An Introduction to Scientific Data Mining

Chandrika Kamath
Center for Applied Scientific Computing
Lawrence Livermore National Laboratory
http://www.llnl.gov/casc/people/kamath

January 14, 2002

Overview of tutorial

- Motivation for data mining
- Introduction to the data mining process
- Examples of data mining in science and engineering
- Common themes and issues in scientific data mining
- Pre-processing data for mining
- Challenges and opportunities

Advances in technology enable us to collect ever increasing amounts of data

- Scientific data can be obtained from
  - experiments
  - observations
  - simulations
- Collection of data made possible by advances in
  - sensors (telescopes, satellites,…)
  - computers (faster, more memory, parallel,…)
  - storage (disks, tapes,…)

➔ We need fast and accurate data analysis techniques to realize the full potential of our enhanced data collecting abilities.

Manual data exploration techniques are not suitable for massive data sets

1 Terabyte = 750,000 floppies
= 300 million pages of text
= 100,000 medical X-rays
= 250 movies

1 Petabyte = 1024 TB

20-25 Terabytes !

➔ Visual data analysis for moderate-sized data is impractical given its subjective nature and human limitations in absorbing detail - it is impossible for massive data sets.

Science data sets are not only massive but are also very complex

- Multi-sensor, multi-spectral, multi-resolution data
- Spatio-temporal data
- High-dimensional data
- Mesh data from simulations
  - structured and unstructured meshes
- Data contaminated with noise
  - sensor noise, clouds, atmospheric turbulence,...

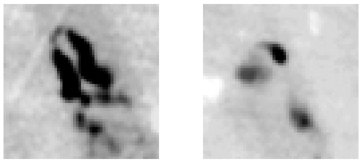
➔ We need something better than the traditional data analysis techniques for science and engineering data.

Overview of tutorial

- Motivation for data mining
- Introduction to the data mining process
- Examples of data mining in science and engineering
- Common themes and issues in scientific data mining
- Pre-processing data for mining
- Challenges and opportunities

MIT’s Technology Review (Jan’01) - data mining is a ‘top ten’ emerging technology

- Data mining: The semi-automatic discovery of patterns, associations, anomalies, and statistically significant structures in data
- Pattern recognition: The discovery and characterization of patterns
- Pattern: An ordering with an underlying structure
- Feature: Extractable measurement or attribute



FIRST images (sundog.stsci.edu)

Pattern: Radio galaxy with a bent-double morphology
Features: Number of “blobs”
Maximum intensity in a blob
Spatial relationship between blobs (distances and angles)

Scientific data mining brings together work being done in several disciplines

- Artificial intelligence, Machine learning
- Computer vision
- High performance computing
- Image understanding
- Mathematical optimization
- Pattern recognition
- Electrical engineering
- Statistics
- ….

➔ **Data mining brings together the mature offshoots of technologies at a time when we are ready to exploit them.**

I limit the scope of what is data mining and what I will discuss in this tutorial

- Data mining is not (Thearling '97)
  - data warehousing
  - ad-hoc query and reporting
  - on-line analytic processing (OLAP)
  - data visualization
  - software agents
- In this tutorial, I will not discuss issues related to
  - collecting, storing, or accessing data, though they may form part of the data mining infrastructure
  - parallel or distributed data mining techniques, though they may be essential for massive datasets

Overview of tutorial

- Motivation for data mining
- Introduction to the data mining process
- Examples of data mining in science and engineering
- Common themes and issues in scientific data mining
- Pre-processing data for mining
- Challenges and opportunities

Data mining is being applied to problems in several scientific domains

- Often complements existing data analysis techniques
  - statistics
  - exploratory data analysis
  - domain-specific techniques
- Techniques developed in the context of one domain can easily be applied or extended to another domain

➔ The diversity of scientific applications provides a rich environment for the practice of data mining.

First, a few caveats ….

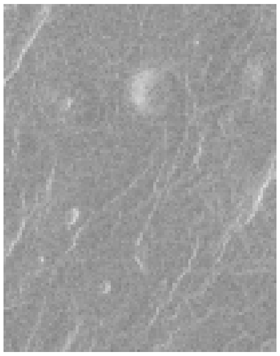
- Brief overview, not a comprehensive survey
- Focus is on the breadth, not the depth
- Intent is to highlight the diversity of applications and identify the similarities
- Some overlap in the applications, so a problem may appear in different application domains

Data mining in astronomy and astrophysics: “look up” data

- Astronomers have long used data analysis techniques
  - FOCAS: Faint object classification and analysis system (Jarvis/Tyson, ‘81)
  - star/galaxy discrimination using neural networks (Odewahn ‘92)
  - morphological classification of galaxies using neural networks (Storrie-Lombardi ‘92)
- Data can be obtained from observations and simulations

Data miners have found astronomy to be a rich source of problems

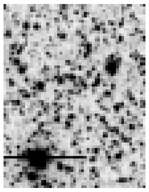
- SKICAT: star/galaxy classification using decision trees (Fayyad ‘96)
- JARTool: detecting volcanoes on Venus (Burl ‘98)
- Diamond Eye: find, analyze, and catalog spatial objects (Burl ‘01)
- Sapphire: identifying useful information in scientific data e.g. bent-double galaxies in the FIRST survey (Kamath ‘01)



JARTool

Characteristics of astronomy data

- Large size: MACHO (8 TB), SDSS (15TB)
- Usually not in databases
- Observations at different wavelengths
- Variable quality of the data
- Uncertainty of measurement
- Noisy data, missing values
- No ground truth
- Can have a temporal aspect
- Relatively easily accessible: National Virtual Observatory

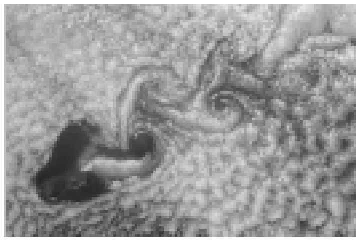


MACHO

FIRST

Data mining in remote sensing: “look down” data

- A very rich source of data
- Long history of data analysis techniques
- Diverse set of applications
  - oceanography and marine resources
  - mineral and oil resources
  - land use and mapping
  - geology
  - water resources
  - environment
  - agriculture and forestry



Atmospheric vortices near Guadalupe Island (http://eos.nasa.gov)

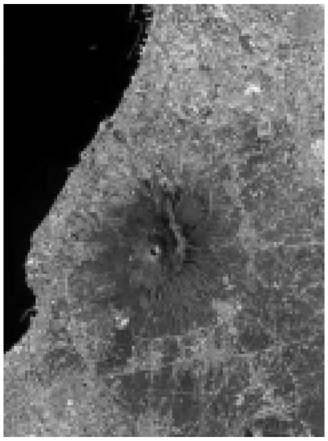
Several projects that currently collect and mine remote sensing data

- IKONOS 1m resolution (www.spaceimaging.com)
- Earth Observing System (eospso.gsfc.nasa.gov)
- NASA Goddard Earth science enterprise (www.earth.nasa.gov)
- ADAM (Algorithm development and mining) (www.itsc.uah.edu)
- Various efforts in Geographic Information Systems (GIS)
- State and national agencies

Jefferson Memorial
www.spaceimaging.com

Characteristics of remote sensing data

- Multi-sensor, multi-resolution, and multi-spectral
- Spatial and temporal aspect
- Noisy
- Data fusion needed
- Comparisons important
- MASSIVE! (EOS = 11000 TB)



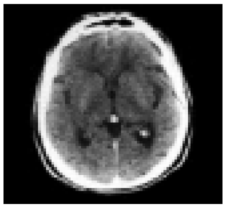
Mount Vesuvius (http://eos.nasa.gov)

Data mining in biology: bio-informatics

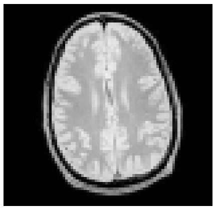
- Bioinformatics
  - bridge between biology and information technology
  - analysis of gene sequences, understanding higher order structure of proteins ….
  - Use neural networks, hidden Markov models,….
  - human genome effort (Venter et. al in Science 2001, also article in Nature Feb., 2001)

Data mining in biology: medical imaging

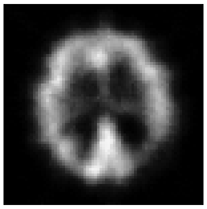
- Different types of data
  - MRI and PET scans, mammograms, ultrasound, …
  - protein crystallography, DNA microarrays, ….
- Different tasks
  - identifying tumors, detecting changes, …
  - protein structure, genomics, …



CT scan



MRI scan



SPECT scan

Characteristics of biology data

- Genomics data is often in databases
  - infrastructure issues
  - integration of databases
  - flexible access to the data
  - data changing and being updated
- Image data
  - can be noisy, with unclear features
  - need image registration
  - can be 3-dimensional
  - possible privacy issues

Data mining in chemistry

- Data analysis techniques being used to
  - analyze molecular patterns
  - identify relationships between compounds
  - drug discovery
- Sources of data
  - computer simulations
  - combinatorial chemistry: react a set of starting chemicals in all possible combinations

➔ The potential payoff for success is enormous!

Data mining in non-destructive evaluation

- Used to study the inside of an object without affecting it
  - contents might be dangerous
  - more cost-effective
- Several applications
  - bridge inspection
  - land mine detection
  - materials characterization
  - flaw/damage in components

Before

After

Data mining in security and surveillance

- Several applications
  - human face recognition
  - signature recognition
  - military applications
  - automated target recognition
  - fingerprint/retinal identification

Original

Compressed 26:1

- Characteristics of the application
  - real time turnaround
  - security and privacy issues
  - massive amounts of data (FBI: 200 million fingerprint cards at 10MB each=200TB)

Data mining in high energy physics experiments (www.star.bnl.gov)

- Accelerating sub-atomic particles to nearly the speed of light and forcing their collision
- A few particles collide and produce a large number of additional particles
- Interested in special events and signatures of particles
- Each collision (event) generates 1-10 MB of raw data
- $10^7 - 10^8$ events/year = 300 TB/year
- An experiment may run for 3 years
- Process the data to extract 100-200 summary elements (features) for each event

➔ Science data can be high dimensional.

Data mining in fields using computer simulations

- Computer simulations - the third mode of science complementing theory and experiment
- Understand complex phenomena by analyzing mathematical models on high performance computers
- Qualitative and quantitative insights into phenomena
  - too complex to be solved by analytical methods
  - too expensive, impractical, or dangerous to study using experiments
- Provide results of comparable accuracy to experiments and fill the gap between analytical approaches and physical experiments

Computer simulations are used in several scientific and engineering fields

- Astrophysics - modeling how stars evolve
- Computational fluid dynamics - flow around an airplane, understanding turbulence
- Combustion - interaction between turbulent flow fields and chemical reactions
- Structural mechanics - car crash tests, stability of structures such as bridges
- Climate - modeling El Niño and global warming
- Chemical engineering - understanding pharmaceutical processes, studying mixtures of chemicals

A *very simplistic* introduction to computer simulations

- Partial differential equations (PDE) form a cornerstone of numerical simulations
- Physical phenomena (fluid flow, heat transfer) depend in complex ways on space and time
- Interesting physical phenomena usually arise from the nonlinear interactions of different length and time scales
- Use fundamental principles (conservation of mass, energy, momentum) to create a mathematical model

A *very simplistic* introduction to computer simulations (contd.)

- Model will typically have several independent and dependent variables

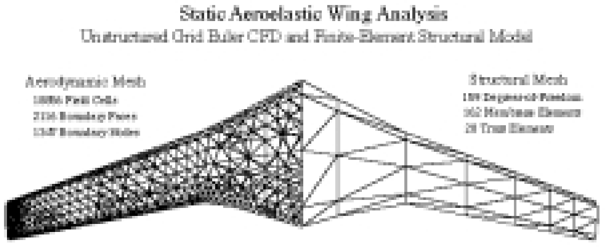
Linear, variable coefficient PDE: $u_{xx} + 3x^2 u_{xy} + u_{yy} + u_x - u = 0$

Nonlinear PDE: $u_{xx} + 3u_{xy} + u_{yy} - u_x^2 - u = e^{x-y}$

- In addition to the PDE, we need
  – the region of space and time on which the PDE must be satisfied
  – the boundary and initial conditions that must be met

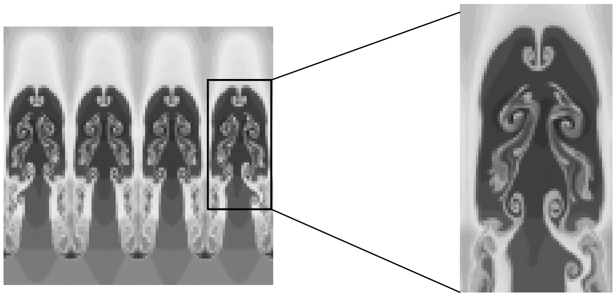
How do we go from a PDE to a computer simulation?

- The region of space and time must be “discretized”
- Solve the PDE at the grid or mesh points through techniques such as finite elements, finite differences,...



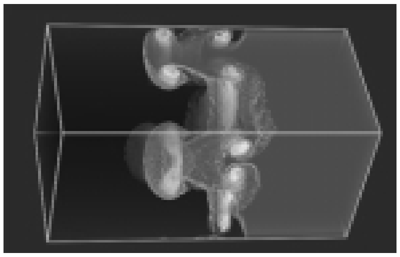
Characteristics of complex computer simulations

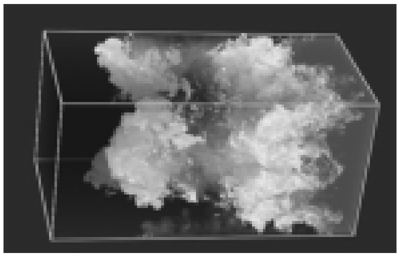
- Run for days/weeks on a massively parallel system
- Produce terabytes of output at each time step
- Typically store output for analysis later on
- Often model a well understood physical phenomena (important exception: turbulence)



Details on one of the largest simulations run on an ASCI machine at LLNL

- Mesh size: 2048 x 2048 x 1920
- 960 nodes of the IBM system
- 27,000 time steps
- 173 hours of machine time, 226 hours of wall clock time
- 3 TB of graphics data, spread over 275,000 files



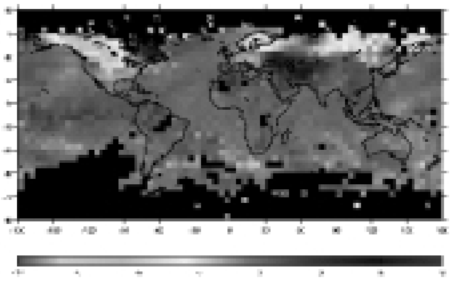


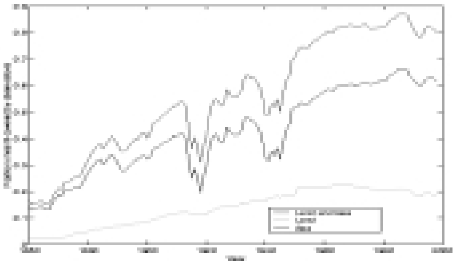
Applications of data mining in computer simulations

- Analysis of simulation output
  - data mining can complement visualization
- Identification of coherent structures in turbulence
- Understanding the design parameter space
  - dependence of output data on input parameters
  - “mining the minds of the experts”
- Verification and validation
  - comparison between simulations
  - comparison of experiment to simulation
- Refining the physics model

Data mining in atmospheric sciences

- Climate simulations - understand weather, model effects of volcanoes and El Niño, ….
- Combine simulations with observations
- Observations can be missing in space and time





Data mining, computer vision, and robotics

- Similarities between techniques used in data mining and those in computer vision and robotics
- Several applications
  - industrial tasks: detecting errors in widgets on assembly line, or semiconductor masks
  - tracking eye movement, or gestures
  - medical imaging in surgery
  - robot motion control

Overview of tutorial

- Motivation for data mining
- Introduction to the data mining process
- Examples of data mining in science and engineering
- Common themes and issues in scientific data mining
- Pre-processing data for mining
- Challenges and opportunities

Common themes across science and engineering data sets

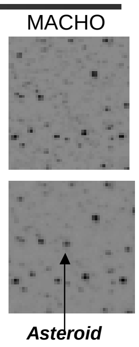
- Data in the form of images or meshes, not features
- Spatial and temporal aspect
- Sizes ranging from gigabytes to terabytes, petabytes, and beyond…
- Desire to exploit data from different sources
- Comparisons are important
- Different data formats and data output options even within a single domain
- Often see structure at different scales
- Can be noisy with missing values
- High dimensional
- Data may be compressed

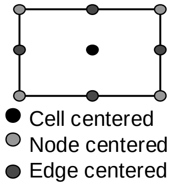
Science data comes in different types

- Different storage formats in an application area
  - FITS, AIPS in astronomy
  - netCDF, GRIB (grid in binary) in climate
- Different ways of generating output
  - sea surface temperatures for each month in a file
  - sea surface temperatures for each year in a file
- Depending on the problem, data can be
  - one-dimensional, usually time series, from sensors or processing of other data
  - two-dimensional (spatial) + time
  - three-dimensional (spatial) + time

Two-dimensional scientific data is available as images or as meshes

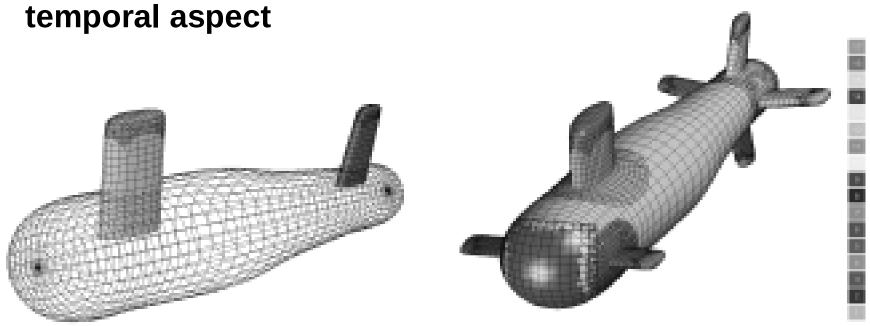
- Typically have spatial and temporal aspects
- Images
  - pixel values can be gray-scale or real
  - images of a scene obtained using different sensors, at different times, at different resolutions
  - images can be noisy, with noise varying from image to image and within an image
- Mesh
  - values at a mesh point are real
  - values can be "cell centered", "node centered" or "edge centered"





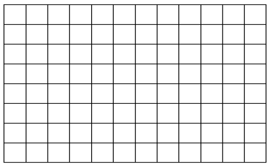
Three dimensional scientific data comes from modeling objects in 3-D

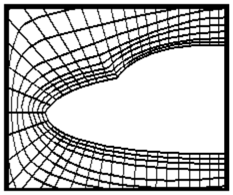
- Values at a mesh point are real
- Values can be "cell centered", "node centered", "edge centered", or "face centered"
- Often have a series of meshes in time: spatial and temporal aspect



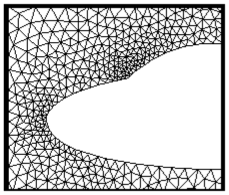
The complexity of meshes makes it difficult to extract features

Cartesian Structured

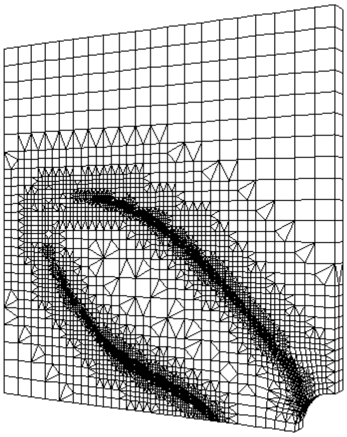




Structured

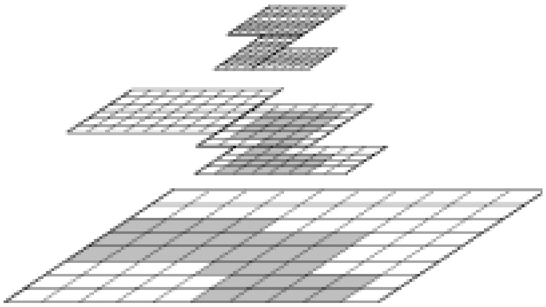


Unstructured

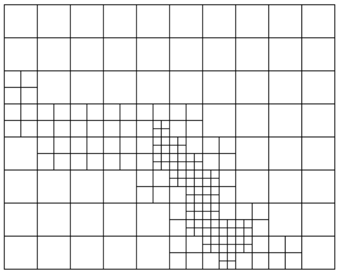


Unstructured

The distribution of mesh points can change with time - need feature tracking



Hierarchy of regular meshes



Composed
‘Unstructured’ mesh

Composite meshes - locally structured, globally unstructured

Science data is not often in a form ready for pattern recognition

- Data available as pixels or variables at mesh points
- But, patterns (e.g. bent doubles) are at a higher level

↓

The raw data must be transformed into features before we can apply pattern recognition.

Extracting features that are robust, relevant to the problem, and invariant to scaling, rotation, and translation is non-trivial and time consuming - but, essential to the success of the pattern recognition algorithm.

Most of the work in data mining focuses on pattern recognition, BUT….

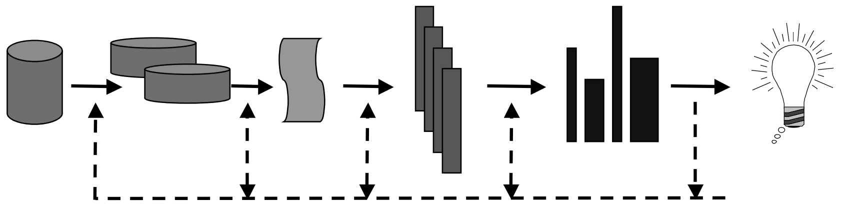
- … it is the data pre-processing which is
  - more influential and time consuming
  - domain specific and therefore less general
- “perhaps as little as 10% effort was spent on classification aspects of the problem.” (Burl ‘98)
- Langley/Simon ‘95: “… much of the power comes not from the specific induction method, but from proper formulation of the problems and from crafting the representation to make learning tractable.”
- Brodley/Smyth ‘95: “… in practical applications, it is often the data and human issues which ultimately dictate success or failure of a project rather than algorithmic and model issues.”

Overview of tutorial

- Motivation for data mining
- Introduction to the data mining process
- Examples of data mining in science and engineering
- Common themes and issues in scientific data mining
- Pre-processing data for mining
- Challenges and opportunities

The Sapphire view of data mining - from a Terabyte to a Megabyte

Raw Data **Target Data** **Preprocessed Data** **Transformed Data** **Patterns** **Knowledge**



← Data Preprocessing → **Pattern Recognition** **Interpreting Results**

Data Fusion
Sampling
Multi-resolution analysis

De-noising
Object-identification
Feature-extraction
Normalization

Dimension-reduction

Classification
Clustering
Regression

Visualization
Validation

An iterative and interactive process

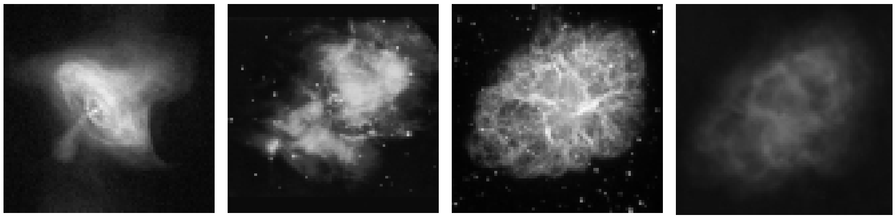
Let’s make a few ‘simple’ assumptions in our discussion of data preparation ….

- We understand the problem and the data
- We have formulated a solution approach
- We have relatively easy access to the data
- We have the software to read, write, and display the data
- We have the software to bring the data into a consistent format

➔To satisfy these ‘simple’ assumptions may require far more time than you expect!

Data fusion may be necessary when data from many sources is available

- Combining information from more than one source to make a more accurate and better informed decision
- Exploit complementary information from different sensors, at different wavelengths, from different viewpoints,….



X-ray Infrared Optical Radio

Images of the Crab Nebula from chandra.harvard.edu

Data registration is an important part of data fusion

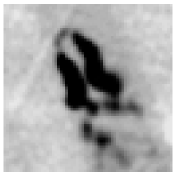
- Obtain a global or local transformation to relate information in one image to information in another image
- Used in data fusion and change detection
- Four major components of data registration
  - feature space
  - search space
  - search strategy
  - similarity metric
- Recent work
  - an excellent survey: Brown 92.
  - wavelet-based multi-resolution techniques
  - evolutionary algorithms as a search strategy
  - Levenberg-Marquardt optimization strategy

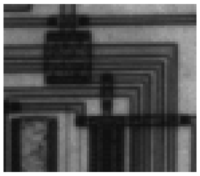
The data may need to be de-noised to better identify the objects

- Noise in the data can be due to the data acquisition process or natural phenomena such as atmospheric turbulence
- De-noising is difficult as cannot always tell what is the signal and what is the noise
- Various techniques
  - spatial filters
  - simple thresholding
  - wavelet-based thresholding
  - non-linear isotropic and anisotropic diffusion

Once the data has been de-noised, we need to identify the "objects" in it

- Identifying the objects is non-trivial
  - tremendous variability of object shapes: man-made vs. natural objects
  - denoising may have smoothed the edges
  - variations in image quality (noise, boundary gaps)





Identifying objects in data is difficult, both in 2 and 3-D images and meshes

- Challenges in traditional image algorithms
  - need many parameters for optimal performance
  - interactions between parameters are complex and non-linear
  - no universally accepted measure of quality of the segmented image
  - no single method can handle variations between images
- Identifying "objects" in mesh data
  - mesh may move/change over time
  - in two/three spatial dimensions + time
  - irregular meshes
  - "objects" may split or merge

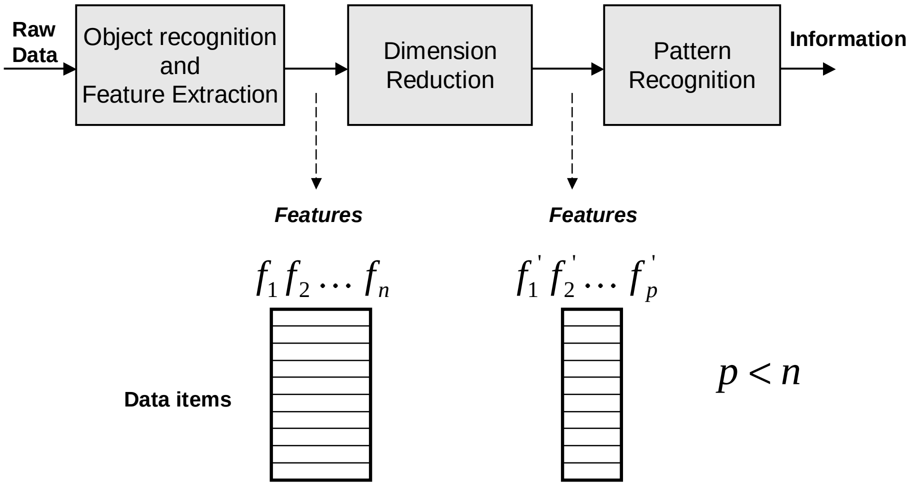
Several techniques are being used in the image processing community

- Thresholding using the image histogram
- Segmentation techniques
  - split and merge (top-down)
  - region growing (bottom-up)
- Edge detection: use a filter to identify an edge
- Combine traditional techniques with evolutionary algorithms to make them more adaptive
- Deformable models for segmentation
  - parametric approach: snakes or active contours
  - geometric approach: level set methods

Once the objects have been identified, the features must be extracted

- Features dependent on the problem
  - identifying relevant features
  - extracting robust features
  - extracting features invariant to scale, rotation, and translation
- Features may include
  - distances, angles, areas
  - histograms
  - fourier or wavelet coefficients
  - various moments
  - ….

May need to reduce the dimension or the number of features



There are several reasons why dimension reduction may be helpful

- **Fewer features may make pattern recognition algorithms computationally tractable**
- **Less time is spent in extracting features**
- **Can minimize correlations between features, which may be a requirement of some algorithms (e.g. GLMs)**
- **Dimension reduction techniques**
  - **exploratory data analysis**
  - **principal component analysis**
  - **independent component analysis**
  - **….**

Other issues that differentiate science data from its commercial counterpart

- **Science data is rarely in databases**
  - **genomics data is an exception**
  - **meta-data may be in a database**
- **Privacy and security issues not always a major concern**
  - **exceptions: security & surveillance, medical data**
  - **astronomy: data frequently on the web**
- **Shortage of labeled data**
  - **generated manually**
  - **"labeled" vs. "interesting" data**
- **Real time turnaround in some cases**
  - **medical applications**
  - **observe interesting phenomena as it occurs**

Overview of tutorial

- **Motivation for data mining**
- **Introduction to the data mining process**
- **Examples of data mining in science and engineering**
- **Common themes and issues in scientific data mining**
- **Pre-processing data for mining**
- **Challenges and opportunities**

Scientific data mining: opportunities abound!

- Mining data sets which are
  - massive (petabytes)
  - spatio-temporal
  - multi-scale
  - multi-sensor
  - multi-dimensional
  - ….
- Data mining techniques are being applied in new areas

➔ The diversity of applications, the richness of problems faced by practitioners, and the opportunity to borrow ideas from other fields make scientific data mining an exciting and challenging field!

Data mining techniques have the potential to solve a problem that has vexed scientists in the last few decades, namely, the sheer size and complexity of their data has resulted in a loss of serendipitous discoveries that were vital to scientific progress in the past.

You can be part of the solution!

Acknowledgements

- The Sapphire project team: Erick Cantú-Paz, Imola K. Fodor, and Nu Ai Tang
- Sisira Weeratunga (LLNL) for insights on simulations and PDEs

http://www.llnl.gov/casc/sapphire

UCRL-PRES-146653: This work was performed under the auspices of the U.S. Department of Energy by the University of California Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48

References

Jarvis, J. and J. Tyson, “FOCAS: Faint Object Classification and Analysis System”, The Astronomical Journal, Volume 86, Number 3, pages 476-495, 1981

Odewahn, S., E. Stockwell, R. Pennington, R. Humphreys, and W. Zumach, “Automated Star/Galaxy discrimination with neural networks”, The Astronomical Journal, Volume 103, Number 1, pages 318-331, 1992.

Storrie-Lombardi, M., O. Lahav, L. Sodre, and L. Storrie-Lombardi, “Morphological classification of galaxies by artificial neural networks”, Mon. Not. R. Astron. Soc., Volume 259, pages 8-12, 1992.

Fayyad, U., P. Smyth, M. Burl, and P. Perona, “A learning approach to objet recognition: applications in science image analysis”, In S. Nayar and T. Poggio (Eds.), *Early Visual Learning,* Oxford University Press, New York.

References

Burl, M., L. Asker, P. Smyth, U. Fayyad, P. Perona, L. Crumpler, and J. Aubele, " Learning to recognize volcanoes on Venus", Machine Learning, Volume 30, pages 165-195, 1998.

JARTool web page: http://www-aig.jpl.nasa.gov/public/mls/mgn-sar/jartool-home.html

Burl, M., Diamond Eye web page: http://www-aig.jpl.nasa.gov/public/mls/diamond_eye

Kamath, C., E. Cantú-Paz, I. K. Fodor, and N. Tang, "Searching for bent-double galaxies in the first survey", in ***Data Mining for Scientific and Engineering Applications***, R. Grossman, C. Kamath, W. Kegelmeyer, V. Kumar, and R. Namburu (eds.), Kluwer 2001.

Sapphire project web page at http://www.llnl.gov/casc/sapphire

References

FIRST web page: sundog.stsci.edu. Includes papers, access to the catalog, and image cutouts

MACHO web page: http://wwwmacho.anu.edu.au/

SDSS web page: http://www.sdss.org/

National Virtual Observatory web page: http://www.srl.caltech.edu/nvo/

HERMES web page: Bridge Inspection Technology for the 21-st century, http://lasers.llnl.gov/lasers/hermes

Venter C. et. al, "The sequence of the human genome", Science, Feb. 16, 2001. http://www.sciencemag.org/content/vol291/issue5507/index.shtml

The Genome Sequencing Consortium, "Initial Sequencing and Analysis of the Human Genome", Nature, February 15, 2001.

References

Langley, P. and H. A. Simon, “Applications of machine learning and rule induction”, Communications of the ACM, Volume 38, Number 11, pages 55-64.

Brodley, C. and P. Smyth, “The process of applying machine learning algorithms”, Workshop on applying machine learning in practice, IMLC 1995 (http://citeseer.nj.nec.com/722.html)

Kamath, C., E. Cantú-Paz, I. K. Fodor, and N. Tang, “Searching for bent-double galaxies in the first survey”, in ***Data Mining for Scientific and Engineering Applications***, R. Grossman, C. Kamath, W. Kegelmeyer, V. Kumar, and R. Namburu (eds.), Kluwer 2001.

Brown, L. “ A Survey of Image Registration Techniques”. ACM Computing Surveys, Vol. 24, Number 4, December 1992.

Le Moigne, J., “Parallel Registration of Multi-sensor remotely senses imagery using wavelet coefficients”, Proc. SPIE Wavelet Applications Conference, Orlando, 1994, pages 423-443.

References

Mandava, V., Fitzpatrick, J., and Pickens, D. (1989). Adaptive search space scaling in digital image registration. IEEE Transactions on Medical Imaging, 8, 251-262.

Thevenaz, P., Ruttimann, U., Unser, M., “A Pyramid Approach to Sub-pixel Registration based on intensity”, IEEE Transactions on Image Processing, Vol 7, Number 1, January 1998.