Lawrence Livermore National Laboratory

Scientific Data Mining: Why is it Difficult?

Chandrika Kamath

June 25, 2008

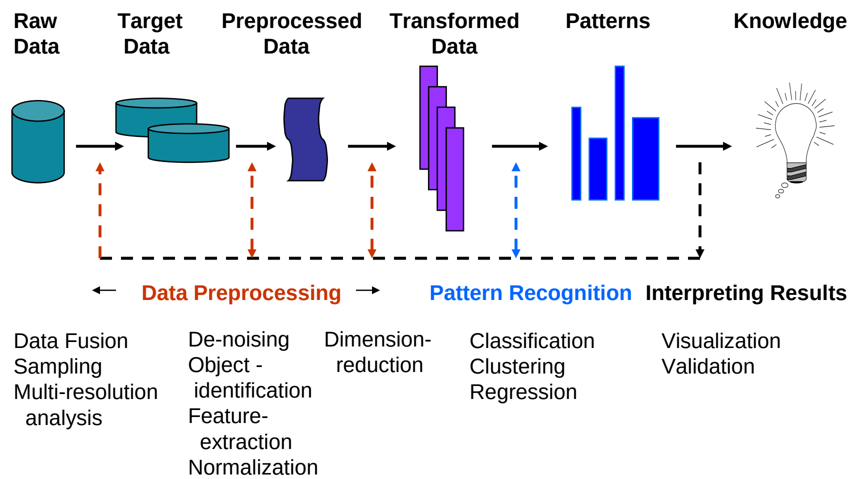
MMDS 2008: Workshop on Algorithms for Modern Massive Data Sets

LLNL-PRES-404920: This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344

Sapphire: using data mining techniques to address the data overload problem

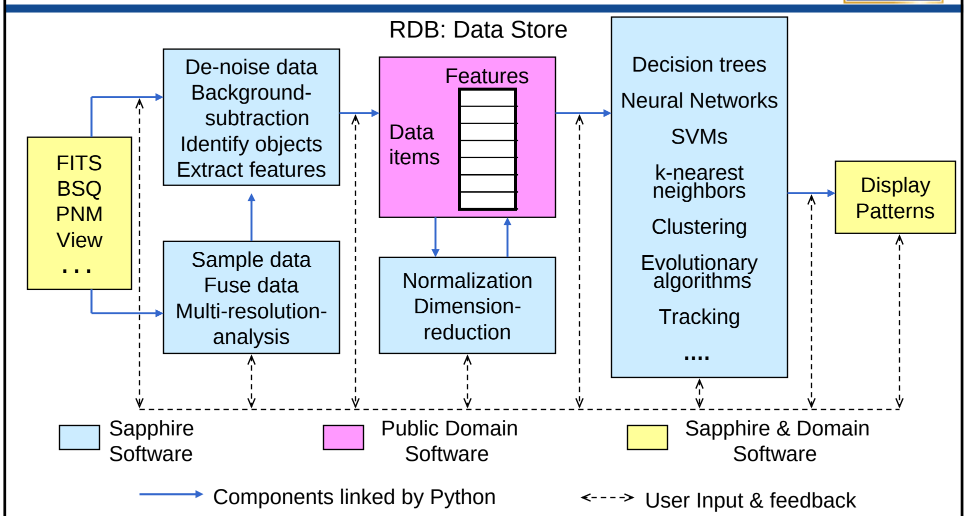
- We analyze science data from experiments, observations, and simulations: massive *and* complex
- Sapphire has a three-fold focus
  - research in robust, accurate, scalable algorithms
  - modular, extensible software
  - analysis of data from practical problems

Scientific data mining - from a Terabyte to a Megabyte



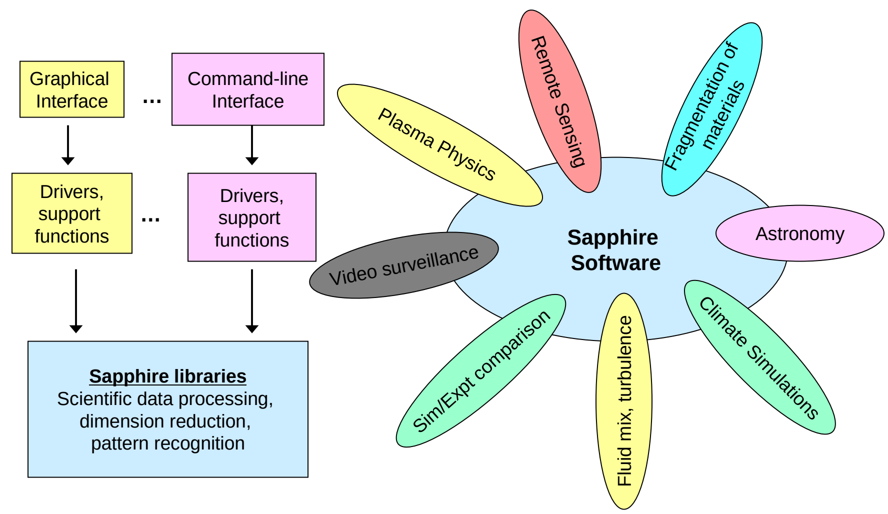
An iterative and interactive process

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)

The modular software allows us to meet the needs of different applications



Classification of Bent-double Galaxies in the FIRST Survey

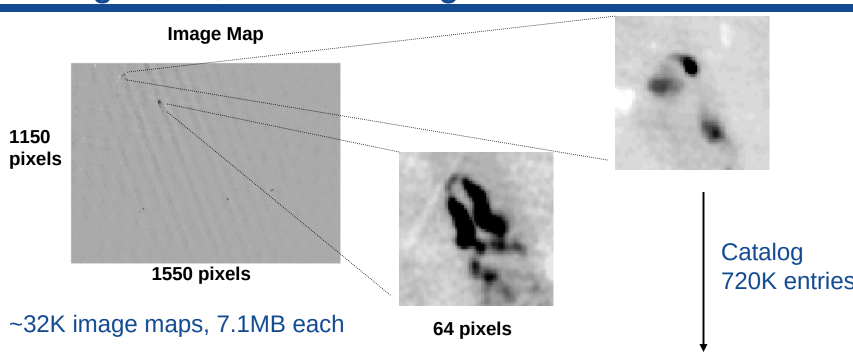
Sapphire: Erick Cantú-Paz, Imola Fodor, Chandrika Kamath, Nu Ai Tang

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 / degree^2
- 1.8'' pixels, resolution 5'', rms 0.15mJy
- Image maps and catalog available

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

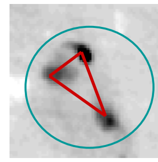


Catalog entry

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 features from the catalog

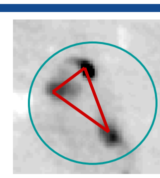


- Group catalog entries to identify a galaxy
  - 1 entry: unlikely to be bent-doubles
  - > 3-entry: all “interesting”
  - classify 2- and 3-entry galaxies separately
- Focus on the 3-entry galaxies
  - 195 training examples; 167 bents
  - extract relevant features
  - build a decision tree
  - use the tree to classify the 15K unlabeled galaxies

Iterate till error < 10%

➔Goal: identify likely bent-double galaxies for further observations by astronomers

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

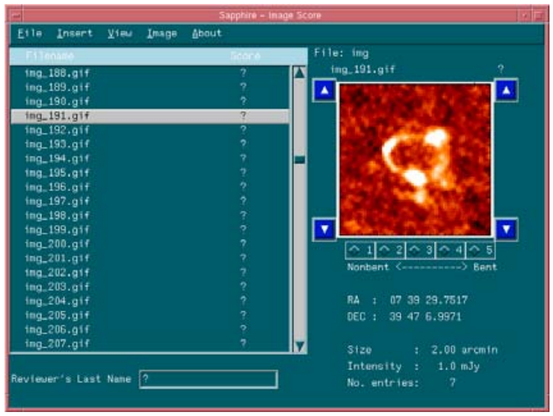


- Group catalog entries to identify a galaxy
  - 1 entry: unlikely to be bent-doubles
  - > 3-entry: all “interesting”
  - classify 2- and 3-entry galaxies separately
- Focus on the 3-entry galaxies
  - 195 training examples; 167 bents
  - extract relevant features
  - build a decision tree
  - use the tree to classify the 15K unlabeled galaxies

Iterate till error < 10%

➔Goal: identify likely bent-double galaxies for further observations by astronomers

Challenge: validation of results is subjective, tedious, and inconsistent



- Original training set: 195 (167 bents, 28 non-bents)
- Validated data: 290 (92 bents, 198 non-bents)

We tried building new models with the larger balanced training set with 485 examples

Error rate (std. error) – 10 runs of 10-fold cross validation

Method	Gini (no pruning)	Gini (pruning)
Single tree	22.79 (0.31)	19.77 (0.18)
Histogram-based (10 trees)	18.69 (0.28)	18.27 (0.30)
Sampling-based (10 trees)	18.21 (0.23)	17.31 (0.17)
Adaboost (10 trees)	21.87 (0.42)	20.40 (0.45)
Bagging (10 trees)	19.40 (0.28)	18.35 (0.34)
ArcX4 (10 trees)	20.48 (0.39)	20.12 (0.20)

➔ The error rate is now ~20% in comparison to 10% with the smaller training data set

Observations: good quality training data is hard to find; interpret accuracy results with caution

- Why did the error rate go up?
  - a more balanced (= different) training set
  - still using features suited for old training set
  - new galaxies added were borderline – therefore, likely to be misclassified
- So, what do we do next?
  - iterate and refine the features for new training data
  - recall: goal - identify galaxies for further observation

➔ We used the different methods to rank-order the galaxies

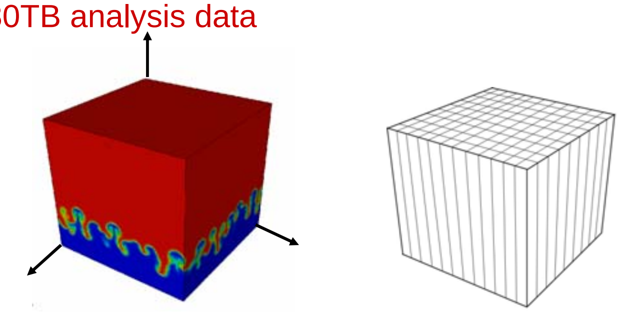
Analysis of Bubbles and Spikes in Rayleigh-Taylor Instability

Sapphire: Abel Gezahegne, Chandrika Kamath

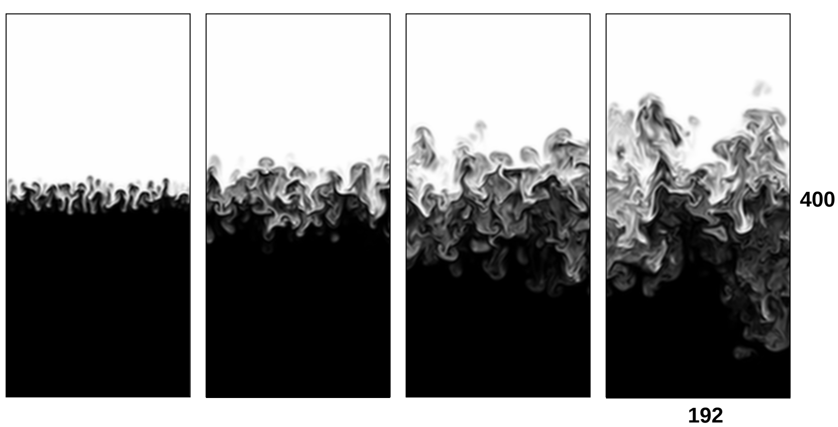
Physicist: Paul L. Miller (LLNL)

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

- DNS simulation of the Rayleigh-Taylor instability
  - regular Cartesian grid: 3072**3 grid points
  - 5 variables per grid point
  - 249 time steps
  - 80TB analysis data



The first step is to define a bubble…

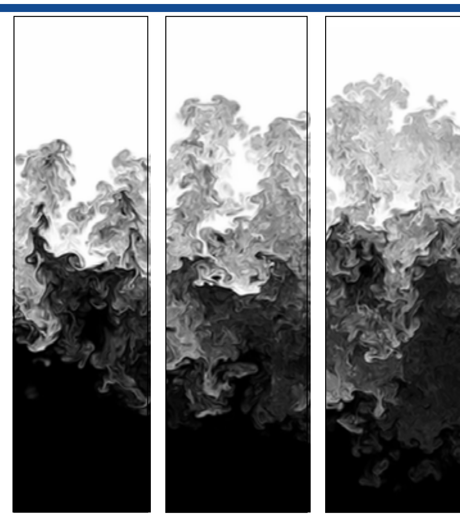


A slice through the density variable at time steps 100, 200, 300, 400

Convention: Smaller values are darker in image.

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

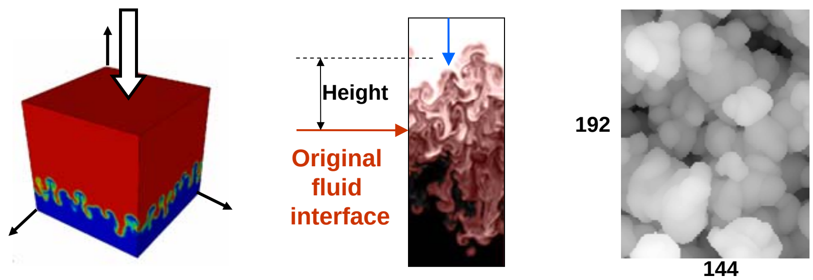
Density variable at time steps 500, 600, 700



Challenges: no precise definition of bubbles, range of scales, massive data, distributed data

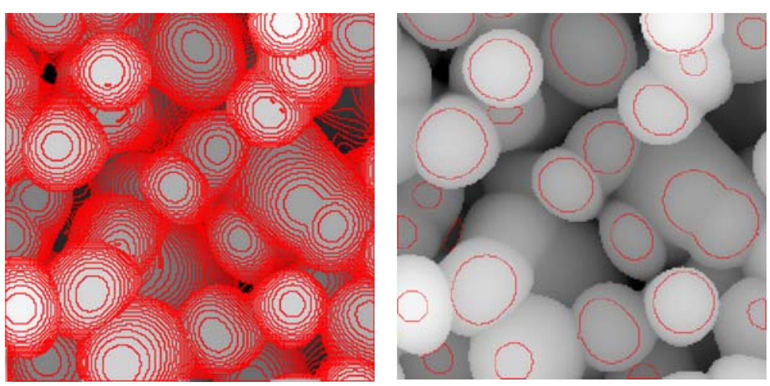
- We used a progressive approach to the analysis
  - a small subset of the data at every 50-th time step
  - all data at every 50-th time step
  - all the data – only once!
- We focused on algorithms which
  - were computationally inexpensive
  - applicable to distributed data
  - had few parameters
  - were relatively insensitive to choice of parameters

We used the density to find the bubble boundary and considered its height as a 2-D image



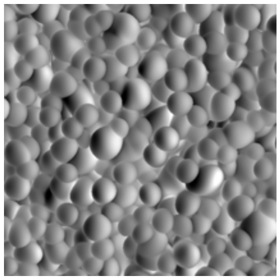
Height-depth map

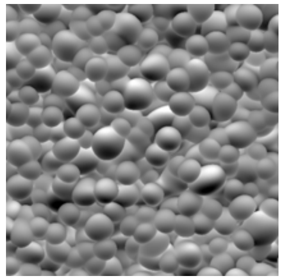
Bubble counting – Method 1: traditional 2D region growing (time step 50)

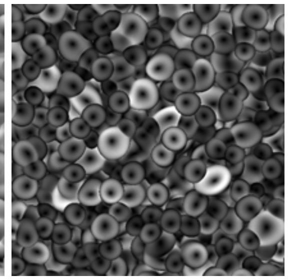


2800 seconds to process a 3072x3072 image

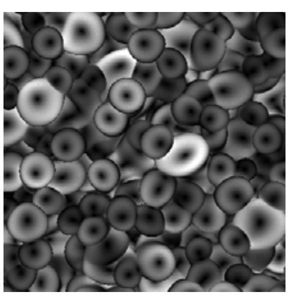
Bubble counting – Method 2: domain-specific approach using the mag-X-Y velocity (time step 50)

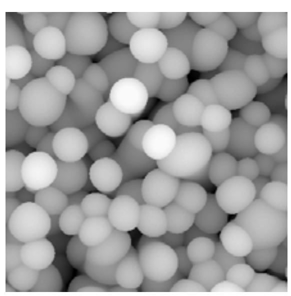

X velocity

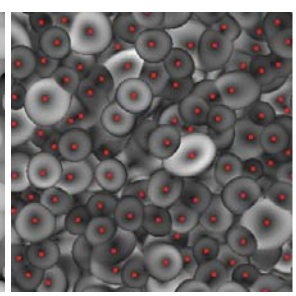

Y velocity


Mag X-Y velocity

Bubble counting – Method 2: identifying the bubble tips

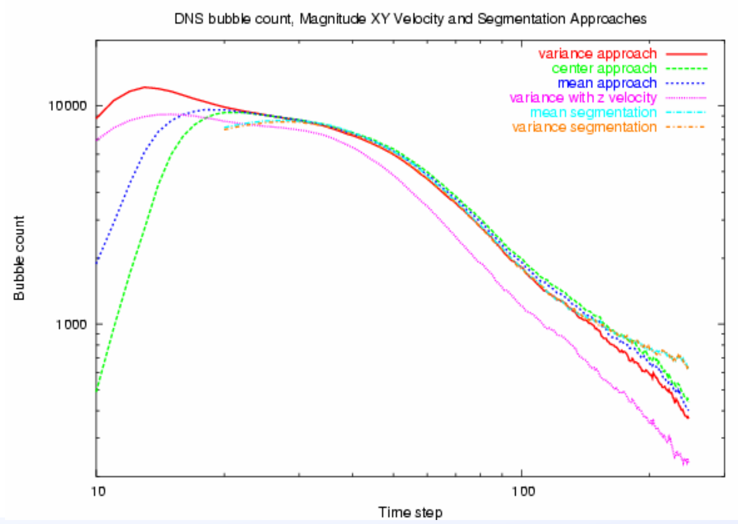

Mag X-Y velocity


Height-depth map

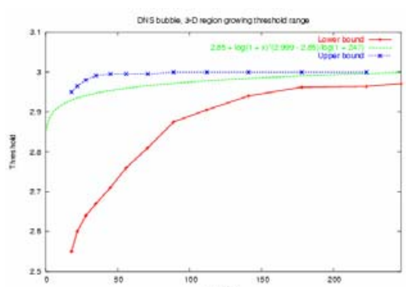

Bubble tips

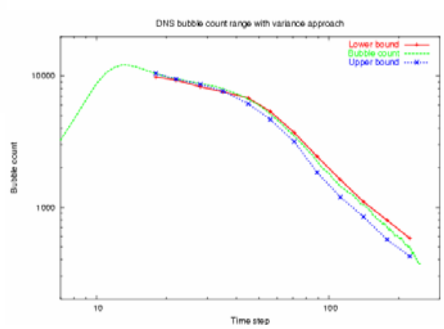
8 seconds to process a 3072x3072 image

How do we know we have the right results? use different methods + domain expertise to verify…



… and investigate the sensitivity of the results to changing the 3-D region-growing threshold





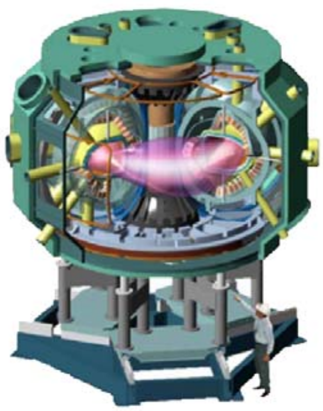
Observations

- Try to exploit domain-specific characteristics of data
- To gain confidence in results
  - try different methods
  - conduct studies to observe sensitivity of results to algorithm parameters
- To handle massive data sets
  - try simple algorithms – they often work very well!

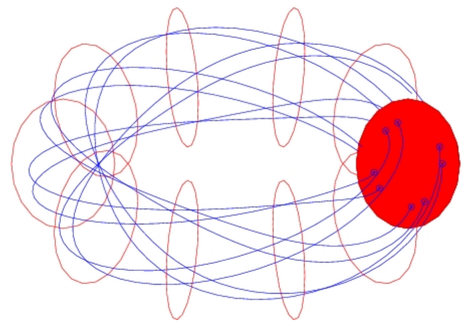
Analysis of Orbits in Poincaré Plots

Sapphire: Chandrika Kamath, Abraham Bagherjeiran, Erick Cantú-Paz, Siddharth Manay
Physicists: Neil Pomphrey, Don Monticello, Josh Breslau, and Scott Klasky (PPPL)

We want to automatically classify orbits in a Poincaré plot

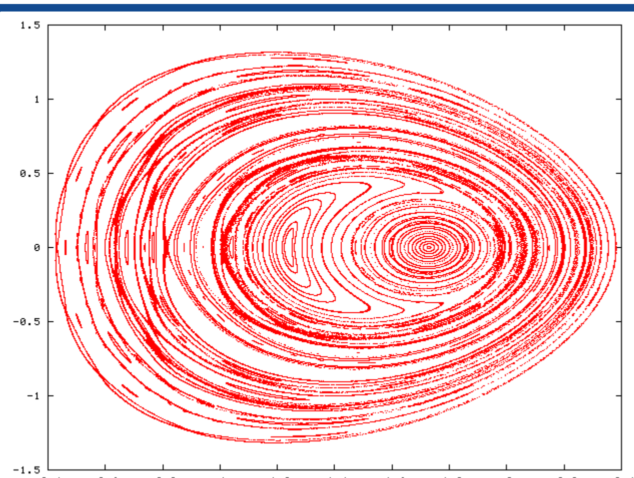


National Compact Stellarator Experiment



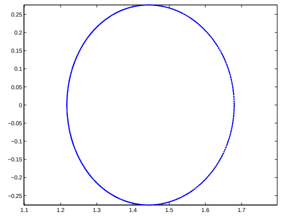
Schematic of a puncture plot

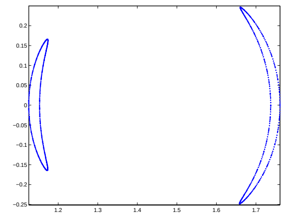
A sample Poincaré plot from computer simulations



We consider four classes of orbits – determined by the location of the initial point

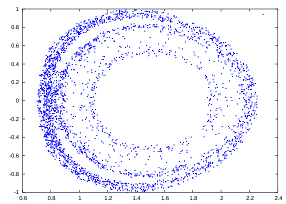
Quasi-periodic

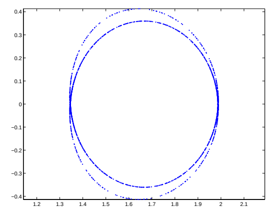




Island chain

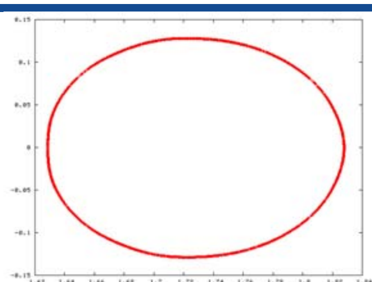
Stochastic

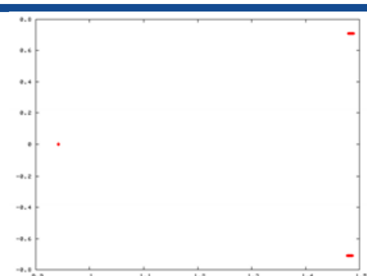


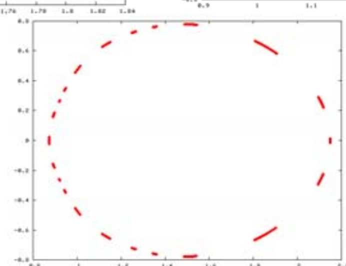


Separatrix

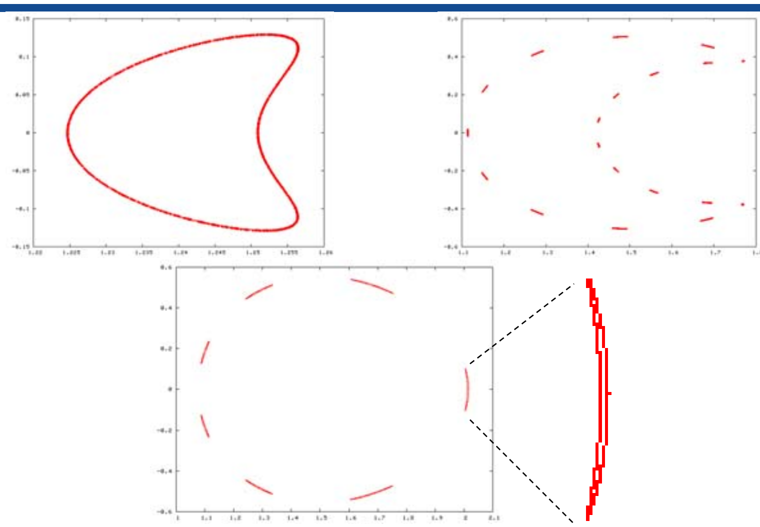
Challenge: There is a large variation in the orbits of any one class, e.g. quasiperiodic orbits



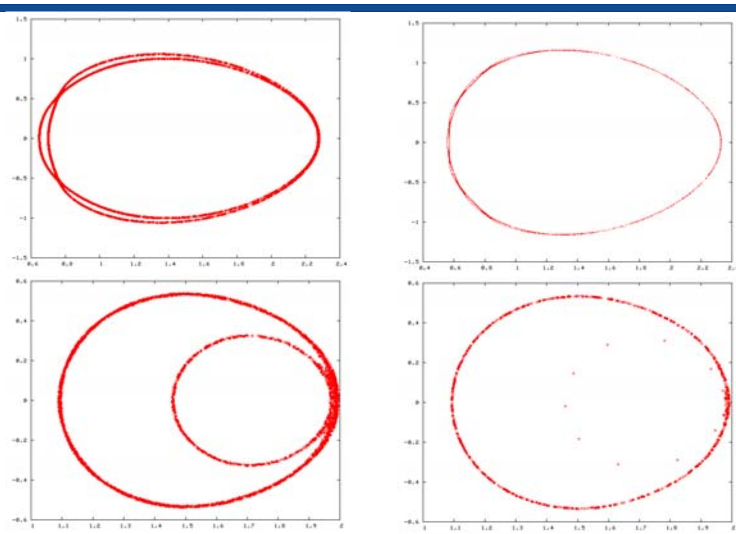




Variation in island-chain orbits



Variation in separatrix orbits

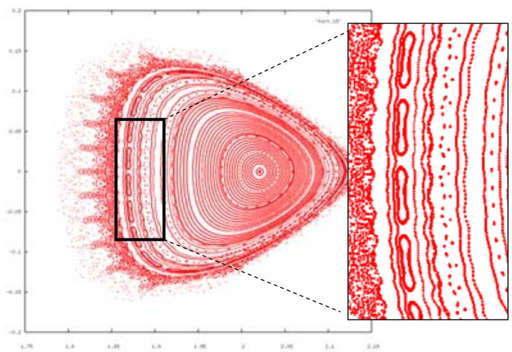


5000 points

1000 points

Observation: feature extraction is difficult, but key to accurate results

- Variation in the data may make it difficult to
  - identify good features
  - extract them in a robust way



Identifying missing orbits

Summary: challenges to mining scientific data

- Quality of the data – noise in data, small and unbalanced training data, …
- Massive size of the data
- Identification and extraction of good features
- Variation in the data: challenge to algorithms
- Lack of understanding of the scientific phenomena
- Need to verify results
- Reasoning in the presence of uncertainty
- …

Acknowledgements

- Key members of the Sapphire project team
  - Research: Erick Cantú-Paz, Samson Cheung, Chandrika Kamath
  - Software: Erick Cantú-Paz, Samson Cheung, Chandrika Kamath, Abel Gezahegne, Cyrus Harrison, Nu Ai Tang
  - Applications: Erick Cantú-Paz, Samson Cheung, Imola Fodor, Abel Gezahegne, Chandrika Kamath
- Our collaborators for sharing their data and domain expertise
- Funding: NNSA ASC, LLNL LDRD, SciDAC (SDM, GSEP)

Contact: kamath2@llnl.gov
https://computation.llnl.gov/casc/sapphire