Conference on Automated Learning and Discovery

June 11-13, 1998
Carnegie Mellon University, Pittsburgh, PA

Sponsored by Carnegie Mellon University's newly created
Center for Automated Learning and Discovery
with support from the National Science Foundation

Information about the conference is available via
the World Wide Web at
http://www.cs.cmu.edu/~conald

Scientific Program Committee

Sebastian Thrun (chair)
Christos Faloutsos
Stephen Fienberg
Tom Mitchell
Andrew Moore
Roni Rosenfeld
Richard Scheines
Teddy Seidenfeld
Kannan Srinivasan
Larry Wasserman

Local Arrangements Committee

Diane Stidle
Michael Bowling
Frank Dellaert
Kunnayut Eiamsa-Ard
Thomas Nichols
Nicholas Roy
Stella Maris Salvatierra
Ashish Sanil
Sean Slattery
Joseph O'Sullivan
Belinda Thom

Conference Venue

The Conference on Automated Learning and Discovery will bring together leading researchers from scientific disciplines concerned with learning from data. It will cover scientific research at the intersection of statistics, computer science, artificial intelligence, databases, social sciences and language technologies. The goal of this meeting is to explore new, unified research directions in this cross-disciplinary field.

The conference features eight one-day cross-disciplinary workshops, interleaved with seven invited plenary talks by renowned statisticians, computer scientists, and cognitive scientists. The workshops will address issues such as: what is the state of the art, what can we do and what is missing? what are promising research directions? what are the most promising opportunities for cross-disciplinary research?

CONALD differs from other meetings in the field in its broad, interdisciplinary scope. The goal of CONALD is to characterize the state-of-the-art in automated learning and discovery, and to identify promising cross-disciplinary research directions. The format will be very much tailored towards open discussions and free exchange of ideas. This meeting will be summarized by a written report that will be made available to the scientific community and National Science Foundation.

Schedule of Events

	Day One Thursday, June 11	Day Two Friday, June 12	Day Three Saturday, June 13
8:30 AM	Coffee & Donuts (8:55) Opening	Coffee & Donuts	Coffee & Donuts
9:00 9:30	Invited Speaker *Robert Tibshirani*	Invited Speaker *Herb Simon*	Invited Speaker *Stuart Geman*
10:00	Invited Speaker *Tom Dietterich*	Invited Speaker *Daryl Pregibon*	Workshop Wrap-up
10:30	Break	Break	Break
11:00 11:30	Workshops W1, W2, W3 & W4	Workshops W5, W6, W7 & W8	Workshop Wrap-up con't
12:00 PM			Invited Speaker *David Heckerman*
12:30	Lunch	Lunch	Lunch
1:00			
1:30			
2:00 2:30 3:00	Workshops continued	Workshops continued	Post-Conference Local Activities
3:30	Break	Break	
4:00 4:30 5:00	Workshops continued	Workshops continued	
5:30 6:00	Invited Speaker *Michael Jordan*		
6:30		River Boat Banquet (6:15) Leave CMU (6:40) Boat Dep. (10:15) Return	

On-site registration: Thursday & Friday, 8:30AM-5:00PM & Saturday, 9:00AM-Noon.
See page 5 for Invited Speaker abstracts. See page 8 for Workshop descriptions.

Invited Speakers

Robert Tibshirani
Learning from Data: Statistical Advances and Challenges

In this talk I survey some recent trends and developments in the field of Statistics, and try to identify some important research problems in the general area of "Learning from Data." The fields of statistics, computer science and artificial intelligence (at least) are key players in this area: I will try to pinpoint some of the things that statisticians can teach CS and AI, and what we hope to learn from them.

ROBERT TIBSHIRANI is a Professor in the Departments Public Health Sciences, and Statistics at the University of Toronto, where he has been since 1985. He received a Ph.D. in Statistics from Stanford University in 1984.

Tom Dietterich
Learning for Sequential Decision Making

Many important problems involve making a sequence of decisions. The field of "reinforcement learning" studies algorithms for learning near-optimal policies for sequential decision making. This talk will review reinforcement learning and then discuss methods for scaling up reinforcement learning algorithms to solve very large problems. These methods include some new techniques for hierarchical reinforcement learning.

THOMAS G. DIETTERICH is Professor of Computer Science at Oregon State University. He received the Ph.D. in computer science from Stanford University in 1984, the M.S. from the University of Illinois in 1979, and the A.B. (in mathematics) from Oberlin College in 1977. He is the Executive Editor of the journal *Machine Learning* and the editor of the MIT Press Series on Adaptive Computation and Machine Learning.

Michael Jordan
Graphical Models and Variational Approximation

Graphical models (also known as Bayesian belief networks) provide an elegant formalism for managing uncertainty that unifies much of the literature on stochastic modeling. For sparse networks (e.g., networks in the form of chains or trees, such as Kalman filters, hidden Markov models, and probabilistic decision trees), graphical model algorithms are exact, efficient and practical. For dense networks, however, the exact algorithms are often (hopelessly) inefficient, and this fact has hindered the application of this richer class of models to large-scale problems. I discuss variational methodology, which provides a general framework for approximate graphical model inference. The variational methods I present are efficient; moreover, they tend to be more accurate for dense networks than for sparse networks. They can readily be combined with exact techniques to yield a class of algorithms that perform well for a variety of network architectures. I illustrate these ideas with examples of applications of dense networks to problems in diagnosis, prediction, and control.

MICHAEL I. JORDAN is Professor in the Department of Computer Science and the Department of Statistics at the University of California, Berkeley. He received his Masters in Mathematics from Arizona State University, and earned his PhD in Cognitive Science from the University of California, San Diego. He has worked on a variety of topics in the area of machine learning, focusing on neural networks and graphical models.

Herb Simon
Using Machine Learning to Understand Human Learning

An important application of AI methods has been to model human learning processes. Human learning probably employs several different, and complementary mechanisms: to differentiate and recognize stimuli, to learn new procedures and compile them for greater efficiency, to build associative information stores. The talk will survey a range of methods, including serial (symbolic) and parallel (connectionist) methods, that have been proposed, and the evidence linking them to empirical data on human learning.

HERBERT A. SIMON's research has ranged from computer science to psychology, administration, and economics. The thread of continuity has been human decision-making and problem-solving processes, and extensive use of the computer for simulating human thinking. Born in 1916 in Milwaukee, Wisconsin, Simon was educated at the University of Chicago. Since 1949, he has been on the faculty of Carnegie Mellon University, where he is Richard King Mellon University Professor of Computer Science and Psychology. In 1978, he received the Alfred Nobel Memorial Prize in Economic Sciences, and in 1986 the National Medal of Science.

Daryl Pregibon
Realtime Learning and Discovery in Large Scale Networks

In many financial and communications industries, data from networks are becoming increasingly important. Two aspects of network data that are particularly challenging concern the size of the data stream and it's timeliness. Indeed we have been concerned with industrial strength applications that consist of hundreds of millions of transactions (i.e., packets, conversations) that need to be analyzed, summarized and acted upon, within 24 hours. Furthermore some applications (e.g., fraud and network intrusion detection) require real-time analysis and intervention.

We will share our experiences in these applications, stressing statistical and computational issues. Several examples of web-based delivery systems will be presented that convey the scope and scale of what is possible today --- and what might be coming tomorrow. [Joint work with RA Becker, C Cortes, and AR Wilks]

DARYL PREGIBON is Head of Statistics Research at AT&T Labs. His group is focused on providing the computational and theoretical foundation for the application of statistics to very large data sets. He is Past-Chair of the premier datamining conference (KDD) and remains the Chair of the Committee on Applied and Theoretical Statistics at the National Academy of Science.

Stuart Geman
Probabilistic Grammars and their Applications

I will review the theory of formal grammars. The simplest grammars are regular grammars, which become Markov chains when fitted with a probability distribution. More general and more powerful are the context-free grammars, the probabilistic versions of which are equivalent to branching processes. Most applications in language, vision, and neural modeling appear to require even more general grammars: context-sensitive grammars or grammars of still higher order. There is no general theory for how to put probability distributions on context-sensitive grammars, yet it is increasingly apparent that successful applications require a probabilistic framework, with its associated statistical theory and theory of inference. I will suggest a method for placing workable probability distributions on context-sensitive (and more general) grammars, and I will discuss some applications.

STUART GEMAN received his Ph.D. in Applied Mathematics from MIT in 1977. He joined the faculty at Brown in 1977, where he has been Professor of Applied Mathematics since 1985, and the James Manning Professor since 1997. He received the Presidential Young Investigator Award in 1984 and is a Fellow of the IMS. Professor Geman's research has been in nonparametric inference, the theory and applications of Markov random fields, and Bayesian methods for image analysis and vision. He has worked extensively with industrial applications of computational vision. His current interests are in compositional models for vision, coding theory, statistical methods for analyzing multi-unit neuronal recordings, and theories of neural representation.

David Heckerman
A Bayesian Approach to Causal Discovery

DAVID HECKERMAN is a Senior Researcher of Decision Theory & Adaptive Systems Group at Microsoft Research

[Abstract not available at press time; please see CONALD web page: http://www.cs.cmu.edu/~conald]

Workshops

Workshops will be organized in four parallel tracks

W1: Learning Causal Bayesian Networks

Chairs: Richard Scheines and Larry Wasserman

This workshop will focus on several topics in an area of automated learning and discovery that has flourished in the last several years: causal discovery. Interested participants are encouraged to submit papers on work in progress, some of which will be chosen for discussion during the workshop.

Depending on interest, we will select a subset of the topics below to cover in the workshop. For each topic, a few working papers among those submitted will be distributed beforehand, and a discussion leader will be assigned to lead an hour long discussion on the topic.

Topics include:

- Identifying Causal Parameters
- Handling Latent Variables
- Computational Methods, i.e., issues of computational efficiency
- Model Scoring & Model Selection
- Modeling Feedback
- Case Studies

W2: Mixed-Media Databases

Chairs: Christos Faloutsos, Alex Hauptmann and Michael Witbrock

This workshop is intended for researchers with an interest in learning from multiple media. The workshop will emphasize both algorithms and applications of learning with mixed media databases. Papers that describe algorithms should cover either novel approaches designed to benefit from mixed-media data, or modifications of standard algorithms that utilize multiple media data sources. Application papers should clearly demonstrate the benefits of learning from two or more types of media.

Different media areas to be addressed include:

- Vision: Image, Video, and VRML
- Speech and Audio
- Text, including OCR, Closed-Captioning, handwriting, and web-documents
- Olfactory perception
- Haptic and Touch sensing

W3: Machine Learning and Reinforcement Learning for Manufacturing

Chairs: Sridhar Mahadevan and Andrew Moore

In recent years there has been a flurry of research on statistics and machine learning applied to decision making and control. Progress has been made in many academic areas such as Reinforcement Learning, Neural Networks and, Diagnostic Bayesian Networks. Applications are emerging in the control of continuous processes, batch processes (such as wafer fabrication), probabilistic diagnosis, and industrial engineering tasks such as optimal control of transfer lines or production scheduling.

Two kinds of presentations will be particularly encouraged for this workshop:

1. Case studies of learning, applied statistics, datamining or neural networks applied to manufacturing systems.
2. Advances in algorithms or theory that permit learning approaches to scale up to bigger problems.

Interested participants are encouraged to submit papers on work in progress, some of which will be chosen for discussion during the workshop.

W4: Large-Scale Consumer Databases

Chairs: Mike Meyer, Teddy Seidenfeld and Kannan Srinivasan

[Abstract not available at press time; please see CONALD web page, http://www.cs.cmu.edu/~conald/workshop.shtml]

W5: Visual Methods for the Study of Massive Data Sets

Chairs: Bill Eddy and Steve Eick

This workshop is intended for researchers interested in the development and use of visual methods for the exploration and analysis of very large data sets. Visual methods include both static and dynamic displays. Very large data sets can include both standard numerical data and other complex and multi-media data sets such as text and images.

Interested participants are encouraged to submit full or position papers (see Submission Instructions, page 12). Several of the submissions will be selected for discussion during the workshop. Both general methods and specific applications are encouraged.

W6: Learning fromText and the Web

Chairs: Jaime Carbonell, Steve Fienberg, Tom Mitchell and Yi-Ming Yang

An increasing fraction of the world's information and data is now represented in text form. For example, the World Wide Web, online news feeds, and other Internet sources contain a tremendous volume of information, much of which would be useful to computer programs if only it were represented in a more structured, computer readable form.

The goal of this workshop is to explore computer methods for automatically extracting information from text and hypertext sources. Examples might include systems that automatically extract descriptions of corporate mergers by monitoring online newsfeeds, or systems that automatically extract addresses and phone numbers from home pages on the web.

Interested participants are encouraged to submit full workshop paper or position papers. Relevant topics include (but are not restricted to)

- Computer methods for information extraction from text and hypertext
- Automated learning of such methods
- Automatic text summarization
- Automatic text classification

W7: Robot Exploration and Learning

Chairs: Howie Choset, Maja Matarić and Sebastian Thrun

Recently, robot exploration and learning techniques have begun to have a profound impact on the field of robotics. Exploration and learning techniques have been at the heart of many recent successes in the field of robotics, where they have provided additional robustness to robot behavior as well as eased the task of robot programming.

Robotics is highly attractive to researchers in the field of learning and discovery, as robots are a prime example for a variety of embedded application domains that integrate perception and action. Robotic domains stress the importance of a variety of current research challenges, including partial observability, non-stationarity, and uncertainty in perception and action.

This workshop will discuss open questions in robot exploration and learning, including: How can we scale up robot exploration and learning so that more can be learned from less data? How can we devise efficient exploration strategies for dynamic and high-dimensional environments? What role will learning and exploration play in upcoming application domains such as service robotics? How can learning support cooperation of multi-robot teams? How can we bring together and achieve synergy between the various disciplines involved in the study of robot exploration and learning and beyond?

This workshop seeks to bring together leading researchers in the diverse field of robot exploration and learning, to characterize the state of the art, to identify the most promising research opportunities. We also plan to informal discussion between different robotics disciplines, with the goal of leading into novel collaborations.

W8: Discovery in Natural and Social Science

Chair: Raul Valdes-Perez

The practitioners of knowledge discovery par excellence are found in natural and social science. Knowledge discovery tools for scientific research must engage busy and exacting users, and therefore tend to deal squarely with the special tasks of inference that arise there.

This workshop will address methods of and systems for collaborative (scientist/computer) discovery and the issues that these raise. This workshop will only accept position papers (see Submission Instructions, page 12).

Submission Instructions

General participants are strongly encouraged to submit a paper to CONALD. The purpose of the paper submission is two-fold: To give other workshop participants the opportunity to familiarize themselves with a participant's work, and to guide workshop organizers in defining a workshop's scope and agenda. Submissions must be submitted to a particular workshop and the corresponding organizers will select papers to be accepted. Accepted papers will be collected and distributed to all participants of CONALD.

CONALD solicits the submissions of two types of papers: position papers and full papers. Position papers are intended to communicate the participant's experience and interest in a specific workshop to other workshop participants. Full papers will allow participates to describe an interesting problem domain in more depth, or to report interesting preliminary results obtained for a topic relevant to the target workshops. Papers are not intended to present finished work. The workshops will be highly informal, and the workshop chairs might decide not to schedule formal presentations.

Submission format: Unless otherwise specified in the workshop description, position papers are limited to two pages (preferable only one). Full papers are limited to six pages. Papers must be submitted on 8 1/2 x 11 inch or A4 paper, with margins no smaller than 1 inch. The font must be 10 point or larger. Not every workshop will accept full papers. Please check!

Submission instructions: Papers are directly submitted to a particular workshop. Authors are requested to send 4 (four) copies of their paper to

CALD / CONALD'98
Workshop W? (fill in workshop names and numbers)
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA 15213-3891
USA

Papers must arrive at CMU by Monday, February 16, 1998. Papers received after this date will be returned unopened. The name of the workshop must appear on the envelope. Authors may submit papers to two workshops, as long as these workshops take place at different days. Notification of receipt will be mailed to the first author (or designated author) by March 1, 1998. Authors will be informed of the decision by March 30, 1998. Submissions to CONALD workshops may also be submitted to other workshops or conferences.

Please direct all inquiries to conald@cs.cmu.edu.

Air Transportation

The Pittsburgh International Airport is a USAirways hub, and is also served by Delta, NorthWest, Continental, United and American. The airport is located approximately 21 miles from Oakland/CMU. Ground transportation options include:

1. Port Authority Transportation (public transportation) - (412-442-2000). PAT's Airport Flyer 28X Bus only makes 4 stops between the airport and Oakland. The Oakland stop is at South Bellefield Avenue, a short walk to Holiday Inn & University Club. Service on the 28X Airport Flyer is from 5:30 a.m. to 12:30 a.m. daily, at 45 minute intervals. The trip takes 45-60 minutes and costs $1.95 per person, exact change required. Boarding is at the lower level at ground transportation bus stop.
2. Airlines Transportation Co. - (412-471-8900). Purchase ticket from attendant at their location in the baggage claim area of the airport. The cost is $12.50 per person one way or $21 round trip. The bus leaves the baggage claim area on the hour from 9 a.m. - 10 p.m. weekdays. The Oakland stop is the Holiday Inn.
3. Yellow Cab Service - approximately $33 per trip plus gratuity. Follow ground transportation signs in the airport. Cabs are always waiting and available.
4. First Class Limousine Service - (412-462-8000). $37.95 per trip (includes gratuity). Call several days in advance and give them your flight information. They will have a car and driver waiting for you on your arrival.

All prices are accurate as of January, 1998.

Accommodations

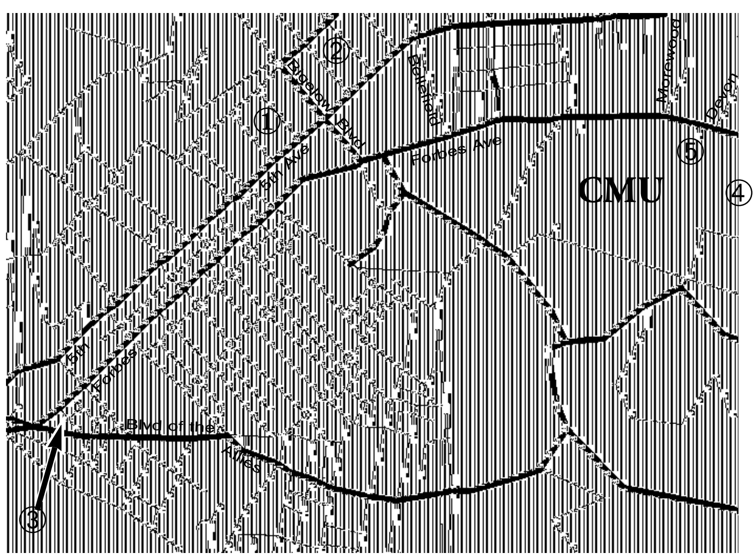
Special rates are available for CONALD participant if reservations are made before the deadlines listed. For other reservation options, try http://www.travelbase.com

Lodging	CONALD Rates
① The University Club 123 University Place 412-621-1890 http://www.universityclub-pgh.com Valet parking Within walking distance to campus	$70.00/single Make reservations by 5/10/98

Lodging	CONALD Rates
② Holiday Inn Select University Center	$103.00/single
100 Lytton Avenue 412-682-6200 $12.00 Parking/day Shuttle service to campus Within walking distance to campus	Make reservations by 5/21/98
③ Hampton Inn	$75.00
3315 Hamlet Street 412-681-1000 Free Airline transportation to hotel Free Parking Shuttle service to campus	Make reservations by 5/21/98
④ Resnick Hall Campus Dorms	$31.59/single
Suites for 5 people, with 1 bathroom, 2 doubles & 1 single room per suite For reservations send email to CONALD@cs.cmu.edu	$23.50/double Make reservations by 4/30/98 Full payment required with reservation

Area Map

⑤ University Center, site of CONALD



Registration Form

Name: ____

Affiliation (for badge only): ______________________________________

Company or Institution: ___________________________________

Address: ____

City, State, Zip Code: ____________________________________

Country: _________________ Email: ___________________

Phone number: (   ) ___________ Fax number: (   ) ___________

The River Cruise Banquet on Friday, June 12 is included in your registration fee.

______ YES, I will be attending and will need transportation

______ YES, I will be attending but I will provide my own transportation

______ NO, I am not able to attend the River Cruise Banquet

Recreational activities are being planned for post conference on Saturday June 13.

______ YES, I am interested in participating

______ NO, I am not interested

For space considerations we need to know which workshop you'll be attending each day. Please choose only one workshop for each day.

Day One Workshops	Day Two Workshops
___ W1:Causal Bayesian Networks	___ W5: Visual Methods for Massive Data
___ W2: Mixed-Media Databases	___ W6: Learning from Text and the Web
___ W3: Machine Learning for Manufacturing	___ W7: Robot Exploration and Learning
___ W4: Large-Scale Consumer Databases	___ W8: Discovery in Natural and Social Sci

	Early Registration (by May 10, 1998)	Late Registration (after May 10, 1998)
Regular	$150	$200
Student	$90	$130
CALD Corporate Members	Free for 10 attendees	

Registration: ____________

Additional River Cruise Banquet ($50/person) ____________

Total Payment Enclosed: ____________

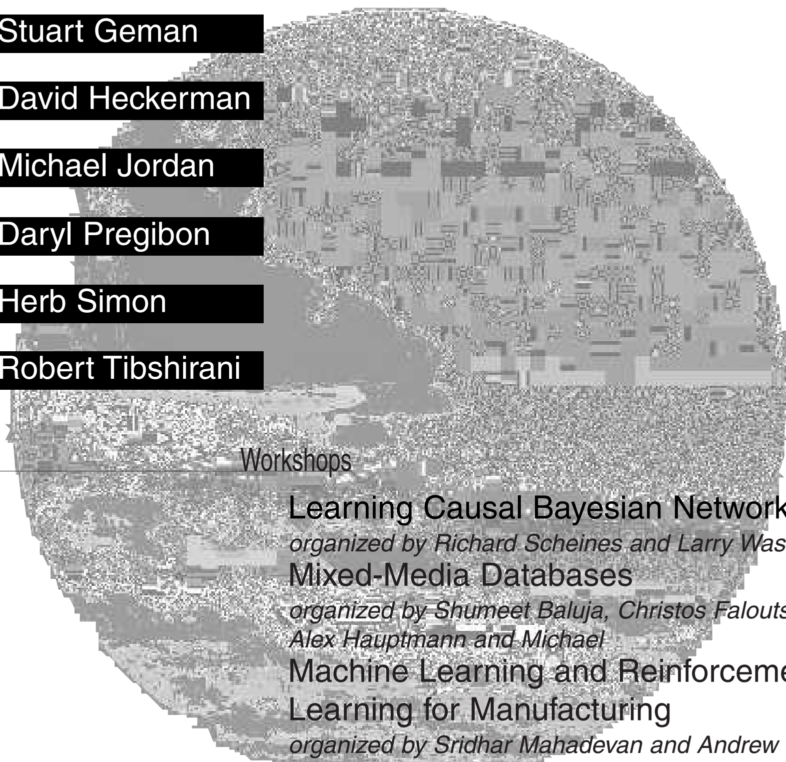
Please make checks payable to: Carnegie Mellon University

Complete and mail to this address: CALD / CONALD'98
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA 15213
USA

Travel Stipends: There are a limited number of travel stipends available. Send curriculum vitae, statement of intent, proof of full-time student status and a letter from your advisor to the above address, attention: CONALD Travel Stipend. Requests must be received by April 1, 1998.

Conference on
Automated Learning and Discovery

June 11-13, 1998 Carnegie Mellon University



Invited plenary speakers

- Tom Dietterich
- Stuart Geman
- David Heckerman
- Michael Jordan
- Daryl Pregibon
- Herb Simon
- Robert Tibshirani

Workshops

Learning Causal Bayesian Networks
organized by Richard Scheines and Larry Wasserman

Mixed-Media Databases
organized by Shumeet Baluja, Christos Faloutsos, Alex Hauptmann and Michael

Machine Learning and Reinforcement Learning for Manufacturing
organized by Sridhar Mahadevan and Andrew Moore

Large-Scale Consumer Databases
organized by Mike Meyer, Teddy Seidenfeld and Kannan Srinivasan

Visual Methods for the Study of Massive Data Sets
organized by Bill Eddy and Steve Eick

Learning from Text and the Web
organized by Jaime Carbonell, Steve Fienberg, Tom Mitchell and Yi-Ming Yang

Robot Exploration and Learning
organized by Howie Choset, Maja Mataric and Sebastian Thrun

Discovery in Natural and Social Science
organized by Raul Valdes-Perez