



Finding Features and Anomalies in Scientific Data

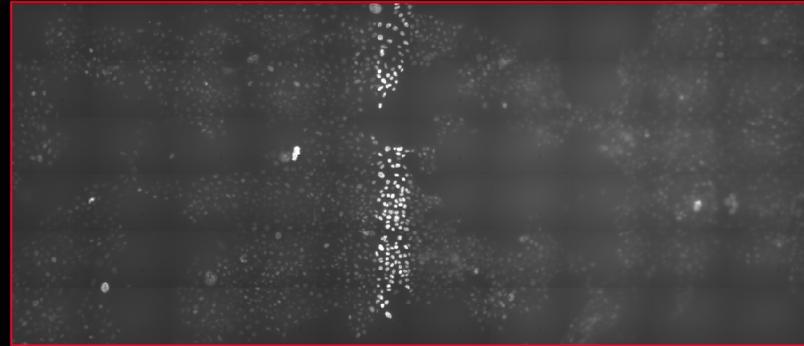
Raquel A. Romano
Visualization Group
Lawrence Berkeley National Laboratory
romano@hpcrd.lbl.gov

October 22, 2005

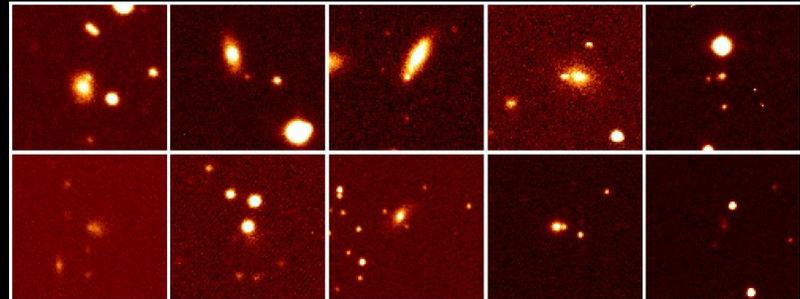
Analyzing Scientific Data



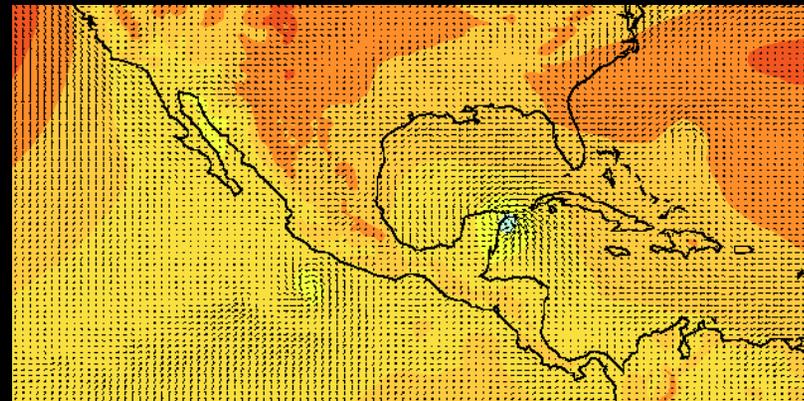
Microscopy
proteins



Astronomy
supernovae



Climate Modeling
hurricanes



Analyzing Scientific Data



unknown
phenomena

*protein
quantification*

*supernova
detection*

known
phenomena

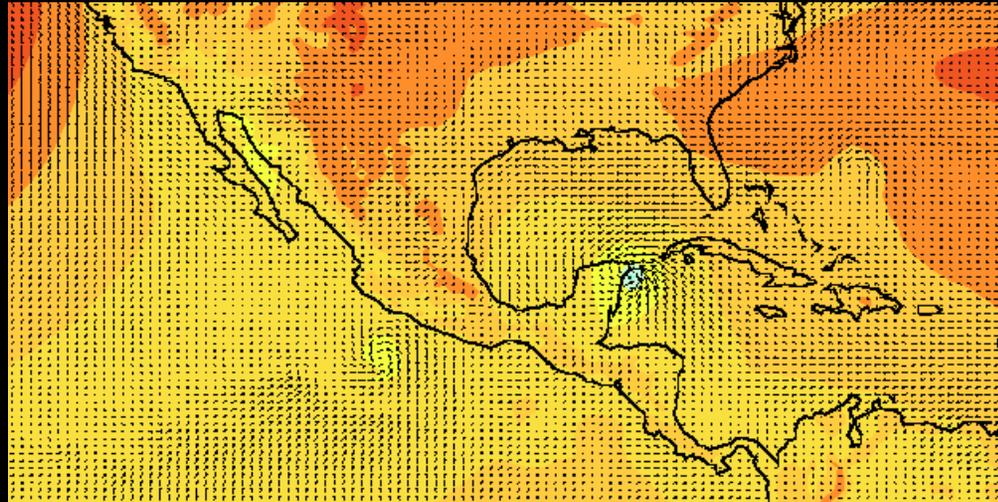
*hurricane
detection*

known
criteria

unknown
criteria

October 22, 2005

Tropical Cyclones from Climate Simulations



Michael Wehner, LBNL Computational Research Division

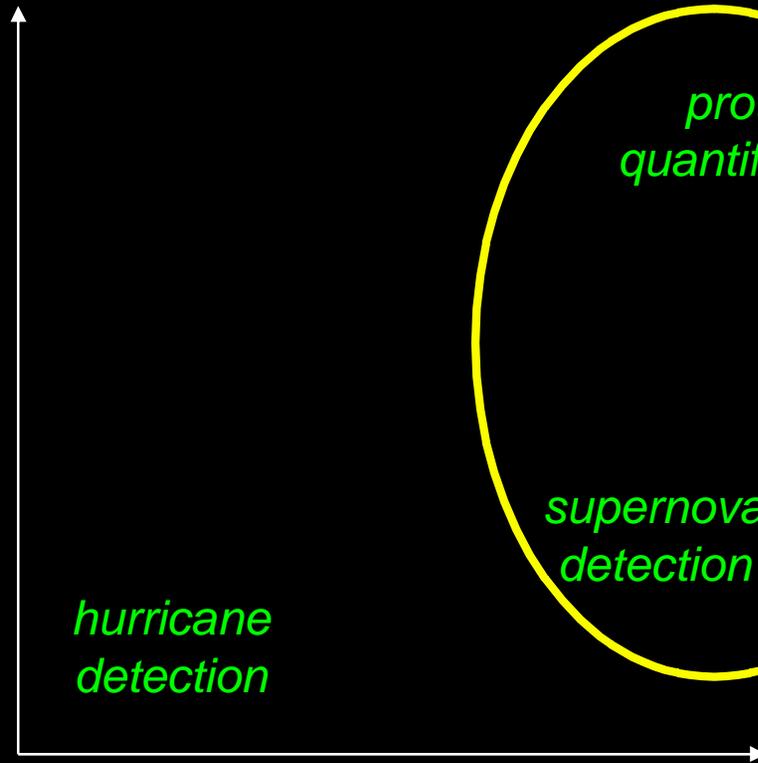
- Define hurricane using carefully designed criteria
 - high wind vorticity
 - low pressure
 - upper air temperature anomalies
- Fixed thresholds give too many false positives
- Spatiotemporal correlations narrow the list of candidate hurricanes

Analyzing Scientific Data



unknown
phenomena

known
phenomena



*hurricane
detection*

*protein
quantification*

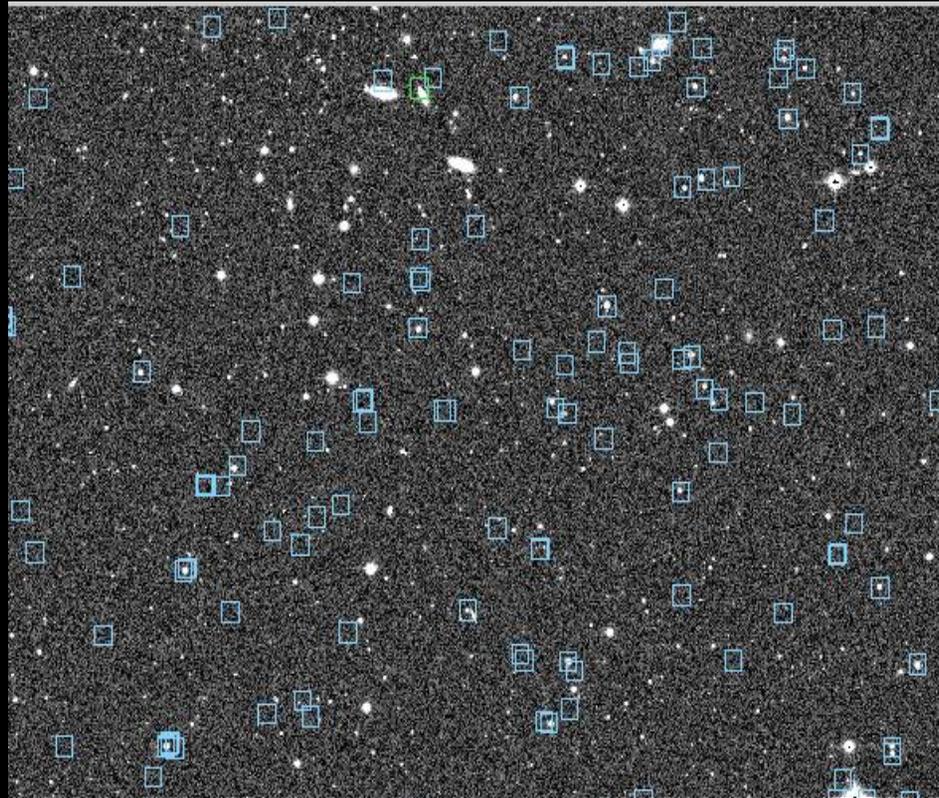
*supernova
detection*

*interesting
machine
learning
problems*

known
criteria

unknown
criteria

Supernovae in Astronomy Images



Nearby Supernova Factory

LBNL

<http://snfactory.lbl.gov>

October 22, 2005

Supernovae in Astronomy Images



galaxy



*galaxy with
supernova*



- Supernovae appear as bright regions near galaxies

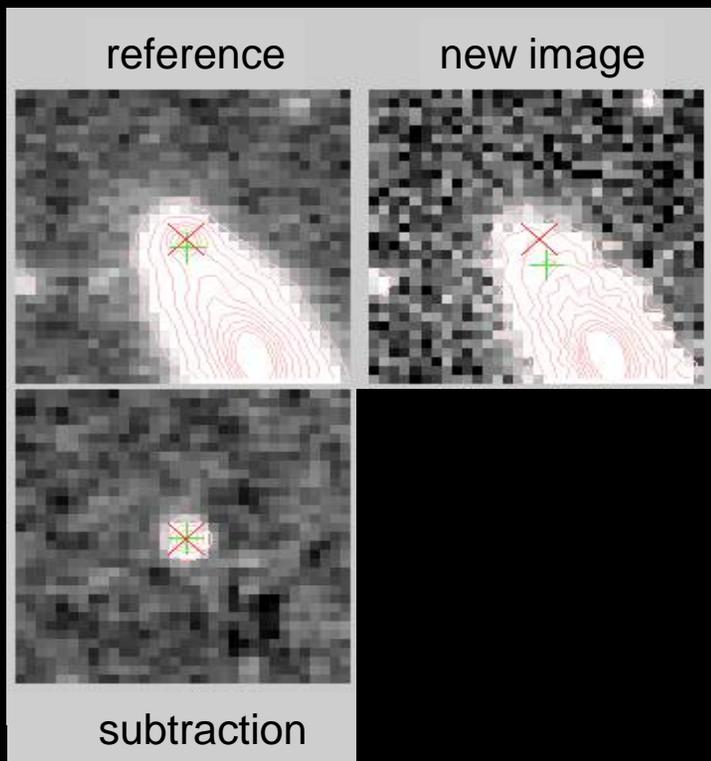
*Nearby Supernova Factory
LBNL*

<http://snfactory.lbl.gov>

Supernovae in Astronomy Images



galaxy galaxy with supernova

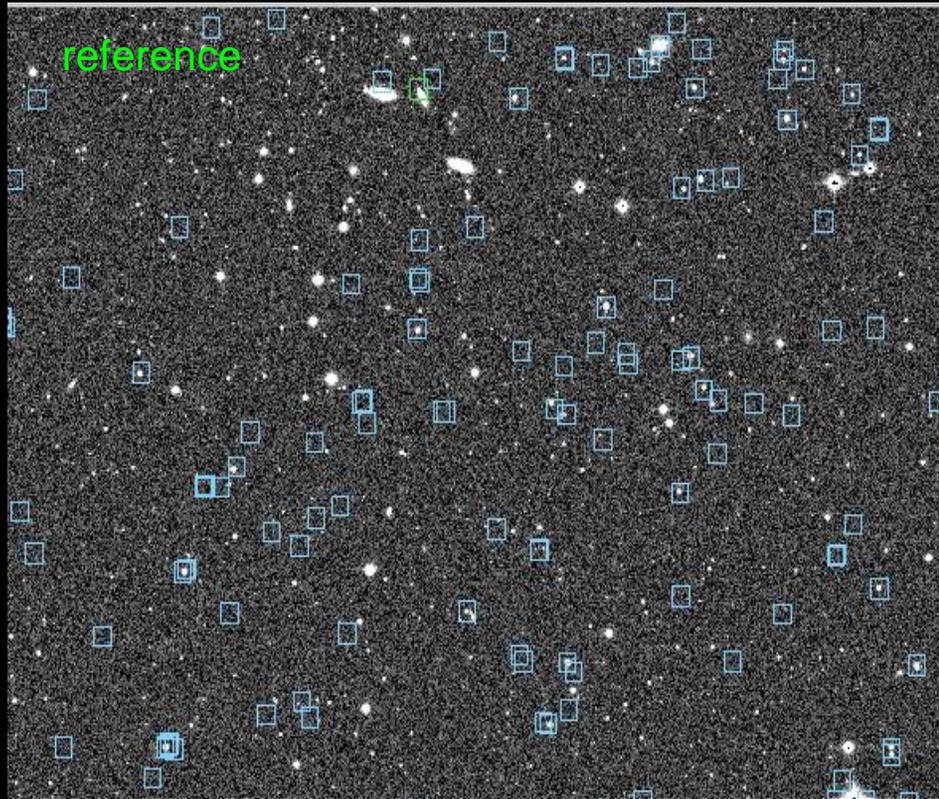


- Supernovae appear as bright regions near galaxies
- Subtract new image from reference and analyze subtraction
- Problem: criteria that define supernovae may also detect variable stars, asteroids, image artifacts

Nearby Supernova Factory
LBNL

<http://snfactory.lbl.gov>

Supernovae in Astronomy Images



More problems:

- extremely noisy imagery
- 30,000 images/night (85 Gb)
- still requires human scanning

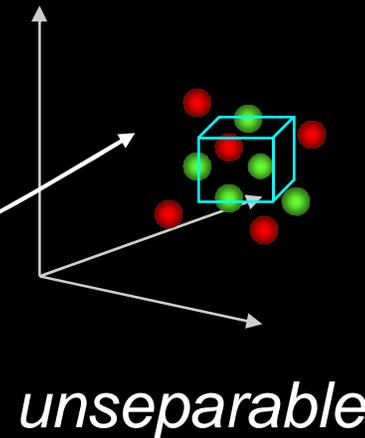
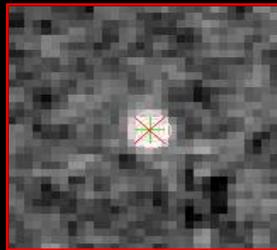
Nearby Supernova Factory
LBL

<http://snfactory.lbl.gov>

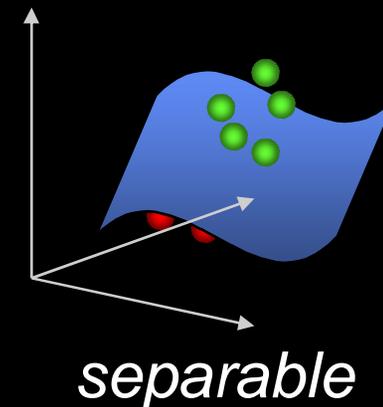
Supernova Detection Criteria



- Compute features from each candidate subimage
- Apply decision criteria in high-dimensional space



- Ripe problem for machine learning!
 - Use human scanning to label positive and negative examples
 - Test existing criteria for separability of true supernovae from others
 - Apply clustering and decision boundary estimation algorithms
 - Analysis of individual and joint feature relevance



Features vs. Anomalies vs. Noise



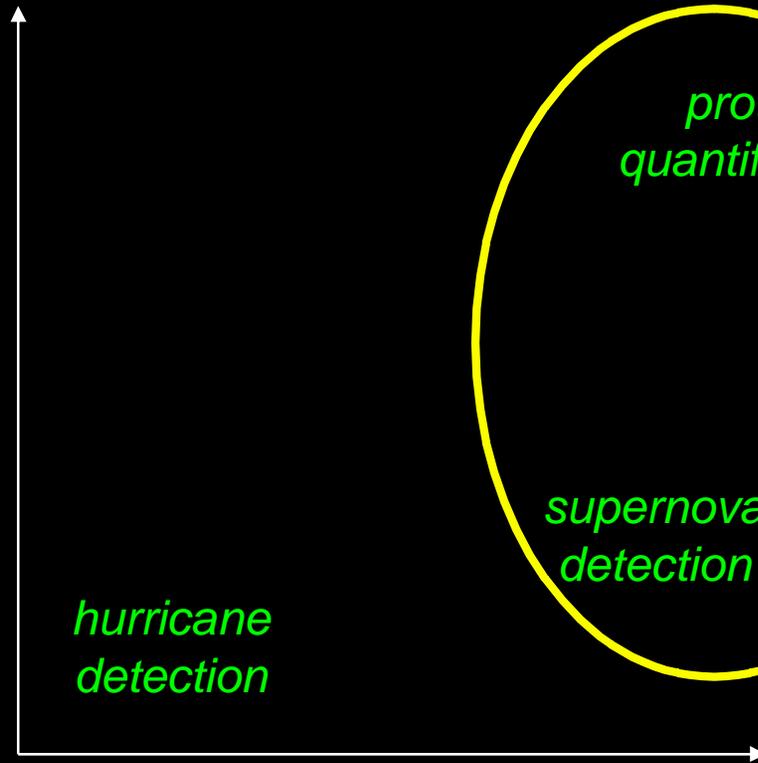
- **feature**
 - a prominent or distinctive aspect, quality, or characteristic
- **anomaly**
 - a deviation or departure from the normal or common order, form, or rule
 - something peculiar, irregular, abnormal, or difficult to classify
- **noise**
 - irrelevant or meaningless data
 - a random and persistent disturbance that obscures or reduces the clarity of a signal

Analyzing Scientific Data



unknown
phenomena

known
phenomena



*hurricane
detection*

*protein
quantification*

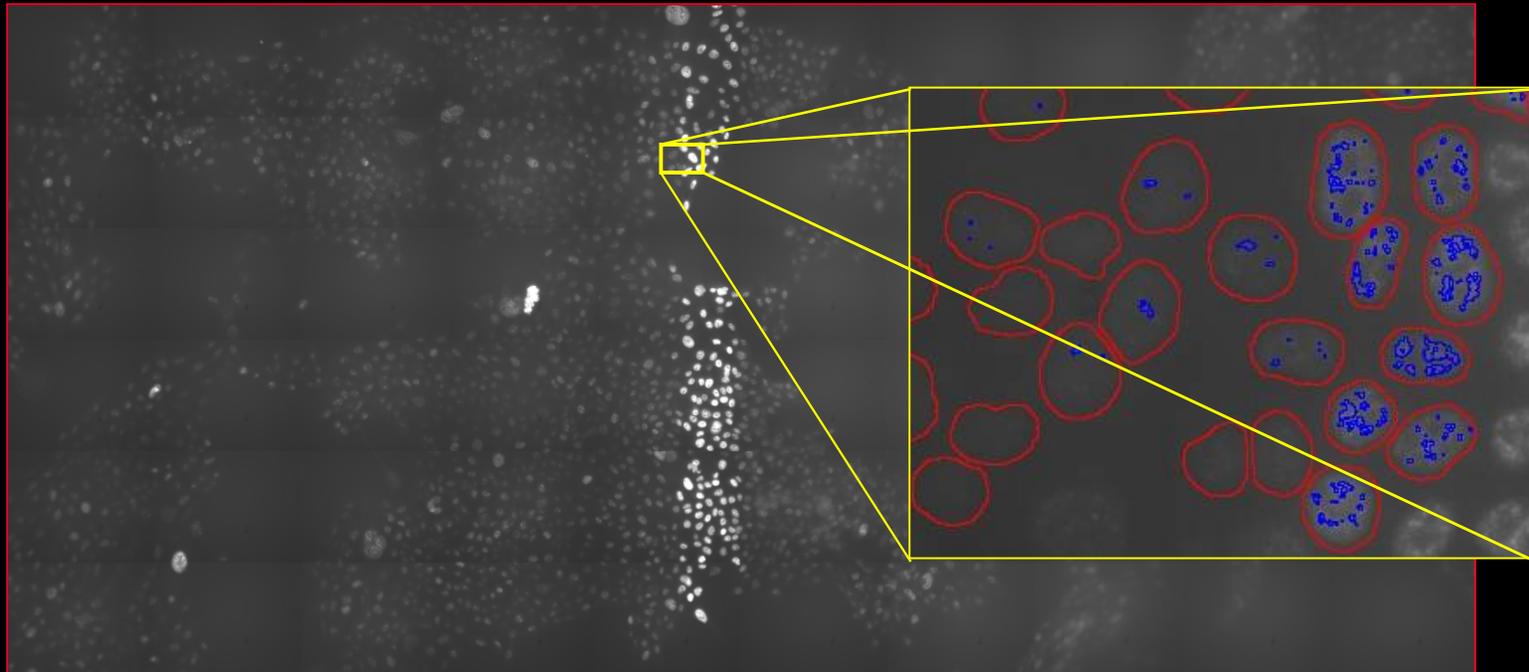
*supernova
detection*

*interesting
machine
learning
problems*

known
criteria

unknown
criteria

Microscopy Images for Radiation Biology

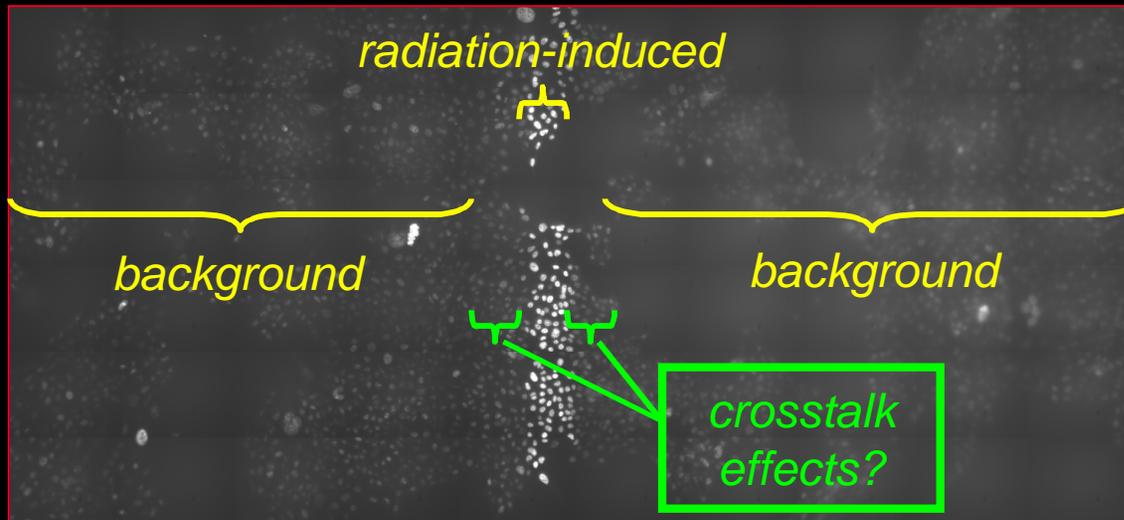


Bright regions are proteins responsible for DNA repair, responding to

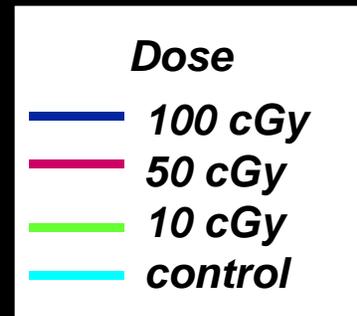
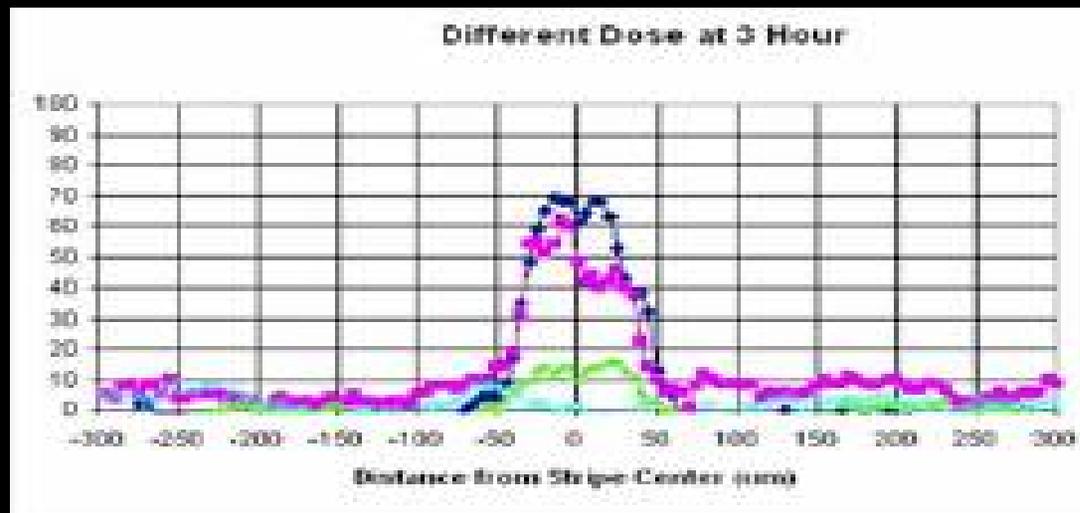
- 1) radiation-induced damage
- 2) background effects
- 3) damage due to multicellular crosstalk: "bystander effect"

- Eleanor Blakely, Cell and Molecular Biology
- Bahram Parvin, Imaging and Informatics

Microscopy Images for Radiation Biology



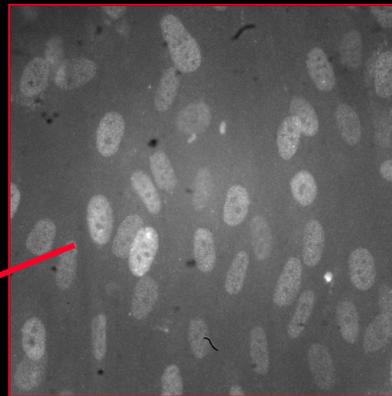
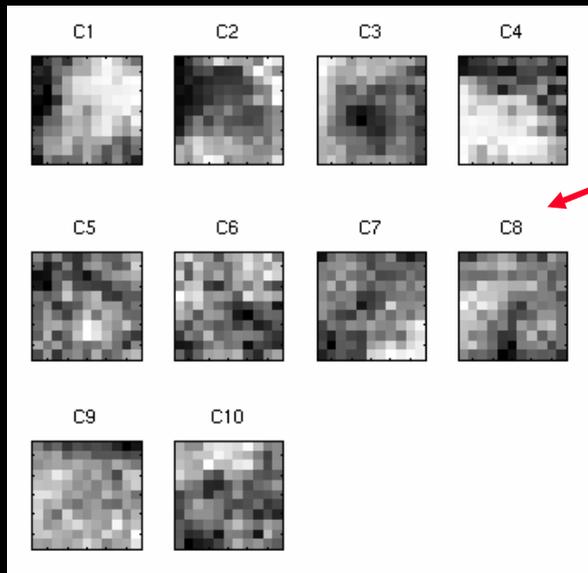
Goal: automatically classify nuclei into categories that correspond to different patterns of protein expression.



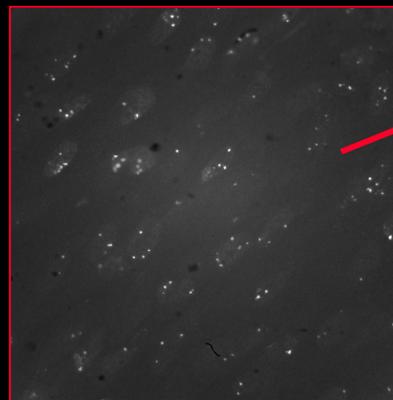
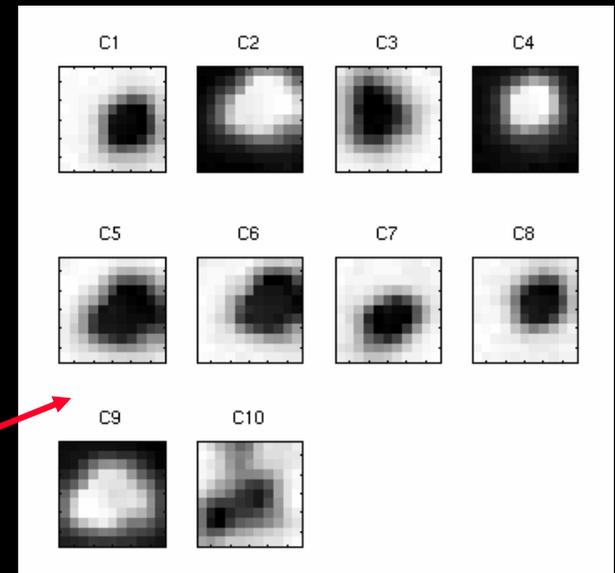
Basis Decomposition by Independent Component Analysis (ICA)



Basis Functions Control Group



Basis Functions Irradiated Group

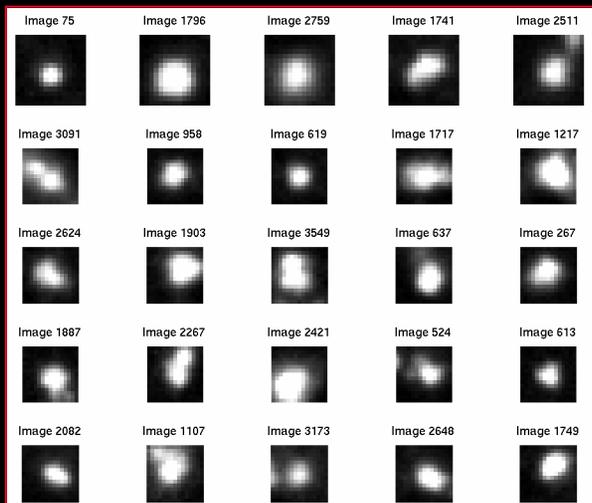


- *Subimages centered and whitened (decorrelated, unit variance)*
- *Basis functions defined up to an unknown sign ambiguity*

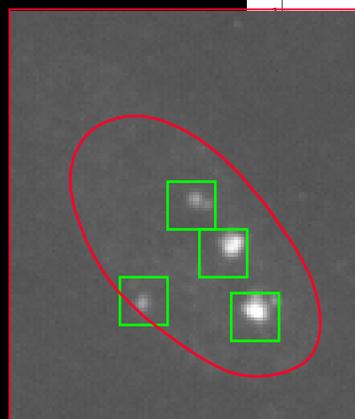
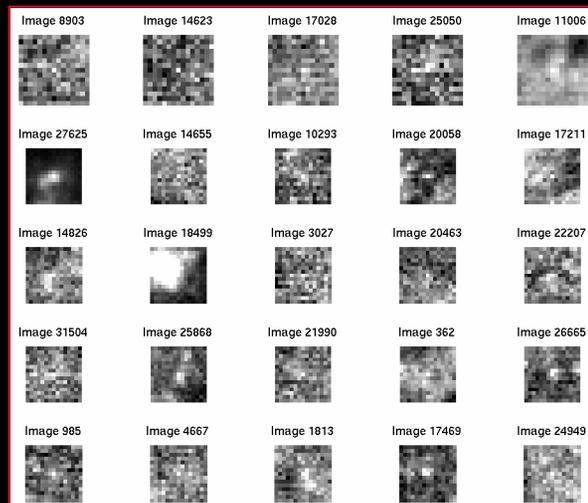
Classification by Maximum Response



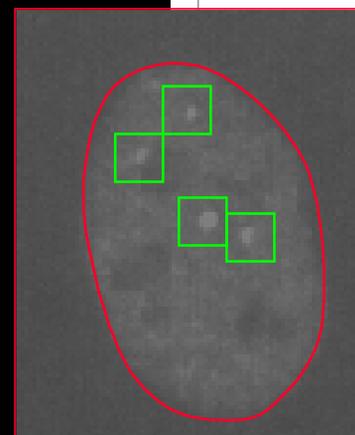
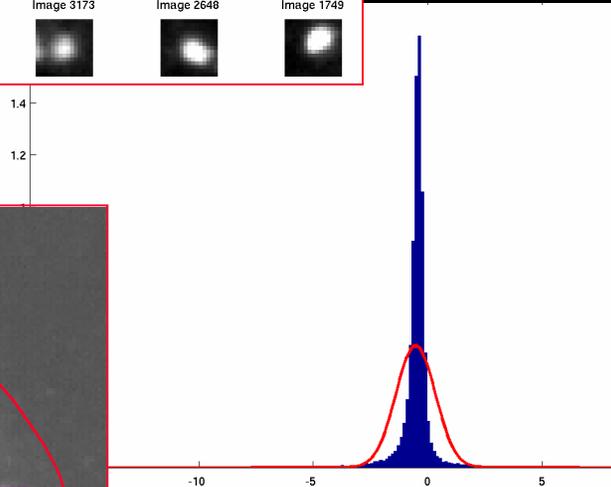
Sample Subimages: FG Class



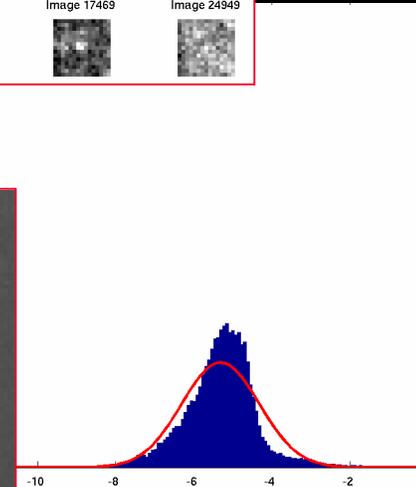
Sample Subimages: BG Class



Sample IR Nucleus



Sample Sham Nucleus



Current Directions



How to include prior knowledge and to what degree?

- Preprocessing of raw data
 - Normalization
 - Noise reduction
- Learning from labeled examples
 - Clustering
 - Boundary detection
- Explicit vs. implicit modeling of interesting features
 - Manually designed models
 - Data-driven models



THE END

October 22, 2005

SNFactory Cuts



- APSIG The signal-to-noise in the candidate aperture (ap.)
- PERINC The percent increase from REF!NEW in the candidate ap.
- PCYGSIG Normalized: flux in $2 \times \text{FWHM}$ ap. - flux in $0.7 \times \text{FWHM}$ ap.
- MAXPIXSIG Limit on maximum pixel value (unused)
- MXY The X-Y moment of the candidate
- FWX The FWHM of the candidate in X
- FWY The FWHM of the candidate in Y
- NEIGHBORDIST Distance to the nearest object in the REF
- NEIGHBORMAG Magnitude of the nearest object in the REF
- MAG Magnitude of the candidate
- THETA Angle between the candidate and nearest object in the REF
- NEW1SIG Signal-to-noise of candidate in NEW1
- NEW2SIG Signal-to-noise of candidate in NEW2
- SUB1SIG Signal-to-noise of candidate in SUB1
- SUB2SIG Signal-to-noise of candidate in SUB2
- SUB2MINSUB1 Weighted signal-to-noise difference between SUB1 and SUB2
- DSUB1SUB2 Difference in pixel coordinates between SUB1 and SUB2
- HOLEINREF Signal-to-noise in aperture in the REF
- BIGAPRATIO Ratio of larger aperture to smaller aperture of candidate
- OFFSET Correlation with neighbor distance and angle on subtraction
- RELFWX Candidate FWHM in X divided by NEW image FWHM in X
- RELFWY Candidate FWHM in Y divided by NEW image FWHM in Y