

DyDAN: The Center for Dynamic Data Analysis: Applications of DyDAN Research

What is Discrete Science?

Discrete science is built on a foundation of mathematical sciences, including discrete mathematics, computer science, statistics and operations research.

Discrete science deals with collections of data and information that are massive in size, widely varied in content, and constantly changing. Computer algorithms we are developing seek to identify patterns in seemingly random or unrelated data, or to identify departures from expected patterns. Our goal is to build and refine methods that provide early warning of terrorist activity, disease outbreaks, or other threats to our safety.

Methods of discrete science have become important tools for homeland security, especially when combined with powerful, modern computer methods for analysis and simulation.

Examples of DyDAN Applied Research Areas

Example 1: Intelligence Analysis

Challenge: Can we find unexpected but important patterns in text?

- We develop algorithms to sift through huge amounts of text (news stories, blogs, etc.) to automatically detect pattern changes and “significant” events.
 - Get early warning of terrorist plots
 - Help “connect the dots”
- Our methods try to decide whether new “events” are present in a flow of messages.
 - Suppose messages have been classified as corresponding to events A or B
 - Does a new message correspond to event A? Event B? Or does it (and related messages) define a new event of interest?
- We apply a variety of *machine learning* methods
 - Simplest approaches use word counts
 - More sophisticated methods “learn” how to classify better over time
 - Discrete science research is building and refining methods that allow a machine to “learn” from past data

Example 2: Disease Event Detection

Challenge: Can we reduce threats to the population from bioterrorism and from new and emerging diseases such as avian flu?

- Our algorithms analyze large data sets to detect “bioterrorist events” or “emerging diseases” (SARS, pandemic flu, etc.).
- We are developing new methods in *syndromic surveillance* that analyze data on symptoms and other early indicators to detect outbreaks earlier than would be possible with more traditional methods based on diagnoses.
- We are investigating a variety of new discrete sciences tools for disease event detection:
 - Spatial-temporal “scan statistics”
 - Statistical process control (SPC)
 - Bayesian applications
 - “Market-basket” association analysis
 - Text mining
 - Rule-based surveillance
 - Change-point techniques
 - Entropy measures

Example 3: Author Identification

Challenge: Can we identify the authors of documents in large collections of textual artifacts (e-mails, tapes, transcribed speech, etc.)?

- We are examining a variety of authorship questions and developing related algorithms:
 - Which of a set of authors wrote a document/message?
 - Were two documents written by the same author?
 - Did Osama Bin Laden write this?
- We are addressing new challenges for discrete sciences that are more complicated than in conventional text classification:
 - Large number of possible authors
 - Not much “training data”
 - Authors write on multiple topics
 - Authors write in different styles for different purposes
- We are building on classical work in the statistics literature:
 - Who wrote the disputed Federalist papers, Hamilton or Madison?
 - Who wrote Shakespeare’s plays?
- We are applying methods of machine learning that are related to those used in the text analysis example given above.

Example 4: Bioterrorism and Nuclear Sensor Location

Challenge: Can we develop strategies for sensor network management and data analysis that will provide early warning of nuclear and bioterrorist threats?

- We are addressing the ***Sensor Location Problem:***
 - Decide where to locate sensors for best protection and early warning
 - Determine what type of sensor to place at each location
- We are also studying the related ***Pattern Interpretation Problem:***
 - Analyze sensor data to decide whether an event has occurred:
 - Use data to determine whether additional monitoring is needed to reach a decision
 - Use data to pinpoint the extent and location of an attack
 - Use observations to choose an appropriate response
- We are developing new algorithmic methods that build on facility location methods of operations research and statistical methods for handling sensor errors and false alarms.

Example 5: Port of Entry Inspection for WMDs

Challenge: Can we find ways to efficiently intercept illicit nuclear materials and weapons destined for the U.S. via the maritime transportation system?

- We are developing new algorithms to determine which tests to apply as cargo arrives at U.S. ports of entry.
- Our goal is to find inspection schemes that minimize the total “cost” of inspection, which includes the “cost” of false positives, the cost of false negatives, and the cost of delays to commerce.
- We model inspection as a ***sequential decision making problem*** where a stream of containers is arriving at a port, and we apply techniques from combinatorial optimization to determine how to inspect.
 - We must decide which inspection test to perform next, based on previous results
 - If there are more than 4 types of tests, the problem rapidly becomes computationally intractable

Example 6: Protection Against Invasive Species

Challenge: Can we apply discrete sciences to help the U.S. Coast Guard identify potentially invasive species entering with cargo from abroad? (The Coast Guard is our nation’s first line of defense against invasive species.)

- We are developing methods to search and navigate various data types including images, audio, video, 3D shapes, scientific sensor data and documents.

- New methods for image analysis and efficient database searching will help address scenarios such as the following:
 - A spider is discovered in the hold of a ship arriving with a load of bananas from South America:
 - What kind of spider is it?
 - Is the spider dangerous?
 - Is it safe to unload the cargo of bananas?
- Special challenges that arise in this work include:
 - The difficulty of developing practical methods for dimension reduction in “feature-rich” image data
 - The need to search based on inexact matches

Example 7: Response to Natural Disasters

Challenge: Can we design strategies that improve responses to natural disasters such as hurricanes, floods, earthquakes?

- We apply methods of discrete sciences to assure better planning prior to a disaster and an improved response as events unfold.
- Sample applications include:
 - Location and allocation of key resources prior to an emergency situation
 - Redistribution of resources in response to emergency situations
 - Evacuation planning and execution
- Relevant discrete science methods include:
 - Optimization under uncertainty, including stochastic optimization and approximate dynamic programming
 - Traffic analysis
 - Situational awareness as an aid to emergency management

Example 8: Infrastructure Protection and Risk Mitigation

Challenge: How do we invest available funds to yield the greatest reduction in risk to the fundamental infrastructure of our society (bridges, buildings, economic systems, communication systems, transportation systems, power plants, etc.)?

- We must capture assessments of vulnerability, threat, and consequences (human, economic, psychological, etc.) of attack on our key infrastructure.

- Given these assessments and the multitude of potential uses for homeland security funds, we must decide how to allocate funds to achieve the greatest reduction in risk.
- We apply a variety of methods from the “decision sciences” that include classical risk assessment, assessment of probabilities, “portfolio analysis”, and fundamental principles of measurement and scaling – all topics studied by discrete scientists.

Example 9: Privacy-Preserving Data Analysis

Challenge: How do we gather and share information while protecting privacy?

- Privacy issues arise at all phases of data analysis, including collection, preprocessing, modeling, and testing.
- Sample problems include:
 - How can we share data between different agencies while protecting privacy?
 - How can we make a simultaneous query to two datasets without compromising information in those data sets? (E.g., is individual xx included in both sets?)
- Our work builds on approaches under development in the computer science community that include “Secure Multiparty Computation” and “Privacy-preserving Data Analysis.”

Privacy-preservation underlies all the research at DyDAn.