Mining Science Data

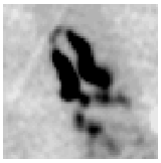
Chandrika Kamath
Lawrence Livermore National Laboratory

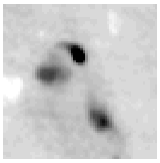
SIAM Conference on Computational Science and Engineering
February 23, 2007

www.llnl.gov/casc/sapphire

Data mining terminology

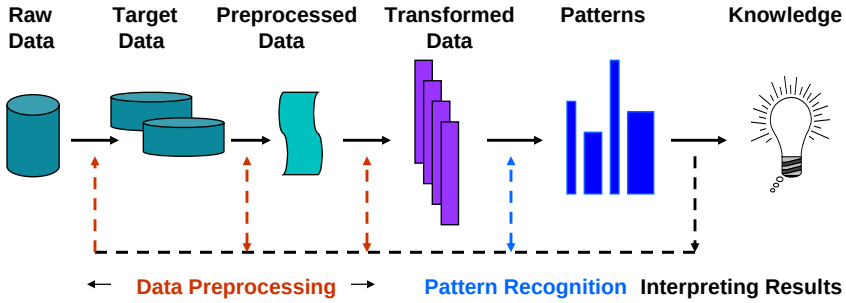
- **Data mining: the semi-automatic discovery of patterns, associations, anomalies, and statistically significant structures in data**
- **Pattern recognition: the discovery and characterization of patterns**
- **Pattern: an ordering with an underlying structure**
- **Feature: extractable measurement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





Pattern: radio galaxy with a bent-double morphology
Features: number of "blobs"
maximum intensity in a blob
spatial relationship between blobs

Large-scale data mining - from a Terabyte to a Megabyte



Data Fusion
Sampling
Multi-resolution analysis

De-noising
Object - identification
Feature-extraction
Normalization

Dimension-reduction

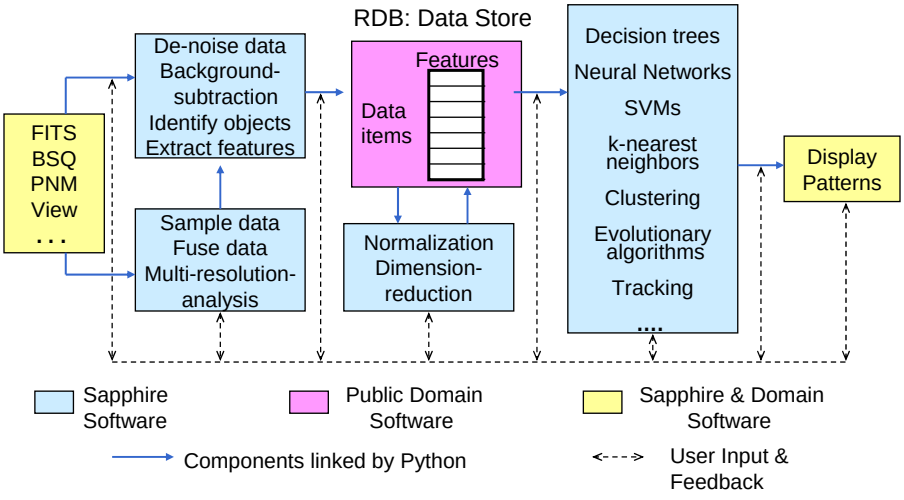
Classification
Clustering
Regression

Visualization
Validation

An iterative and interactive process

CASC

The Sapphire system architecture: flexible, portable, scalable



US Patents 6675164 (1/04), 6859804 (2/05), 6879729 (4/05), 6938049 (8/05), 7007035 (2/06), 7062504 (6/06)

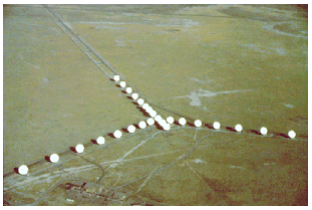
CASC

Classification of Bent-double Galaxies in the FIRST Survey

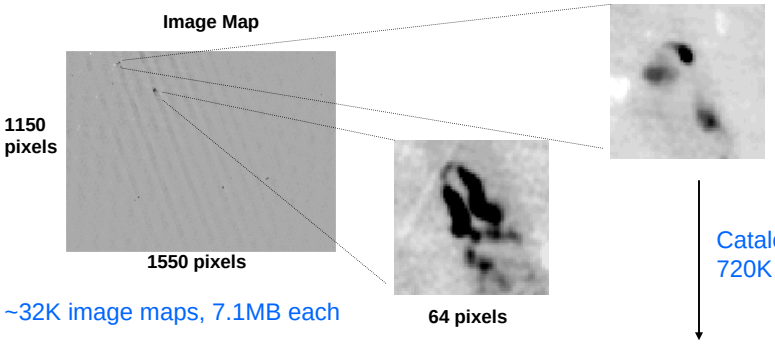
Joint work with FIRST astronomers: Bob Becker, Michael Gregg, Sally Laurent-Muehleisen (LLNL), and Rick White (STScI)

Classifying radio-emitting galaxies with a bent-double morphology

- Faint Images of the Radio Sky at Twenty cm (FIRST)
- Using the NRAO Very Large Array, B configuration
- 10,000 square degrees survey, ~90 radio galaxies / square-degree
- 1.8'' pixels, resolution 5'', rms 0.15mJy
- Images maps and catalog available



FIRST data set: Detecting bent-doubles in 250GB image data, 78MB catalog data

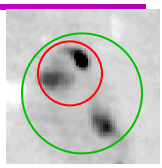


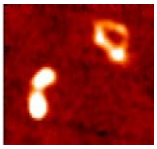
Catalog entry

Radio Galaxy

RA	DEC	Peak Flux (mJy/bm)	Major Axis (arcsec)	Minor Axis (arcsec)	Position Angle (degrees)
00 56 25	-01 15 43	25.38	7.39	2.23	37.9
00 56 26	-01 15 57	5.50	18.30	14.29	94.2
00 56 24	-01 16 31	6.44	19.34	10.19	39.8

Our approach for classifying radio-galaxies using feature from the catalog





- Consider a region of interest
- Group catalog entries within the ROI
- Separate galaxies
  - 1 entry: unlikely to be bent-doubles
  - > 3-entry: all “interesting”
  - classify 2- and 3-entry galaxies separately
  - a small training set becomes smaller (313 ---> 118 + 195)
- Extract features for the 2- and 3-entry galaxies
- Create a decision tree using the training set
- Use the tree to classify the unlabeled galaxies

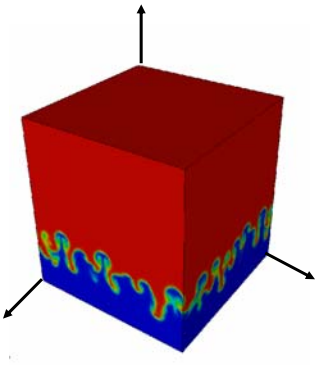
➔ The locations of likely bent-double galaxies were given to the astronomers for further observations

Analysis of Bubbles and Spikes in Rayleigh-Taylor Instability

Joint work with Paul Miller (LLNL)

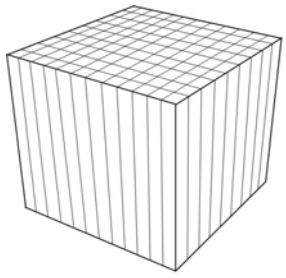
Goal: use image analysis to characterize and track bubbles and spikes

- Two high-fidelity simulations of the Rayleigh-Taylor instability
  - Atwood number: 0.5
- Goals of the analysis
  - bubble counts
  - bubble sizes
  - distances between bubbles
  - bubble dynamics



The data is obtained from the Miranda code on a 3-D regular Cartesian grid

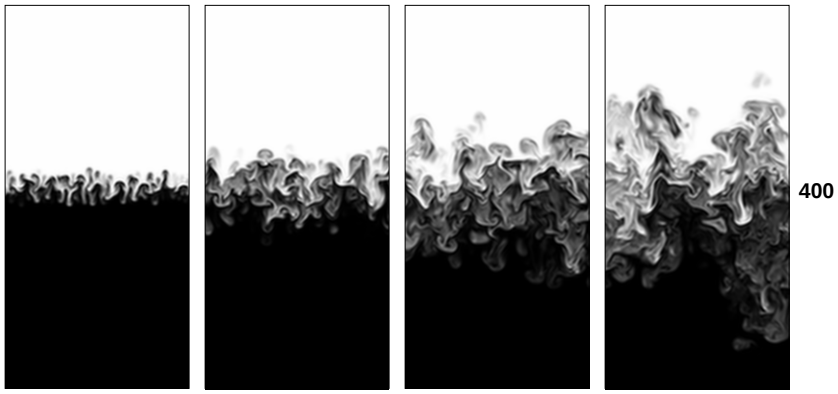
- LES simulation*
  - 1152**3 grid points
  - 7 variables per grid point
  - 758 time steps
  - 30TB analysis data
- DNS simulation**
  - 3072**3 grid points
  - 5 variables per grid point
  - 249 time steps
  - 80TB analysis data



* *Cook, Cabot, and Miller, Journal of Fluid Mechanics, 511, 2004.*

** *Cabot and Cook, Nature Physics, 2, 2006.*

The first step is to define a bubble…

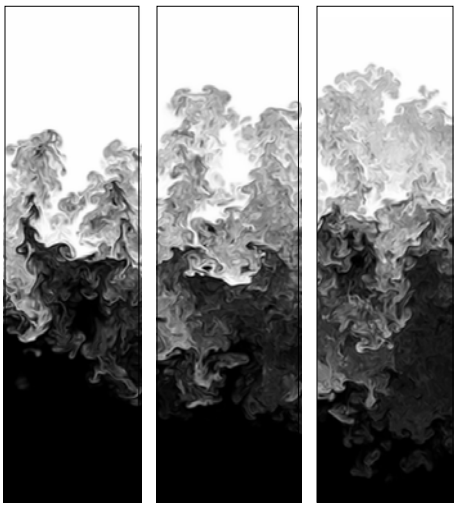


A slice through the density variable: LES data at time steps 100, 200, 300, 400

Convention: Smaller values are darker in image.

… which can be a challenge, especially at the later time steps

Density variable, LES data
time steps 500, 600, 700



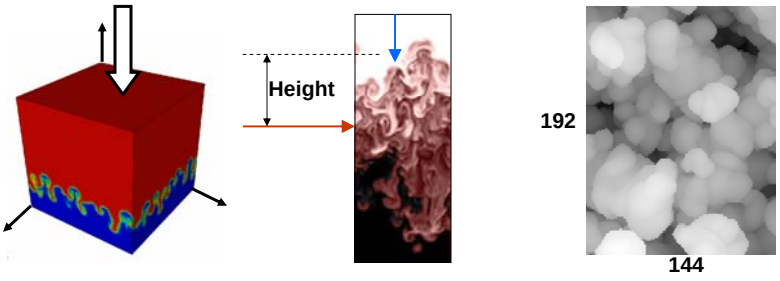
There are several challenges to the analysis

- Lack of a precise definition of a bubble
  - range of scales of the structures of interest
- Massive size of the data
  - distributed nature of output at each time step
- Requirements of the analysis algorithms
  - low computational cost
  - applicable to distributed data
  - few parameters
  - relatively insensitive to choice of parameters
  - a single algorithm and parameters for all time steps
  - multiple algorithms for verification

Our solution approach: refine bubble definition on a small subset of the data

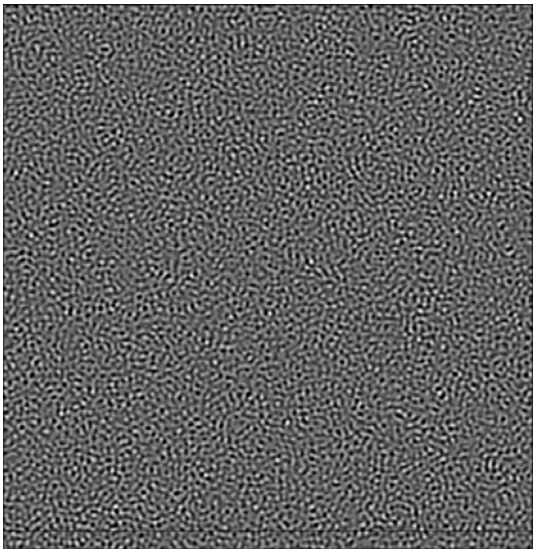
- Start with a small subset (0.04%) of the full data
  - LES data, 6x6 columns (= 144x192x1152)
  - every 50-th time step
- Consider all variables
- Attempt a definition of a bubble
- Iterate to determine parameter settings
  - evaluate on 6x6 columns, every 50-th time step
  - evaluate on full data, every 50-th time step
- Apply to full data – only once

We use the bubble height to generate a 2-D image: the height-depth map (HDM)

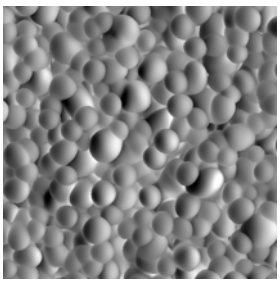


Original fluid interface

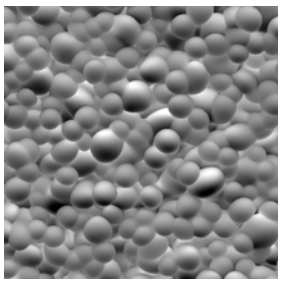
Movie 2: Bubble height map for the DNS data (3Kx3K sub-sampled to 1Kx1K)



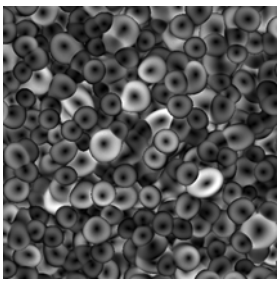
Bubble counting – the mag-X-Y velocity (DNS, time step 50)



X velocity

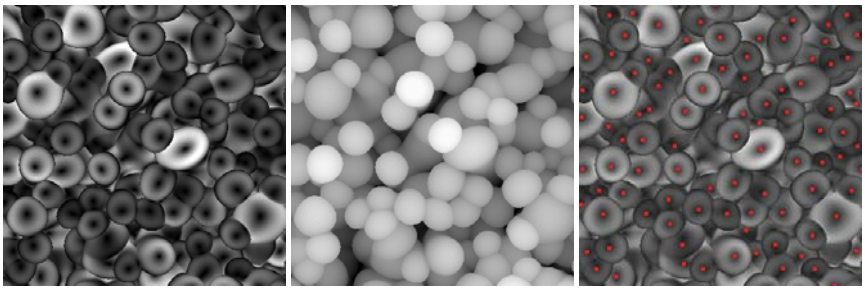


Y velocity



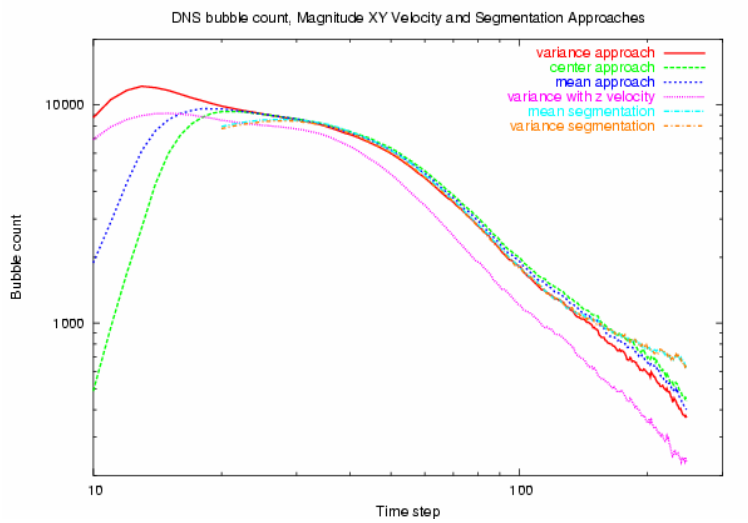
Mag X-Y velocity

Bubble counting – identifying the bubble tips

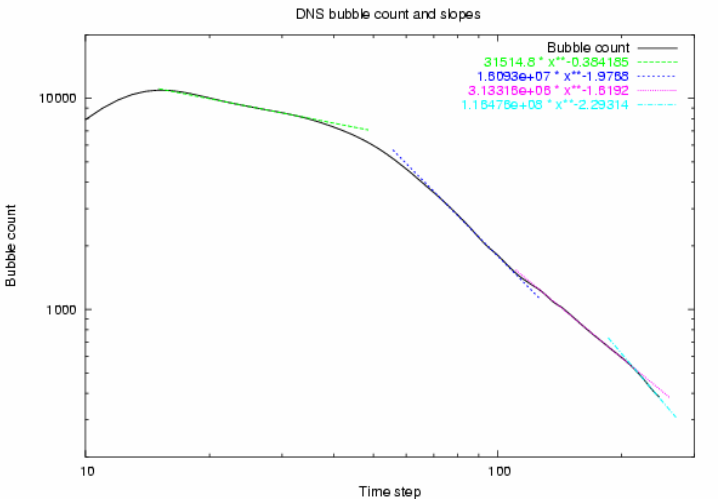


Mag X-Y velocity | Height-depth map | Bubble tips

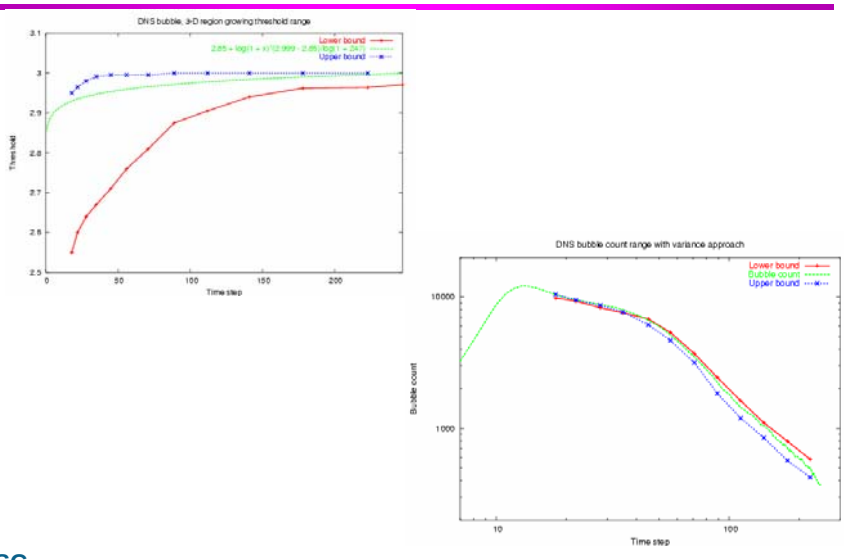
Bubble counts over time: DNS data



The slopes of the bubble count curve – DNS data



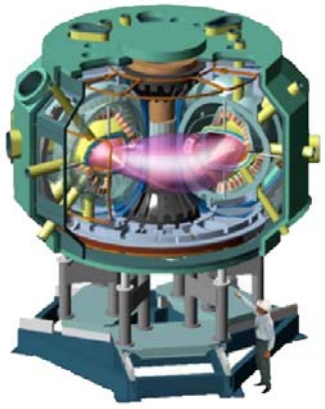
Sensitivity of the DNS bubble count to changing 3-D region-growing threshold



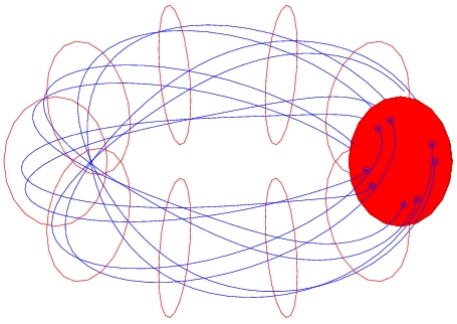
Analysis of Orbits in Poincaré Plots

Joint work with Neil Pomphrey, Don Monticello, and Scott Klasky (PPPL)

We are interested in automatically identifying orbits of various categories

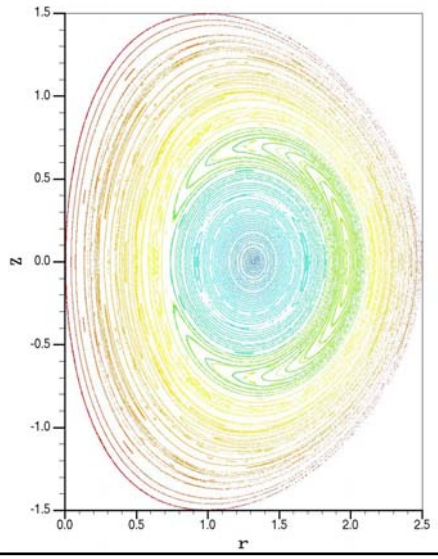


National Compact Stellarator Experiment

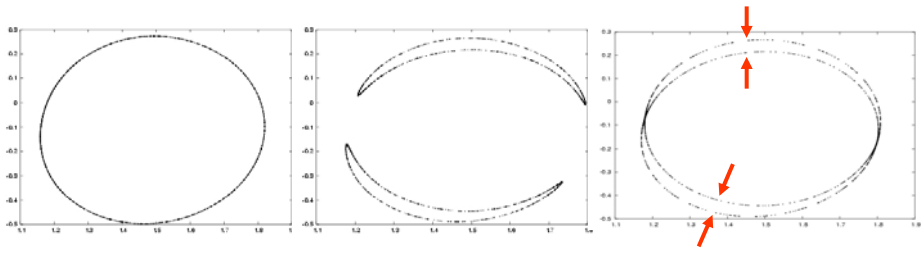


Schematic of a puncture plot

A sample Poincaré plot from computer simulations of the CDX tokamak at PPPL



We consider three categories of orbits

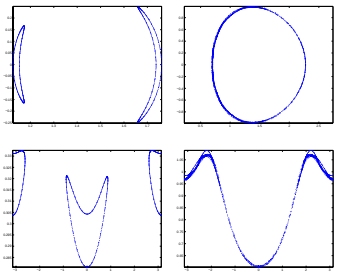


Quasi-periodic | Island chain | Separatrix

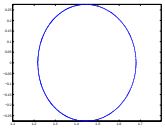
- Goals
  - assign a class label to an orbit
  - find key characteristics of certain orbits
- Orbit class determined by the location of the initial point – a 'missing' orbit if the initial point is not selected.

We are exploring two different methods to classify an orbit

- **Piecewise-polynomial approach**
  - convert to polar form
  - fit polynomials to the data
  - classify using simple rules



- **Graph-based approach**
  - create the MST of the points
  - extract features from the graph
  - use rules or machine learning methods

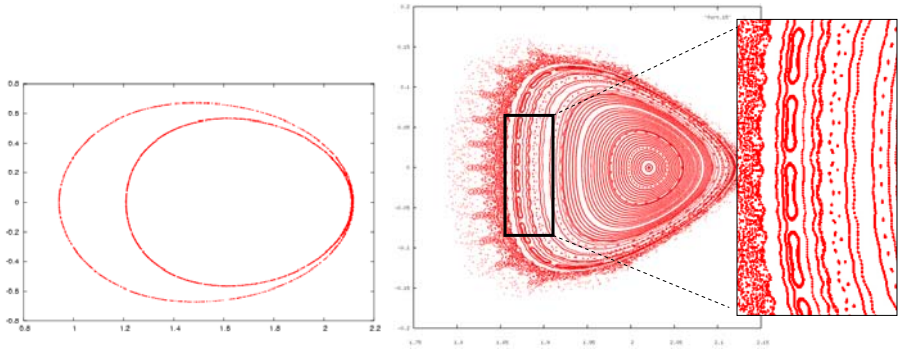


$$\#\,clusters = 1,$$

$$Distance\ between\ endpoints\ of\ \ diameter\ is\ small.$$

Challenges in orbit classification and analysis

- **Identifying the shape of an orbit with few points**
- **Extracting characteristics for multi-orbit plots**
- **Extraction of orbit characteristics for "missing" orbits**



Acknowledgements

- The Sapphire project team
- Our collaborators for sharing their data and domain expertise
- Funding sources
  - DOE NNSA ASC program
  - LLNL LDRD program
  - DOE SciDAC program

www.llnl.gov/casc/sapphire

UCRL-PRES-220610: This work was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract no. W-7405-Eng-48.