Characterization, Modeling, and Simulation of Mouse Microarray Data

David S. Lalush

Bioinformatics Research Center

North Carolina State University

Acknowledgments

- Assistance from:
  - Jeff Tucker (NIEHS)
  - Pierre Bushel (NIEHS)
  - Bruce Weir (NCSU)

- Funded by K01 HG02428, National Human Genome Research Institute

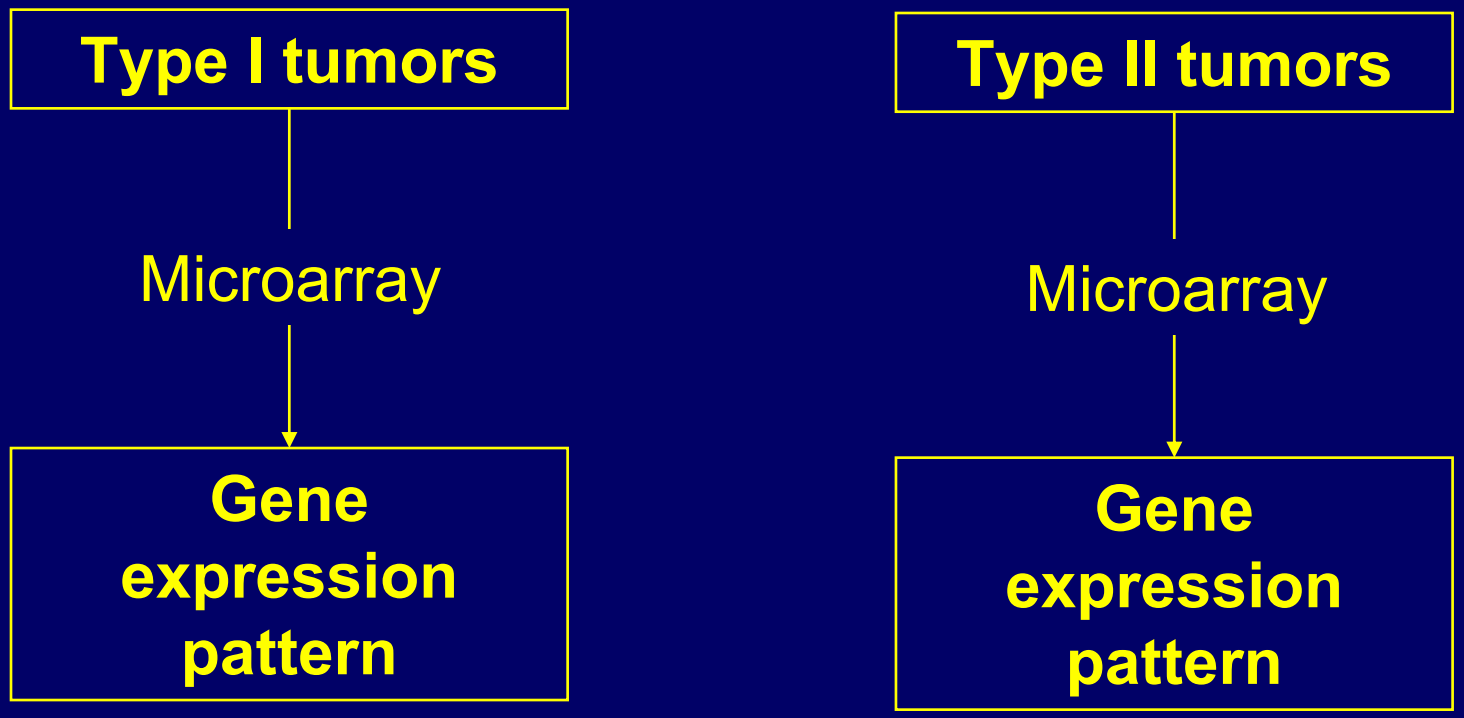
Outline

- Microarray Simulation Project
- Characterization of Microarray Images
- Results of Characterization
- Simulations
- Conclusion

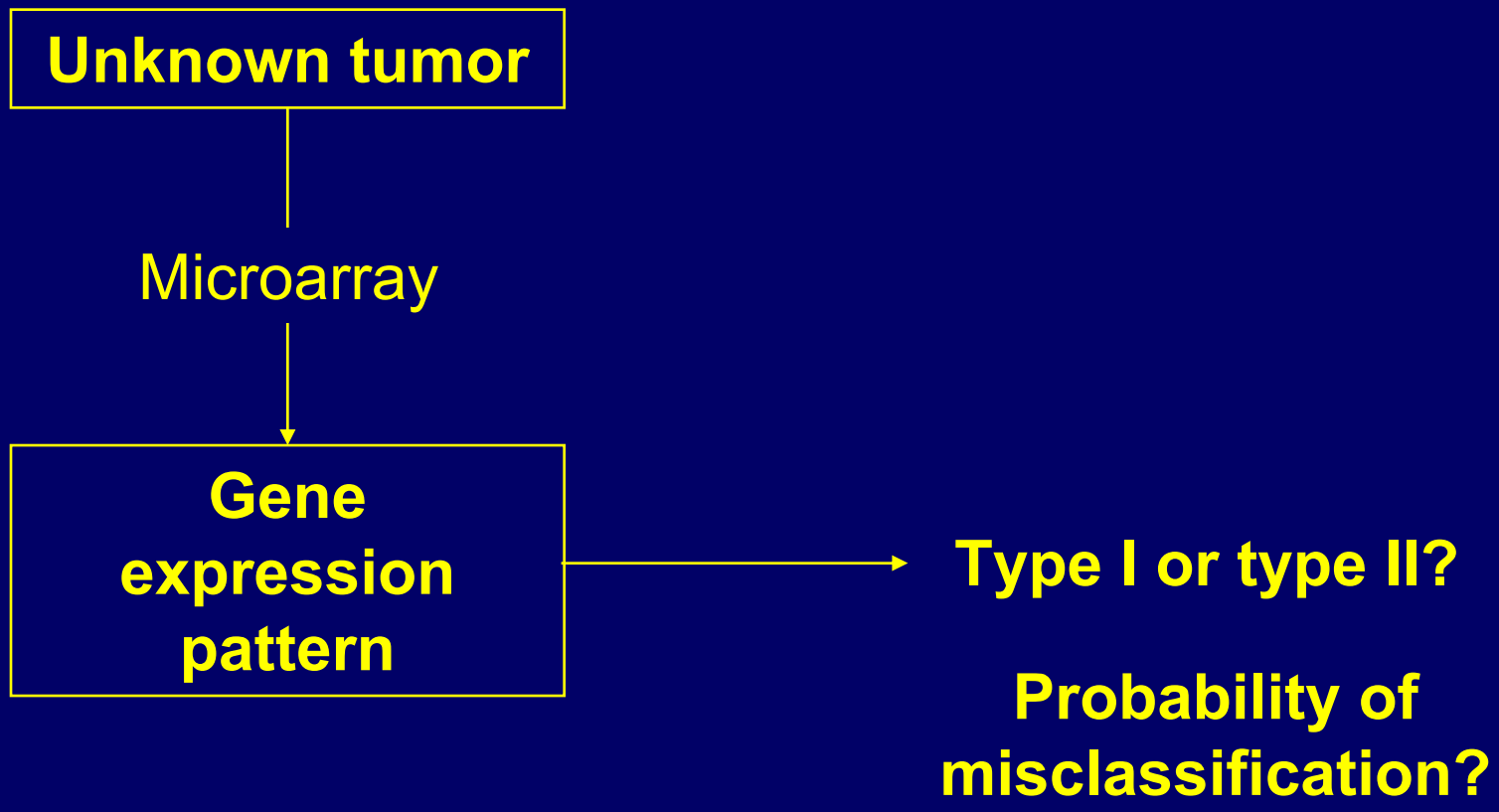
Outline

- Microarray Simulation Project
- Characterization of Microarray Images
- Results of Characterization
- Simulations
- Conclusion

Microarray in Diagnosis



Microarray in Diagnosis



Research Focus

- Evaluating classification methods
- Studying variability in microarray data

Problems:

•Many replications are required to evaluate error rates.

•Microarray experiments are expensive.

•True patterns are unknown in real data.

Microarray Simulation

- Creating a realistic simulation of microarray data
- Accounting for various sources of variability in the system

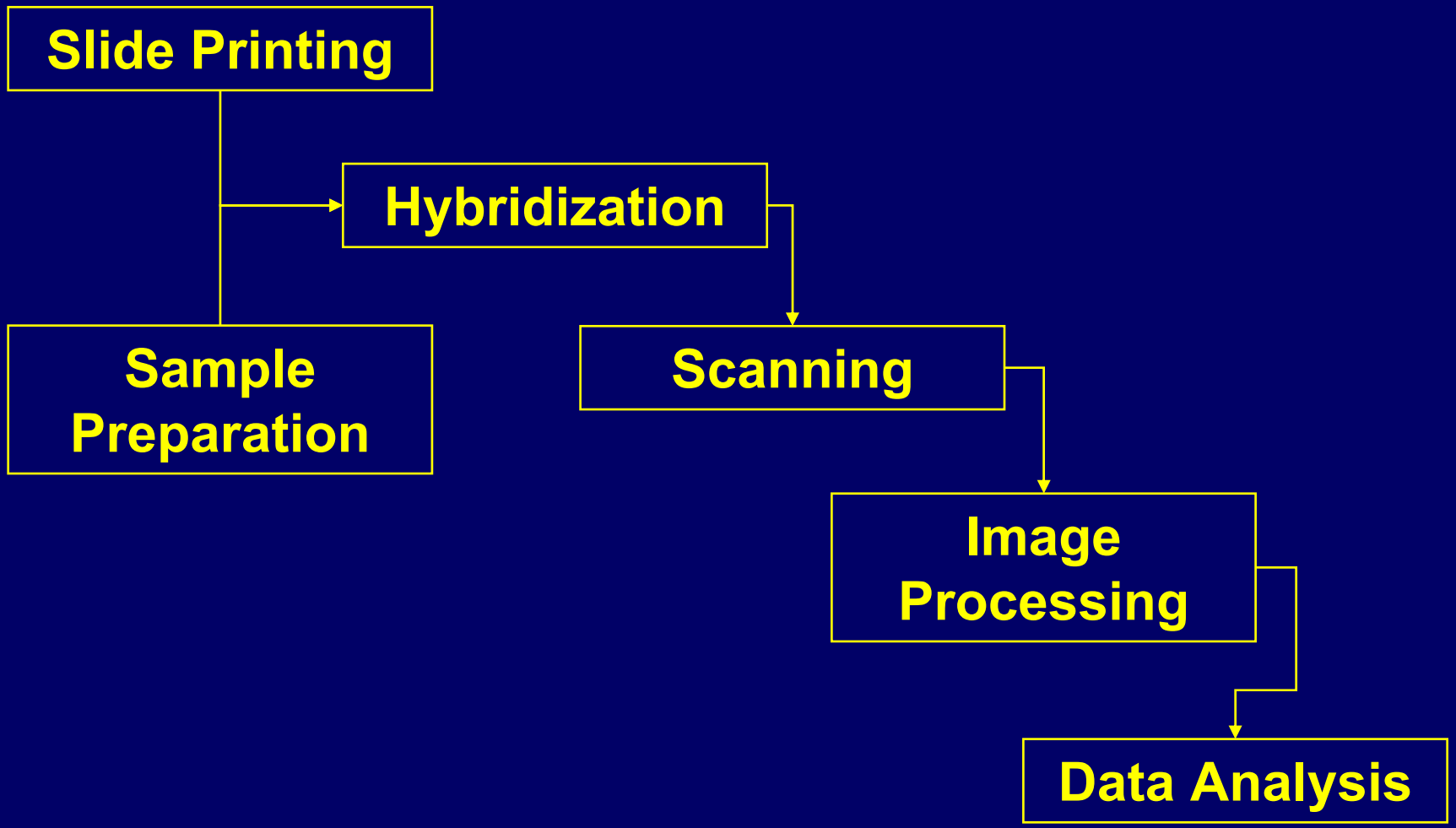
Advantages:

•Generates many replications cheaply.

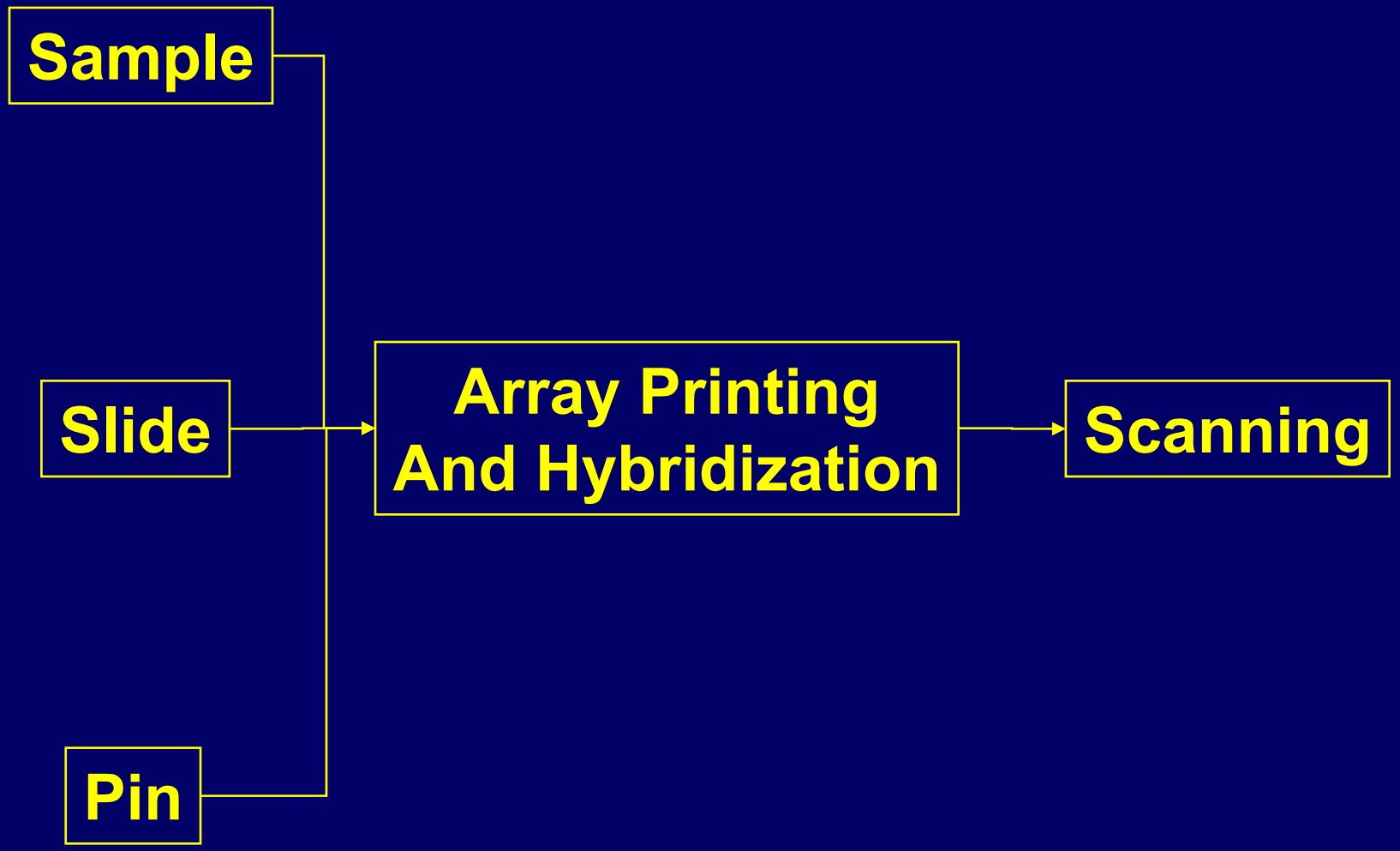
•True patterns are known.

•Can control sources of variability.

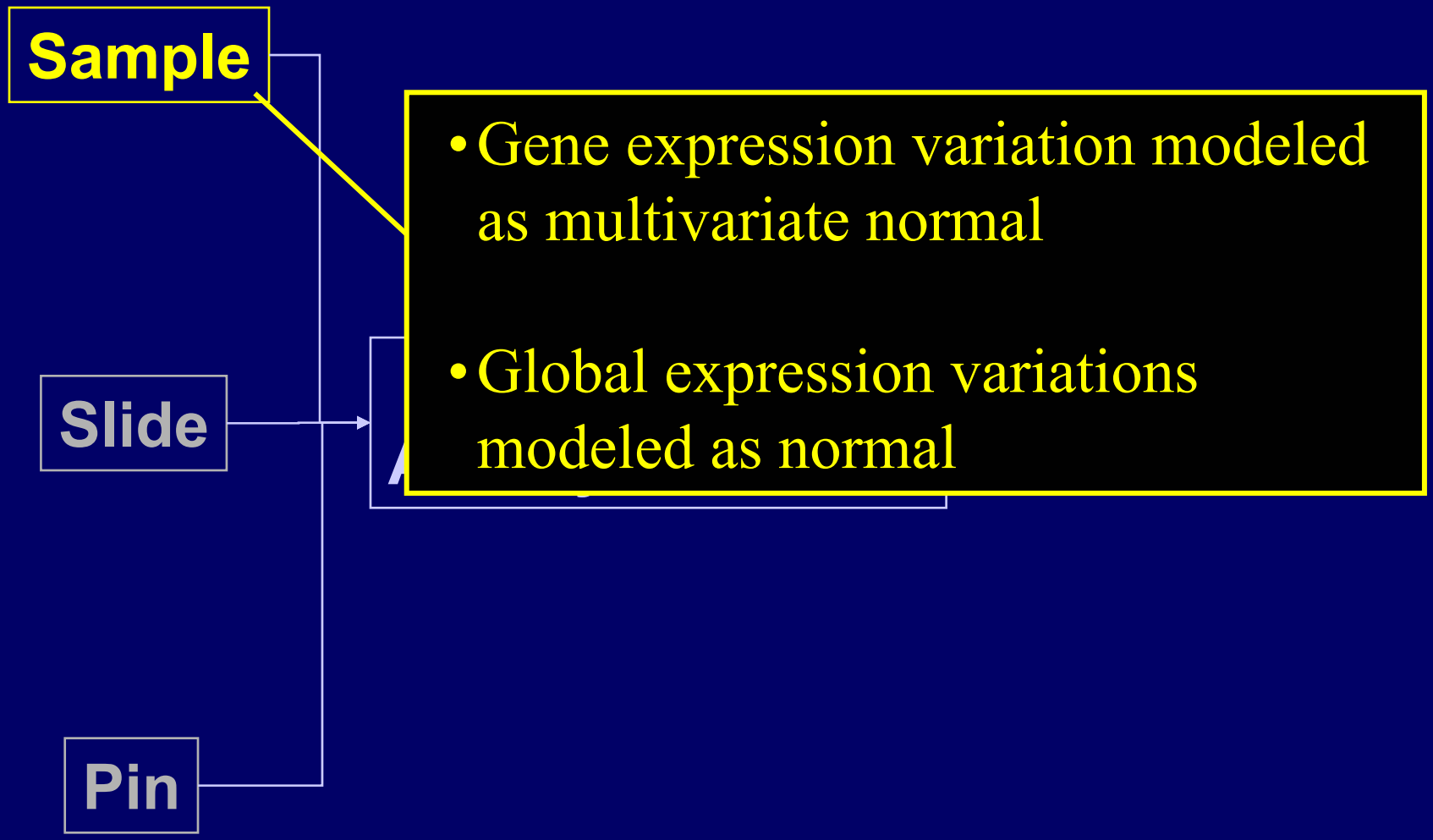
Microarray System



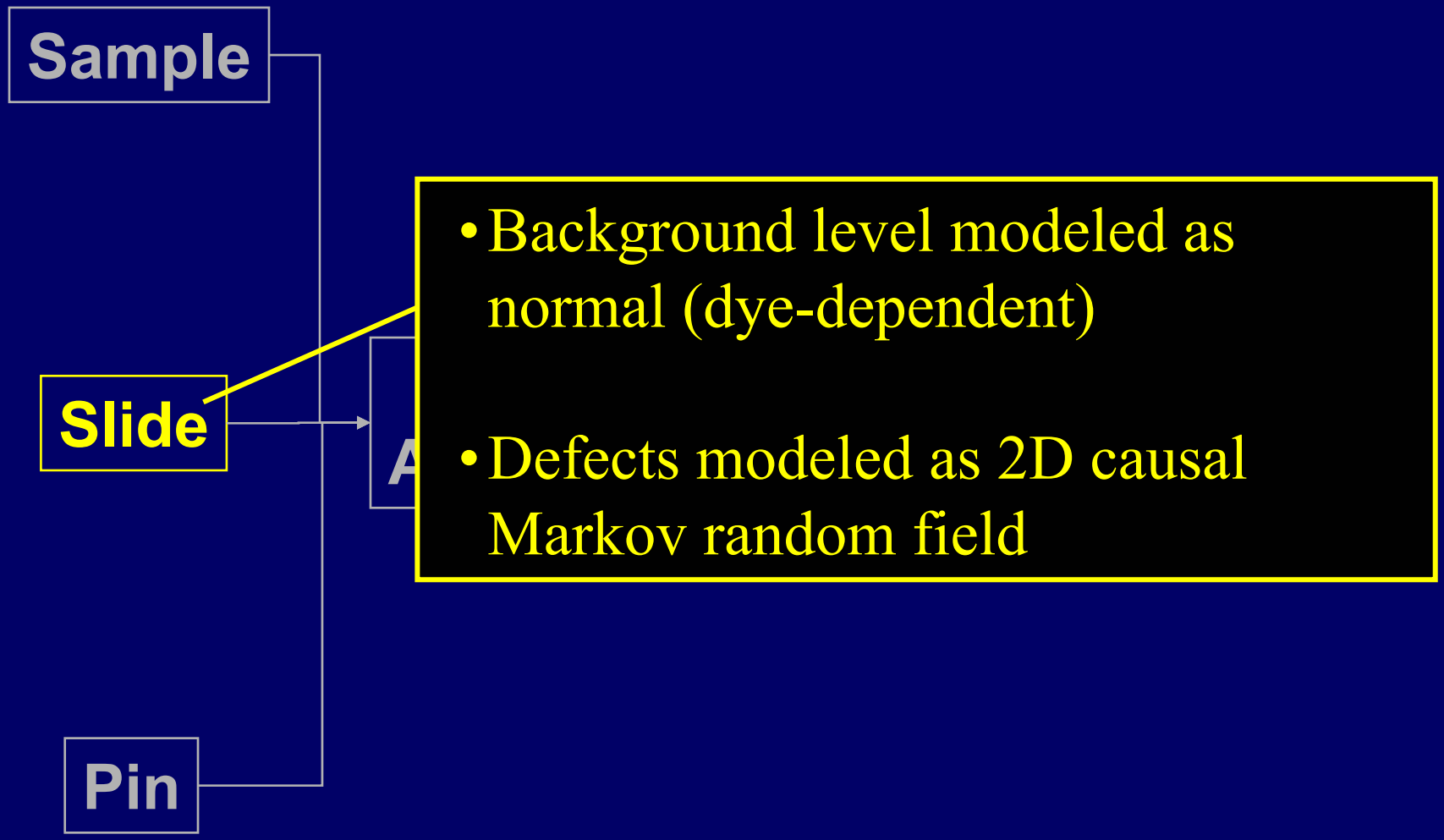
Simulation Model



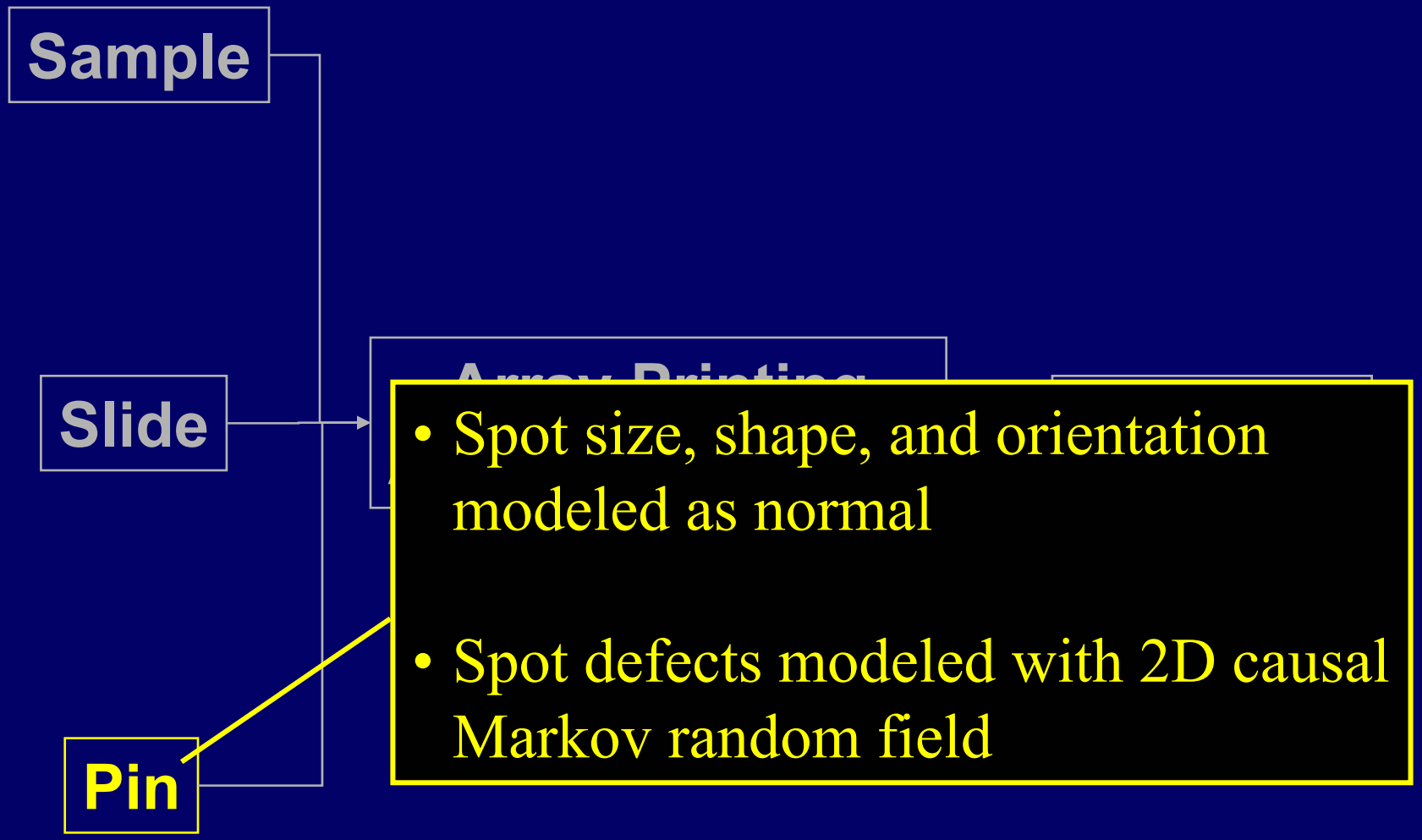
Simulation Model



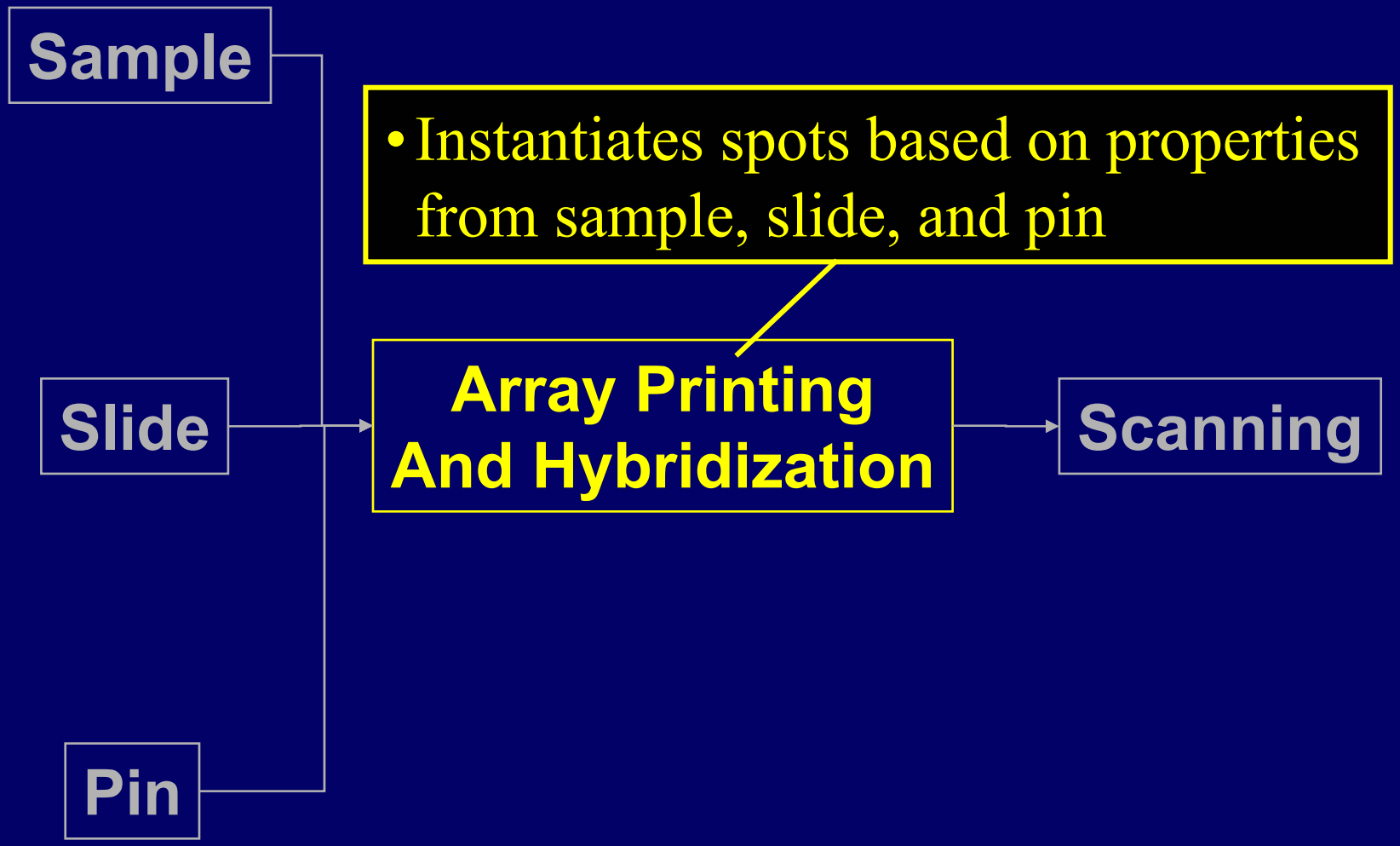
Simulation Model



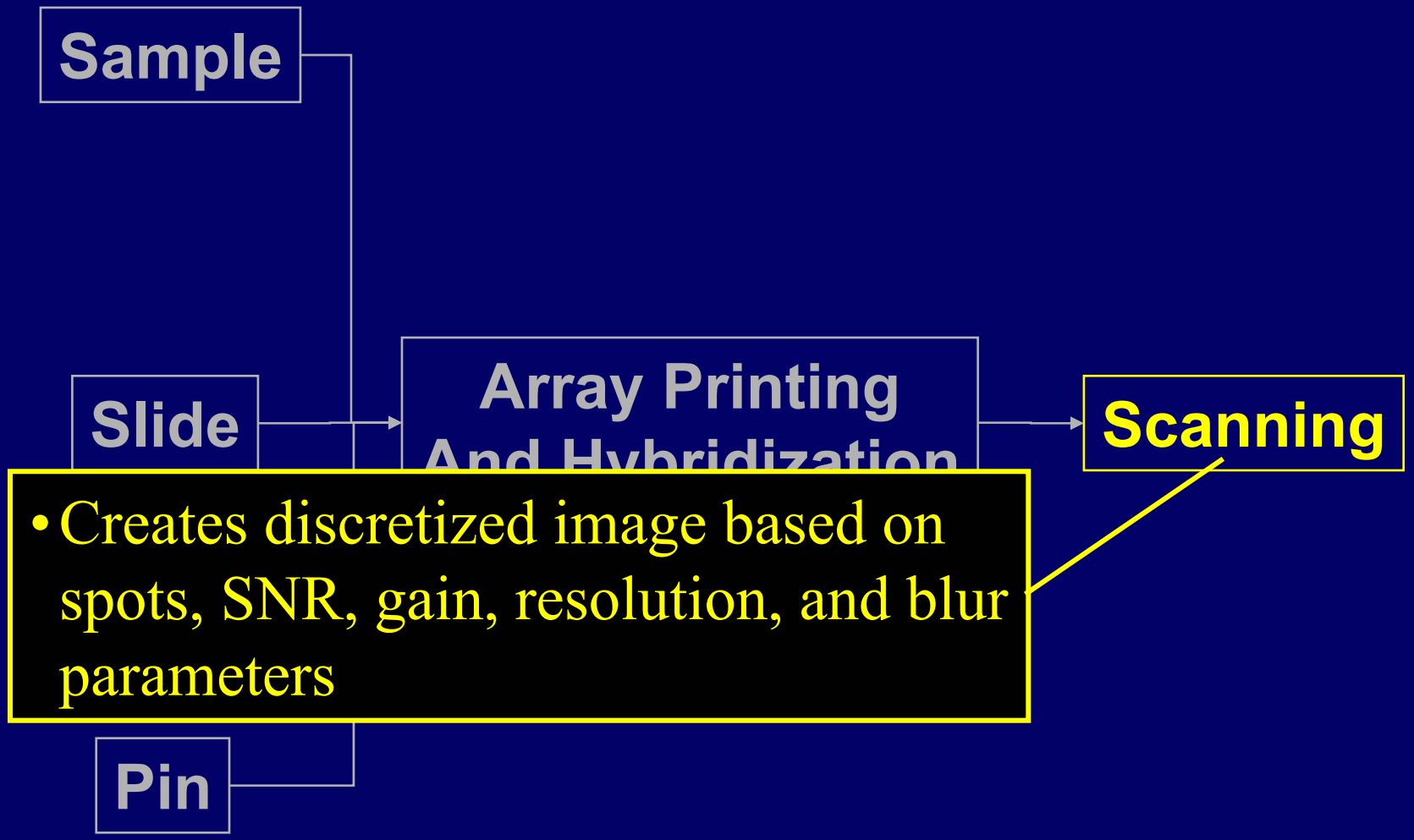
Simulation Model



Simulation Model



Simulation Model



Characterization

- Characterization of existing microarray images
  - Spot properties (size, shape, uniformity)
  - Pin properties (spot uniformity)
  - Slide properties (background, signal-to-noise)
  - Gene properties (mean, variance, covariance)

Outline

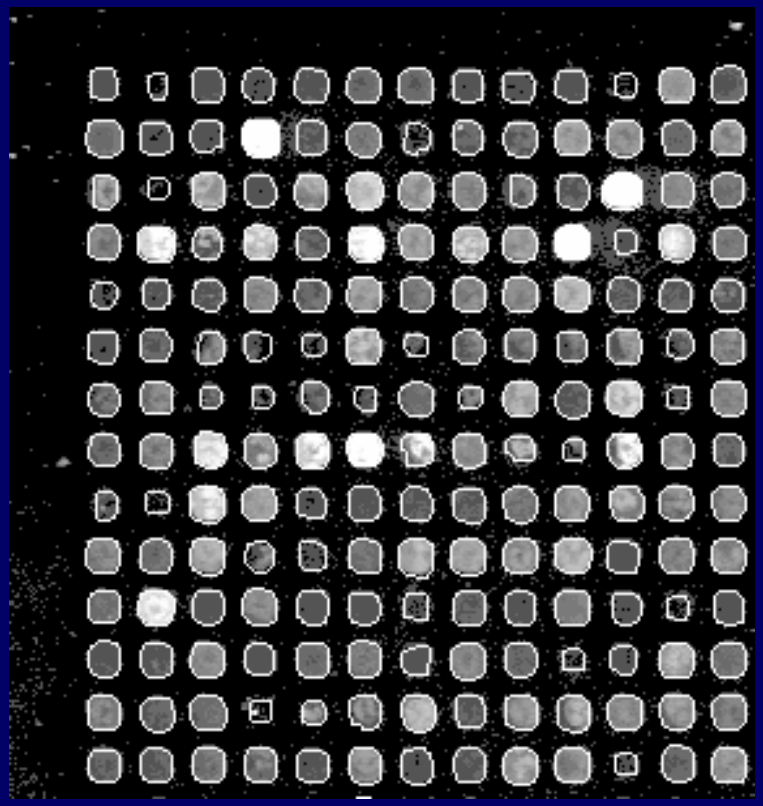
- Microarray Simulation Project
- Characterization of Microarray Images
- Results of Characterization
- Simulations
- Conclusion

Characterization

- Characterization of mouse kidney dataset
  - Six mice
  - Four slides each (2x2 fluor flip)
  - 24 slides in all
  - 5520 spots in 16 blocks, 4x4 block pattern

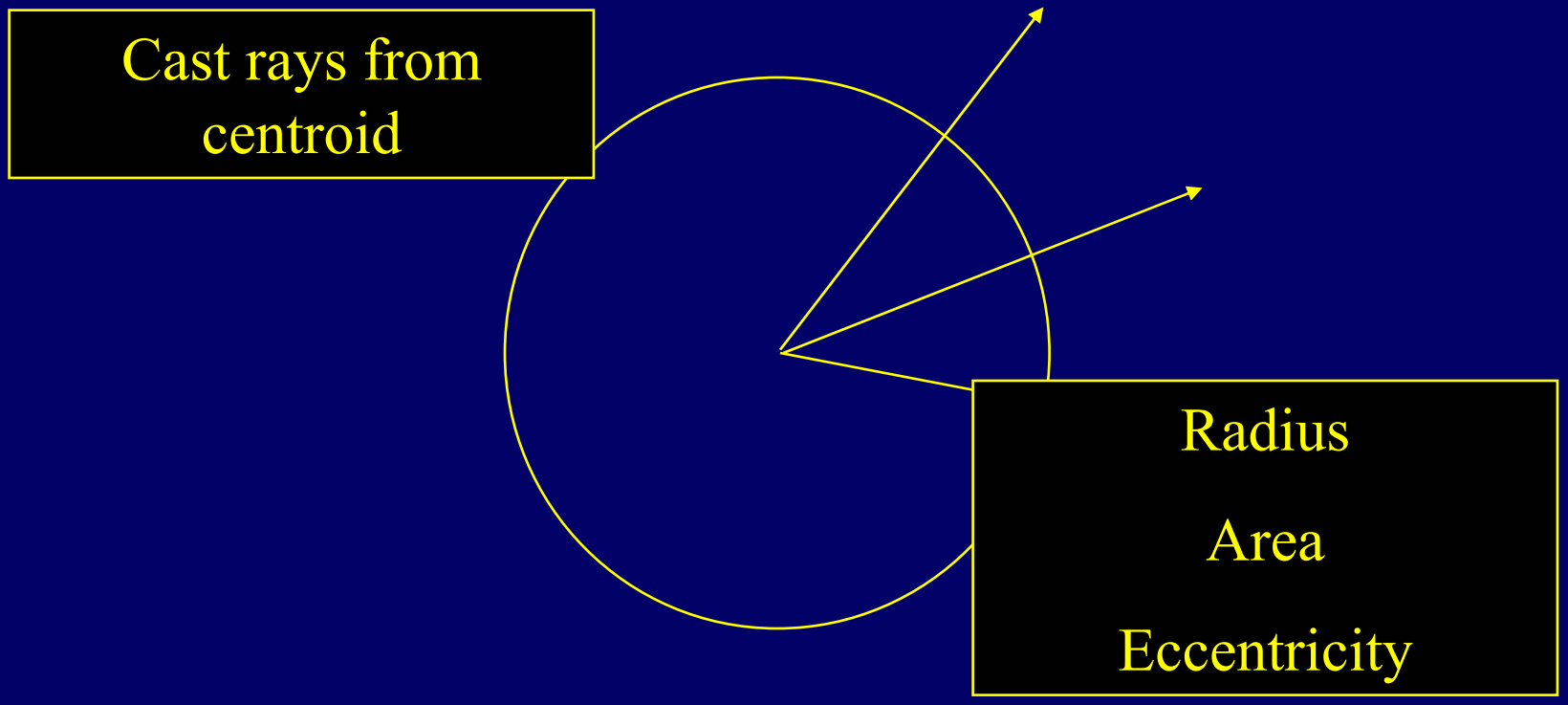
Characterization of Spots

- Step 1: Spot Detection



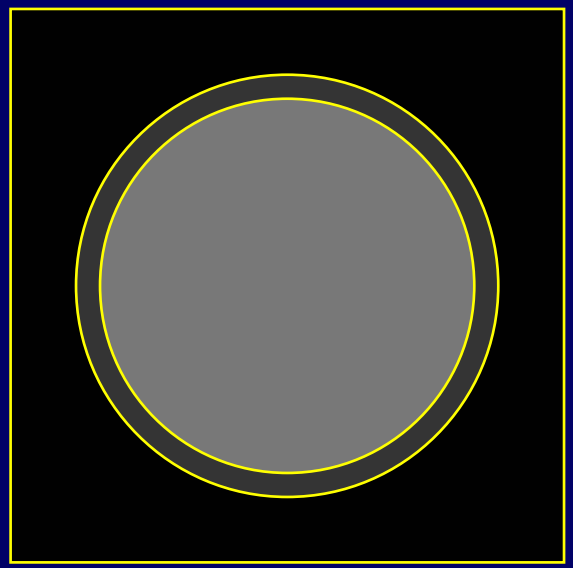
Characterization of Spots

- Step 2: Spot Morphology Measures



Characterization of Spots

- Step 3: Spot Intensity Measures
  - Mean and standard deviation of spot pixels
  - Mean and standard deviation of background pixels

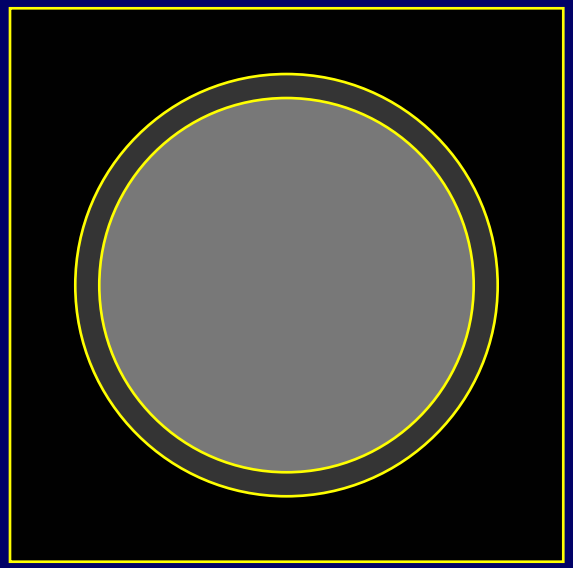


Characterization of Spots

- Step 4: Secondary Intensity Measures

Separability

$$\frac{(signal - background)}{\sqrt{\sigma^2_{signal} + \sigma^2_{background}}}$$

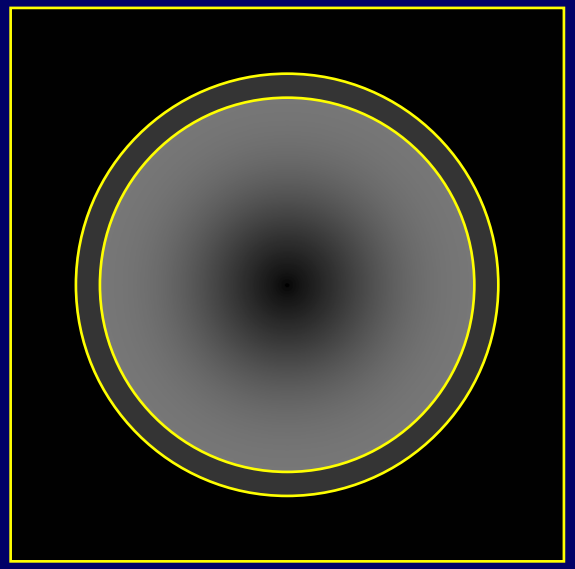


Characterization of Spots

- Step 4: Secondary Intensity Measures

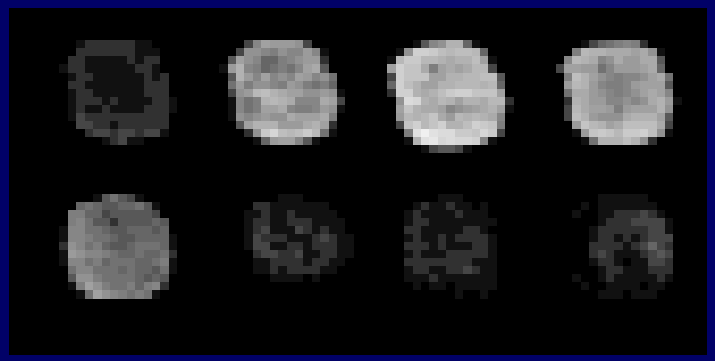
Spot Uniformity

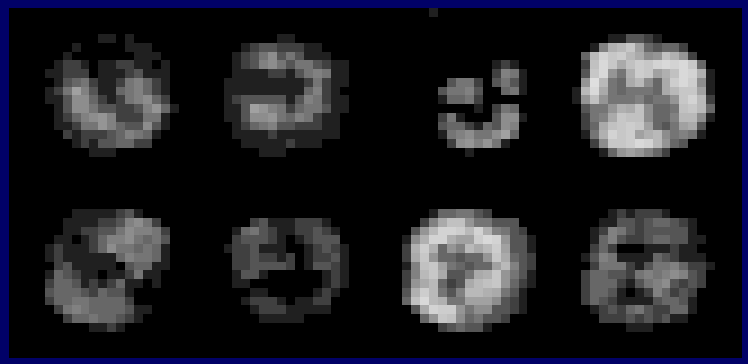
$$\frac{\sigma_{signal}}{signal}$$



Characterization of Spot Defects

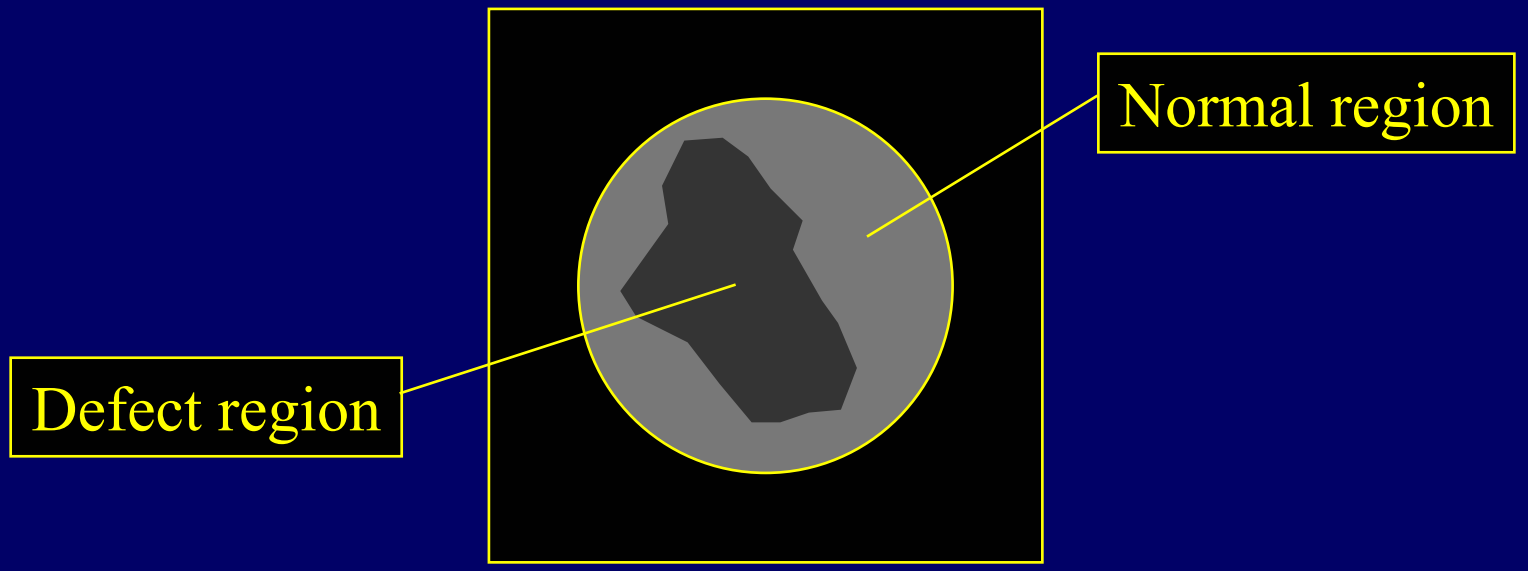
- Spots often exhibit characteristic nonuniformities
  - Low center
  - Spot breaks





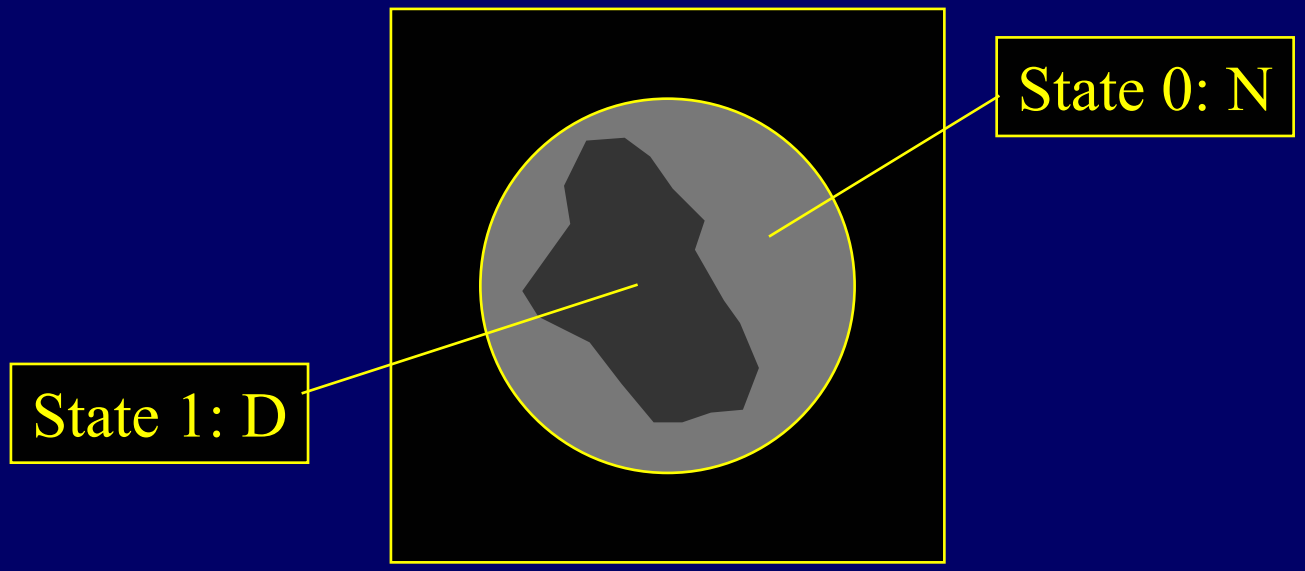
Characterization of Spot Defects

Consider each spot to have two regions



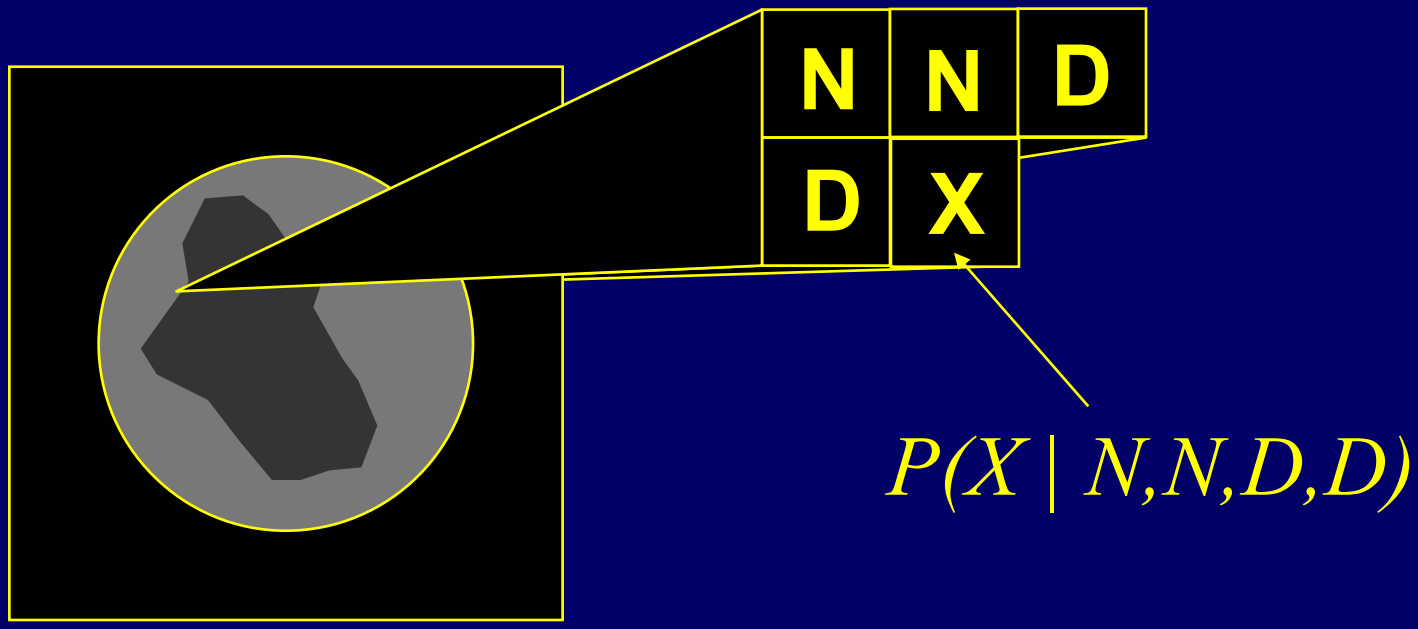
Characterization of Spot Defects

Each region acts as a hidden state.
Each state has its own distribution of emitted intensities.



Characterization of Spot Defects

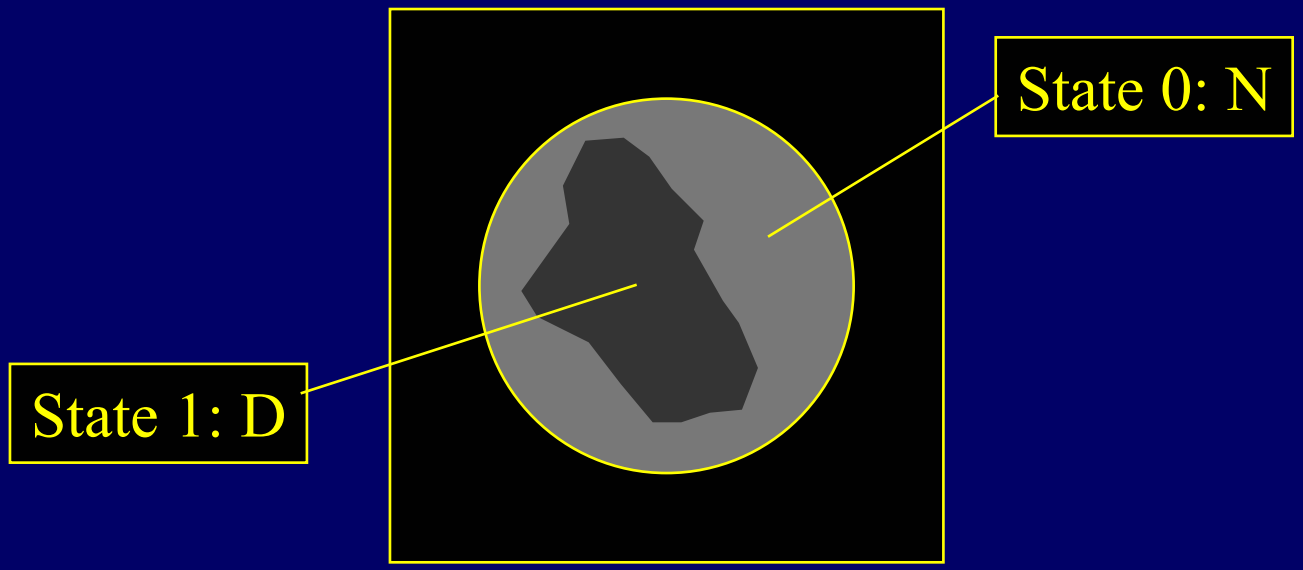
The probability of a pixel being in a given state depends on its neighbors.



Characterization of Spot Defects

Region Model (2D causal MRF):

- 16 parameters for state transition
- 2 parameters for intensity of D region pixels relative to N region (mean, s.d.)

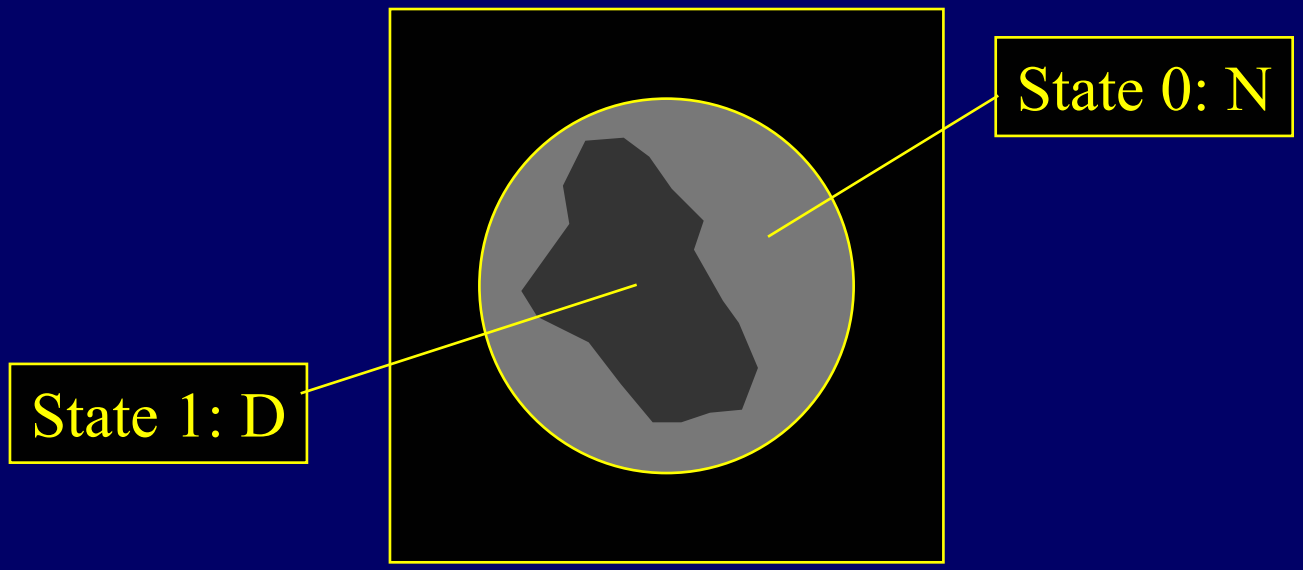


Characterization of Spot Defects

Applying the Region Model

Pixel is in D region if:

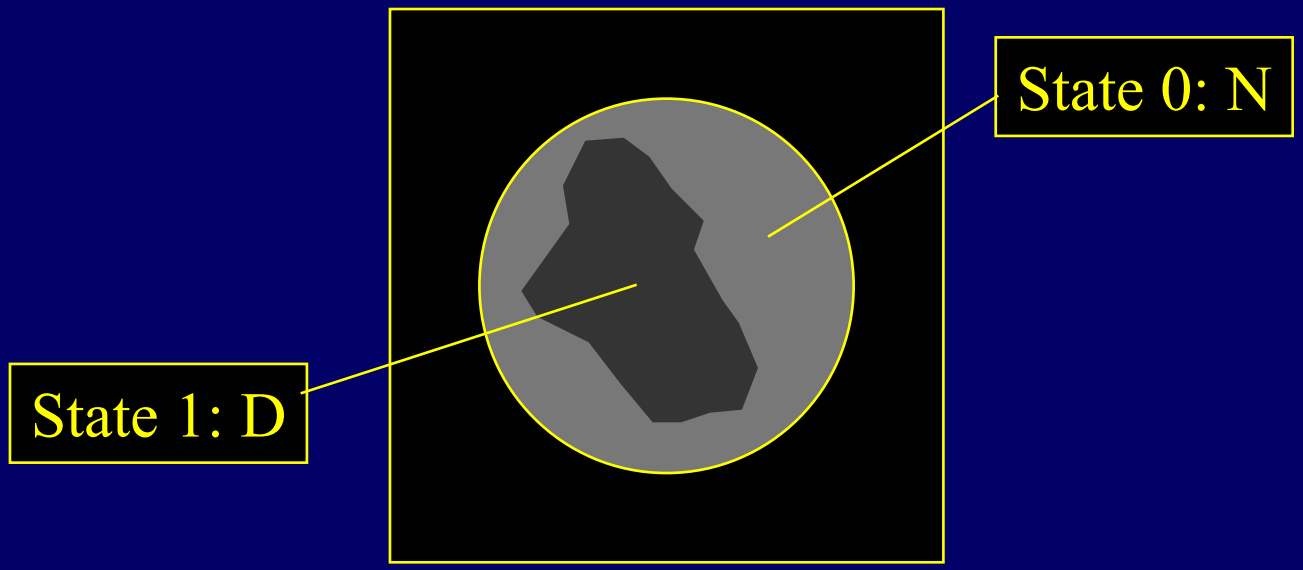
- It is in the spot
- It is below the spot average intensity in BOTH channels



Characterization of Spot Defects

Applying the Region Model

- Smooth region boundary
- Compute the 18 parameters for each spot

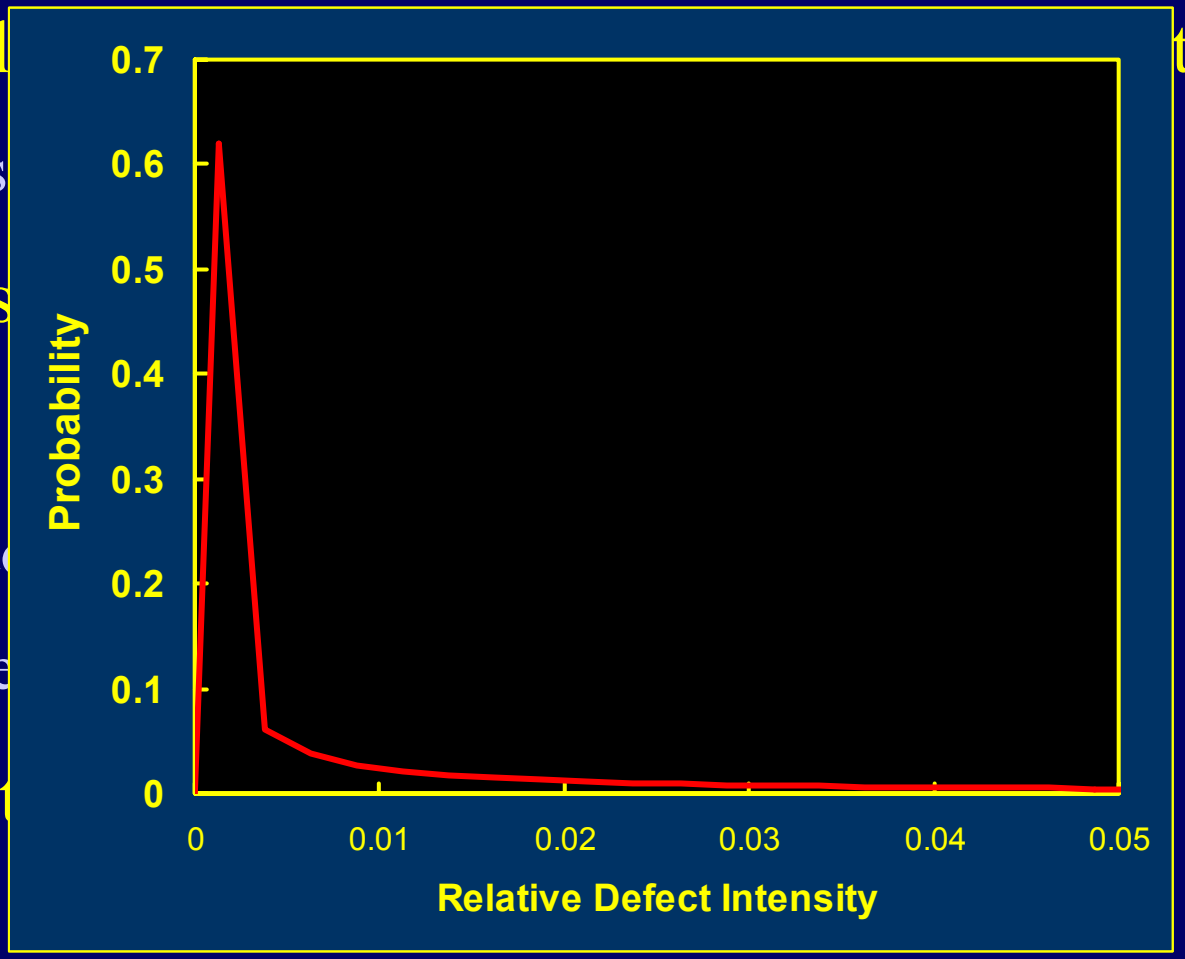


Characterization of Background

- Base level and variation
  - Modeled as stationary across slide
- Background defects
  - Marks, scratches, bright spots, other features
  - Modeled with 2D Markov random field

Characterization of Background

- Classify al
  - Defect is
- Compute s
- Apply 2D
  - Similar t
  - Intensitie
- Measures



Characterization of Gene Expression

- Multivariate normal distribution for each sample (test or reference)
  - Mean vector
  - Covariance matrix
- Linear model to account for global effects from slide to slide and dye effects

Sample = (mean gene expression) + slope * (slide perturbation) + (variable expression)

Characterization of Gene Expression

- Problem: Covariance matrix is BIG (5200x5200)
  - In simulation, we will have to diagonalize it.
- Model the most significant correlations
  - Compute correlations between each pair of genes on each slide
  - Cluster genes by correlation distance
  - Each gene in a cluster has greater than .48 absolute correlation with every other gene in the cluster

Analyzing Characterization Data

- Two-way ANOVA
  - By slide (fixed)
  - By pin (random)
- Which properties varied more?
  - By slide
  - By pin
  - By spot

Analyzing Characterization Data

- Spot morphology measures
- Spot secondary intensity measures
- Spot defect model parameters
- Background defect model parameters (by slide measurement only - no ANOVA)

Only spots with separability > 1 used in ANOVA

Outline

- Microarray Simulation Project
- Characterization of Microarray Images
- Results of Characterization
- Simulations
- Conclusion

Results

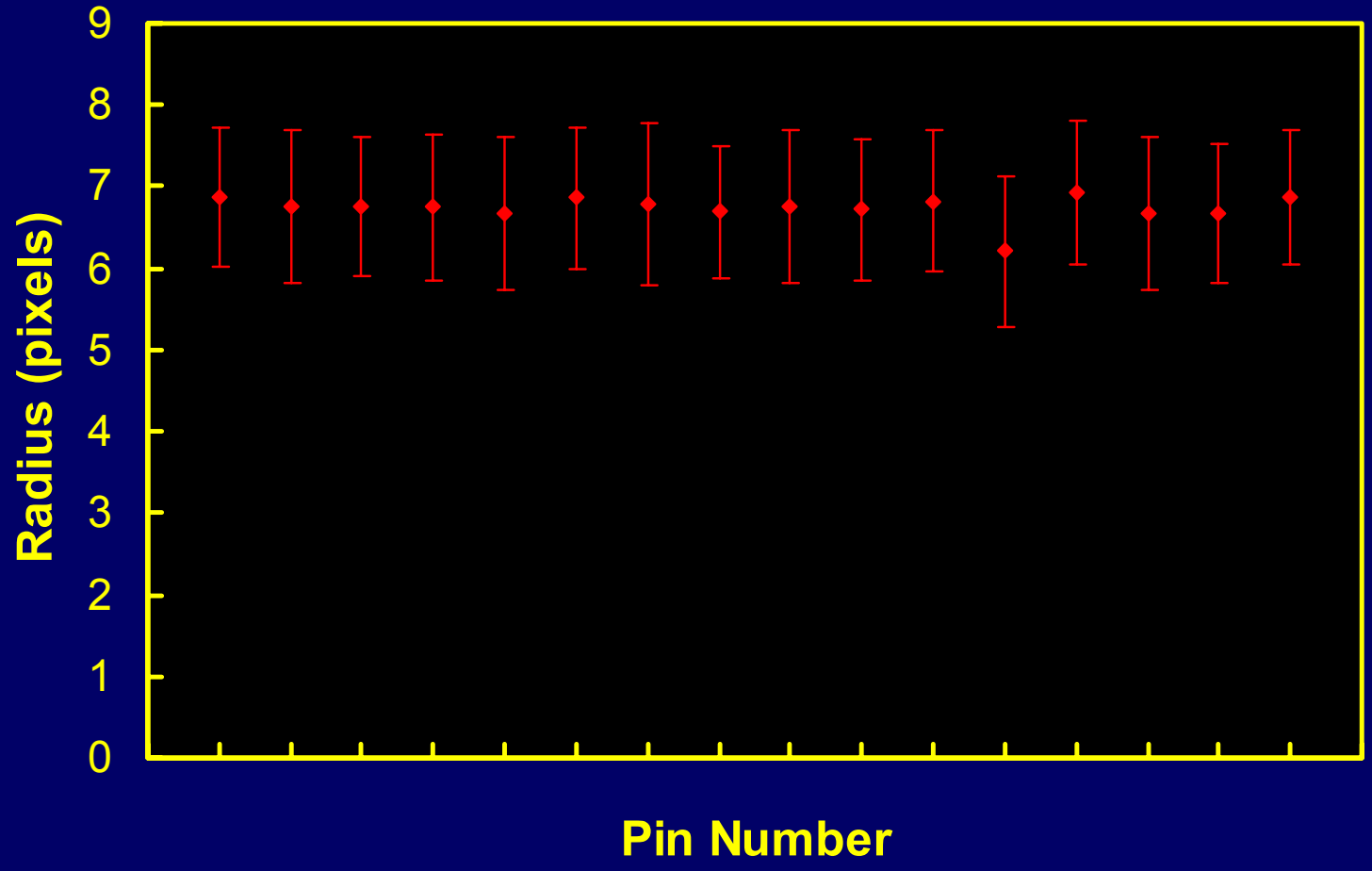
Sometimes the images have their own story to tell.

Results: Spot Morphology

- Most variation (75% for size measures) was attributed to variation by spot
- Pins behaved similarly (mostly)
- Slides showed some differences in last eight slides (mice five and six)

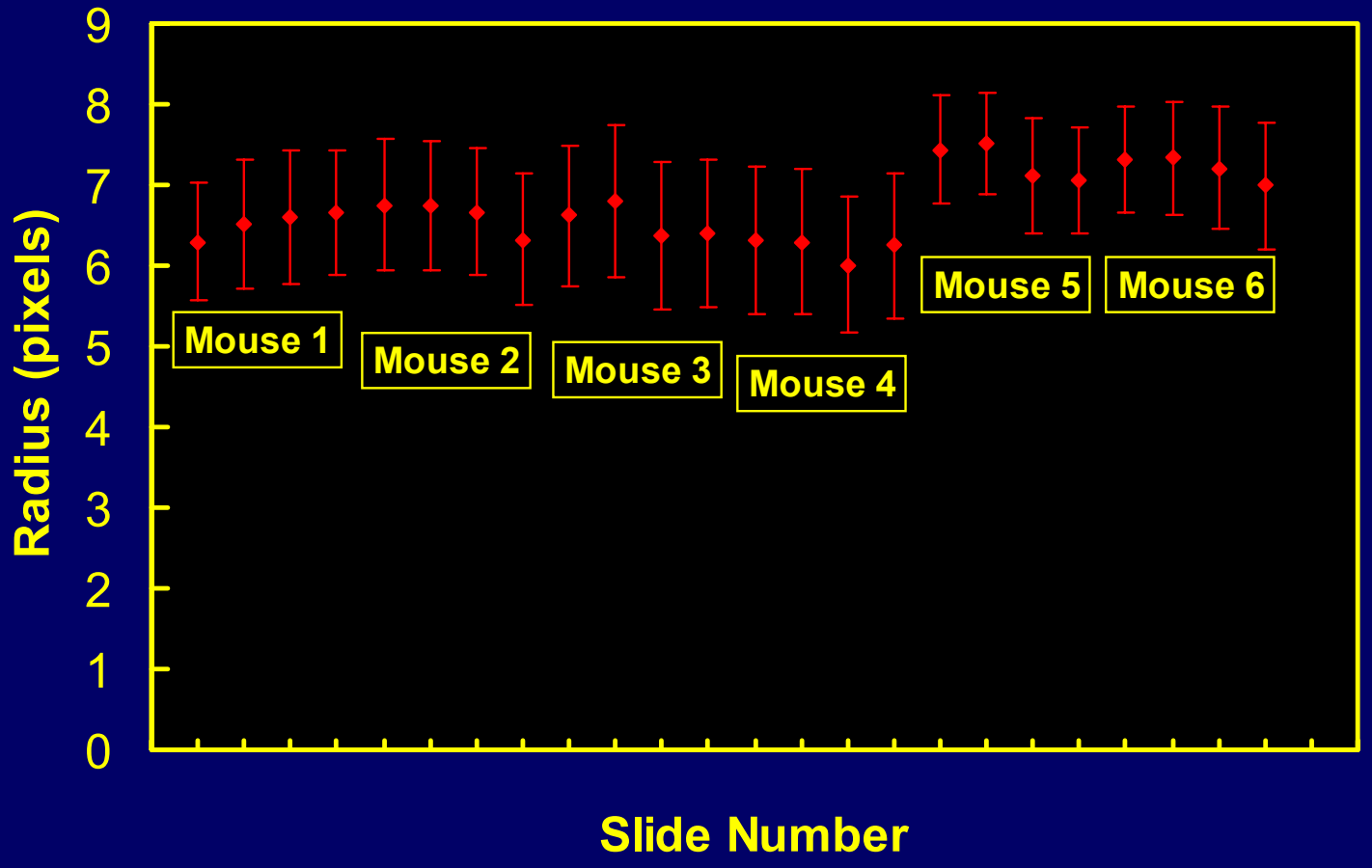
Results: Spot Morphology

Spot size vs. Pin Number



Results: Spot Morphology

Spot size vs. Slide Number



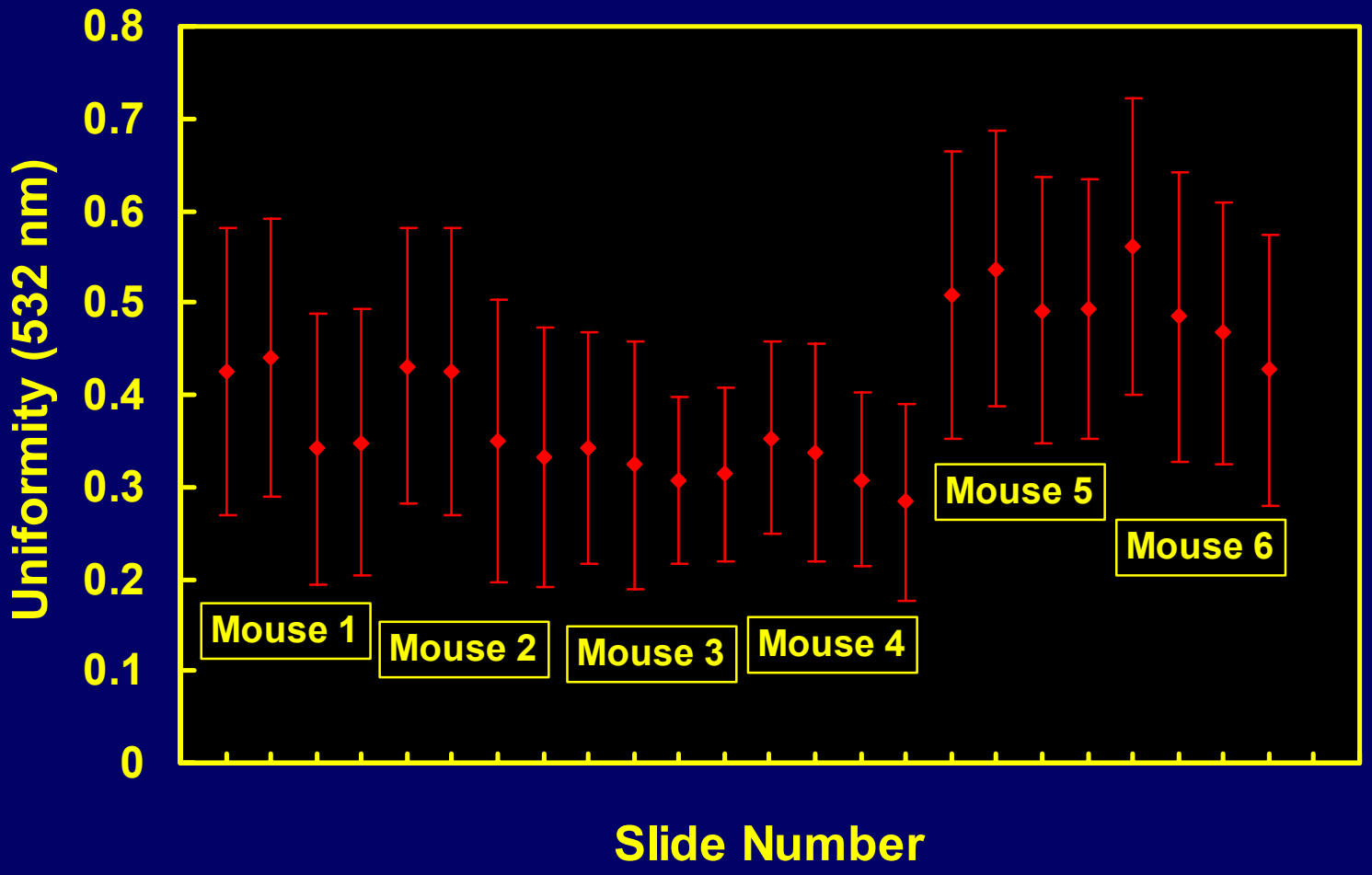
Results: Spot Intensities

- Most variation in separability (83-90%) was attributed to variation by spot

- Spot uniformity varied considerably by slide, mostly due to last eight slides

Results: Spot Intensities

Spot uniformity (532nm) vs. Slide Number

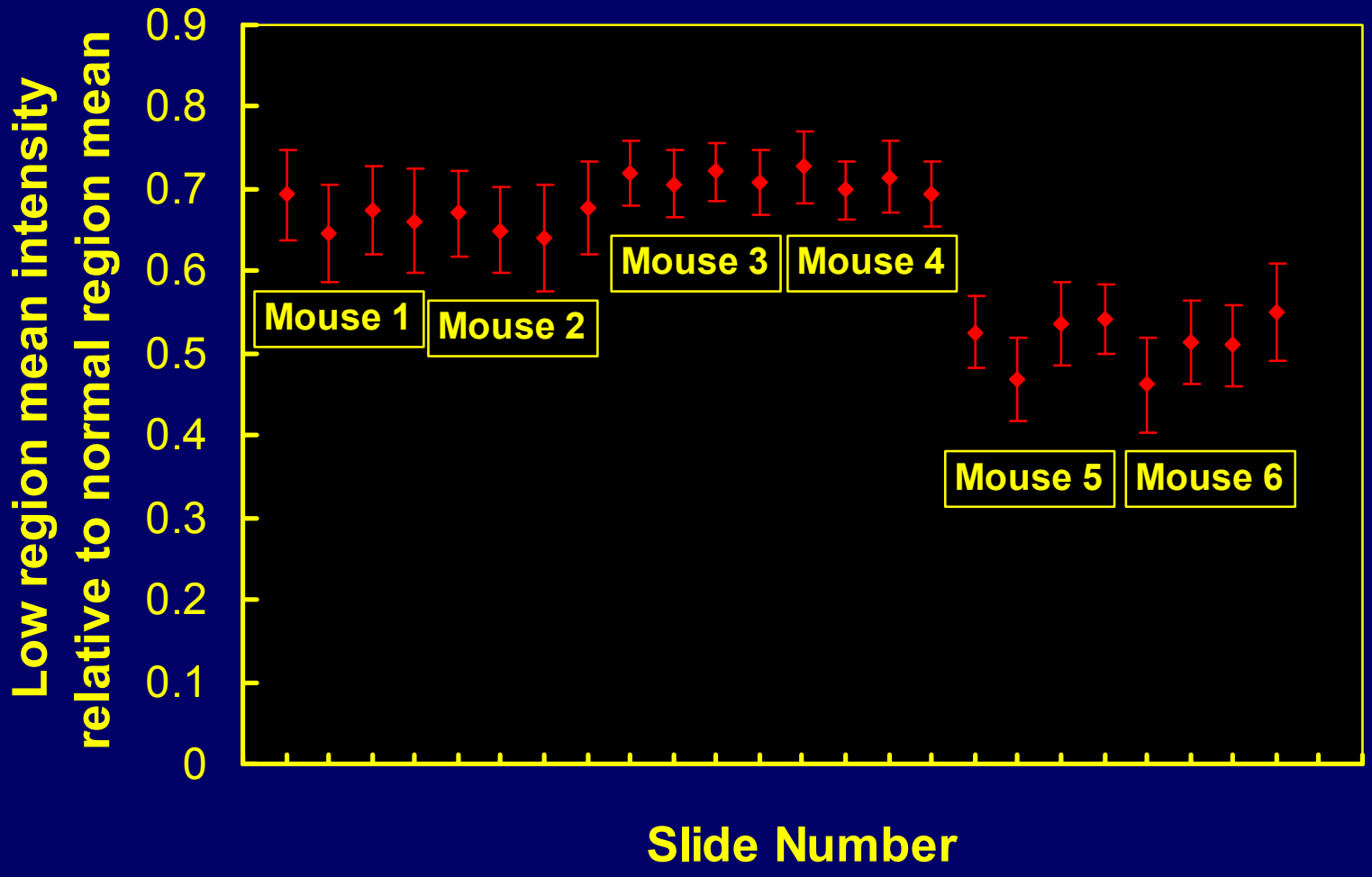


Results: Spot Defect MRF

- The 16 region transition probability parameters varied by pin
  - Model the MRF as a property of a pin, not a slide
- The mean intensity of defect region was strongly dependent on the pin.
- Mean intensity of defect region varied considerably by slide.

Results: Spot Defect MRF

Defect region intensity vs. Slide Number



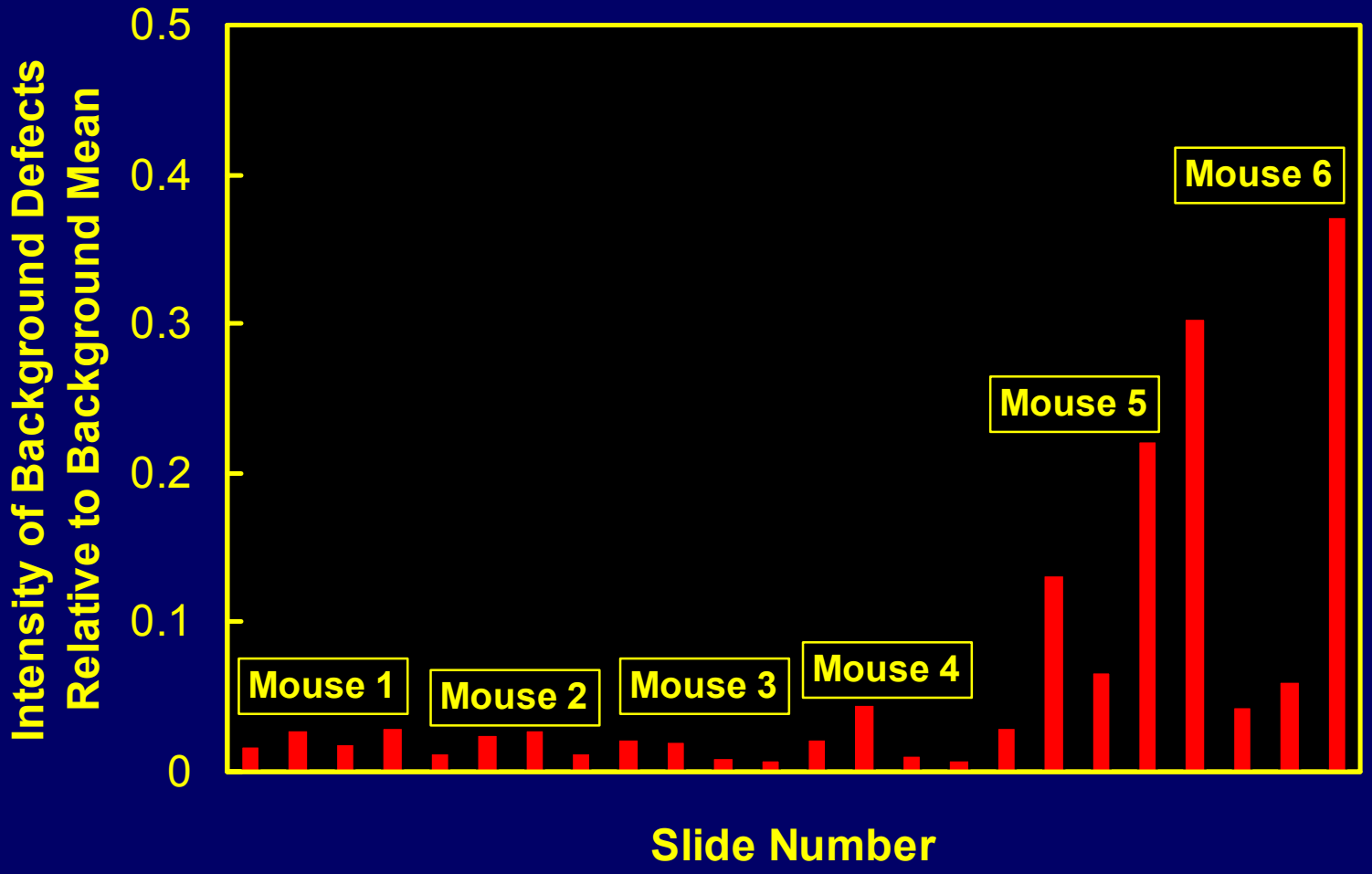
Results: Background MRF

- Last eight slides had more intense background defects

- Last eight also had higher probabilities of generating a defect

Results: Background MRF

Background defect intensity vs. Slide Number



Results: General

- Slide-pin interactions were small (<5% of variance in all cases)

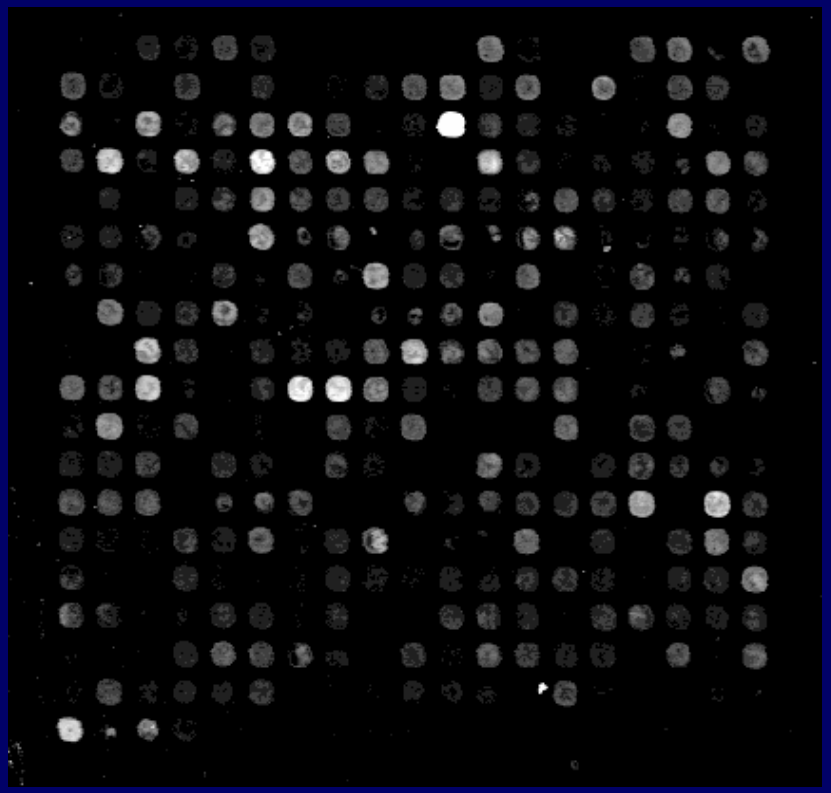
- Therefore, modeling of slide and pin effects separately is justified.

Results: Summary

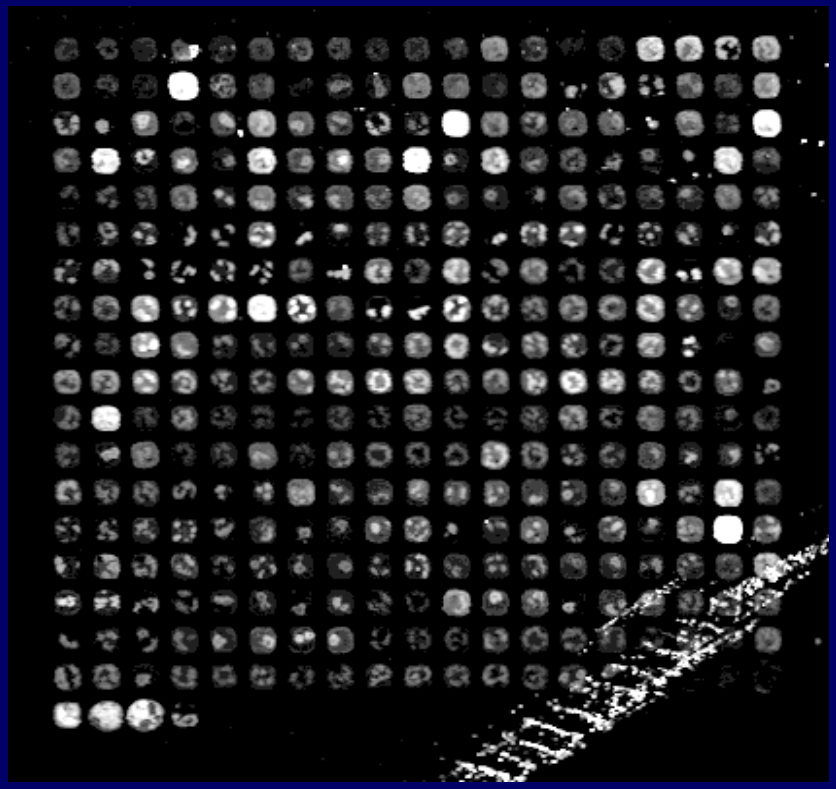
- Characterization shows differences in the properties of slides for mice five and six:
  - Spots were more likely to be broken.
  - Spot breaks were more severe.
  - Background defects were more numerous.
  - Background defects were more intense.

Did this impact the estimated mouse-to-mouse variation?

Results



Slide 2 (Mouse 1)



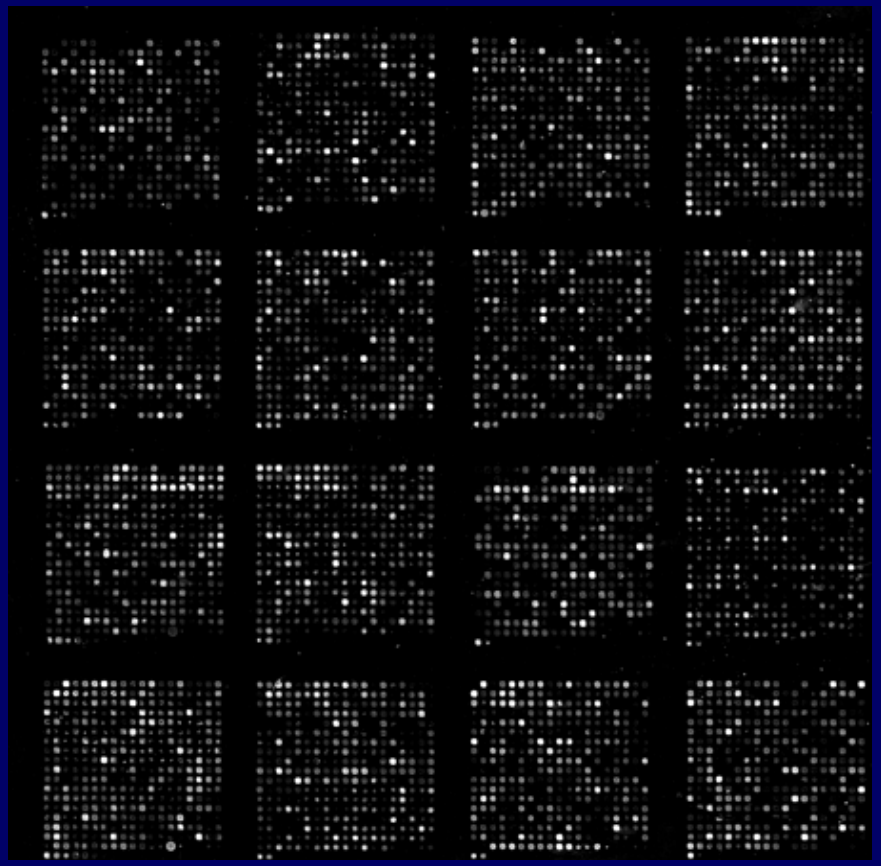
Slide 19 (Mouse 5)

Outline

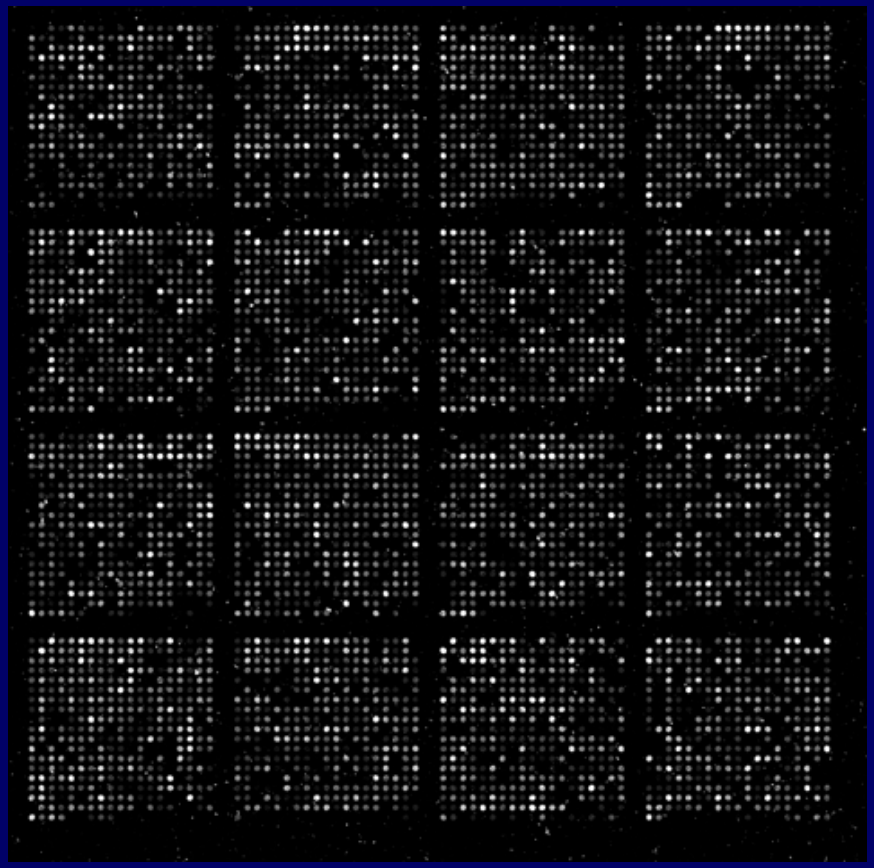
- Microarray Simulation Project
- Characterization of Microarray Images
- Results of Characterization
- Simulations
- Conclusion

Simulations

From mouse 1-4 properties



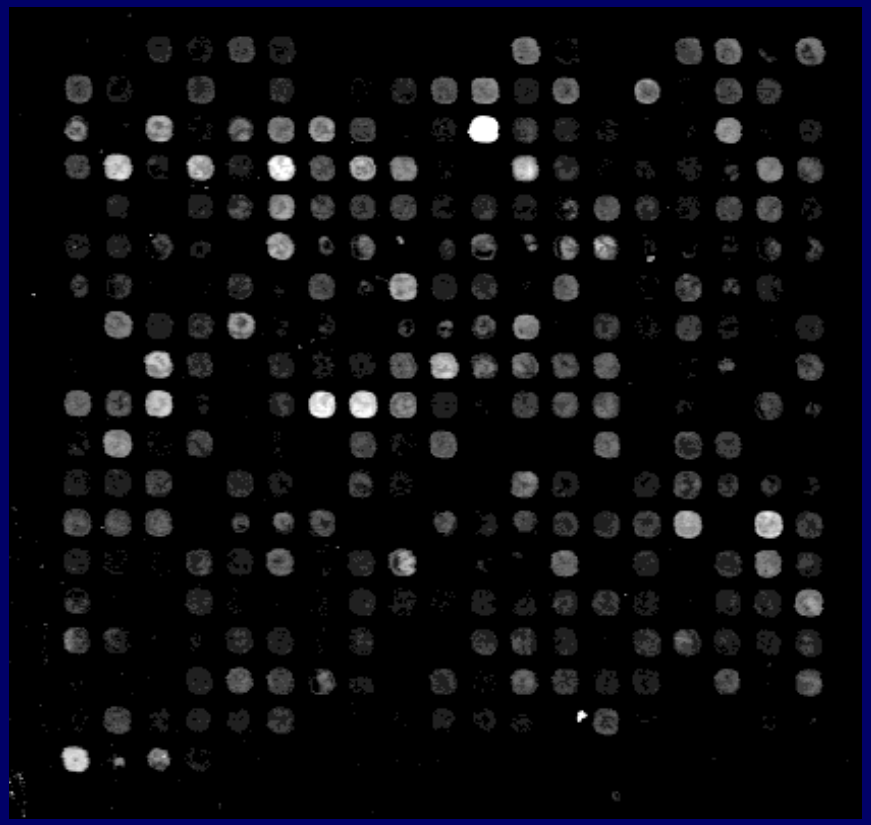
Slide 2 (Mouse 1)



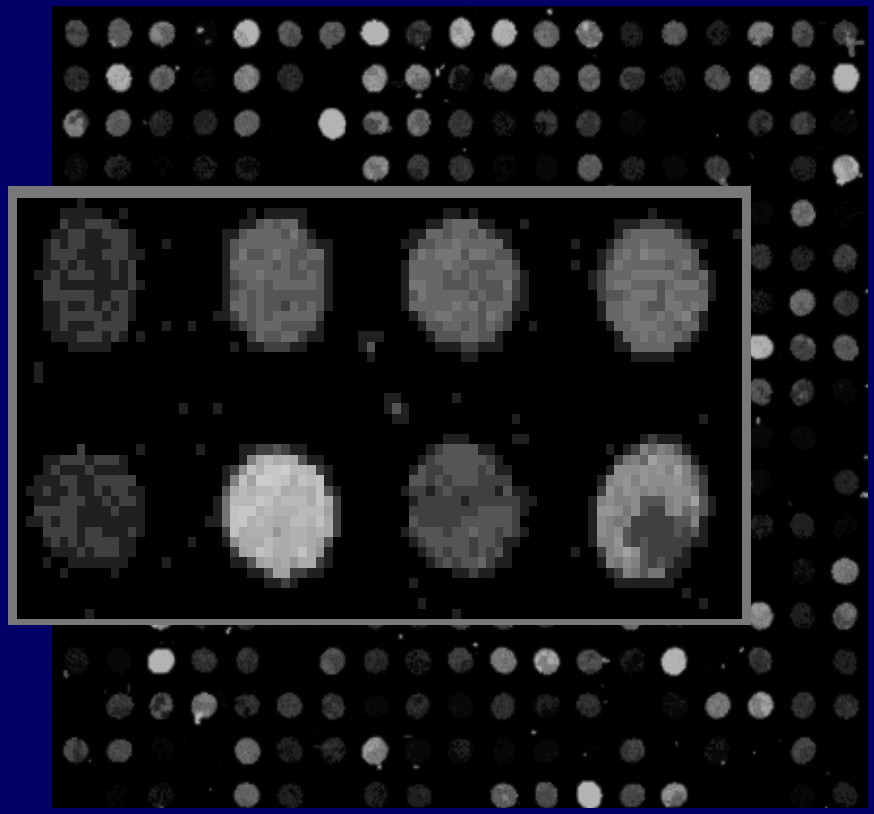
Simulation

Simulations

From mouse 1-4 properties



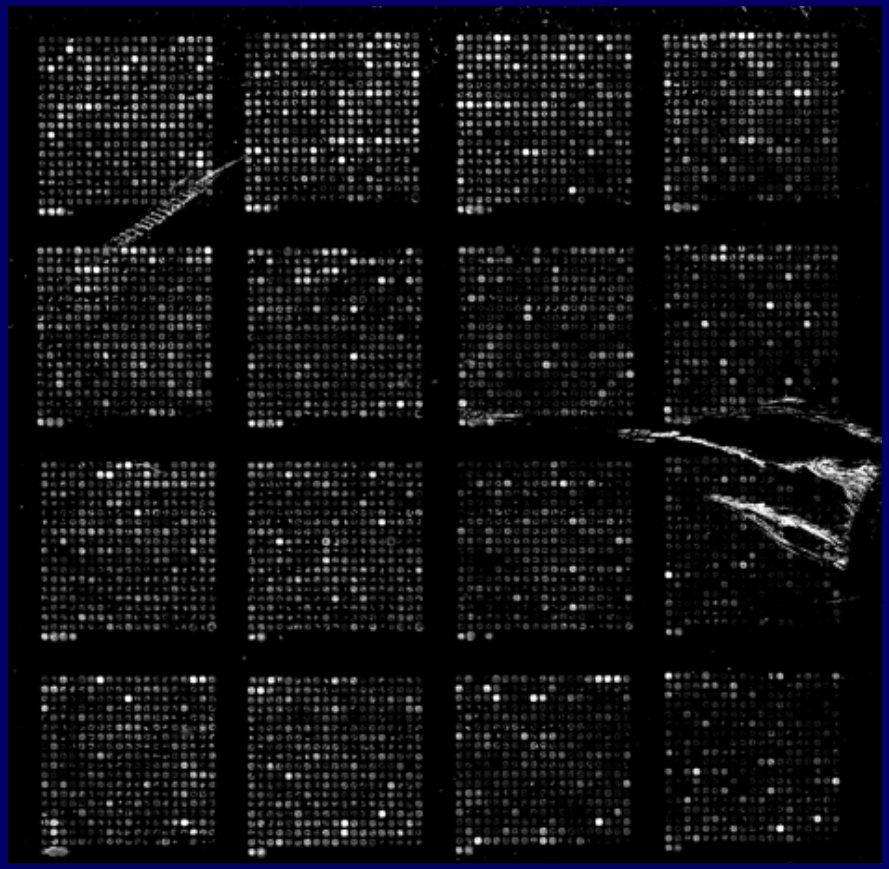
Slide 2 (Mouse 1)



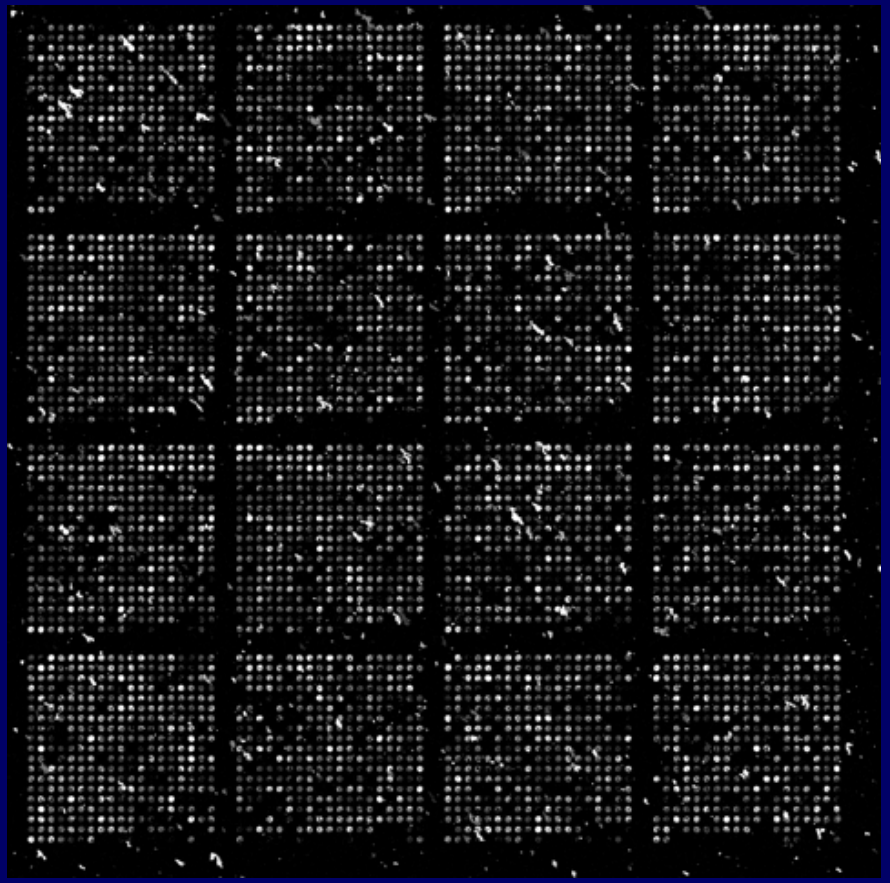
Simulation

Simulations

From mouse 5,6 properties



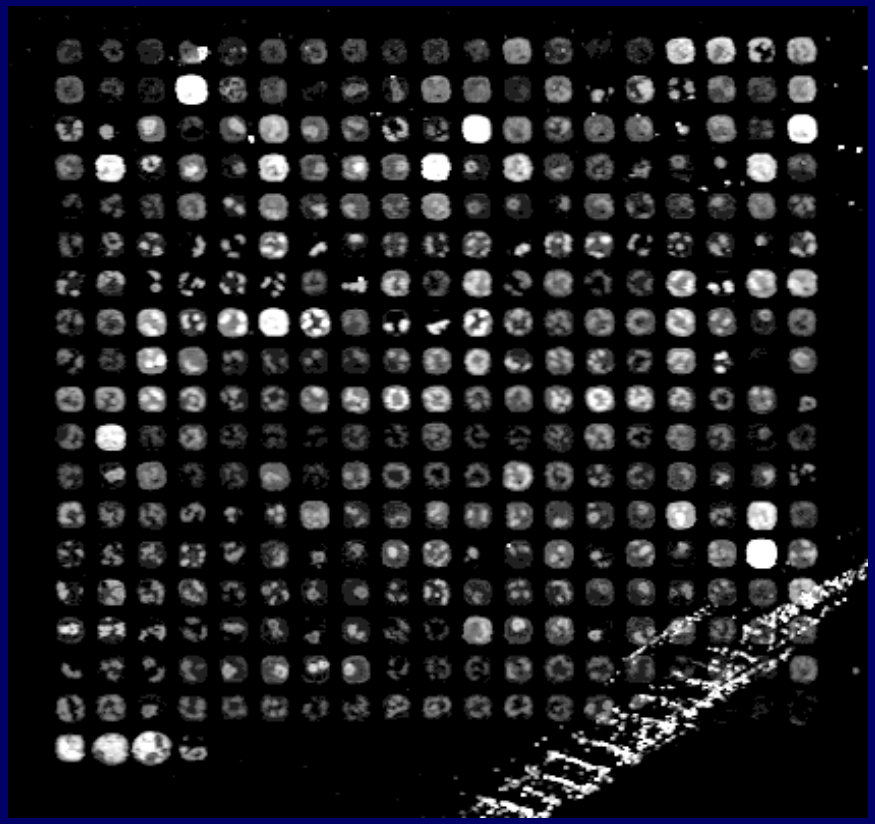
Slide 19 (Mouse 5)



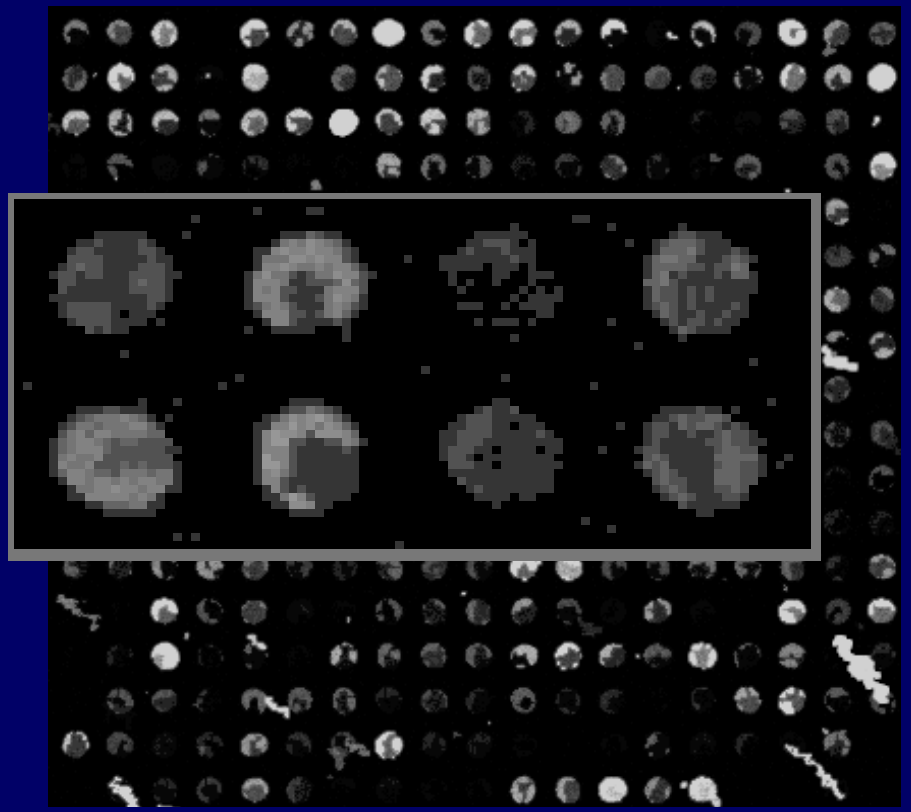
Simulation

Simulations

From mouse 5,6 properties



Slide 19 (Mouse 5)



Simulation

Outline

- Microarray Simulation Project
- Characterization of Microarray Images
- Results of Characterization
- Simulations
- Conclusion

Conclusions

- Characterization of microarray images can reveal important effects
  - In the mouse kidney set, the slides from two mice may have been handled differently.
- Realistic simulation of microarray images may allow us to estimate the effects of variations due to parts of the microarray system.

To Do List

- Noncausal MRF for spot and background defects
- Multiscale modeling of large defects
- Simulation study to estimate effects of spot uniformity and background defects