  - Life or death outcome during hospitalization
  - Survival curve estimation
  - Survival at intervals of 5, 10, 15, 20 years
- The statistical model for any rule depends on unknown values of parameters

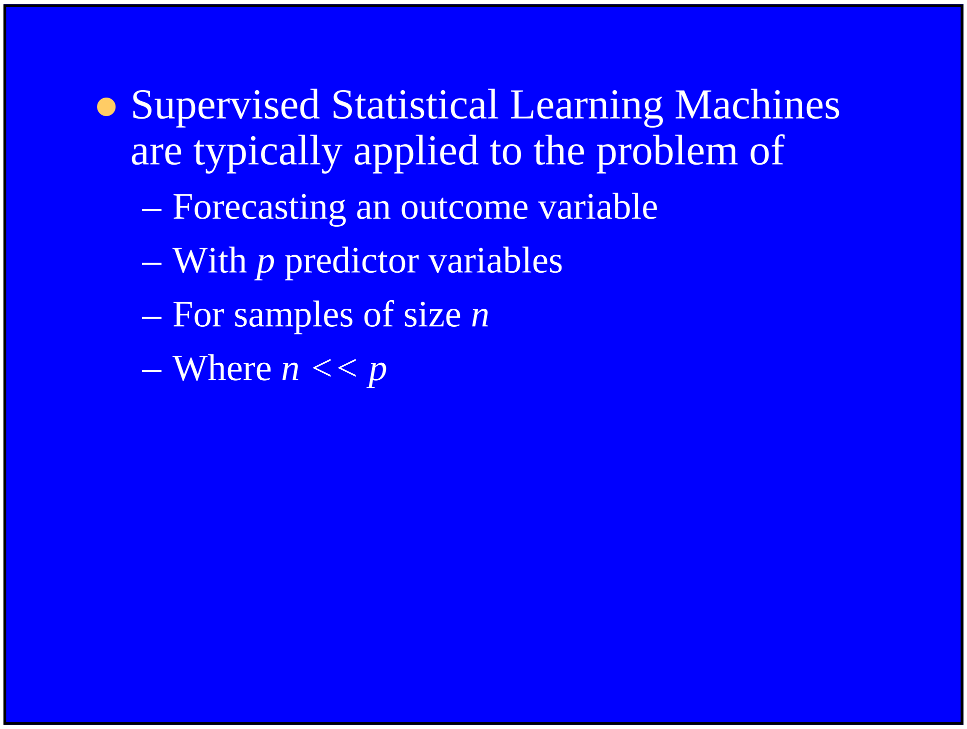


- First Optimization Problem
  - Estimating parameters to minimize error of misclassification for a given rule
- Second optimization problem:
  - finding the rule with minimum error of misclassification



Review of Supervised Statistical Learning Machines

- A *Learning Machine* is an algorithm
- *Statistical* refers to the assumption that the data is not deterministic but stochastic
- *Supervised learning* refers to the existence of *training data* for which the outcome and prediction variables are known



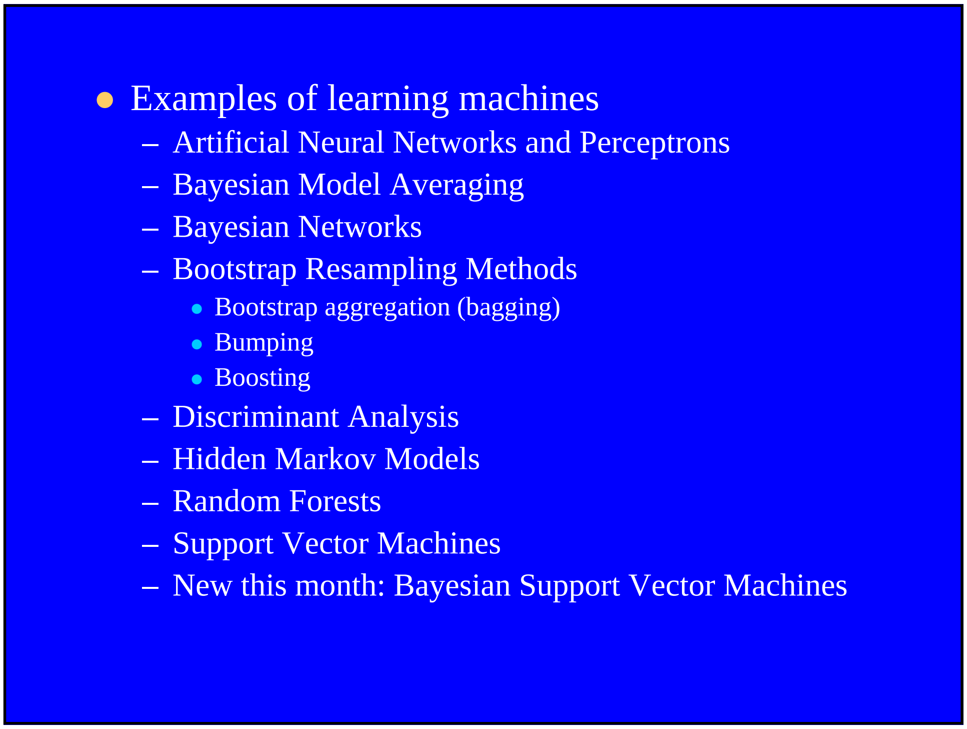
- Supervised Statistical Learning Machines are typically applied to the problem of
  - Forecasting an outcome variable
  - With *p* predictor variables
  - For samples of size *n*
  - Where *n* << *p*

- Conventional statistical methods generally require:

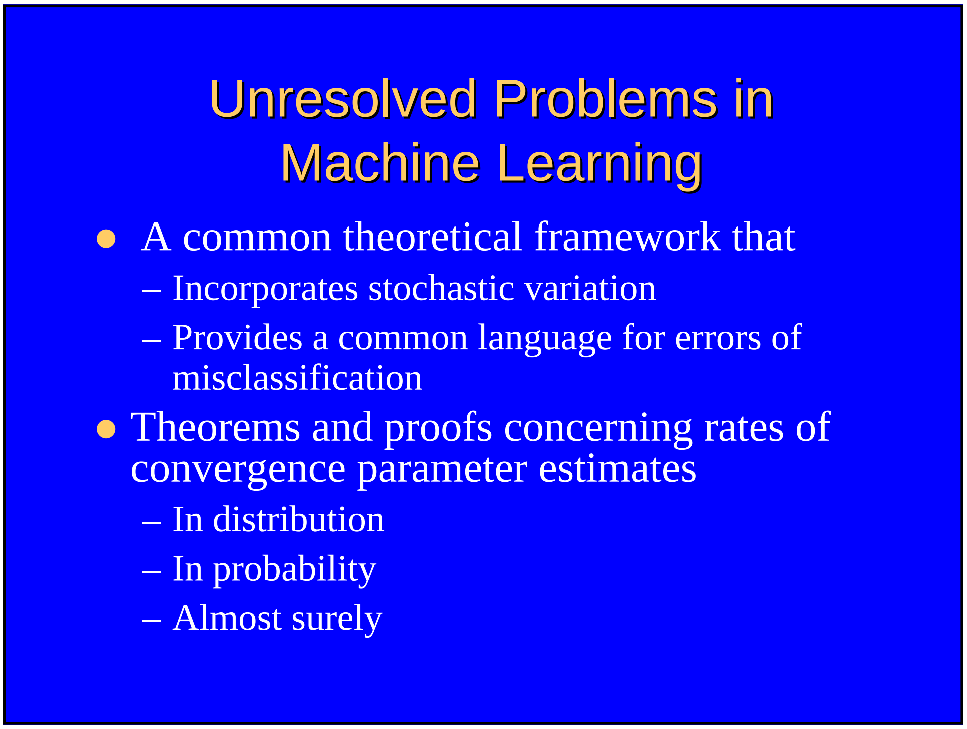
$$n > p^2$$

- This rule-of-thumb generally results in the nonsingularity of the sample variance-covariance matrix of the random vector
- The conventional statistical methods fail if the sample variance-covariance matrix cannot be inverted

- Learning Machines have been developed by
  - Applied Mathematicians
  - Computer Scientists
  - Engineers
  - Statisticians
- Consequently, there is no uniformity in
  - Nomenclature
  - How random variation is modelled



- Examples of learning machines
  - Artificial Neural Networks and Perceptrons
  - Bayesian Model Averaging
  - Bayesian Networks
  - Bootstrap Resampling Methods
    - Bootstrap aggregation (bagging)
    - Bumping
    - Boosting
  - Discriminant Analysis
  - Hidden Markov Models
  - Random Forests
  - Support Vector Machines
  - New this month: Bayesian Support Vector Machines



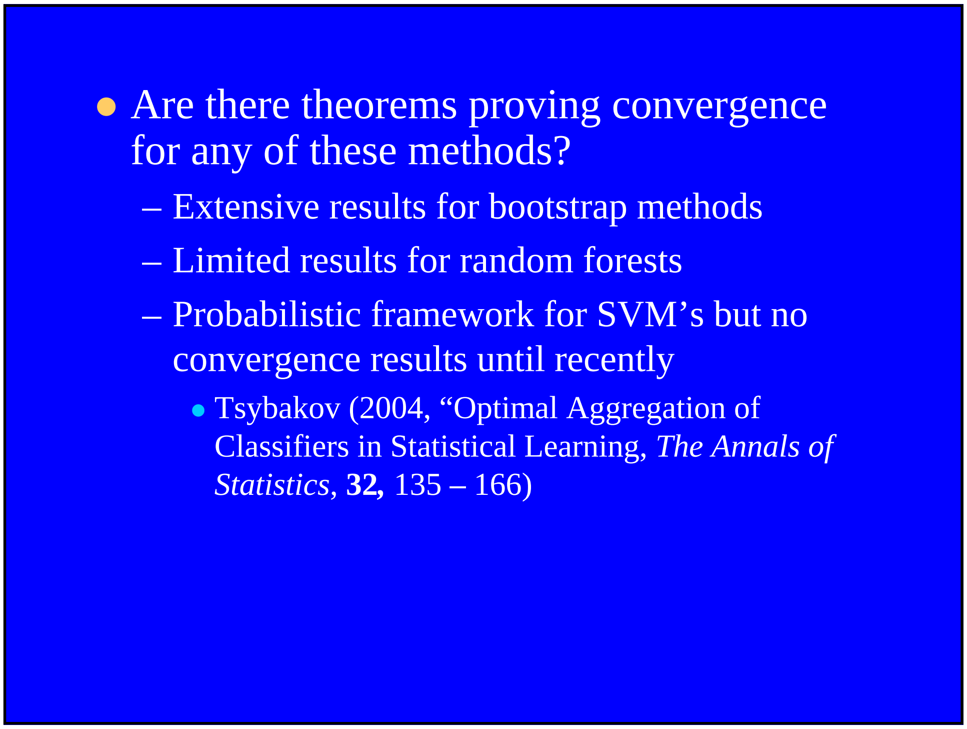
Unresolved Problems in Machine Learning

- A common theoretical framework that
  - Incorporates stochastic variation
  - Provides a common language for errors of misclassification
- Theorems and proofs concerning rates of convergence parameter estimates
  - In distribution
  - In probability
  - Almost surely

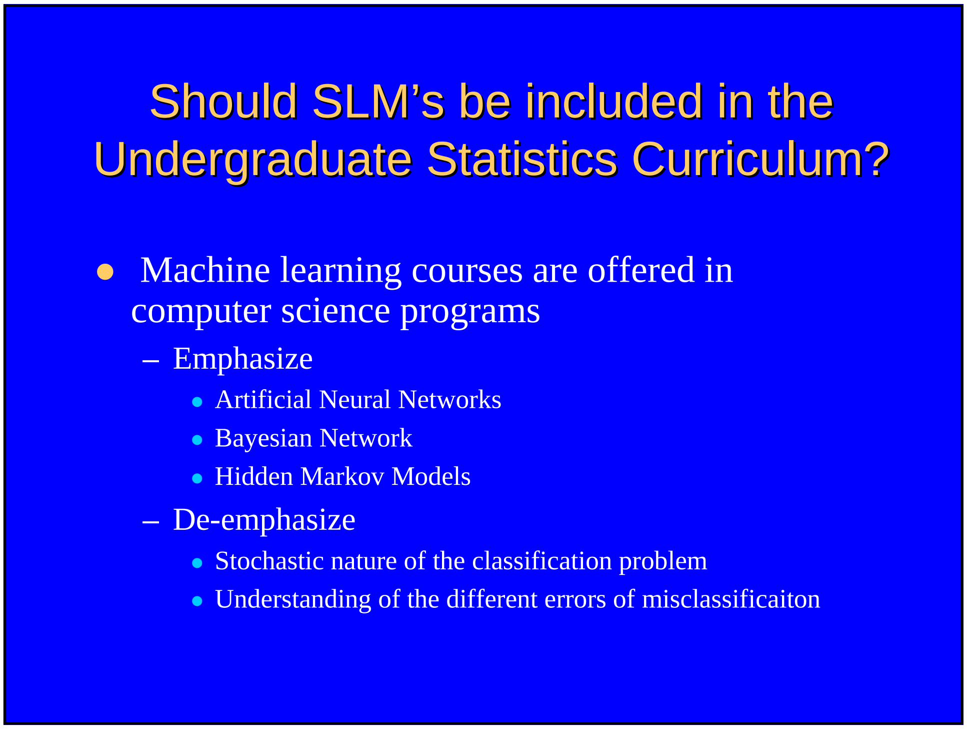
- Theorems and proofs concerning optimality
- Coherent empirical approach to comparison of convergence and optimality among different statistical learning machines

Theoretical Questions

- Can the following be expressed as SVM’s?
  - Bagging
  - Bumping
  - Boosting
  - Random Forests
- Can an SVM be expressed as any one of these?

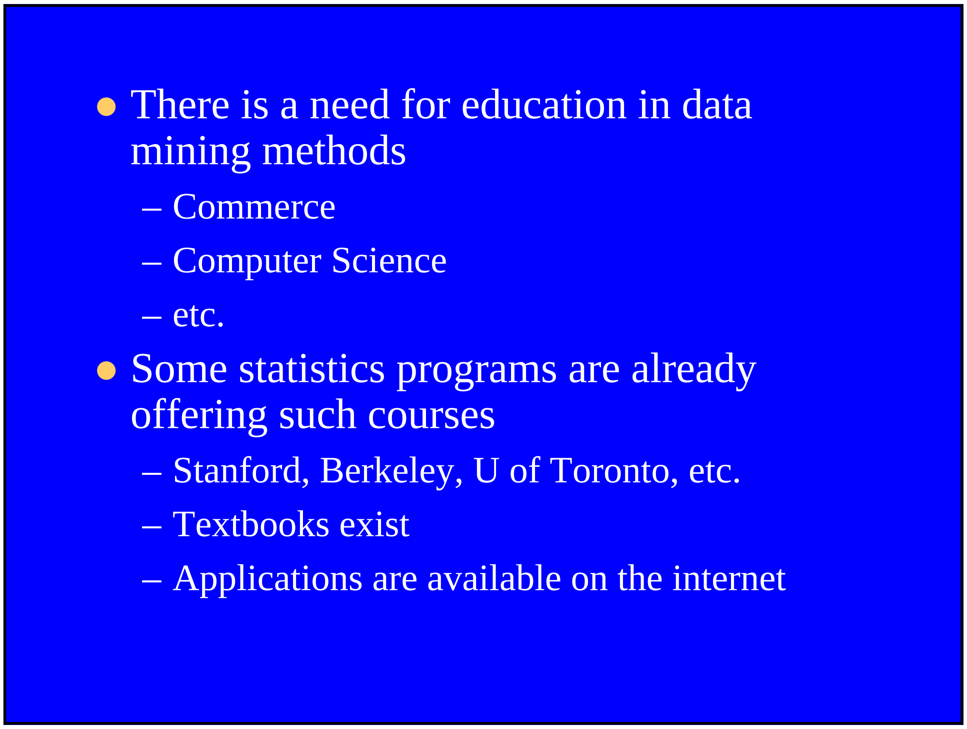


- Are there theorems proving convergence for any of these methods?
  - Extensive results for bootstrap methods
  - Limited results for random forests
  - Probabilistic framework for SVM’s but no convergence results until recently
    - Tsybakov (2004, “Optimal Aggregation of Classifiers in Statistical Learning, *The Annals of Statistics*, **32,** 135 – 166)

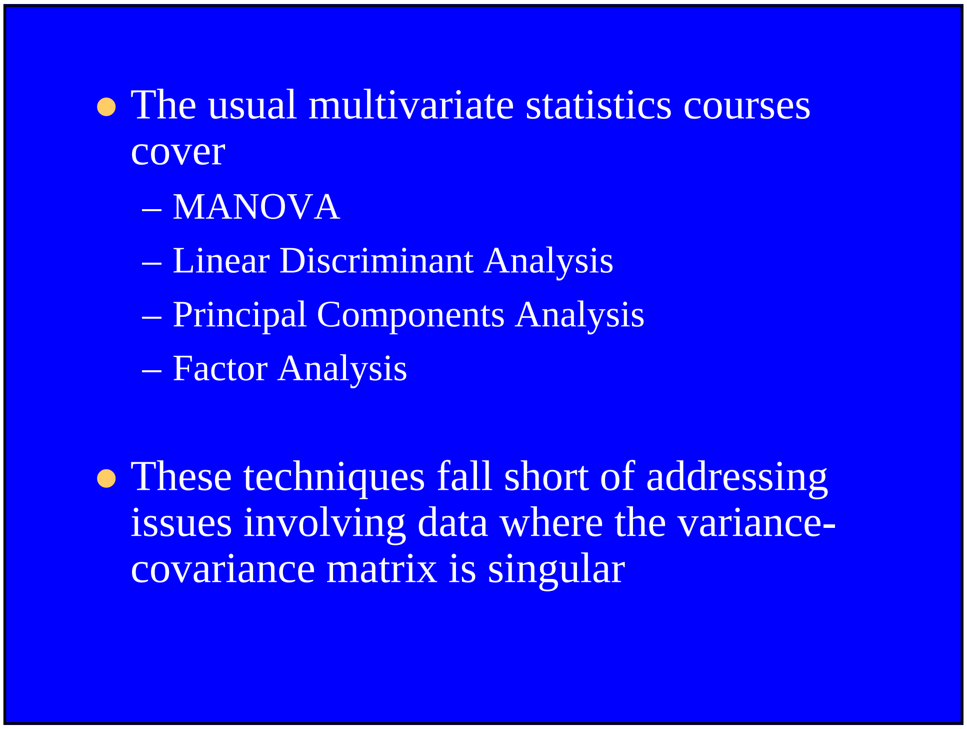


Should SLM’s be included in the Undergraduate Statistics Curriculum?

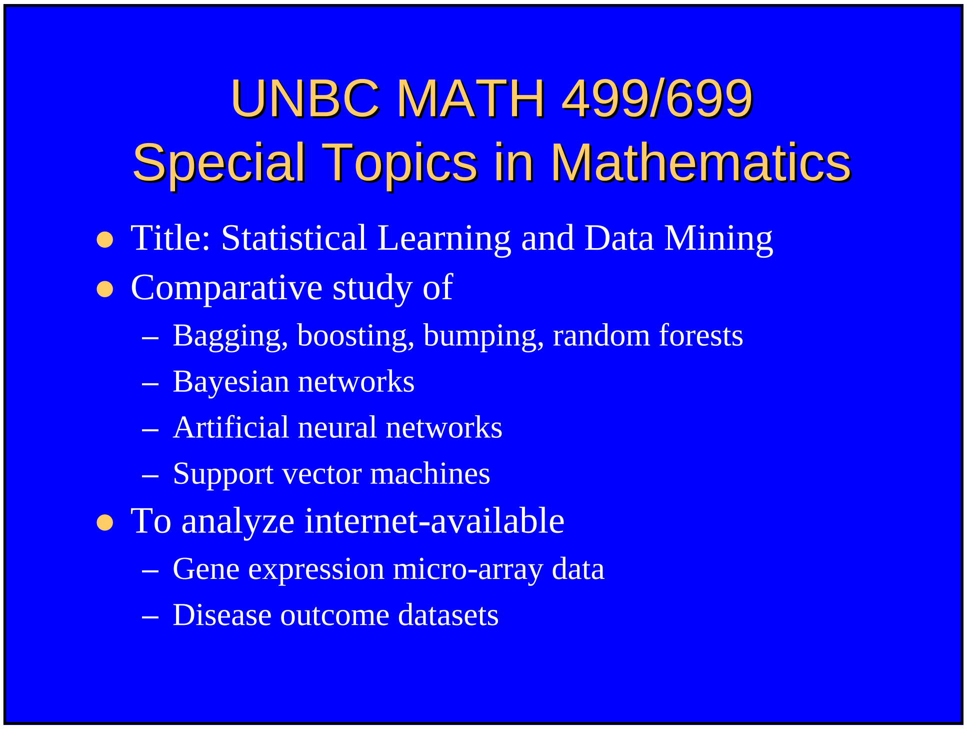
- Machine learning courses are offered in computer science programs
  - Emphasize
    - Artificial Neural Networks
    - Bayesian Network
    - Hidden Markov Models
  - De-emphasize
    - Stochastic nature of the classification problem
    - Understanding of the different errors of misclassificaiton



- There is a need for education in data mining methods
  - Commerce
  - Computer Science
  - etc.
- Some statistics programs are already offering such courses
  - Stanford, Berkeley, U of Toronto, etc.
  - Textbooks exist
  - Applications are available on the internet

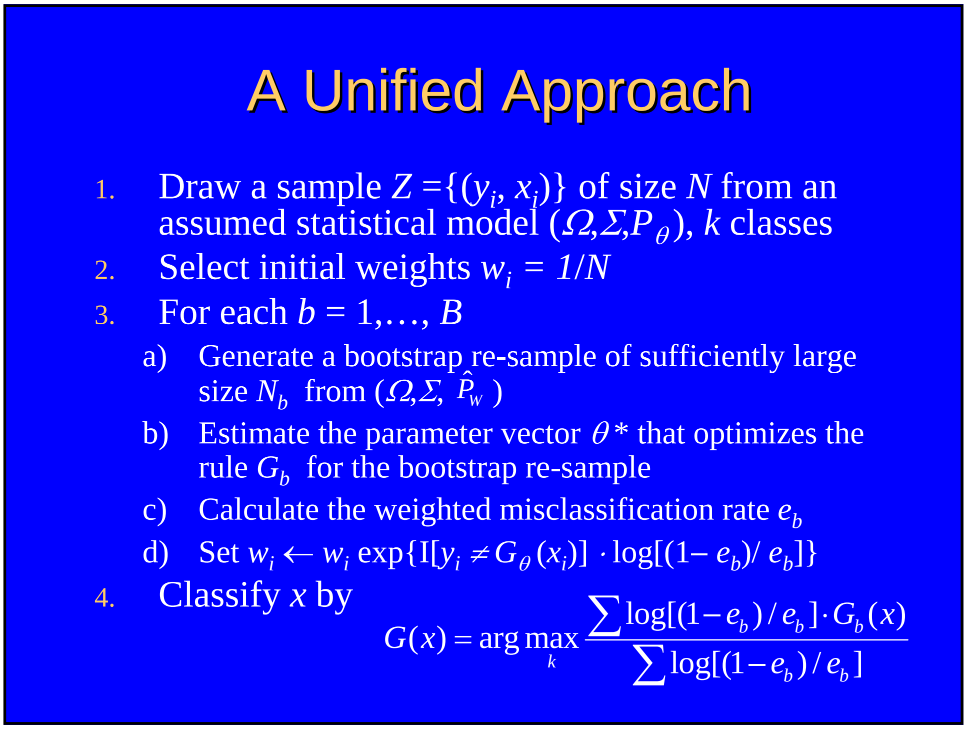


- The usual multivariate statistics courses cover
  - MANOVA
  - Linear Discriminant Analysis
  - Principal Components Analysis
  - Factor Analysis
- These techniques fall short of addressing issues involving data where the variance-covariance matrix is singular



UNBC MATH 499/699
Special Topics in Mathematics

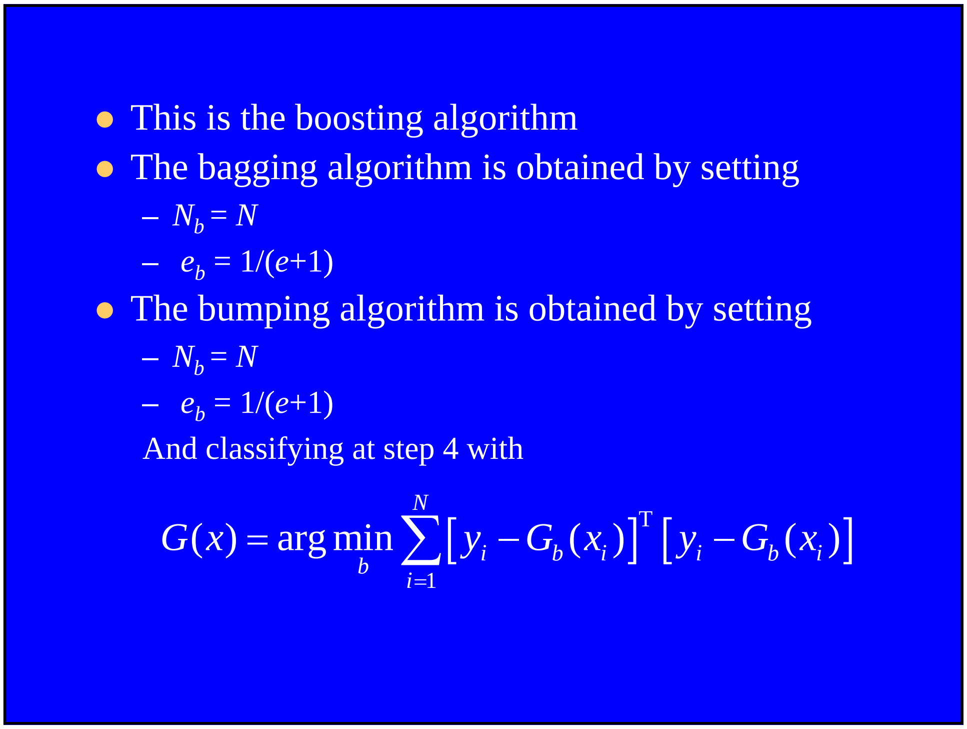
- Title: Statistical Learning and Data Mining
- Comparative study of
  - Bagging, boosting, bumping, random forests
  - Bayesian networks
  - Artificial neural networks
  - Support vector machines
- To analyze internet-available
  - Gene expression micro-array data
  - Disease outcome datasets



A Unified Approach

1. Draw a sample $Z=\{(y_i, x_i)\}$ of size N from an assumed statistical model (Ω,Σ,P_θ), k classes
2. Select initial weights $w_i = 1/N$
3. For each $b = 1,\ldots, B$
   a) Generate a bootstrap re-sample of sufficiently large size N_b from $(\Omega,\Sigma, \hat{P}_W)$
   b) Estimate the parameter vector θ^* that optimizes the rule G_b for the bootstrap re-sample
   c) Calculate the weighted misclassification rate e_b
   d) Set $w_i \leftarrow w_i \exp\{\mathrm{I}[y_i \neq G_\theta(x_i)] \cdot \log[(1- e_b)/ e_b]\}$
4. Classify x by

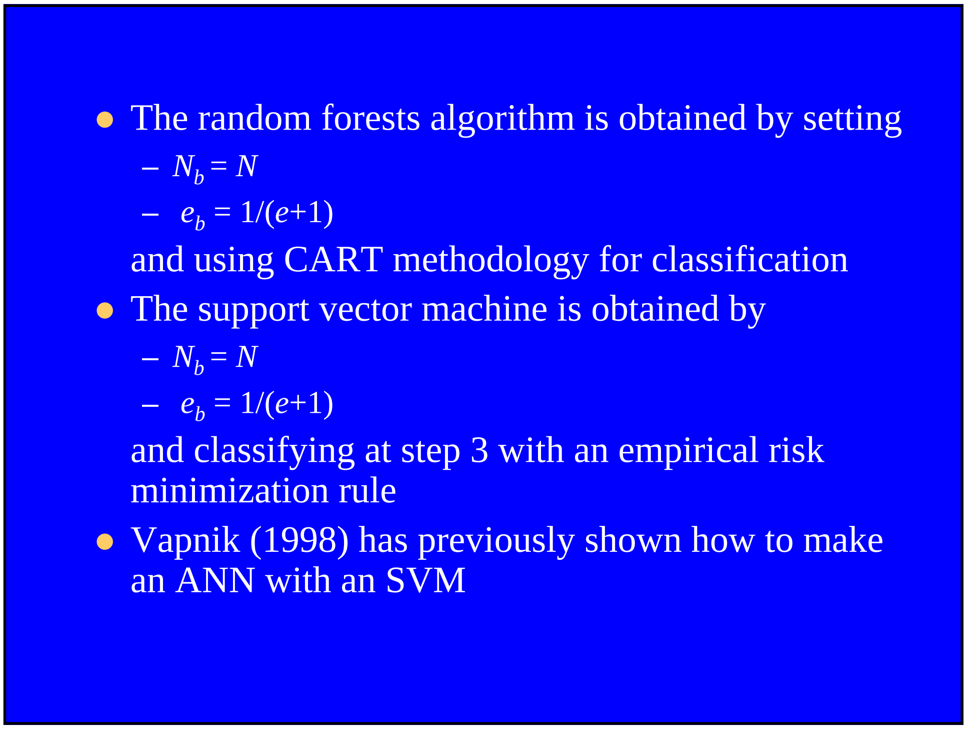
$$G(x) = \arg\max_k \frac{\sum \log[(1-e_b)/e_b]\cdot G_b(x)}{\sum \log[(1-e_b)/e_b]}$$



- This is the boosting algorithm
- The bagging algorithm is obtained by setting
  - $N_b = N$
  - $e_b = 1/(e+1)$
- The bumping algorithm is obtained by setting
  - $N_b = N$
  - $e_b = 1/(e+1)$

  And classifying at step 4 with

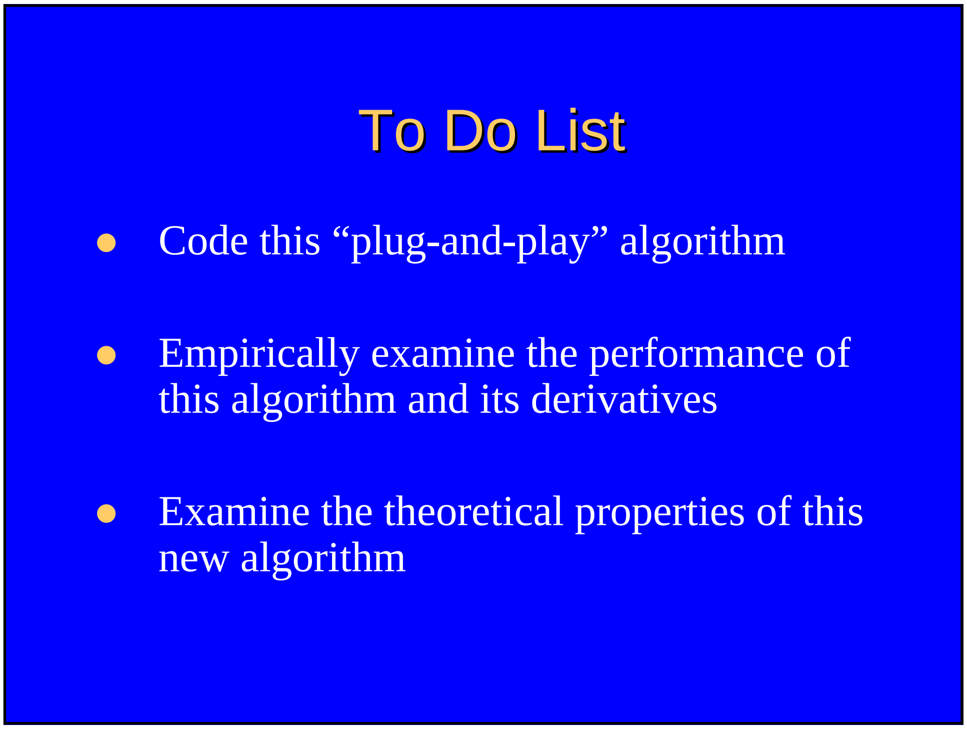
$$G(x) = \arg\min_{b} \sum_{i=1}^{N} [y_i - G_b(x_i)]^{\mathrm{T}} [y_i - G_b(x_i)]$$



- The random forests algorithm is obtained by setting
  - $N_b = N$
  - $e_b = 1/(e+1)$

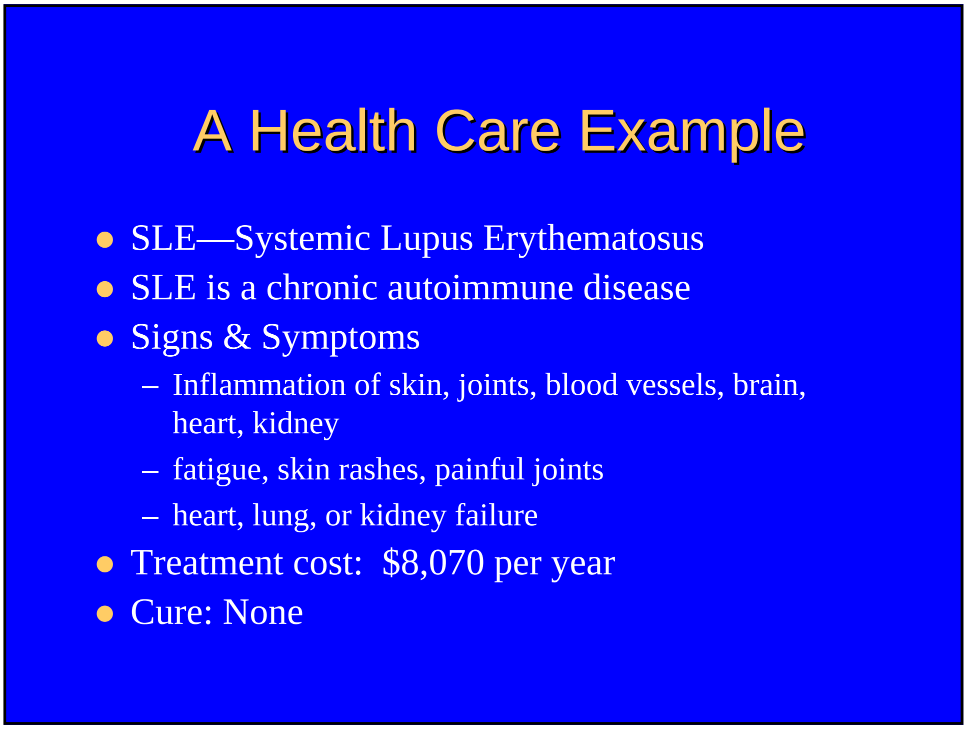
  and using CART methodology for classification
- The support vector machine is obtained by
  - $N_b = N$
  - $e_b = 1/(e+1)$

  and classifying at step 3 with an empirical risk minimization rule
- Vapnik (1998) has previously shown how to make an ANN with an SVM



To Do List

- Code this “plug-and-play” algorithm
- Empirically examine the performance of this algorithm and its derivatives
- Examine the theoretical properties of this new algorithm



A Health Care Example

- SLE—Systemic Lupus Erythematosus
- SLE is a chronic autoimmune disease
- Signs & Symptoms
  - Inflammation of skin, joints, blood vessels, brain, heart, kidney
  - fatigue, skin rashes, painful joints
  - heart, lung, or kidney failure
- Treatment cost: $8,070 per year
- Cure: None

- AS—Ankylosing Spondylitis
- AS is a chronic autoimmune disease
- Signs & Symptoms
  - uveitis, pain and stiffness of the back, hips, shoulders, knees, and ankles
  - fusion of the bones of the spine
- Treatment cost: $3,070 per year
- Cure: None

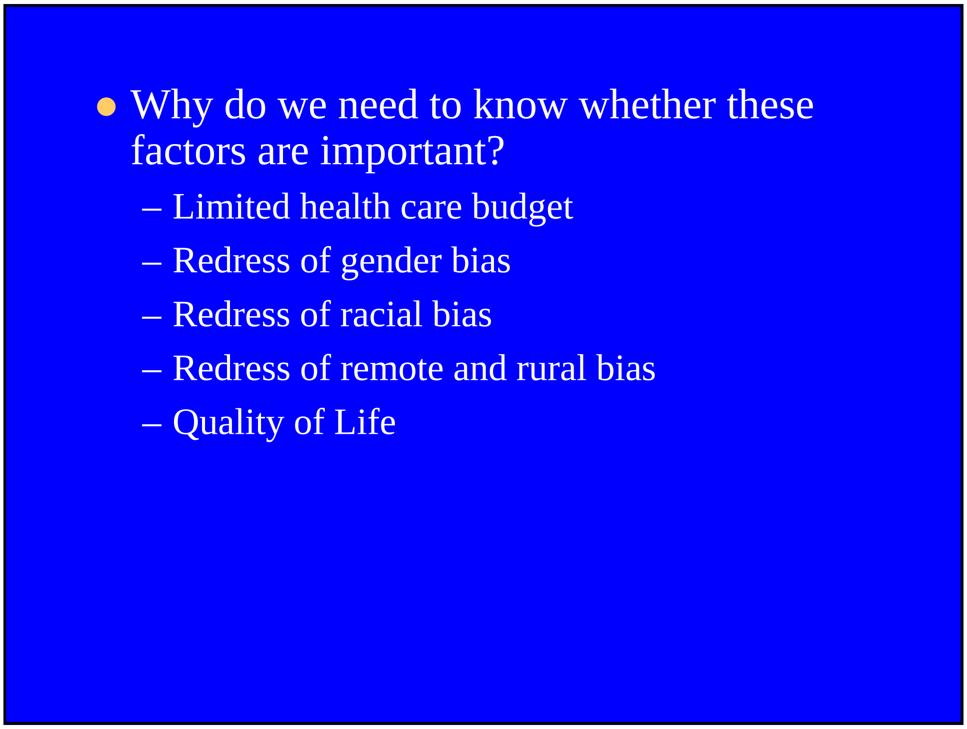
- SSc—Systemic Sclerosis
- SSc is a chronic autoimmune disease
- Signs & Symptoms
  - pain, stiffness, and swelling of joints,
  - skin tightening which restricts movement
  - as well as gastro-intestinal and cardiopulmonary involvement
  - heart, lung, or kidney failure
- Treatment cost: $5,910 per year
- Cure: None

What We Know and What We Don’t

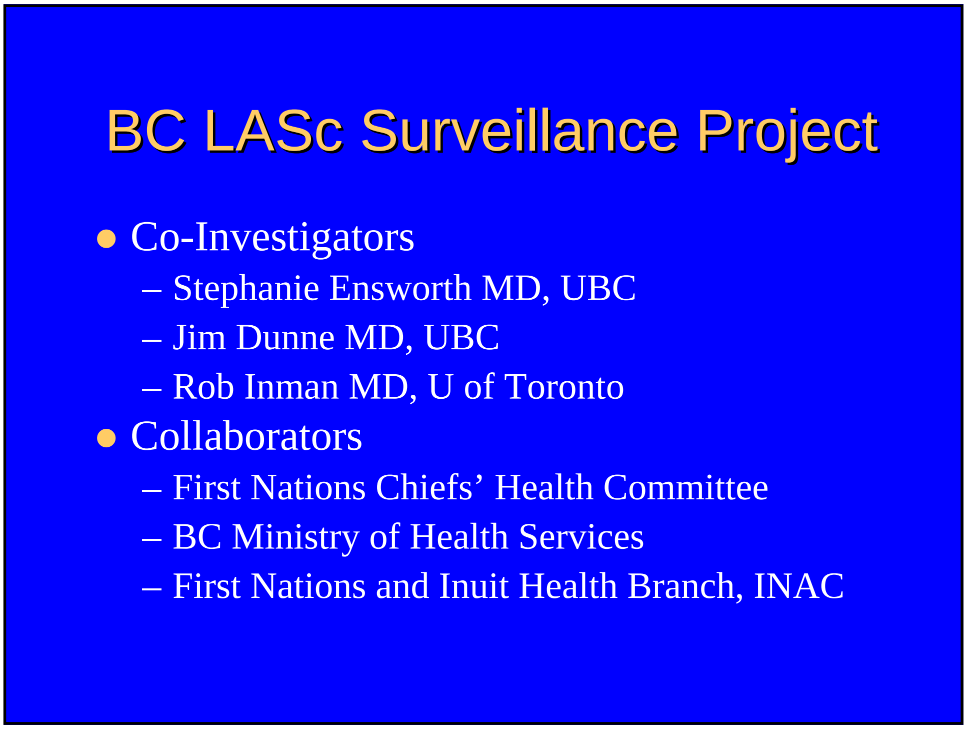
- We don’t know the prevalence and incidence of these diseases in BC
- The evidence suggests that these diseases are more prevalent in BC First Nations
- A comparison:
  - 469 Aboriginal persons in Canada known to have AIDS in 2002
  - We estimate 680 members of BC’s First Nations have SLE

Predicting Health Care Outcomes

- We need to know the important factors
  - Gender
  - First Nations Status
  - Geographical location
    - Toxic exposure
    - Genetics
    - Access to Specialist Health Care
  - Expertise of health care provider

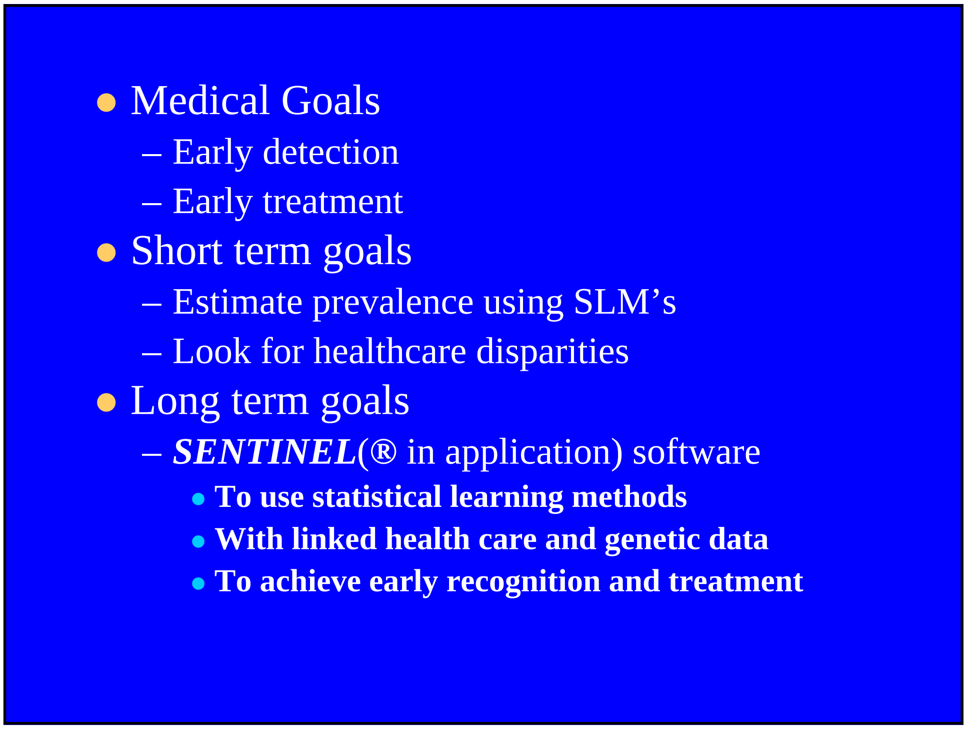


- Why do we need to know whether these factors are important?
  - Limited health care budget
  - Redress of gender bias
  - Redress of racial bias
  - Redress of remote and rural bias
  - Quality of Life

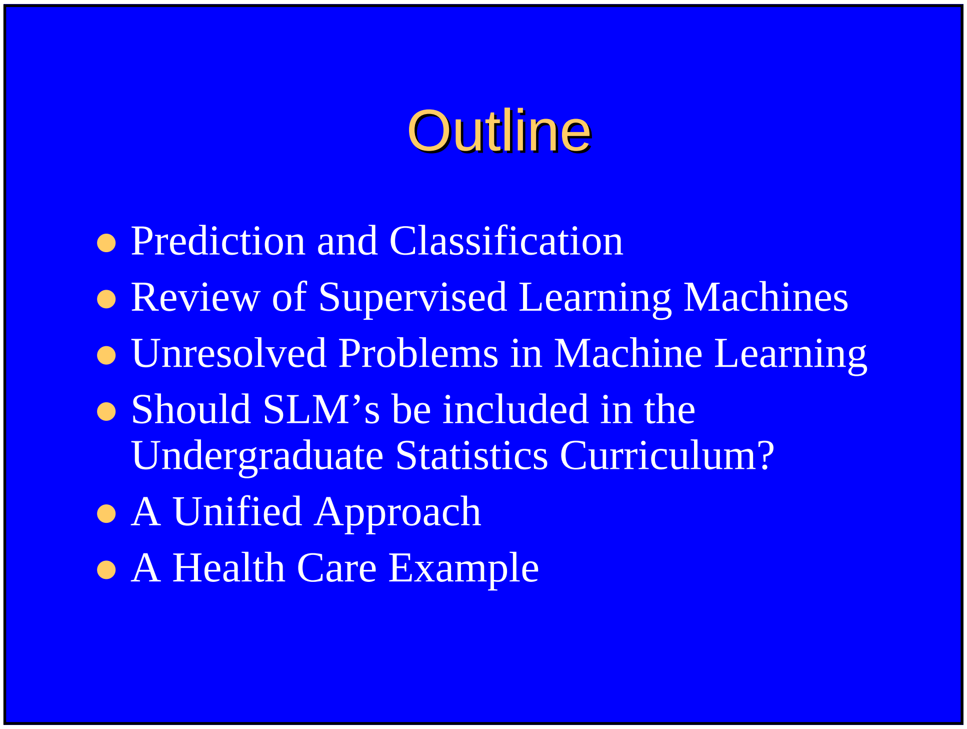


BC LASc Surveillance Project

- Co-Investigators
  - Stephanie Ensworth MD, UBC
  - Jim Dunne MD, UBC
  - Rob Inman MD, U of Toronto
- Collaborators
  - First Nations Chiefs' Health Committee
  - BC Ministry of Health Services
  - First Nations and Inuit Health Branch, INAC

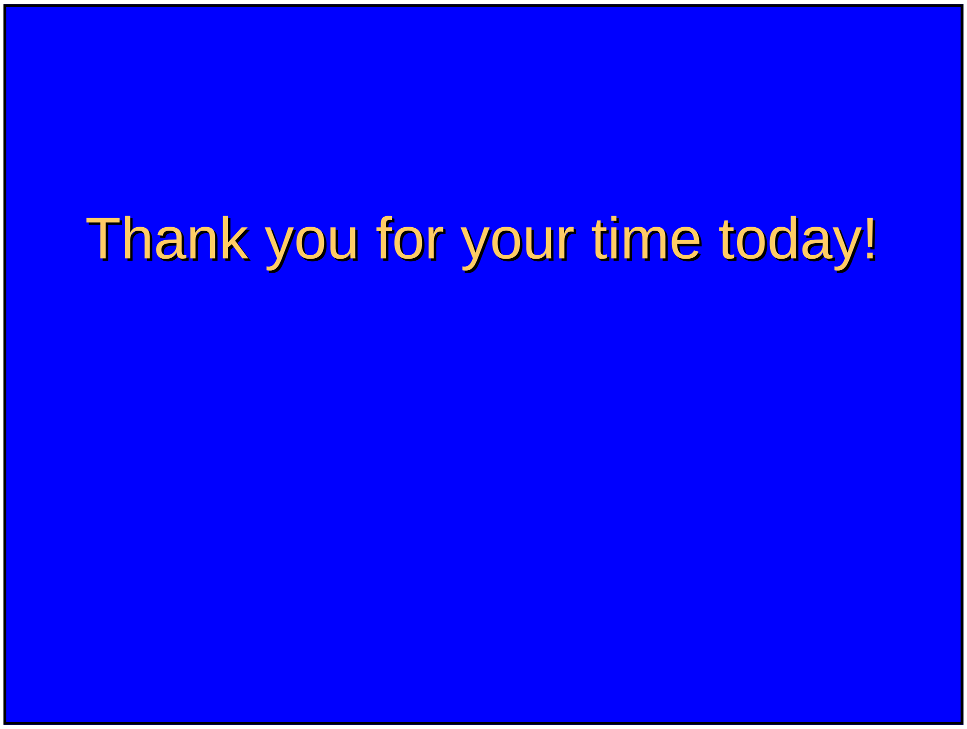


- Medical Goals
  - Early detection
  - Early treatment
- Short term goals
  - Estimate prevalence using SLM’s
  - Look for healthcare disparities
- Long term goals
  - ***SENTINEL***(® in application) software
    - **To use statistical learning methods**
    - **With linked health care and genetic data**
    - **To achieve early recognition and treatment**



Outline

- Prediction and Classification
- Review of Supervised Learning Machines
- Unresolved Problems in Machine Learning
- Should SLM’s be included in the Undergraduate Statistics Curriculum?
- A Unified Approach
- A Health Care Example



Thank you for your time today!