

Chapter 6 - Specific Case Study: Searching for Bent-double Galaxies

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

UCRL-PRES-145087: The work of Chandrika Kamath in Chapters 5, 6, and 7 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.

Mining the FIRST survey for galaxies with a bent-double morphology

- **FIRST: Faint Images of the Radio Sky at Twenty Centimeters**
- **Radio equivalent of the Palomar Observatory Sky Survey (POSS)**
- **10,000 square degrees survey of the North Galactic Cap**
- **Using the NRAO Very Large Array (VLA), B configuration**

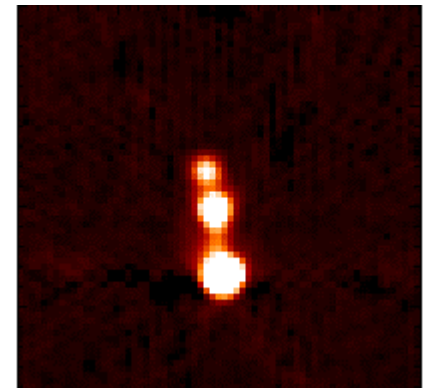
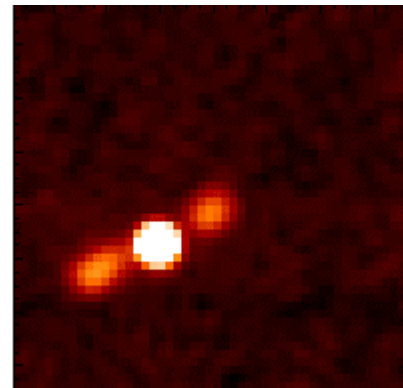
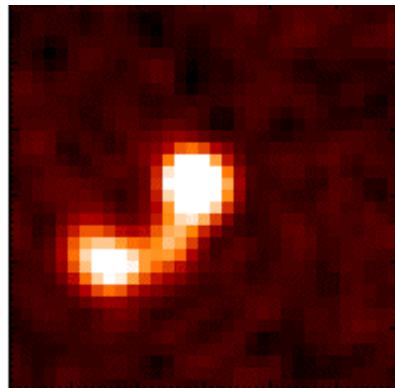
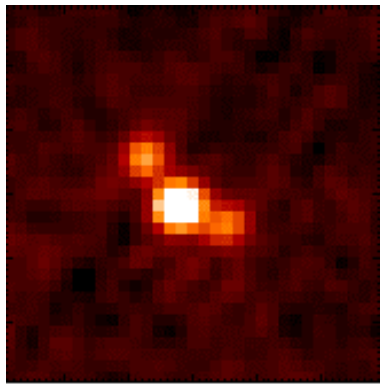
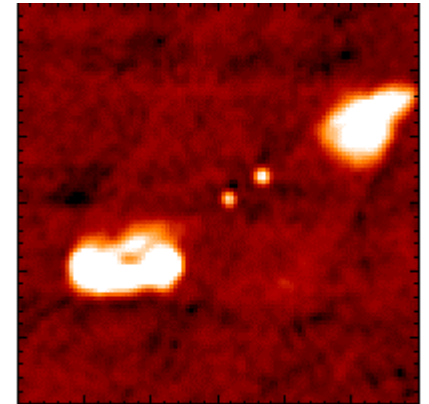
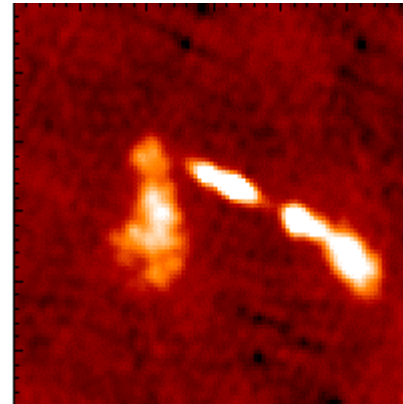
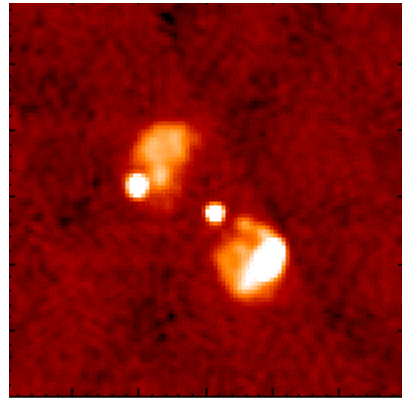
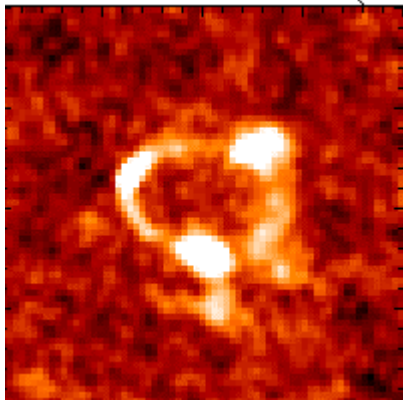


The FIRST data

- **1.8'' pixels, resolution 5'', rms 0.15mJy**
- **~90 radio sources per square-degree at 1mJy threshold**
- **The morphological type of a radio source provides clues to their emission mechanism, source class, and the properties of the surrounding medium**
- **The raw data from the telescopes is extensively processed**
- **Images maps and catalog available (sundog.stsci.edu)**



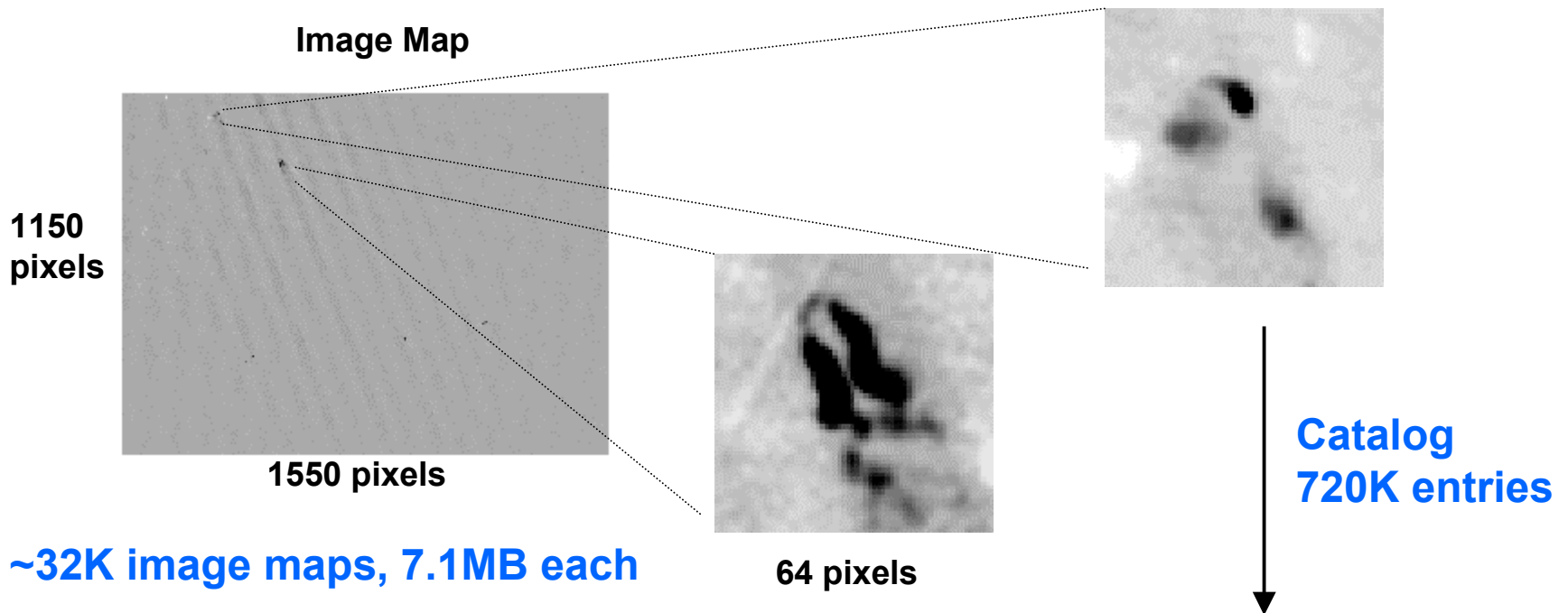
Radio-emitting galaxies (radio-sources) exhibit a wide range of morphological types



We are using data mining to find “bent-doubles” in FIRST

- **FIRST astronomers interested in “bent-doubles”**
 - indicates presence of clusters of galaxies
 - first “identify” using a visual technique
 - followed by optical observations and checks with other surveys
 - **Visual identification is no longer feasible**
 - subjective, tedious, likely to miss cases
 - ~900,000 galaxies in the full survey
- Our goal is to replace the visual identification of bent doubles by a semi-automated one.**
Details at <http://www.llnl.gov/casc/sapphire>

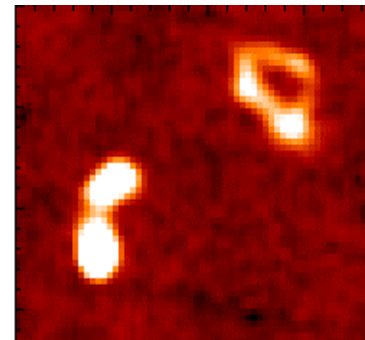
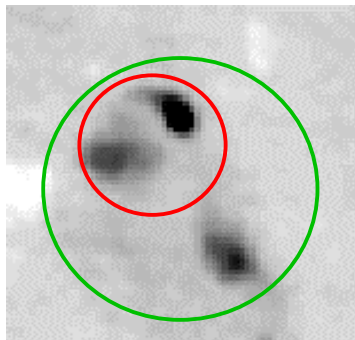
Detecting bent-double galaxies in 250GB image data, 78MB catalog data (as of 7/00)



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

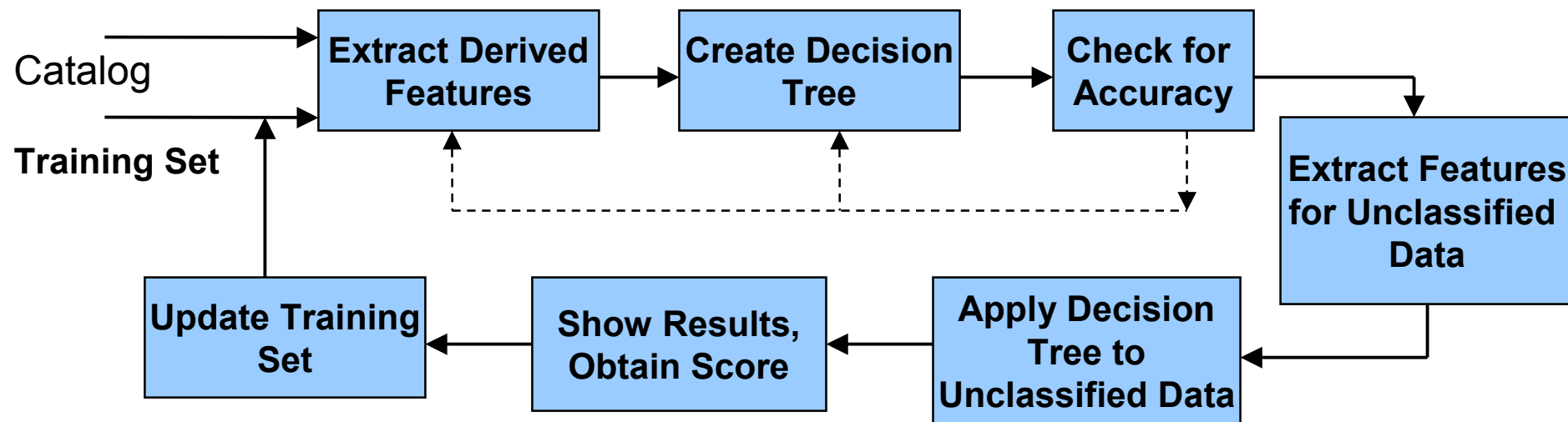
Our approach for finding bent doubles in the FIRST data

- Group the catalog entries into a “galaxy”
- Separate sources based on number of catalog entries
 - 1-entry sources unlikely to be bent-doubles
 - > 3-entry sources all “interesting”
 - study the 2- and 3-entry sources separately
 - results in splitting a small training set (313 ---> 118 + 195)



Our approach for finding bent doubles in the FIRST data

- Calculate features for a galaxy (103 features)
- Use the features to train a decision tree
- Use the tree to classify the unlabeled galaxies and validate the results
- Use validated results to enhance training set



Potential features used to describe a radio galaxy must be carefully selected

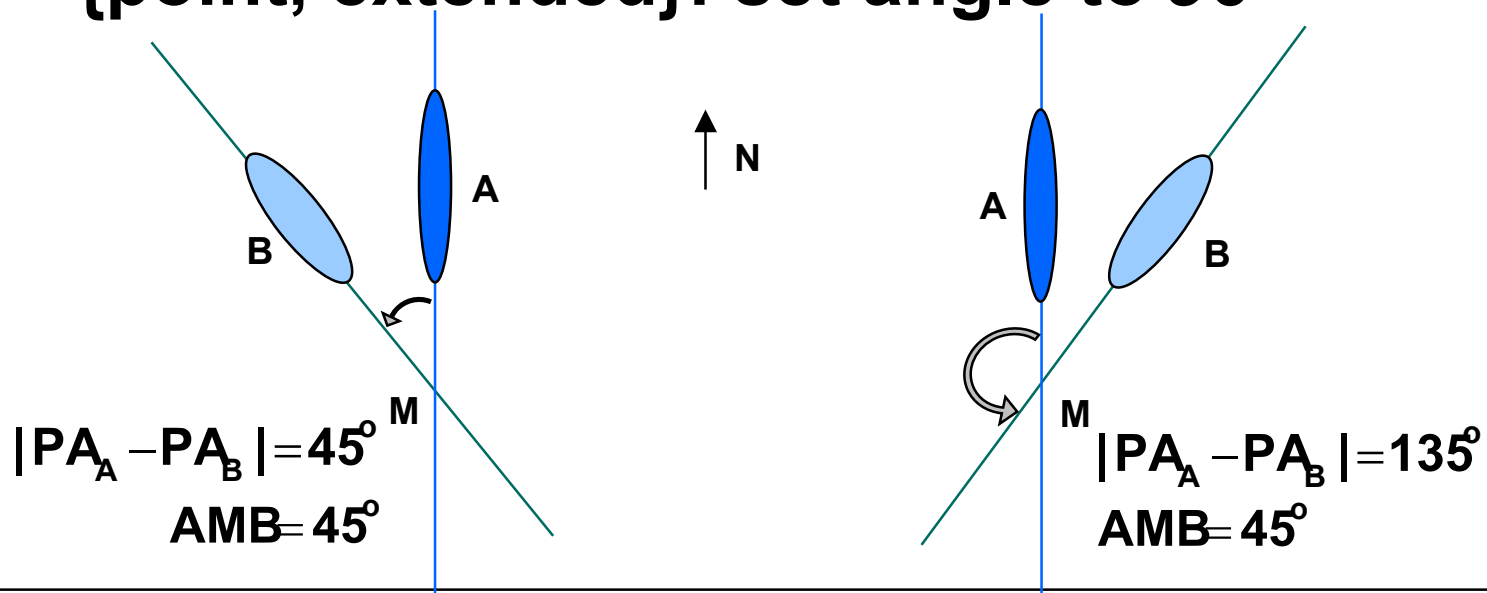
- **Features must be**
 - **relevant to the problem**
 - **robust**
 - **scale, rotation, translation invariant**
- **Feature extraction is an iterative process**
- **Feature extraction requires**
 - **input from the domain scientists**
 - **an understanding on how the data was collected**

Potential features that can be used to describe a radio galaxy

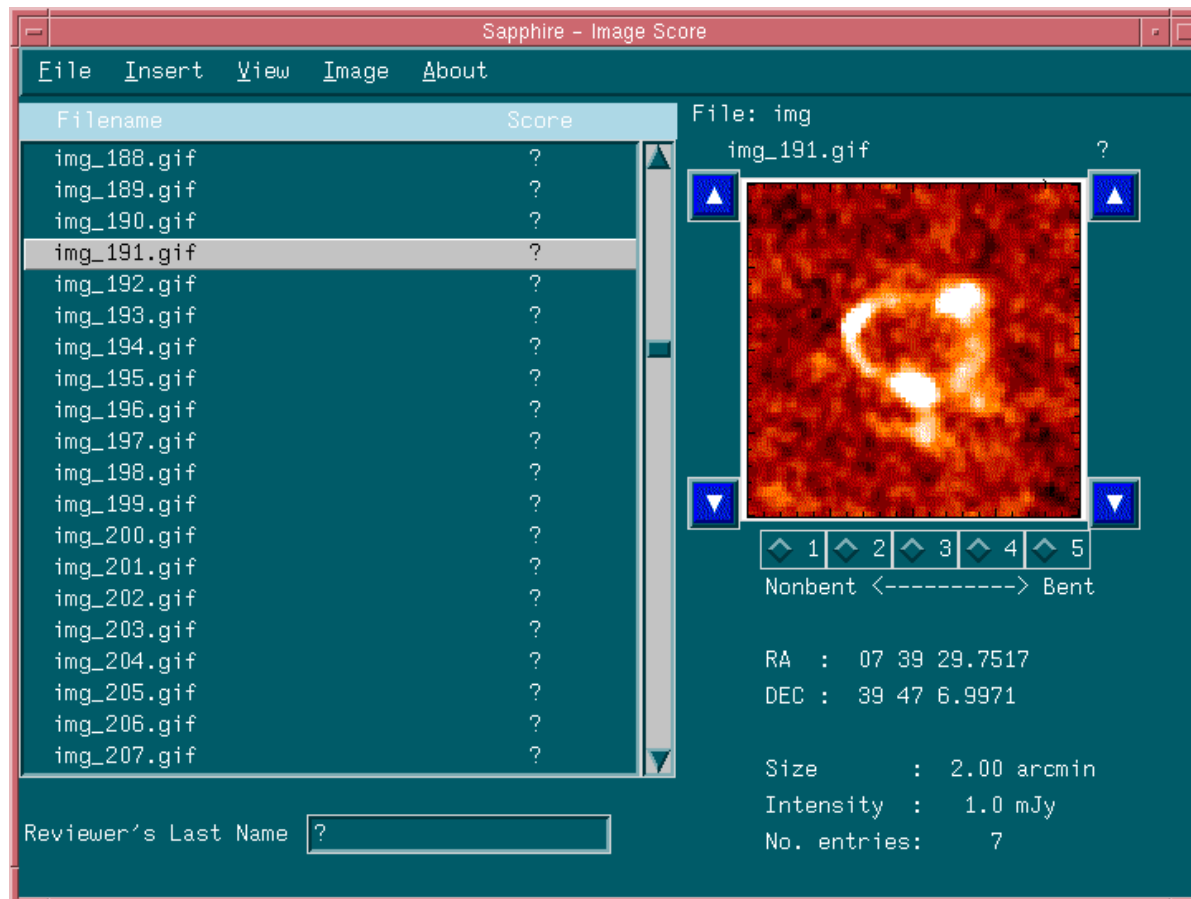
- **Single, for each CE**
 - **peak (integrated) flux, area, major (minor) axis, ellipticity (major/minor), position angle (PA)**
- **Double, for CEs taken 2-at-a-time**
 - **maximum flux, relative flux (ratio of the fluxes), total flux, total area, distance between centers, angle between major axes**
- **Triple, for CEs taken 3-at-a-time**
 - **maximum flux, total flux, total area, distances, angles**

Calculating robust pair-wise angles is difficult

- Position angle differences are sensitive to orientation
- Need to determine angles geometrically
- What is the position angle for point sources?
 - {point, point}: set angle to 180°
 - {point, extended}: set angle to 90°



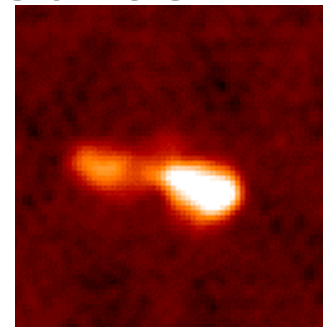
Our Python-based tool allows validation of results and incorporation of user feedback



Our initial results using a single tree for 3-entry sources were satisfactory

- Labeled training set: 167 bents, 28 non-bents
- Performed several inner iterations using pruned trees (c5.0 decision tree software)
- Ten 10-fold cross-validation errors: mean(SE)
 - using all the features: 9.7%(0.3%)
 - using triple features only: 10.7%(0.3%)
- Discriminating features include geometrically calculated angles, relative distances, ellipticity and symmetry measures

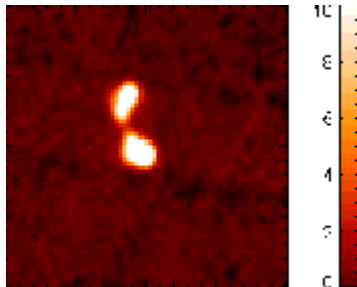
A bent-double missed in a manual search:



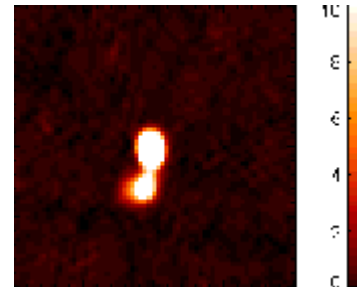
Accurate classification of 2-entry sources needs more refinement

- Labeled training set: 72 bents, 46 non-bents
- Decision trees depend on the training/test set
- Ten 10-fold cross-validation errors: 19.8%(1.0%)
- Reasons why classification is difficult
 - small labeled set
 - lose information in going from images to catalog
 - irrelevant features

Bent-double



Non-bent-double



Our results with ensembles indicate a performance improvement

# trees	Bagging	AdaBoost	ArcX4
2	13.15 (.02)	13.15 (.02)	12.10 (.02)
5	13.68 (.03)	11.05 (.02)	13.15 (.02)
10	12.63 (.02)	10.52 (.02)	10.52 (.02)
20	10.52 (.02)	8.94 (.03)	10.00 (.02)
50	10.52 (.02)	8.94 (.03)	10.52 (.02)
100	11.05 (.02)	8.94 (.03)	8.42 (.02)

% error (SE) using all features

- **Very small ensembles have higher error**
- **Bagging is slightly less accurate than other methods**
- **Very large ensembles may be wasteful**

General observations from mining the FIRST survey

- **Easy (on-line) access to data is a big help**
- **Well-supported public domain tools for reading, writing, and viewing data are very useful**
- **Availability of the catalog allowed a head-start on the classification of the galaxies**
- **Generating a “good” training set is non-trivial**
 - **few bent- and non-bent- doubles identified in a visual inspection**
 - **disagreement among astronomers on cases which are difficult to classify**
 - **lack of ground truth**
- **Extracting relevant features in a robust manner is difficult**

Chapter 6 - References

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

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>