

# A Generic Method for Evaluating Appearance Models and Assessing the Accuracy of NRR

Roy Schestowitz, Carole Twining, Tim Cootes, Vlad Petrovic, Chris Taylor and Bill Crum

# Overview

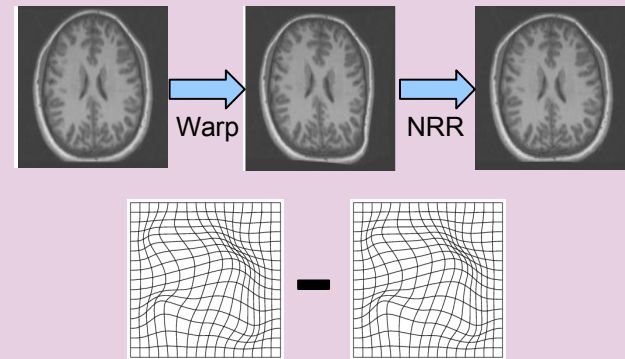
- Motivation
- Assessment methods
  - overlap-based
  - model-based
- Experiments
  - validation
  - comparison of methods
  - practical application
- Conclusions

# Motivation

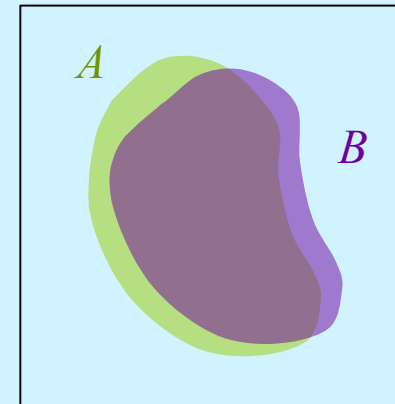
- Competing approaches to NRR
  - representation of warp (including regularisation)
  - similarity measure
  - optimisation
  - pair-wise vs group-wise
- Different results for same images
- Need for objective method of comparison
- QA in real applications (how well has it worked?)

# Existing Methods of Assessment

- Artificial warps
  - recovering known warps
  - may not be representative
  - algorithm testing but not QA



- Overlap measures
  - ground truth tissue labels
  - overlap after registration
  - subjective
  - too expensive for routine QA



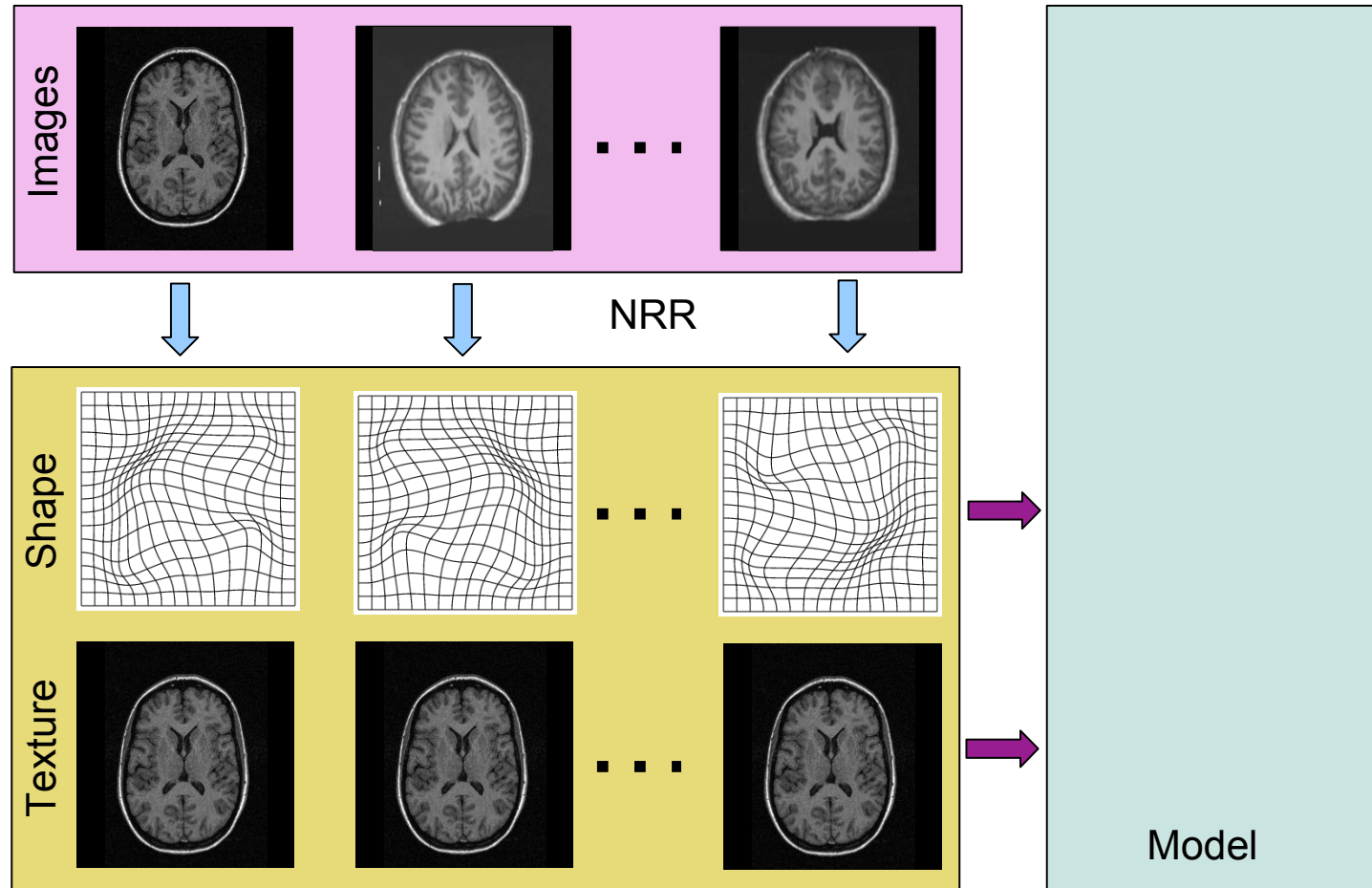
- Need for new approach

# Model-Based Assessment

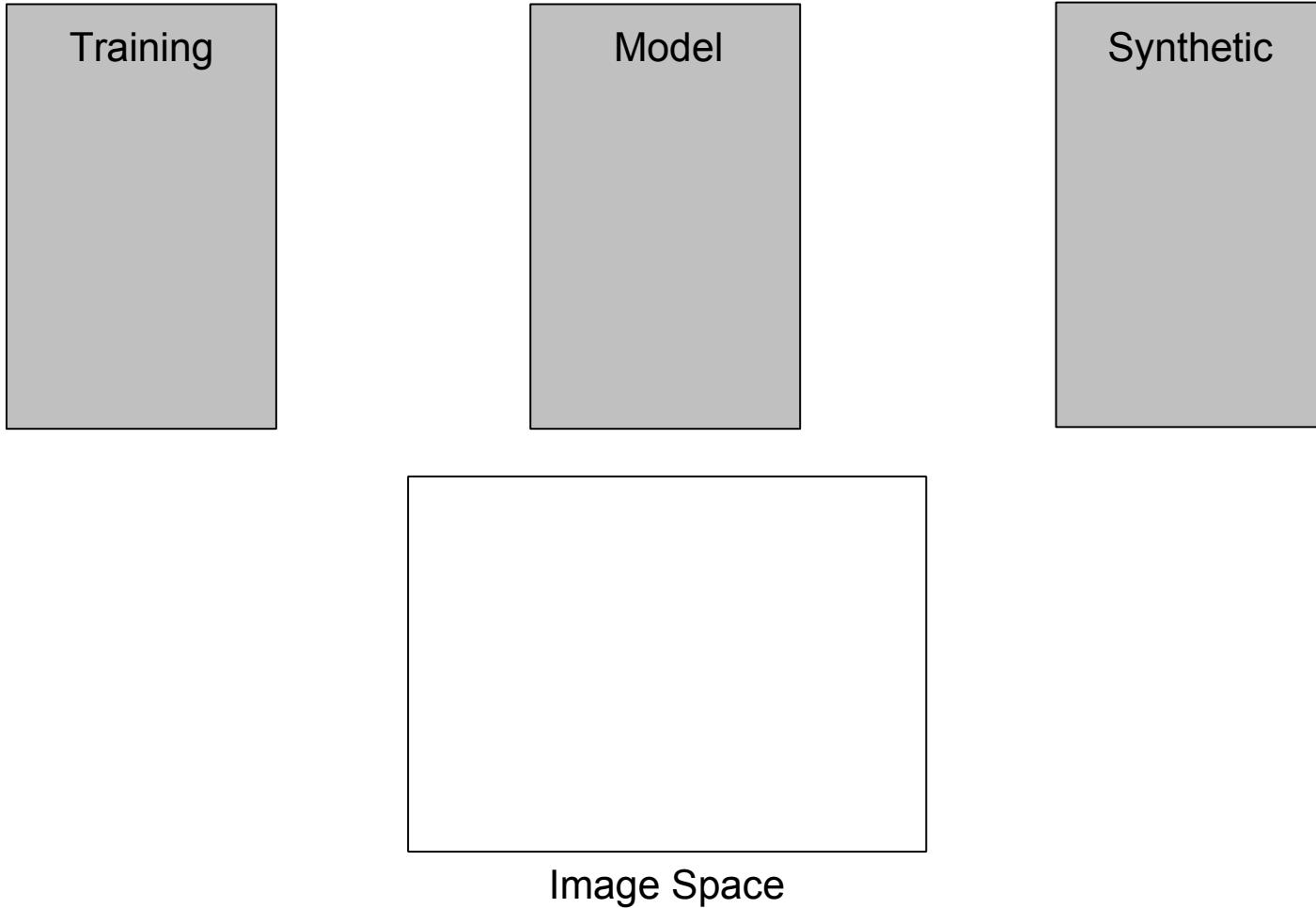
# Model-based Framework

- Registered image set  $\Rightarrow$  statistical appearance model
- Good registration  $\Rightarrow$  good model
  - generalises well to new examples
  - specific to class of images
- Registration quality  $\Leftrightarrow$  Model quality
  - problem transformed to defining model quality
  - ground-truth-free assessment of NRR

# Building an Appearance Model

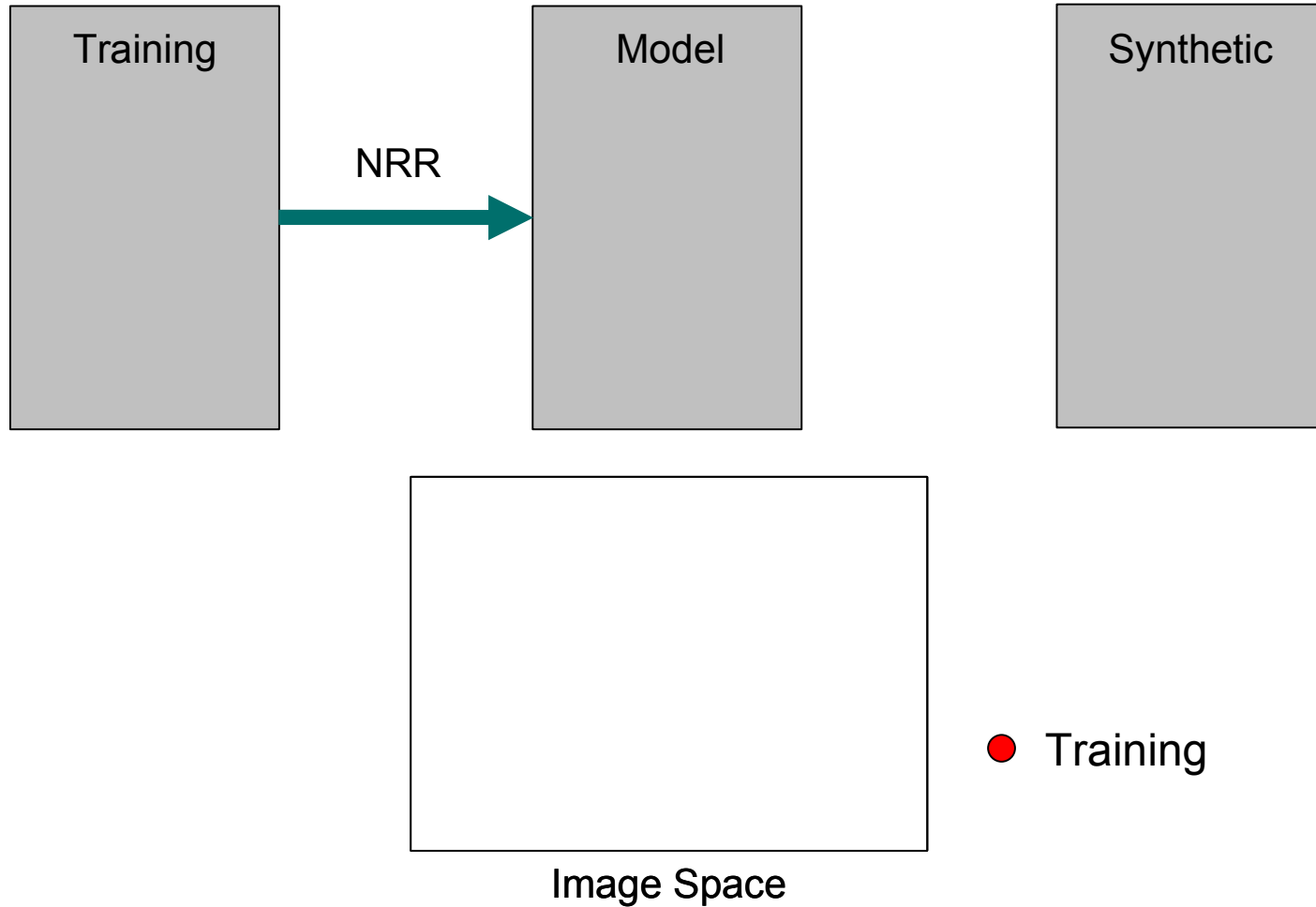


# Training and Synthetic Images

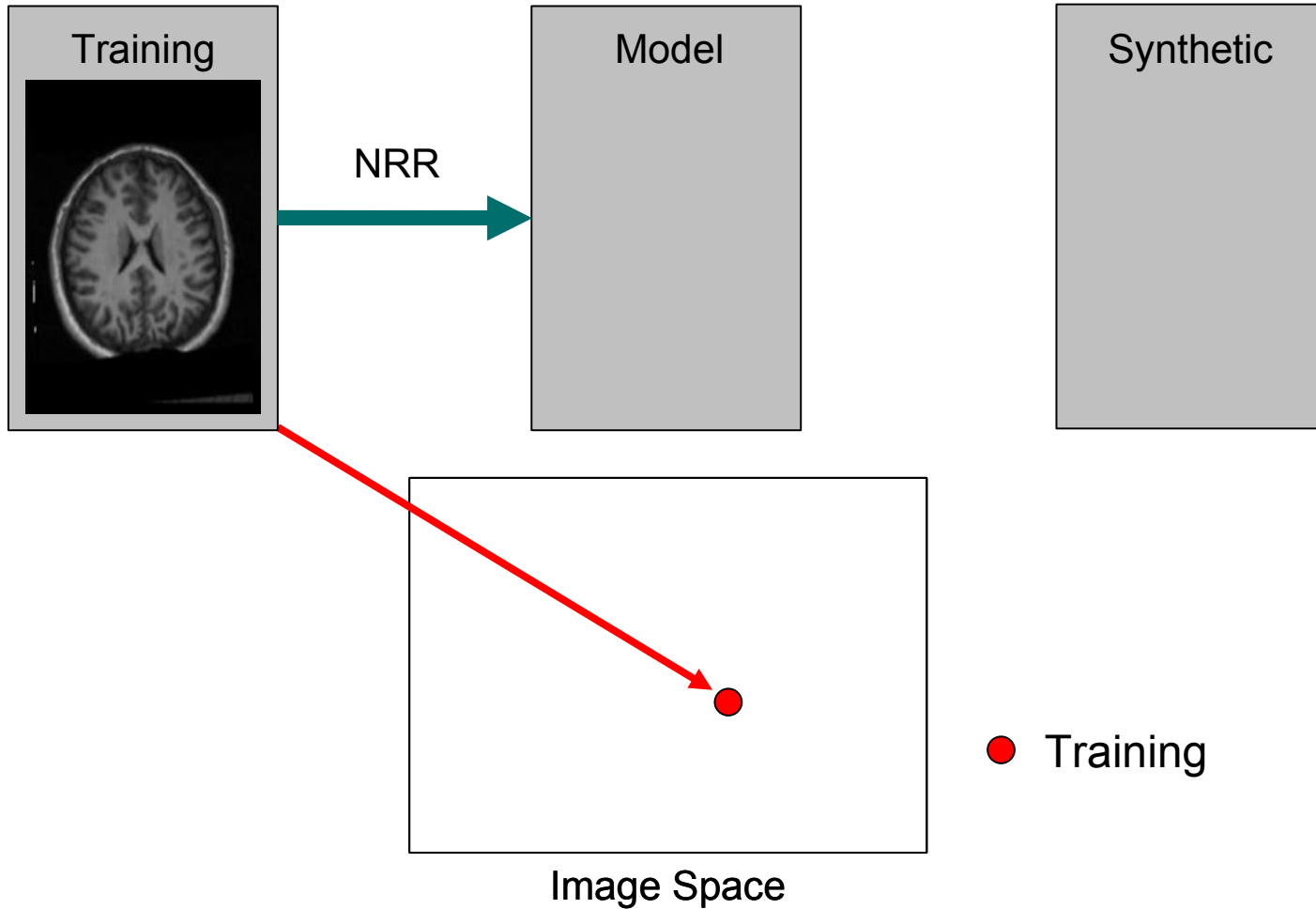




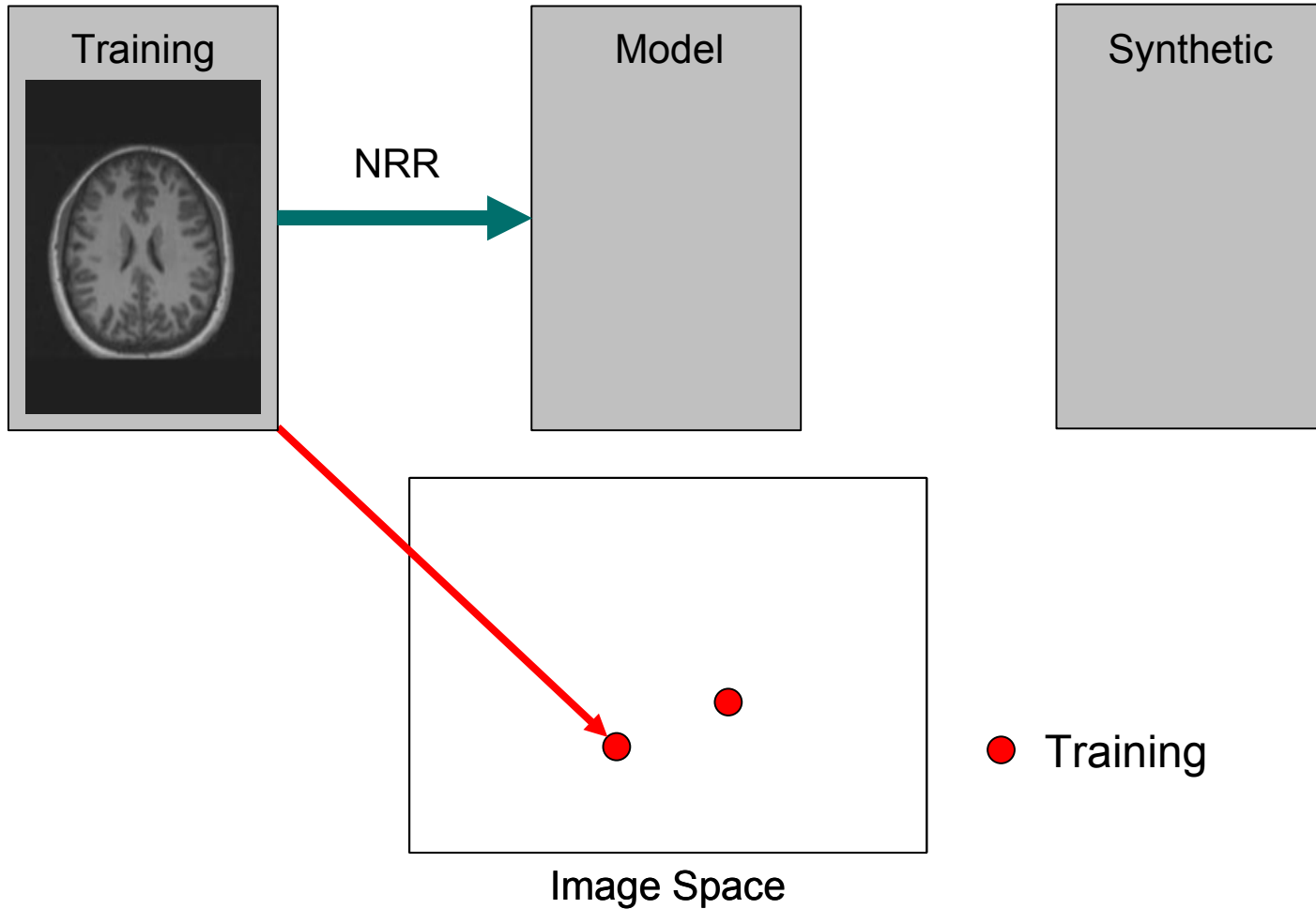
# Training and Synthetic Images



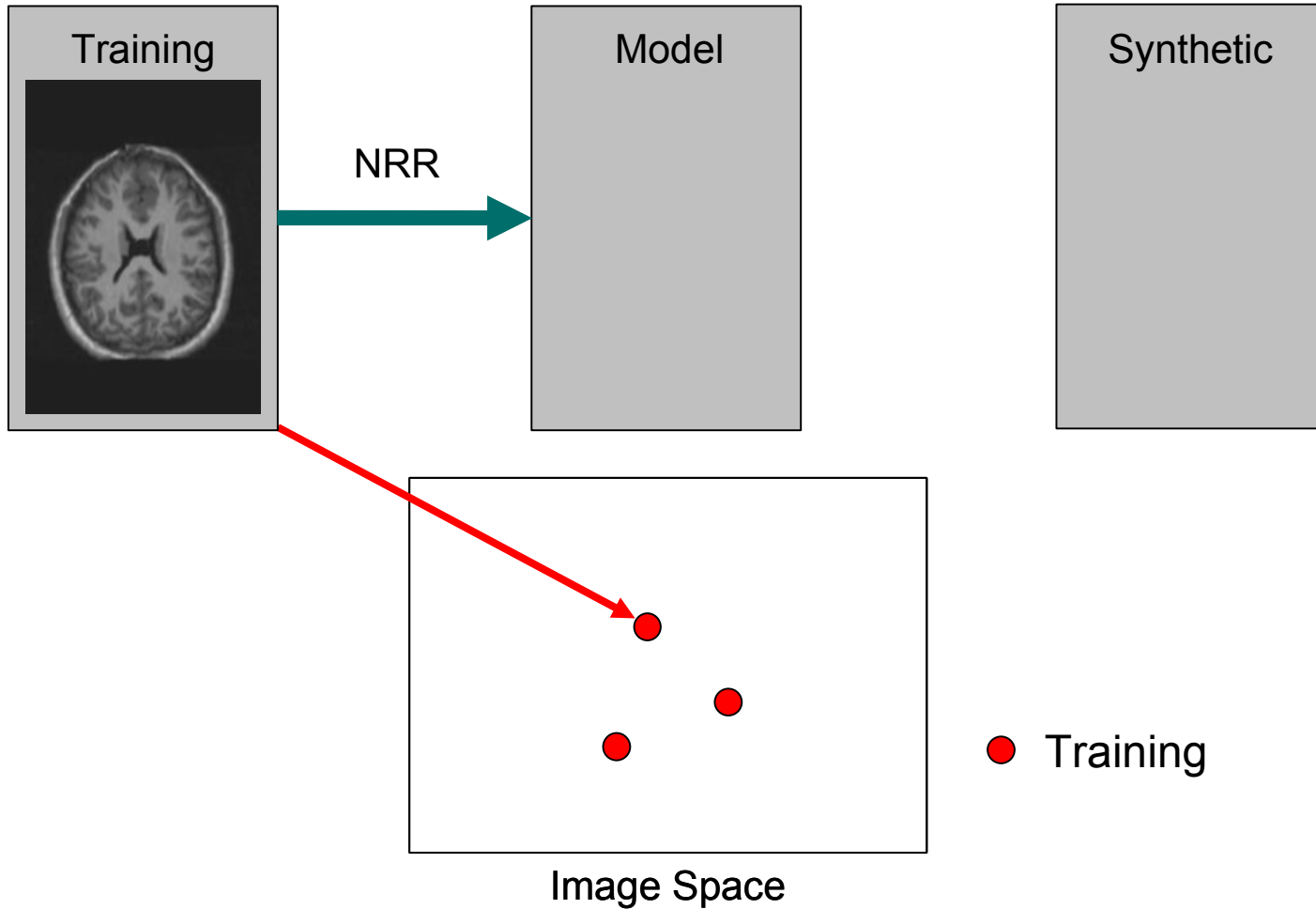
# Training and Synthetic Images



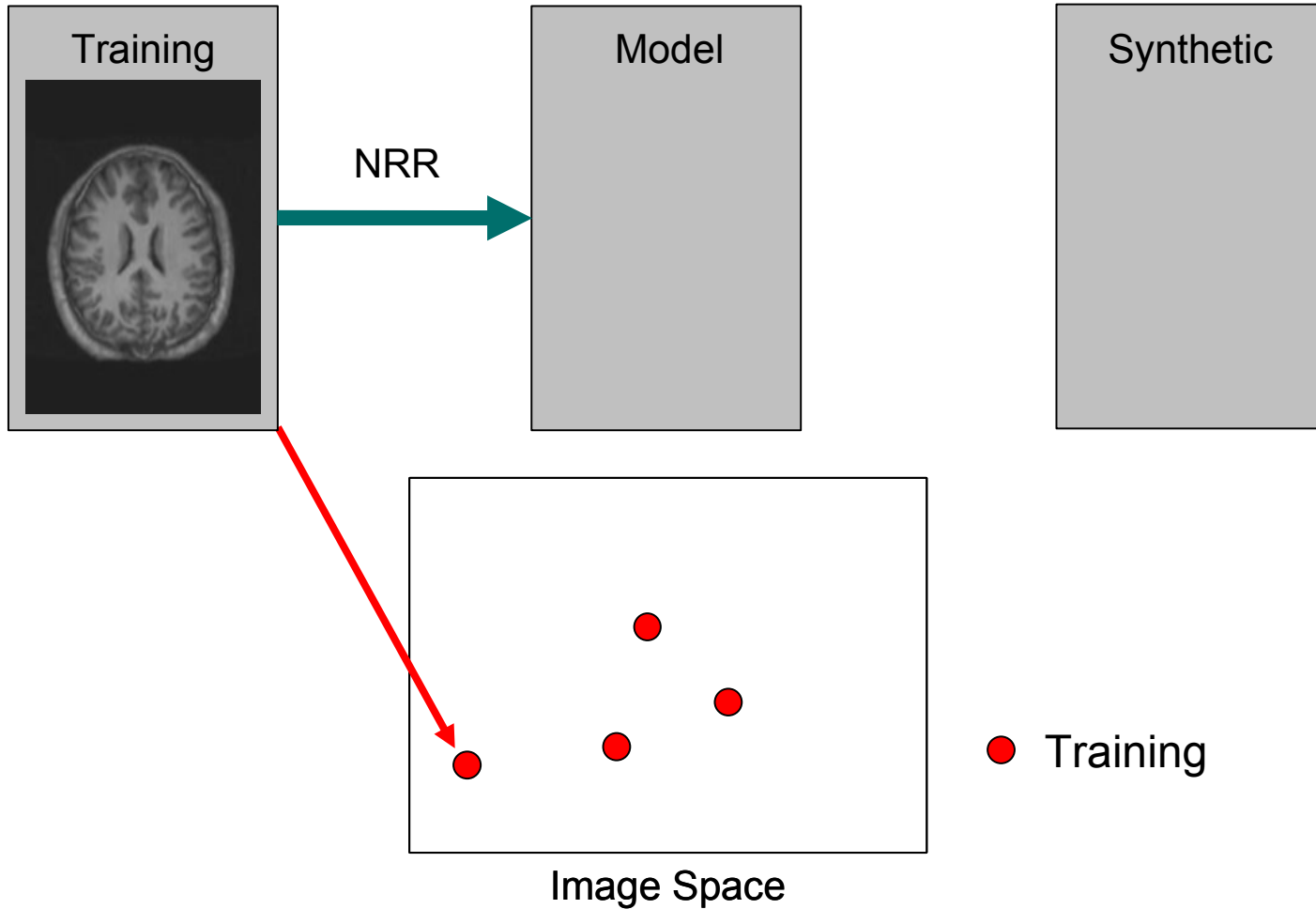
# Training and Synthetic Images



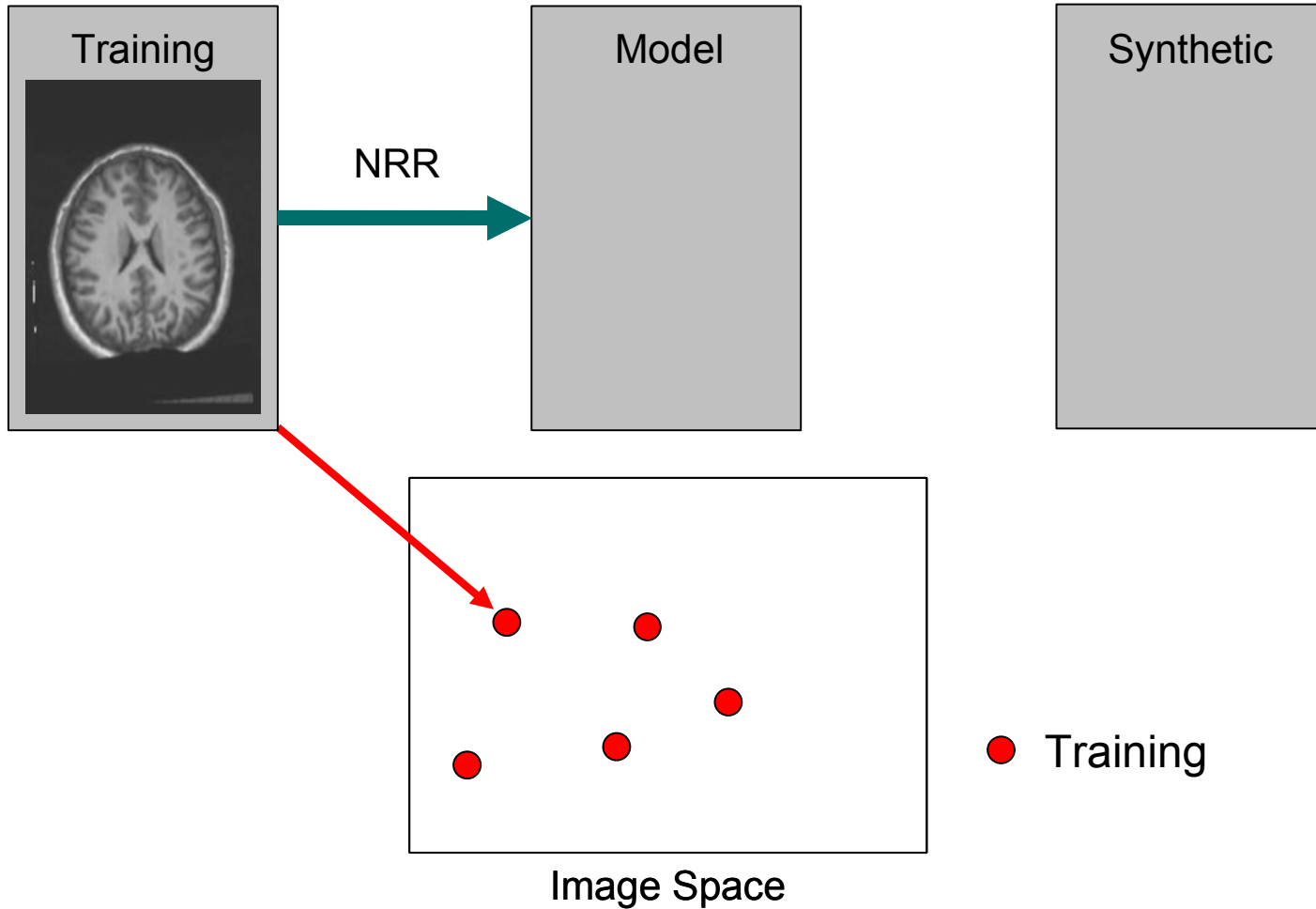
# Training and Synthetic Images



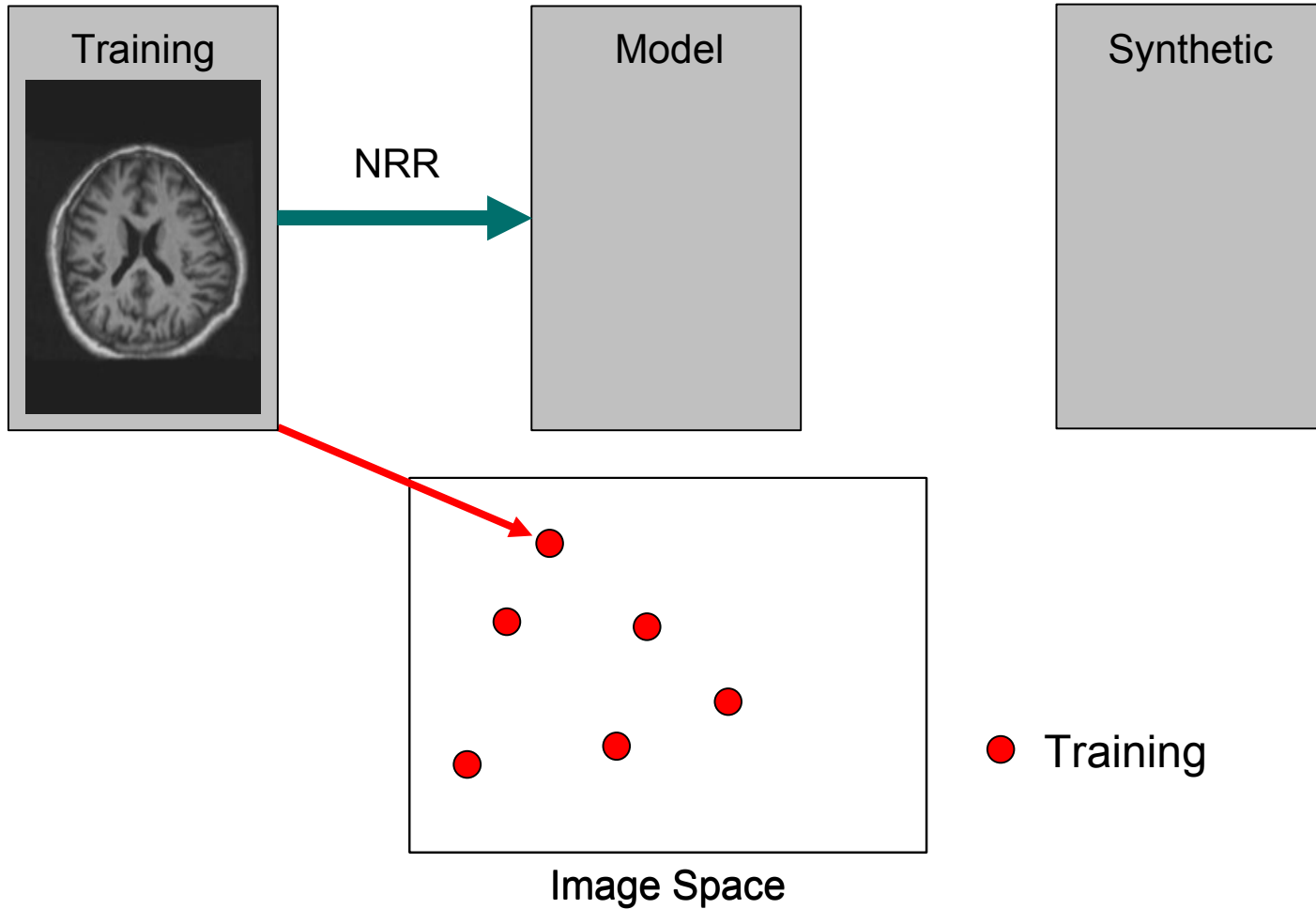
# Training and Synthetic Images



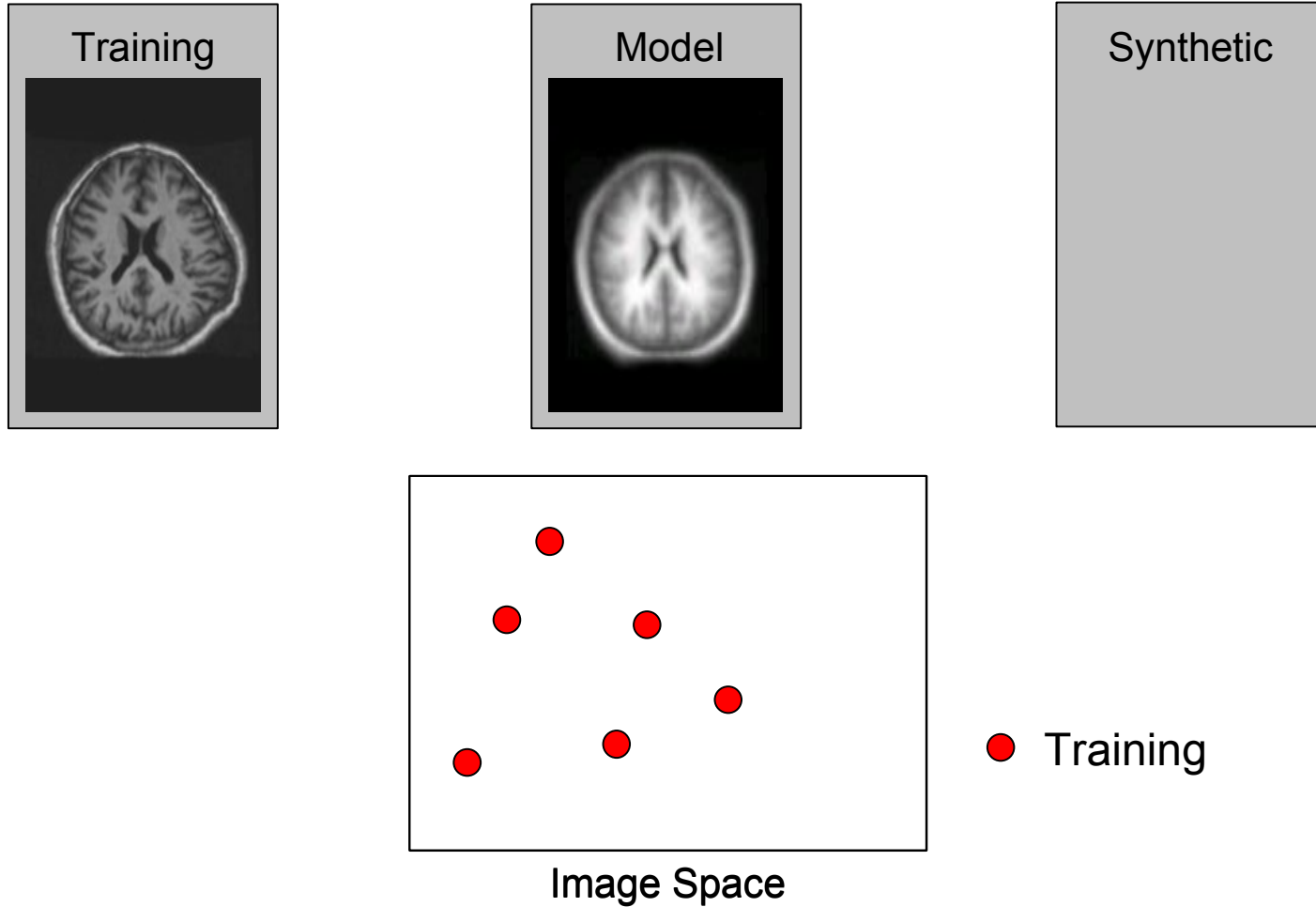
# Training and Synthetic Images



# Training and Synthetic Images

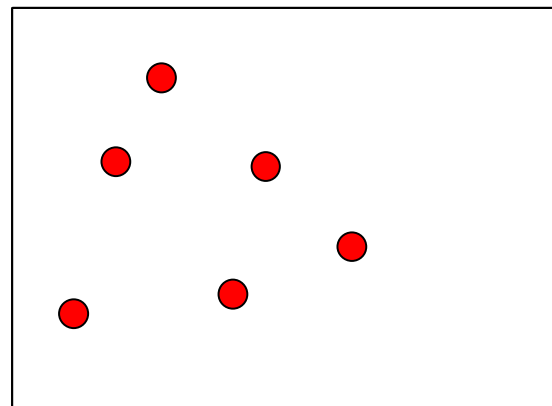
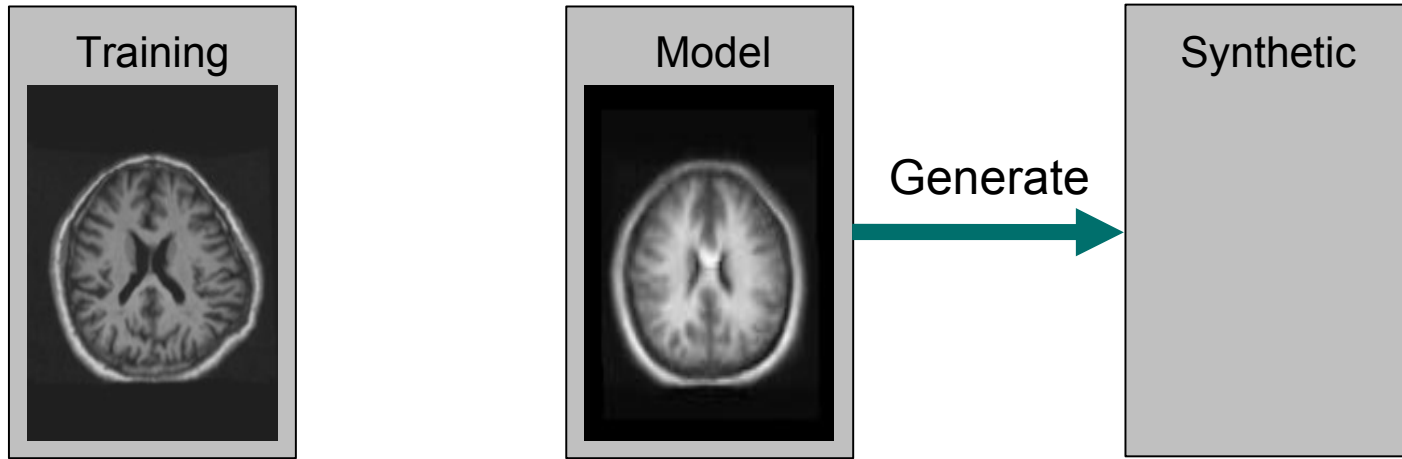


# Training and Synthetic Images





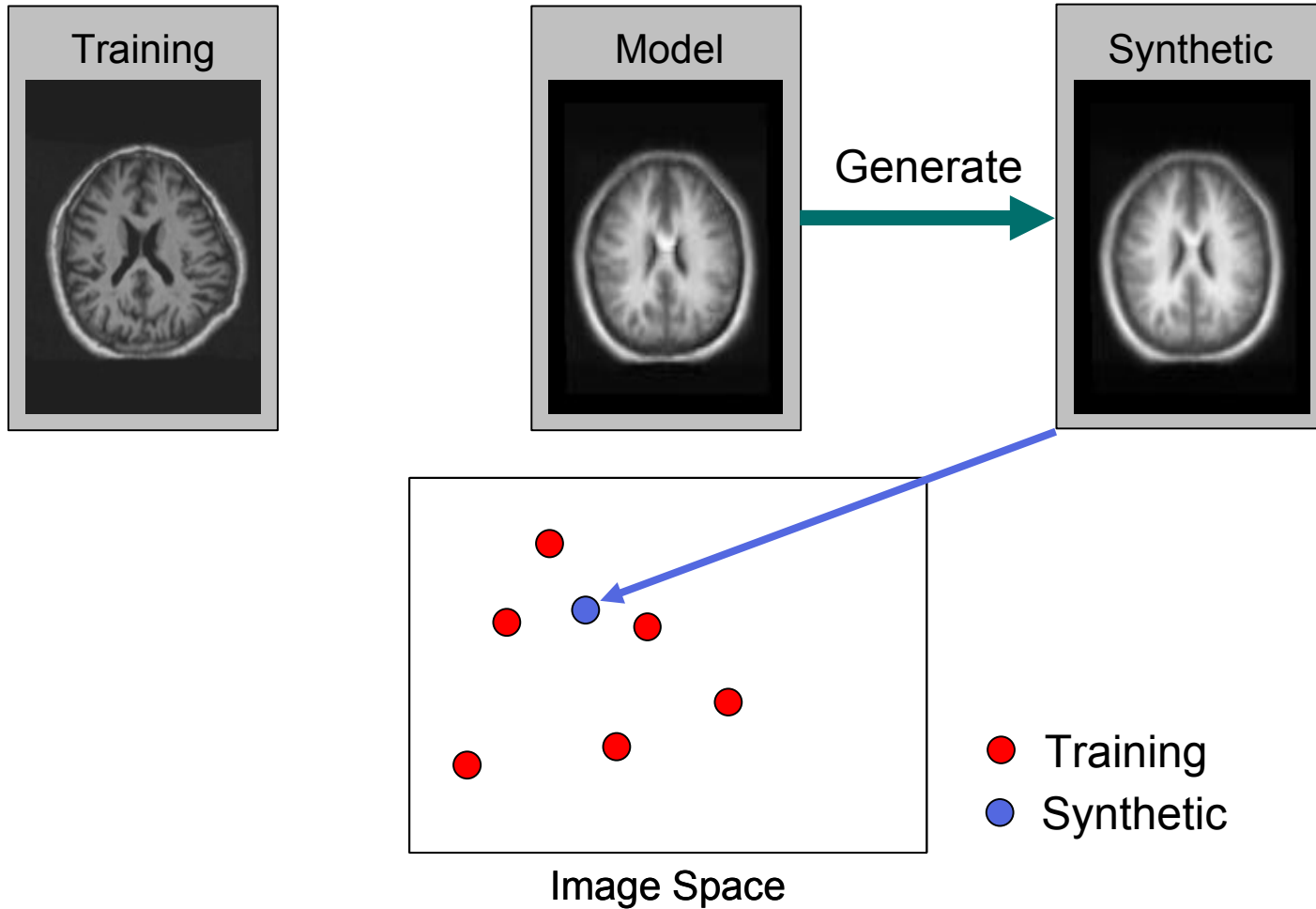
# Training and Synthetic Images



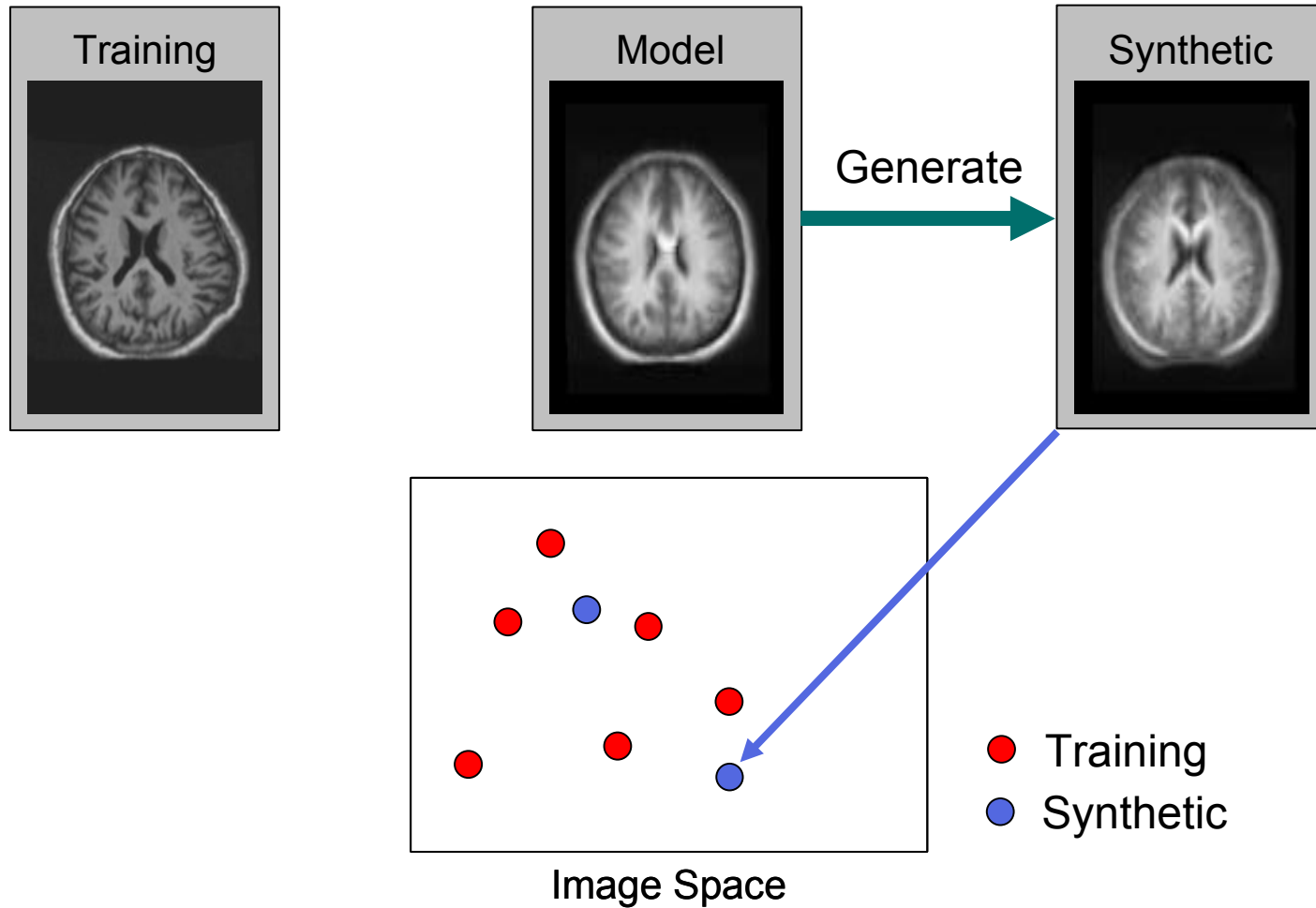
- Training
- Synthetic

Image Space

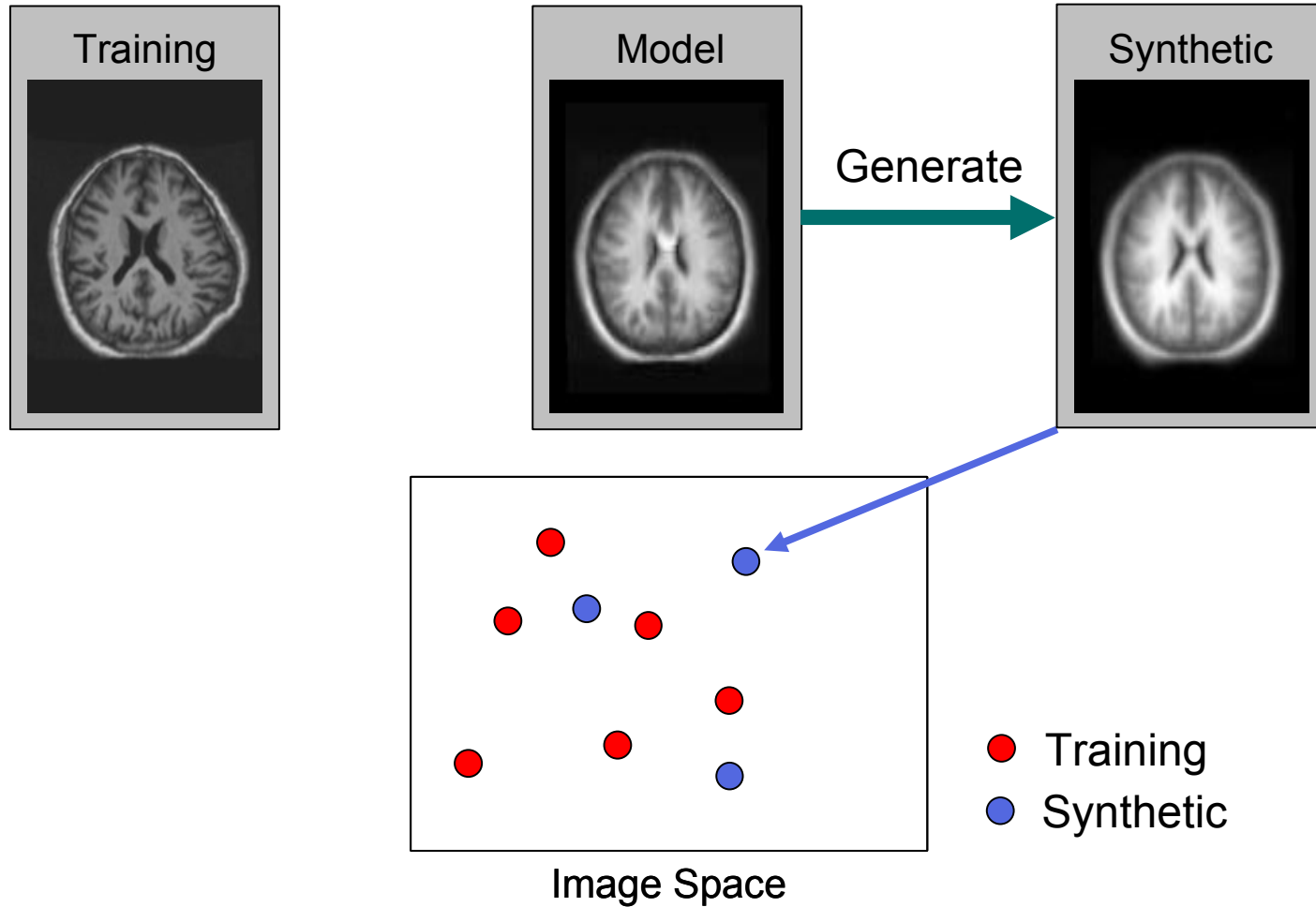
# Training and Synthetic Images



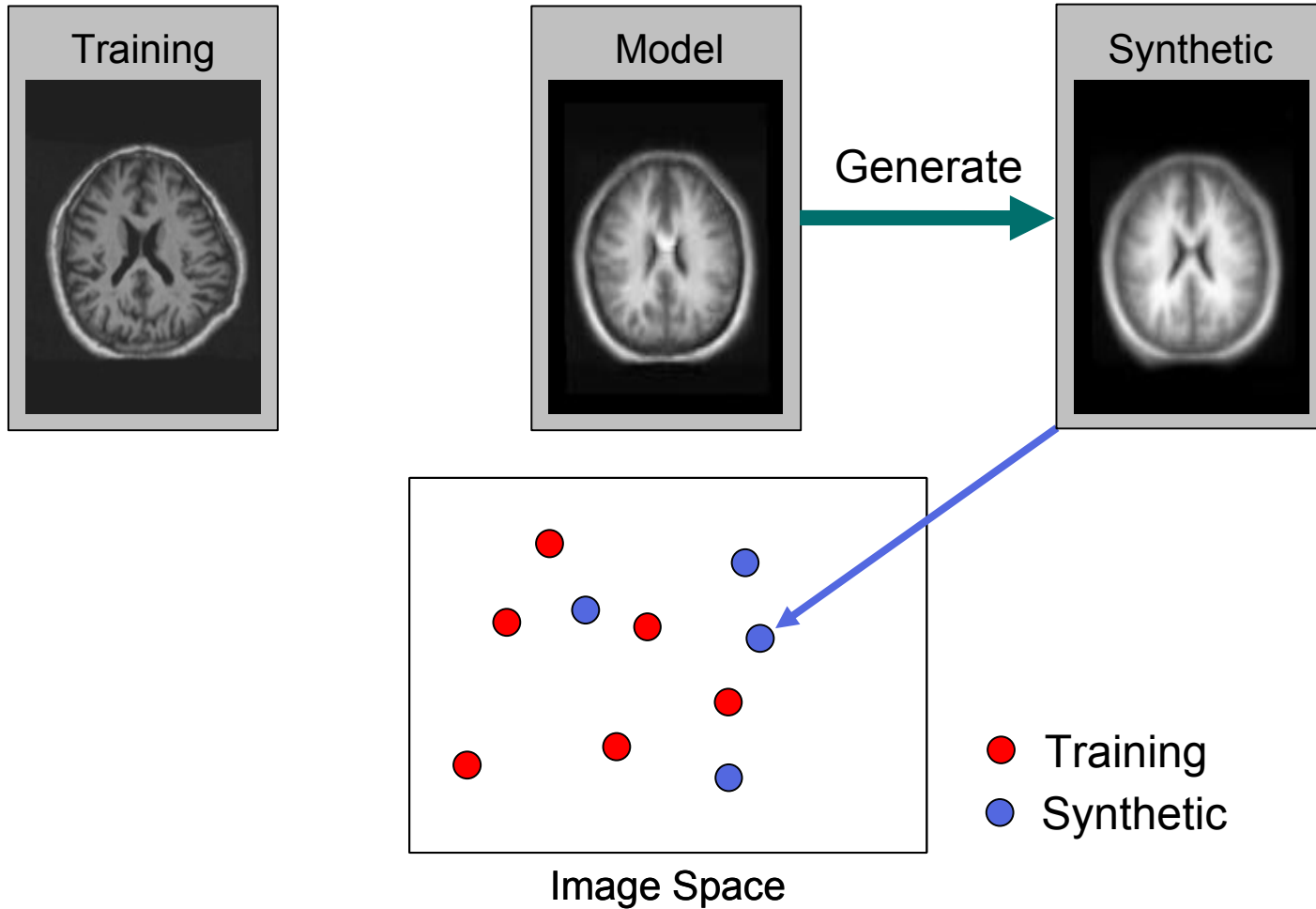
# Training and Synthetic Images



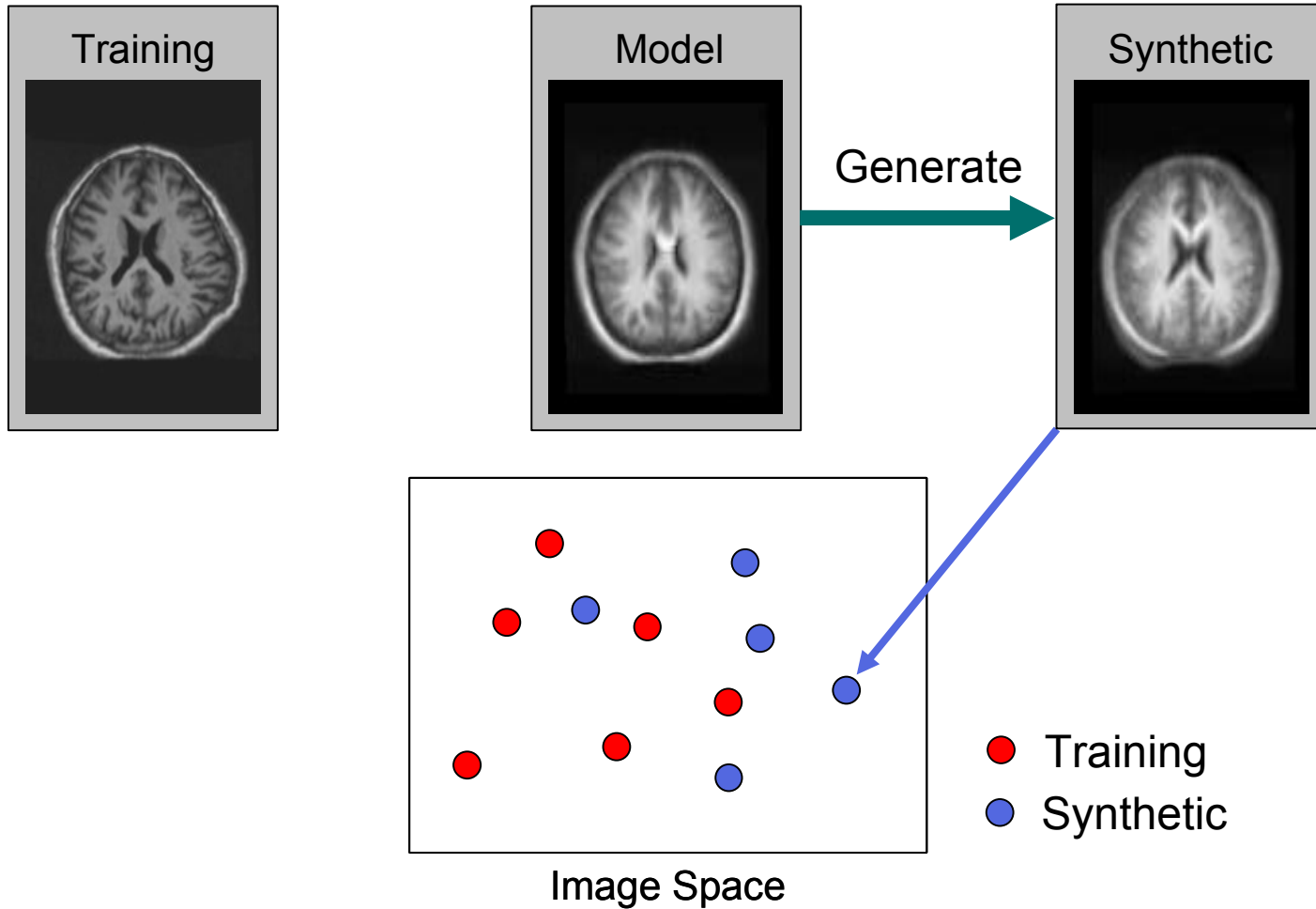
# Training and Synthetic Images



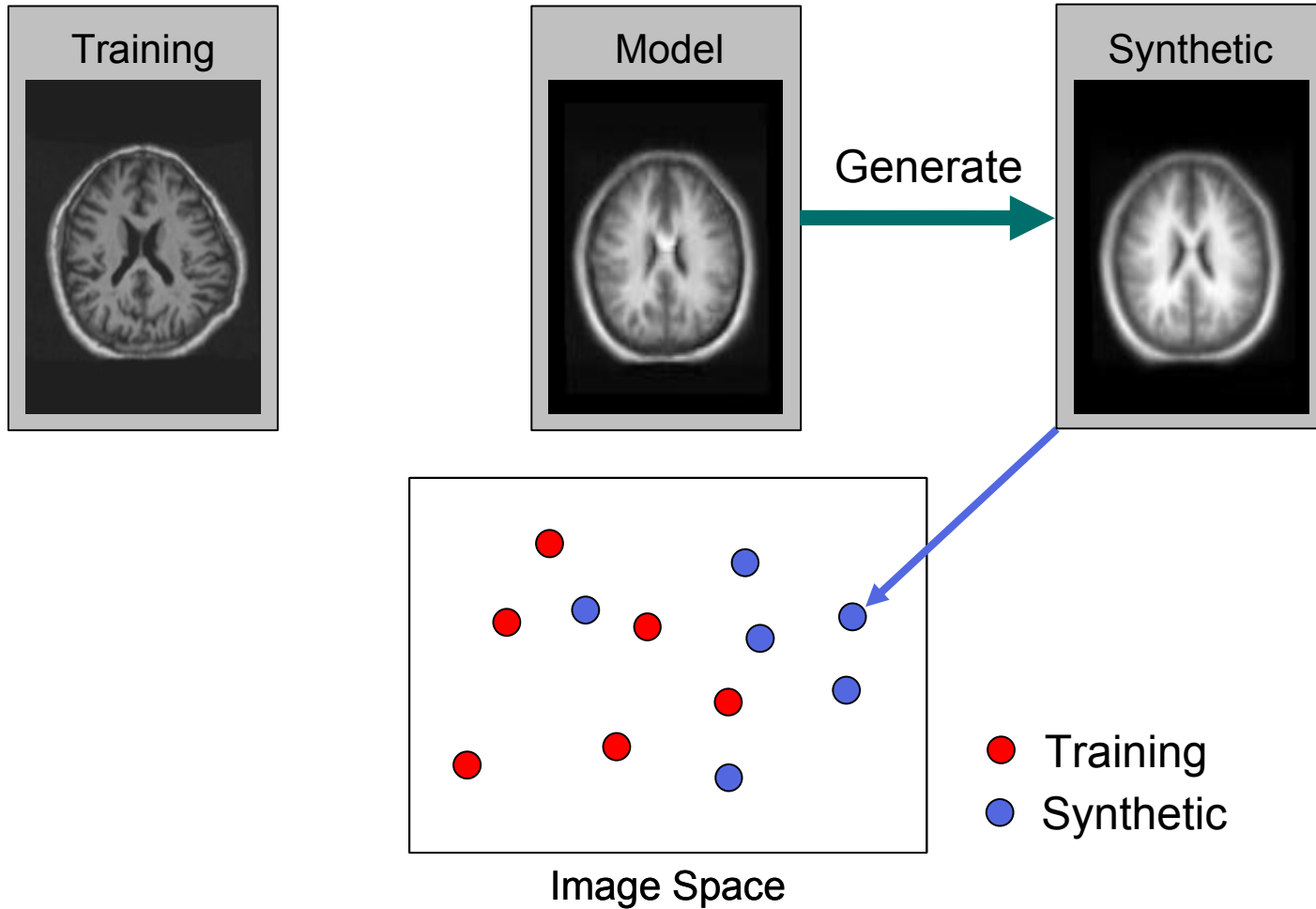
# Training and Synthetic Images



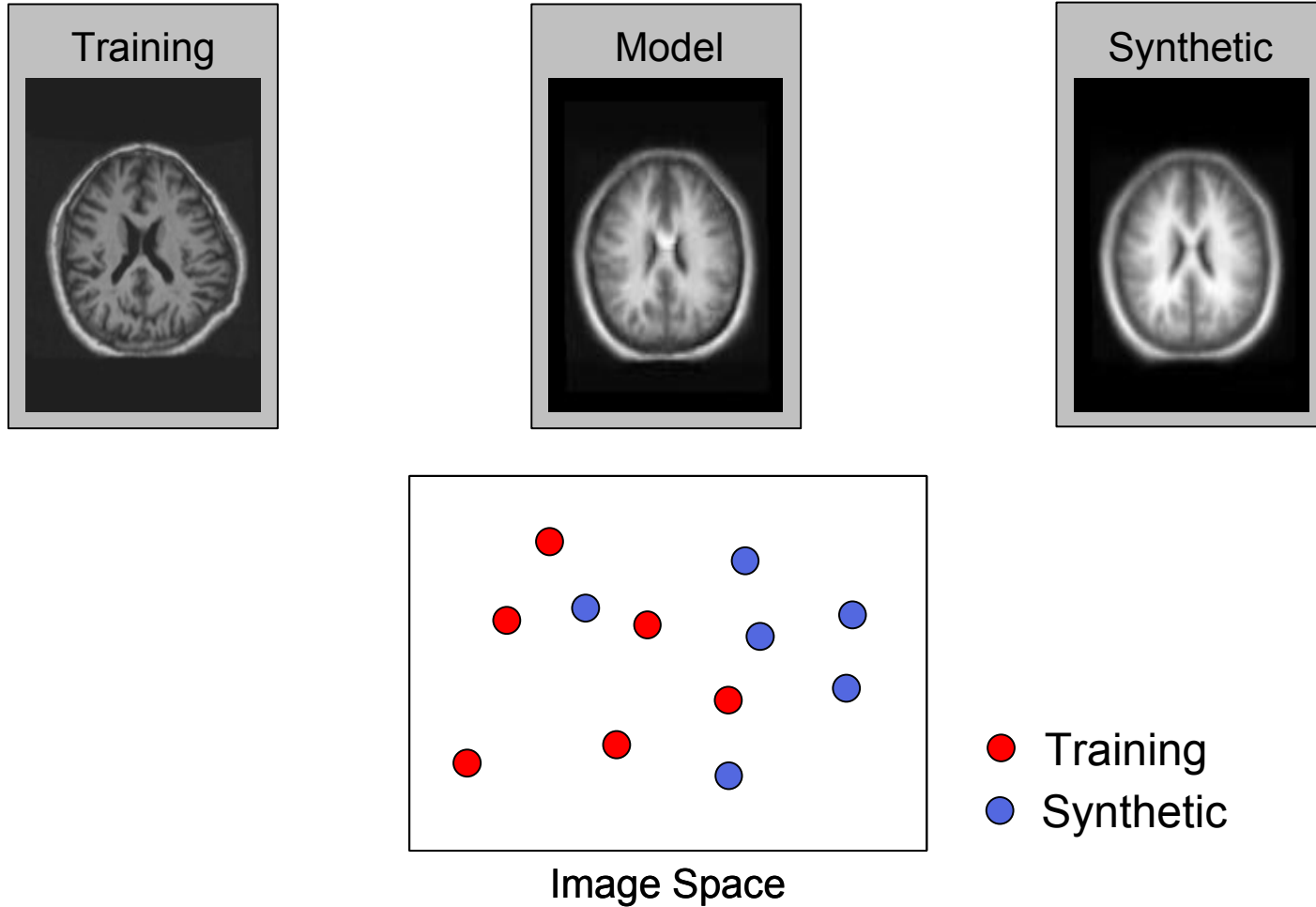
# Training and Synthetic Images



# Training and Synthetic Images

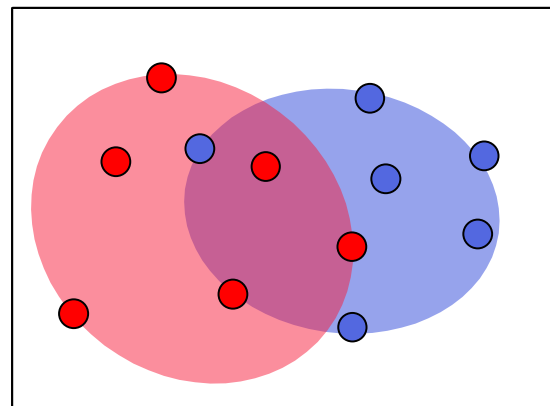


# Training and Synthetic Images





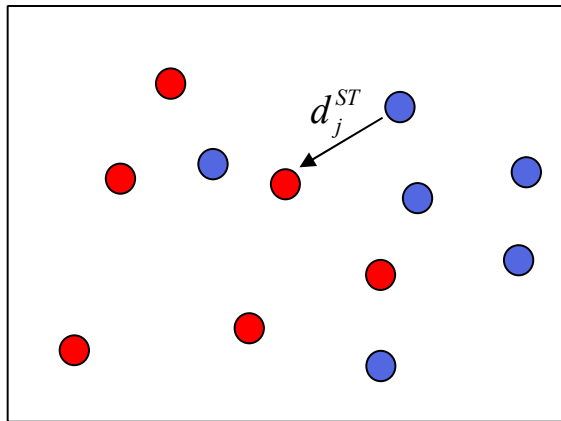
# Training and Synthetic Images



- Training
- Synthetic

Image Space

# Model Quality



- Training
- Synthetic

Given measure  $d$   
of image distance

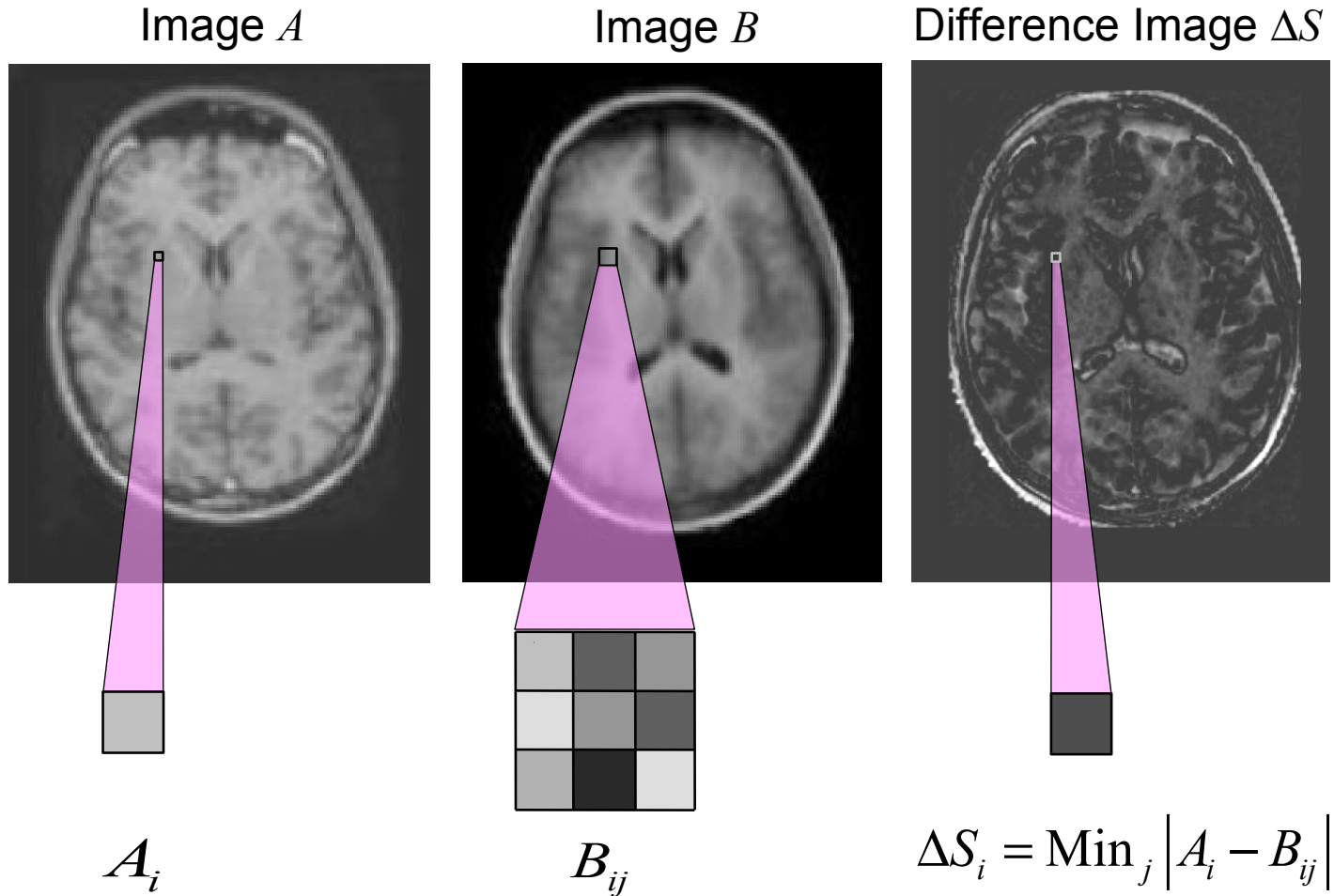
$$Specificity = \frac{\sum_{j=1}^m |d_j^{ST}|}{m} \quad \text{Mean distance to nearest training image}$$

- Euclidean or shuffle distance  $d$  between images
- Better models have smaller distances,  $d$
- Plot  $[-Specificity]$ , which decreases as model degrades

# Measuring Inter-Image Distance

- Euclidean
  - simple and cheap
  - sensitive to small misalignments
- Shuffle distance
  - neighbourhood-based pixel differences
  - less sensitive to misalignment

# Shuffle Distance



# Varying Shuffle Radius

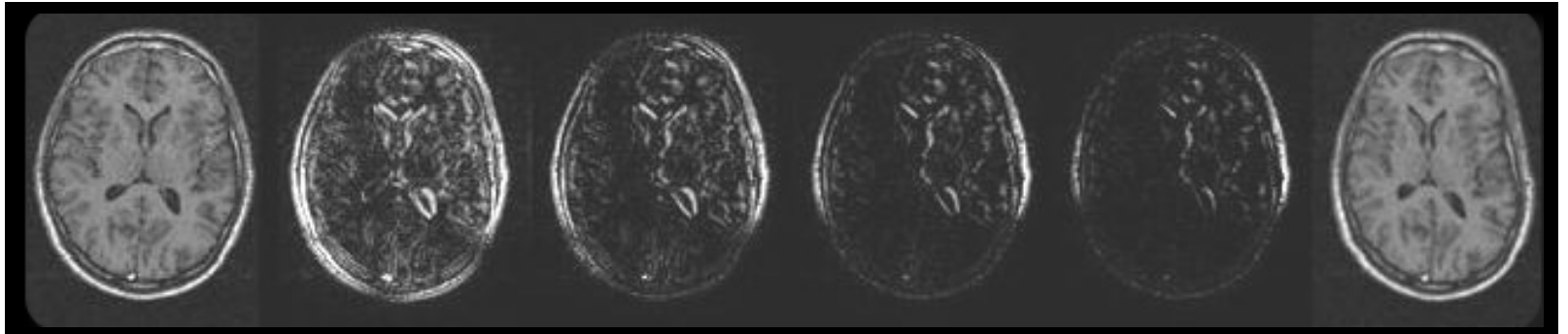


Image *A*

$r = 1$

$r = 1.5$

$r = 2.1$

$r = 3.7$

Image *B*

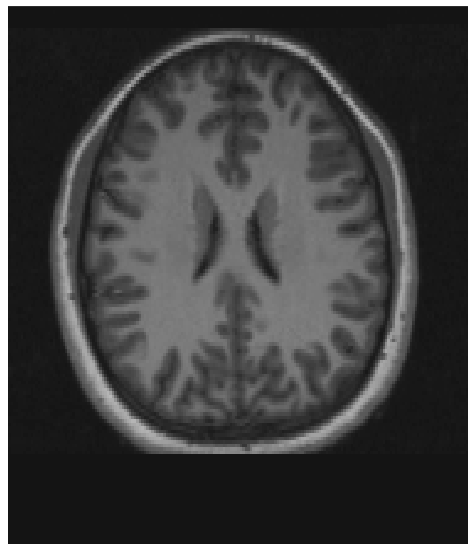
# Validation Experiments

# Experimental Design

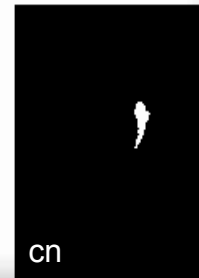
- MGH dataset (37 brains)
- Selected 2D slice
- Initial 'correct' NRR
- Progressive perturbation of registration
  - 10 random instantiations for each perturbation magnitude
- Comparison of the two different measures
  - overlap
  - model-based

# Brain Data

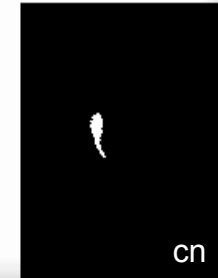
- Eight labels per image
  - L/R white/grey matter
  - L/R lateral ventricle
  - L/R caudate nucleus



Image



LH Labels

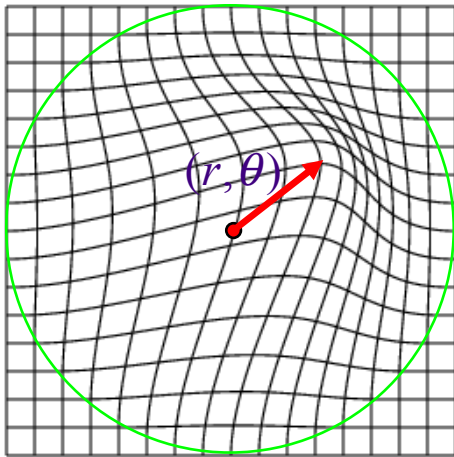


RH Labels

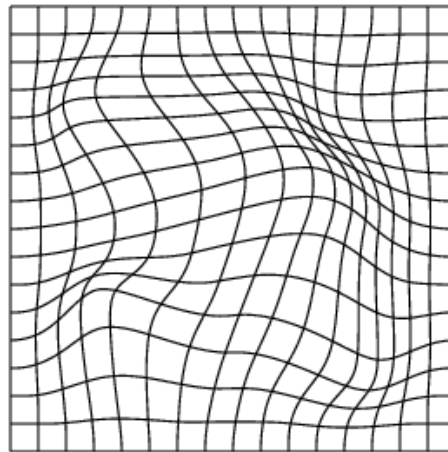


# Perturbation Framework

- Alignment degraded by applying warps to data
- Clamped-plate splines (CPS) with 25 knot-points
- Random displacement  $(r, \theta)$  drawn from distribution

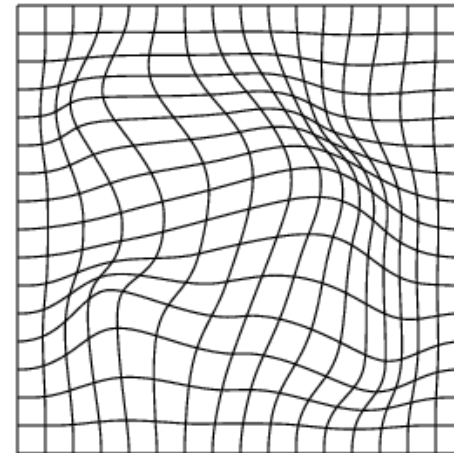
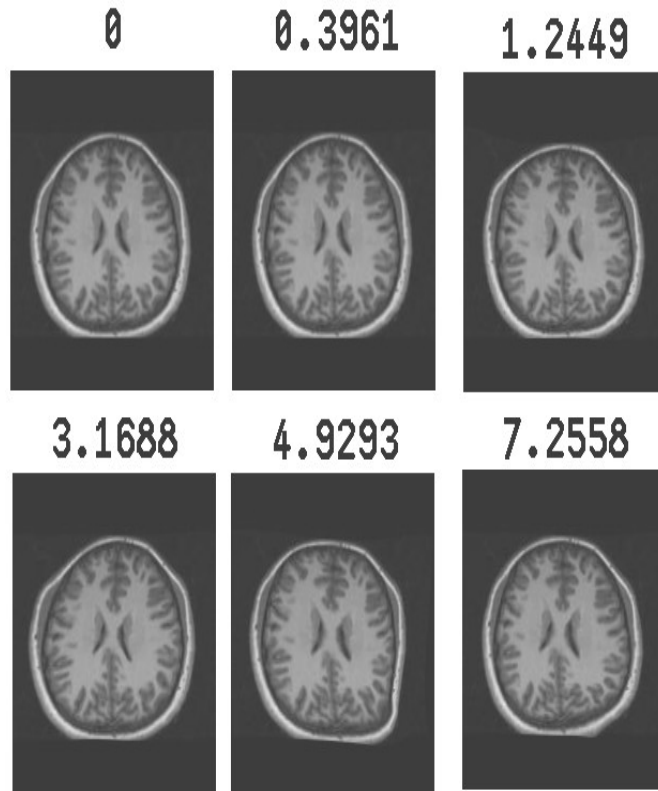


CPS with 1 knot point



Multiple knot points

# Examples of Perturbed Images

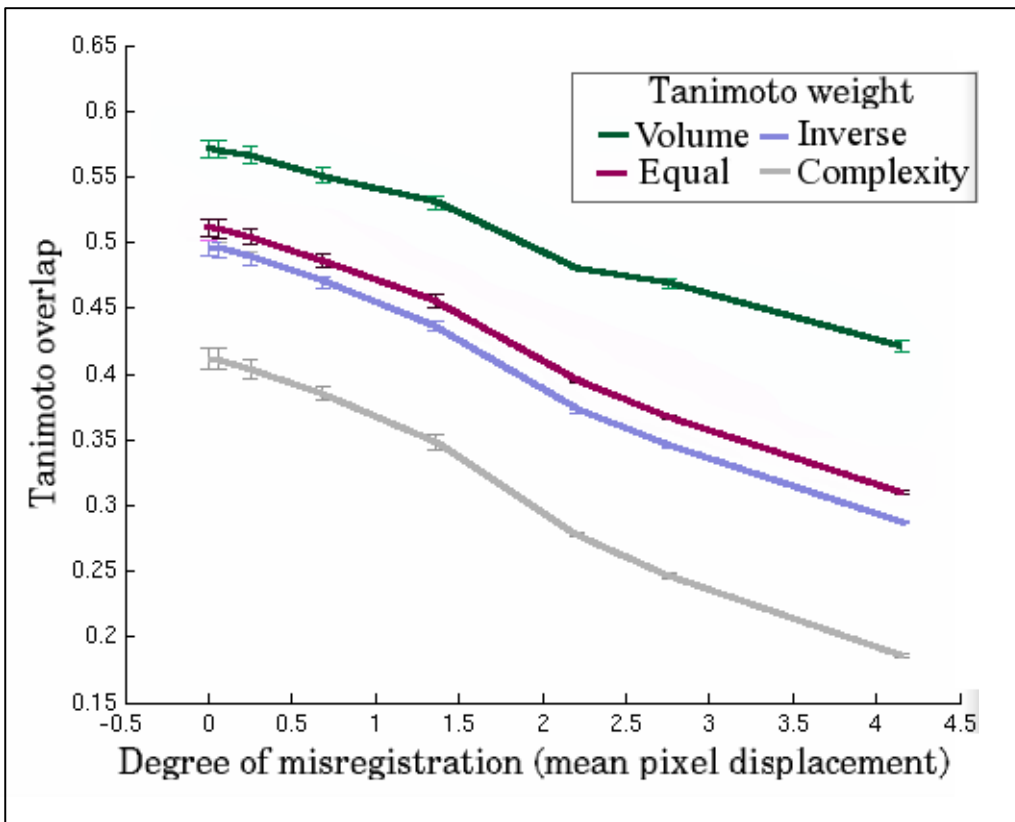


Example warp

Increasing mean pixel displacement

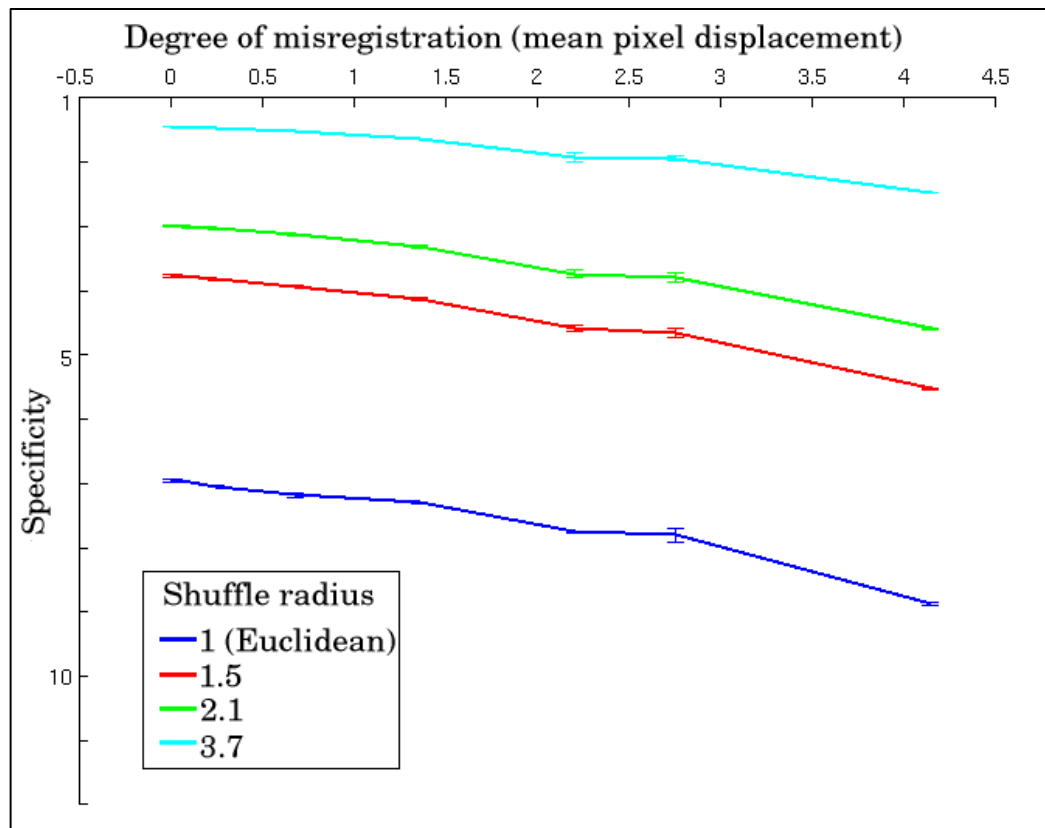
# Results – Generalised Overlap

- Overlap decreases monotonically with misregistration



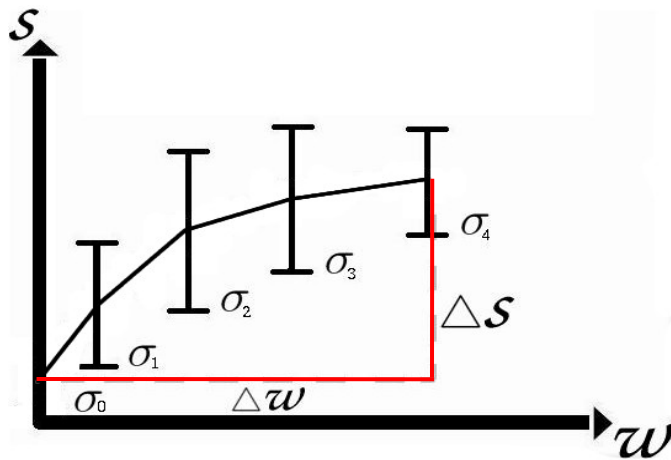
# Results – Model-Based

- [*-Specificity*] decreases monotonically with misregistration



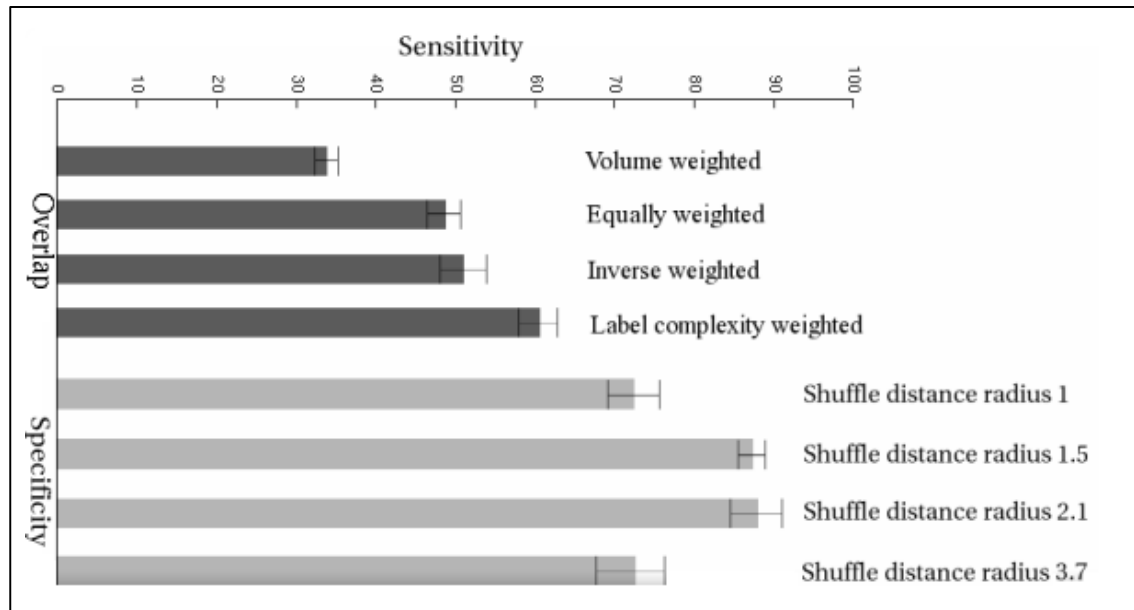
# Results – Comparison

- All three measures give similar results
  - overlap-based assessment requires ground truth (labels)
  - model-based approach does not need ground truth
- Compare sensitivity of methods
  - ability to detect small changes in registration



# Results – Sensitivities

- Sensitivity
  - ability to detect small changes in registration
  - high sensitivity good



- Specificity more sensitive than overlap

# Further Tests – Noise

- A measure of robustness to noise is sought
- Validation experiments repeated with noise applied
  - each image has up to 10% white noise added
  - two instantiations of set perturbation are used
- Results indicate that the model-based method is robust
  - changes in Generalisation and Specificity remain detectable
  - curves remain monotonic
  - noise can potentially exceed 10%

# Practical Application

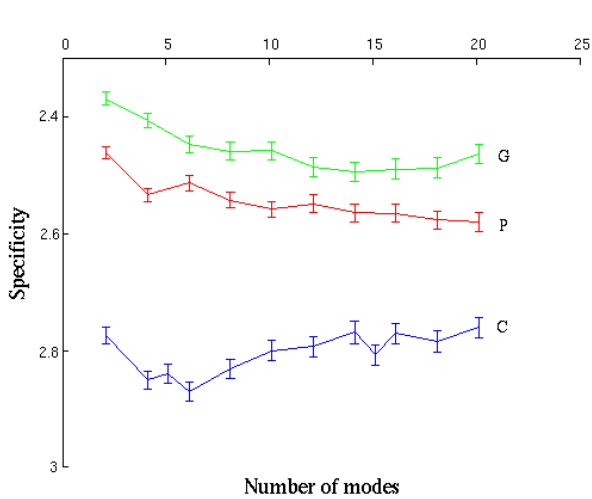


# Practical Application

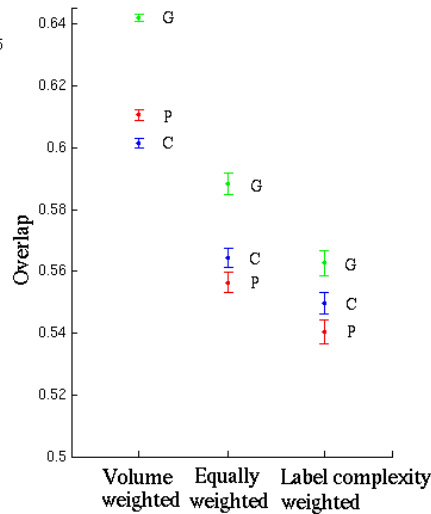
- 3 registration algorithms compared
  - Pair-wise registration
  - Group-wise registration
  - Congealing
- 2 brain datasets used
  - MGH dataset
  - Dementia dataset
- 2 assessment methods
  - Model-based (Specificity)
  - Overlap-based

# Practical Application - Results

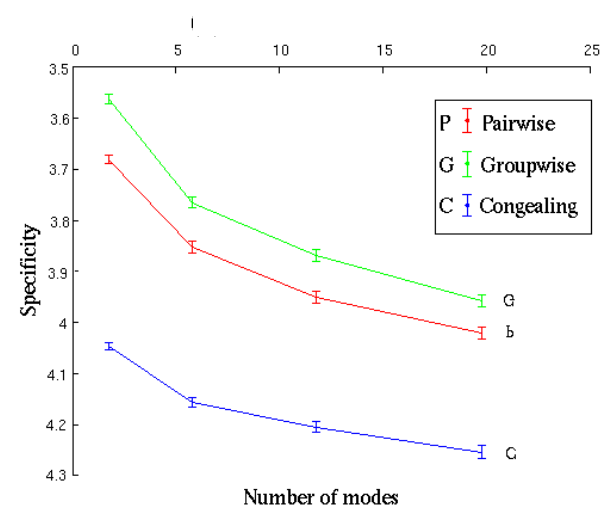
- Results are consistent
- Group-wise > pair-wise > congealing



MGH Data



MGH Data



Dementia Data

# Extension to 3-D

- 3-D experiments
- Work in progress
  - validation experiments laborious to replicate
  - comparison of 4-5 NRR algorithms
- Fully-annotated IBIM data
- Results can be validated by measuring label overlap

# Conclusions

- Overlap and model-based approaches ‘equivalent’
- Overlap provides ‘gold standard’
- Specificity is a good surrogate
  - monotonically related
  - robust to noise
  - no need for ground truth
  - only applies to groups (but any NRR method)