Building Probabilistic Models Around Deterministic Optimization Procedures

Ryan Prescott Adams
School of Engineering and Applied Sciences
Harvard University

Joint work with Danny Tarlow and Rich Zemel

`http://hips.seas.harvard.edu`

Structured Prediction

- **Objective:** Build discriminative models that map from features to high-dimensional, structured discrete spaces.

- **Example: Image Segmentation**

$$f(\boldsymbol{x}\,;\,\theta) \rightarrow \boldsymbol{y}$$

$$\boldsymbol{x} \in \mathcal{X}$$

$$\boldsymbol{y} \in \{1, 2, \ldots\}^K$$

Structured Prediction

- **Objective:** Build discriminative models that map from features to high-dimensional, structured discrete spaces.

- **Example: Image Segmentation**

$$p(\boldsymbol{y} \mid \boldsymbol{x}, \theta)$$

$$\boldsymbol{x} \in \mathcal{X}$$

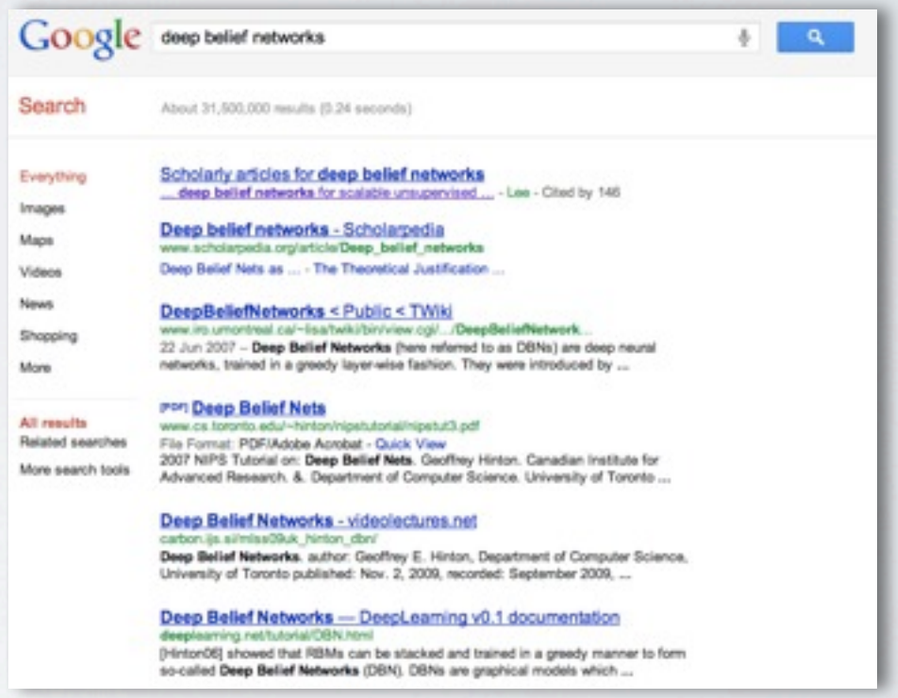
$$\boldsymbol{y} \in \{1, 2, \ldots\}^K$$

Structured Prediction

- **Objective:** Build discriminative models that map from features to high-dimensional, structured discrete spaces.

- **Example: Ranking Documents**

$$p(\boldsymbol{y} \mid \boldsymbol{x}, \theta)$$



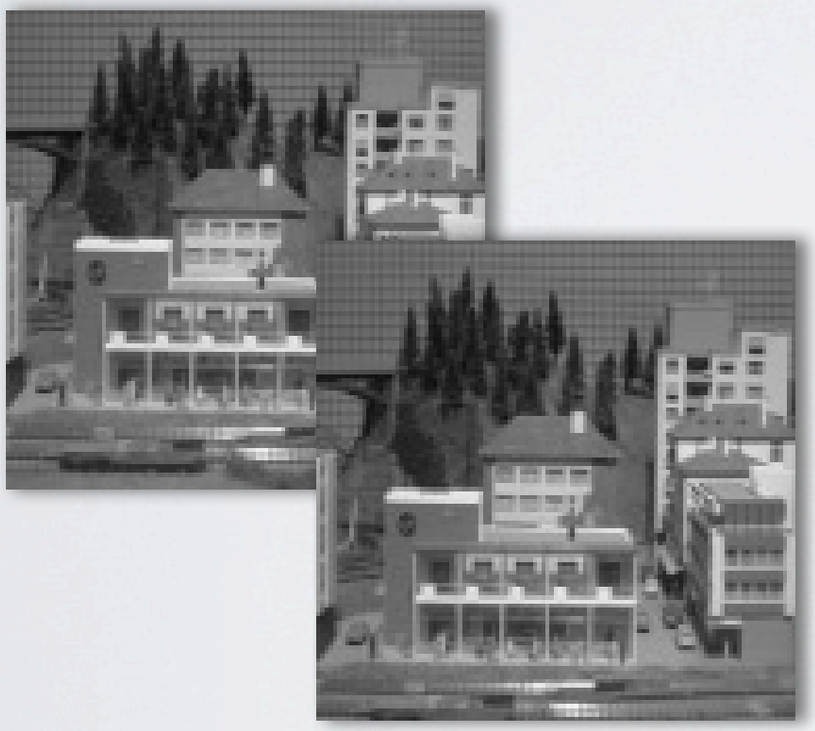
$$\boldsymbol{x} \in \mathcal{X}$$

$$\boldsymbol{y} \in \{1, 2, \ldots\}^K$$

Structured Prediction

- **Objective:** Build discriminative models that map from features to high-dimensional, structured discrete spaces.

- **Example: Matching**



$$\boldsymbol{x} \in \mathcal{X}$$

$$p(\boldsymbol{y} \mid \boldsymbol{x}, \theta)$$

$$\boldsymbol{y} \in \{1, 2, \ldots\}^K$$

Conditional Random Fields

- A very common approach to structured prediction is to construct a feature-dependent energy.

- The resulting Gibbs distribution gives a probabilistic model for the discrete space:

$$p(\boldsymbol{y} \,|\, \boldsymbol{x}, \theta) = \frac{1}{\mathcal{Z}(\boldsymbol{x}, \theta)} \exp\{-E(\boldsymbol{x}, \boldsymbol{y}\,;\, \theta)\}$$

$$E(\boldsymbol{x}, \boldsymbol{y}\,;\, \theta) = \theta^{\mathsf{T}} \psi(\boldsymbol{x}, \boldsymbol{y})$$

- A unique partition function for each training example!

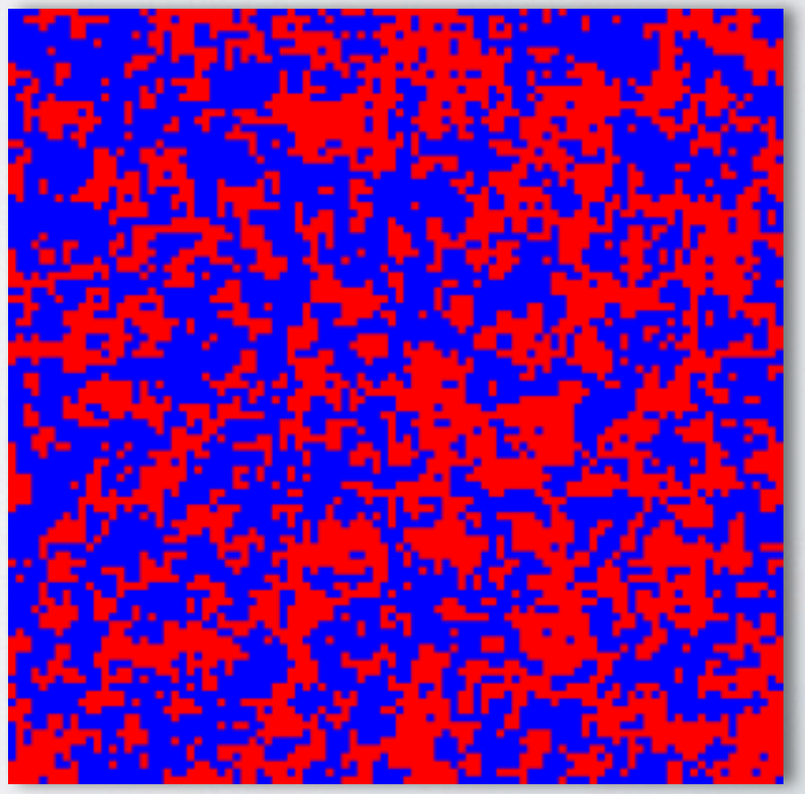
A Computational Discrepancy

- In many MRFs, finding the mode is easy even when finding the partition function is difficult. That is, one may be able to maximize in polynomial time over a set that may be exponentially large.

- **Example: Ising model with positive weights**

- Partition function is a sum over exponentially many configurations.

- Graph cuts finds the most probable configuration in polynomial time.



A Computational Discrepancy

- In many MRFs, finding the mode is easy even when finding the partition function is difficult. That is, one may be able to maximize in polynomial time over a set that may be exponentially large.

- **Example: Bipartite matching**
- Normalization is #P-complete.
- The Hungarian algorithm finds the most probable match in cubic time.

A Computational Discrepancy

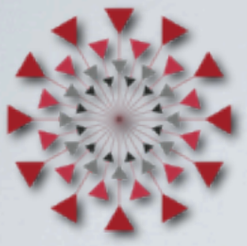
- In many MRFs, finding the mode is easy even when finding the partition function is difficult. That is, one may be able to maximize in polynomial time over a set that may be exponentially large.

- **Example: Shortest paths**

- Normalization is #P-complete

- Djikstra's algorithm finds the most probable solution in

$$O(|E| + |V| \log |V|)$$



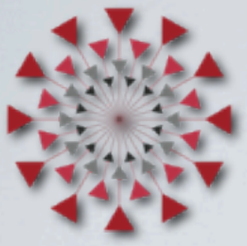
A Simple Generative Model

- Basic idea: Represent the predictive uncertainty in the choice of energy, rather than in the induced MRF.

- Use the efficient mode-finding algorithm as an explicit component of the data generation procedure.

- Use the features to induce a distribution on the parameters of an objective function, sample them and then optimize to generate the observed data.

- **Randomized Optimum Models** (RandOMs) are a generalization of Perturb-and-MAP [Papandreou & Yuille, 2011]



Randomized Optimum Model

- Define a family of objectives on the label space:

$$f_\theta(\boldsymbol{y}) : \mathcal{Y} \to \mathbb{R}$$

- Features of data: $\phi(\boldsymbol{x})$

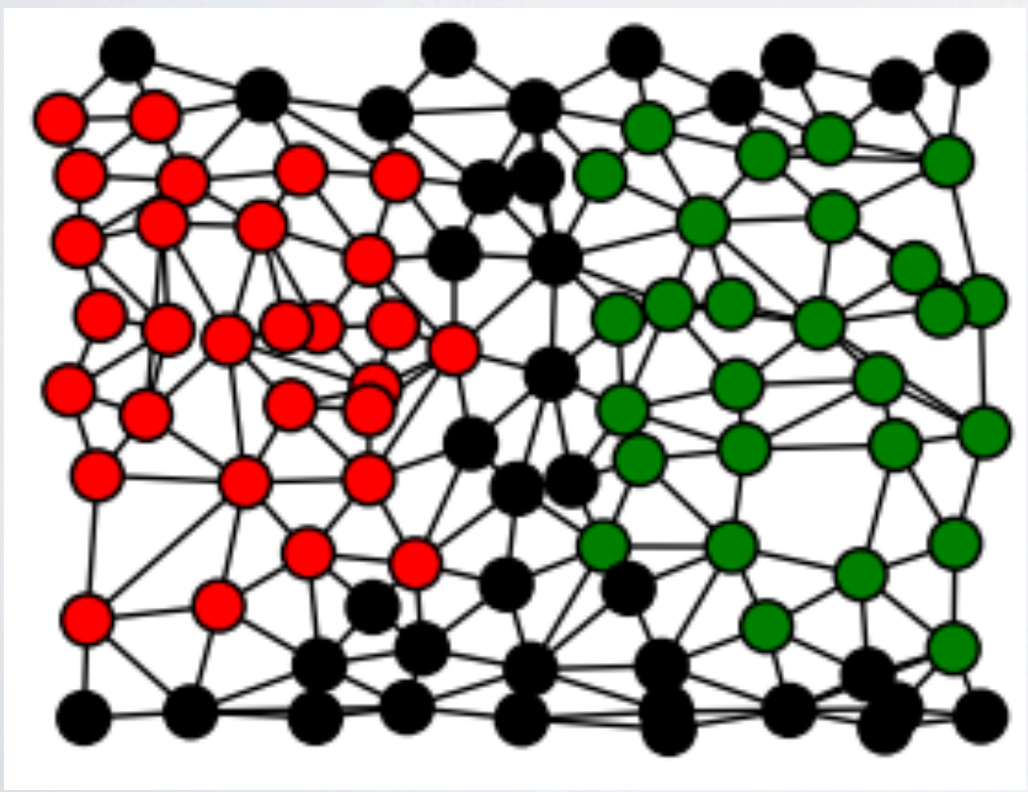
- Parameters of distribution over objectives: γ

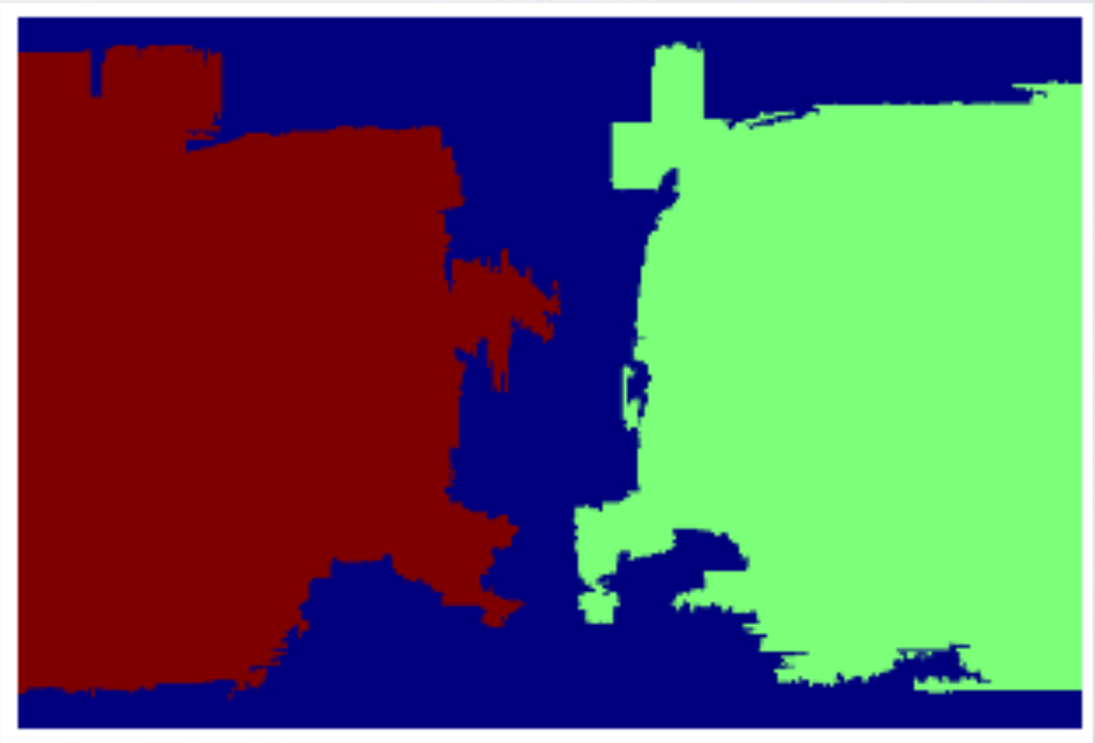
- Draw a random objective: $\theta \sim p(\theta \mid \phi(\boldsymbol{x}), \gamma)$

- Data are the minimizing configuration:

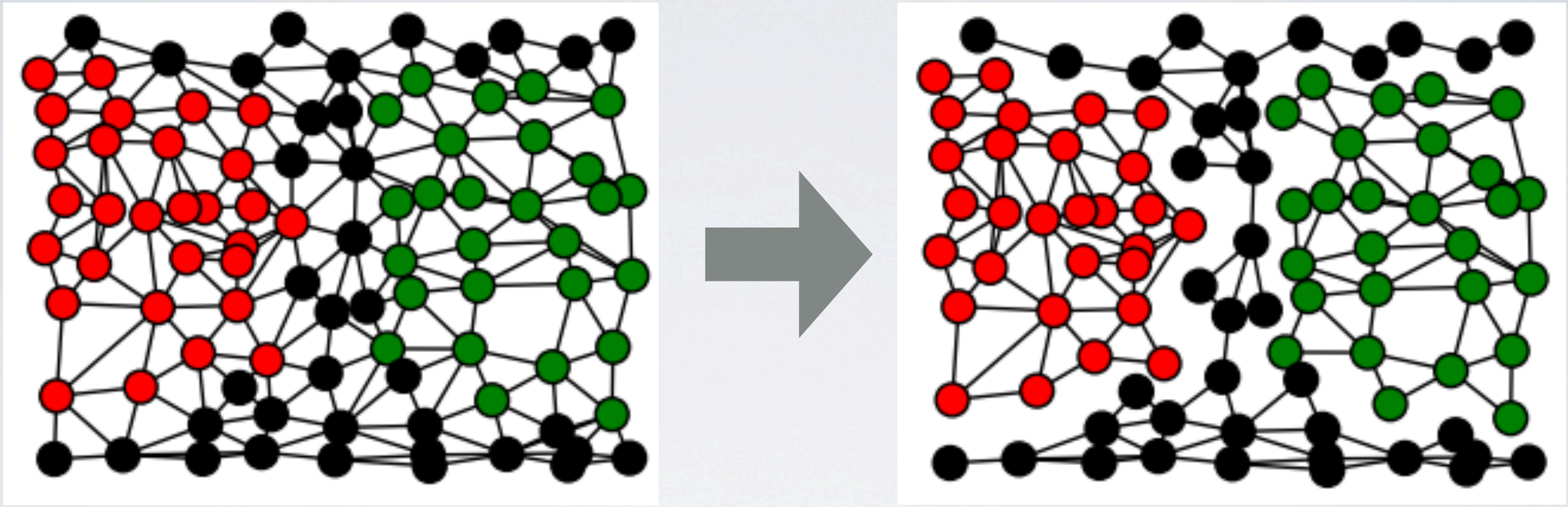
$$\boldsymbol{y} = \operatorname*{argmin}_{\boldsymbol{y}'} f_\theta(\boldsymbol{y}')$$

Non-MRF Example: Connected Components

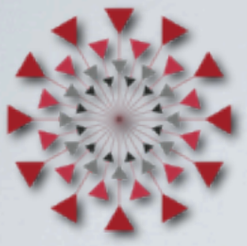




Non-MRF Example: Connected Components



- Draw random edge weights.
- Remove edges beneath threshold.
- Provides random multi-class segmentation.



Learning Randomized Optimum Models

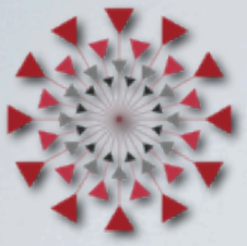
- We learn the maximum likelihood estimates of γ.

- Here, the MLE parameters are the ones that place the greatest mass on the objective functions that result in the observed data.

- Challenge: there are many objectives which are consistent with any given datum.

$$F(\theta) = \operatorname*{argmin}_{\boldsymbol{y}'} f_\theta(\boldsymbol{y}')$$

$$\theta \in F^{-1}(\boldsymbol{y}) \Rightarrow F(\theta) = \boldsymbol{y}$$



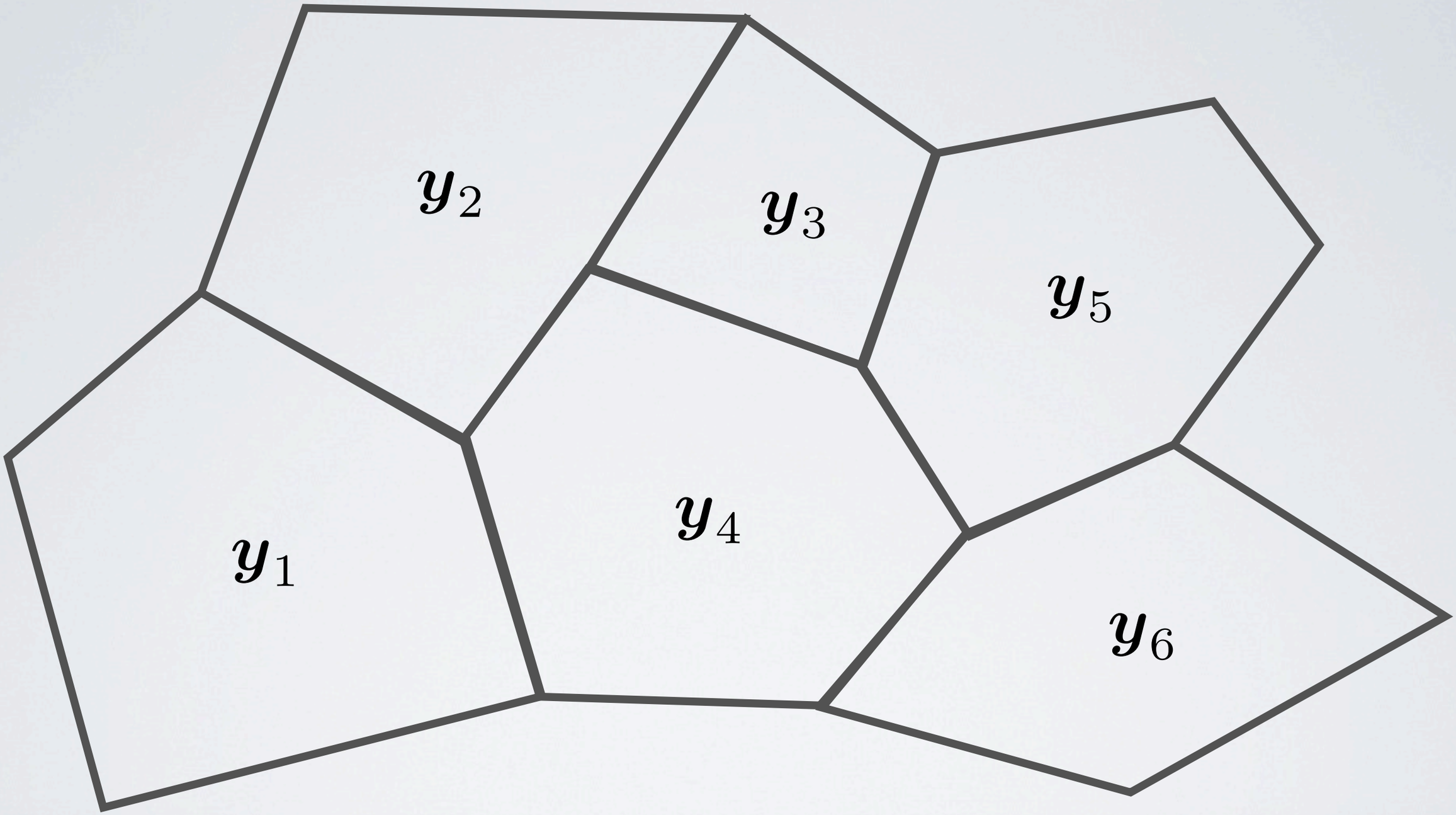
Learning Randomized Optimum Models

$$F(\theta) = \operatorname*{argmin}_{\boldsymbol{y}'} f_\theta(\boldsymbol{y}')$$

$$\theta \in F^{-1}(\boldsymbol{y}) \Rightarrow F(\theta) = \boldsymbol{y}$$

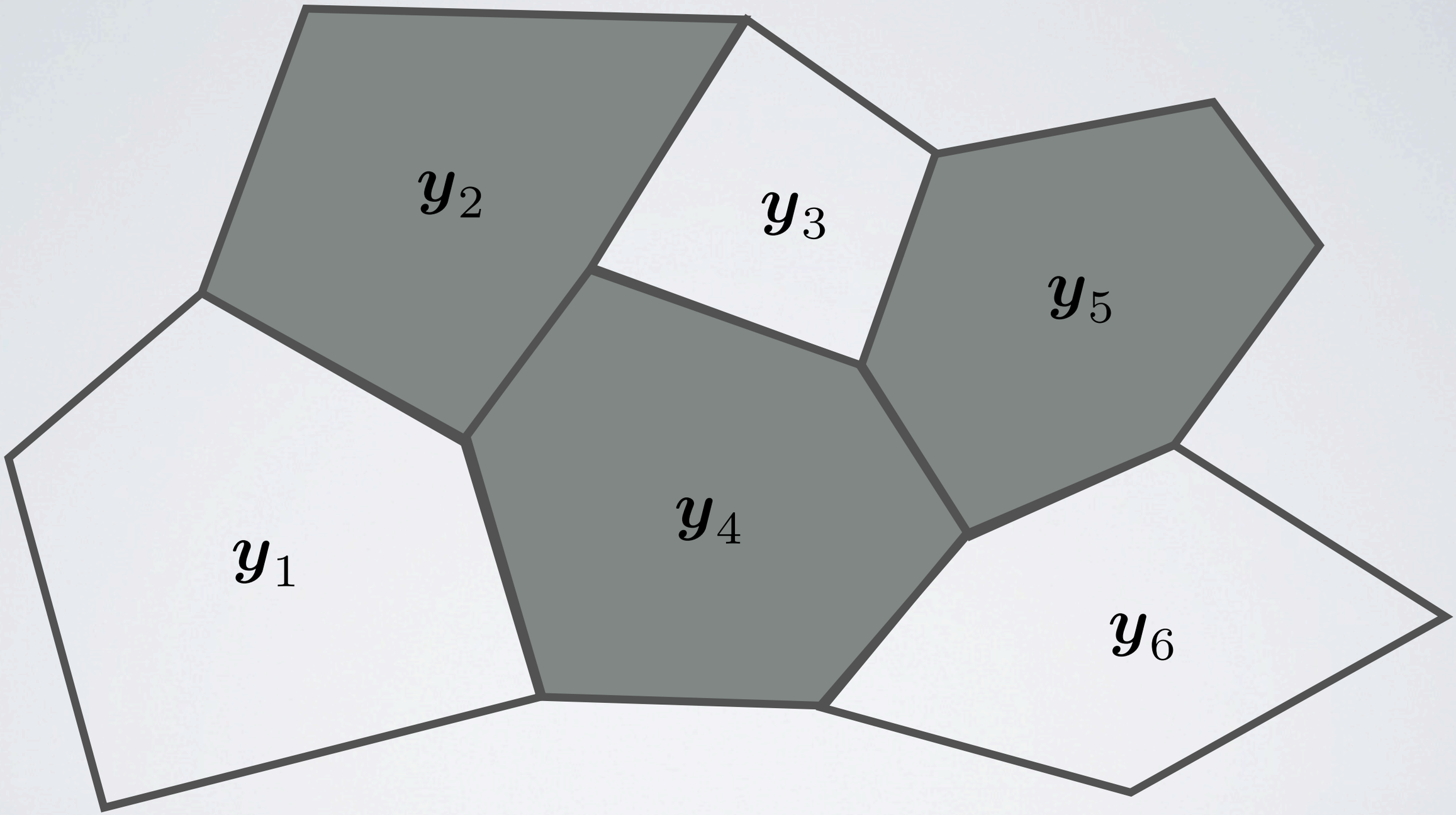
- The likelihood of θ is uniform over $F^{-1}(\boldsymbol{y})$.
- We focus on cases where the sets $F^{-1}(\boldsymbol{y})$ are simple, e.g., convex, star-convex, or easily projected upon.

The Geometry of Inverse Sets



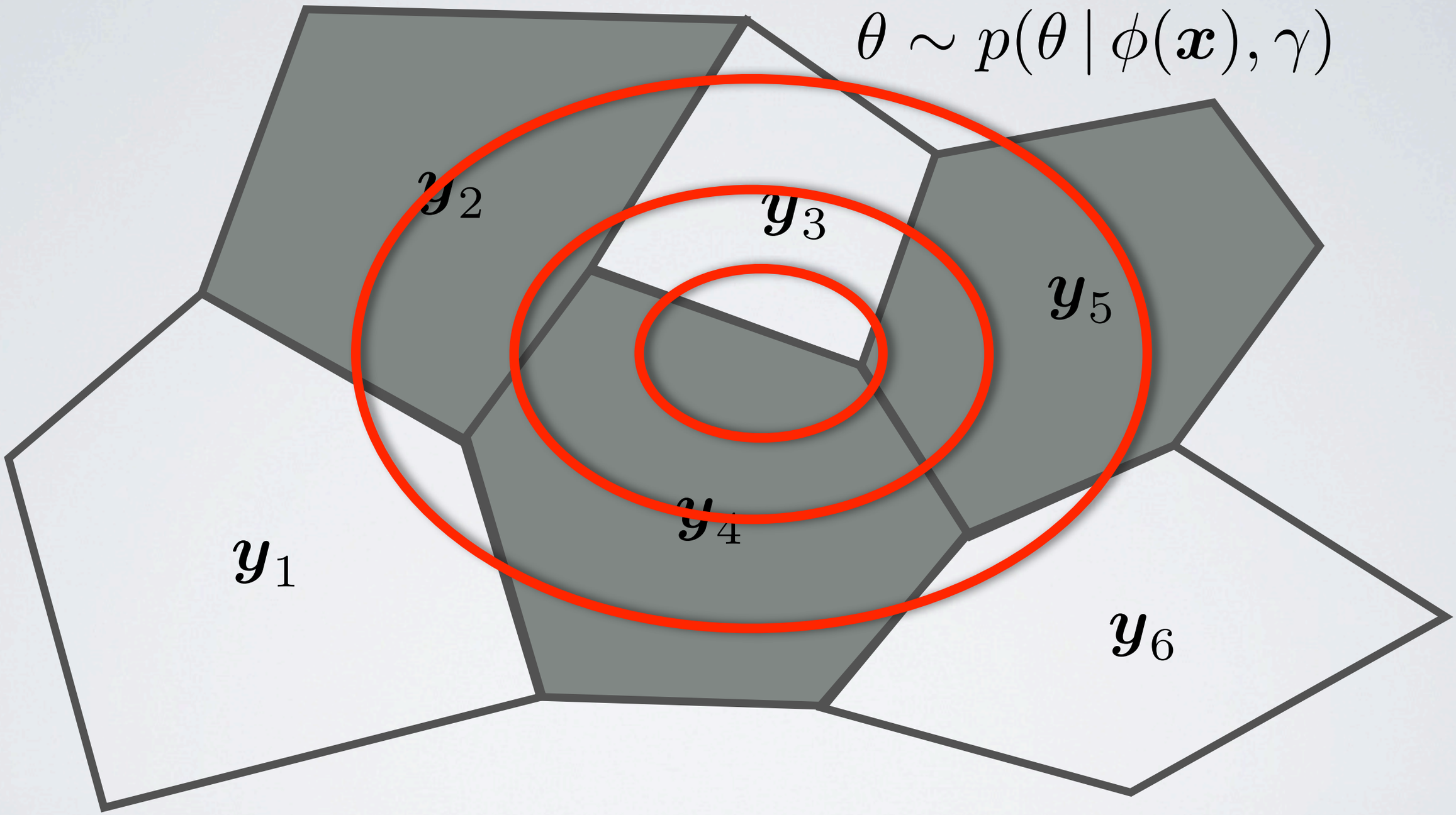
Equivalence classes of objective functions, leading to identical solutions.

The Geometry of Inverse Sets

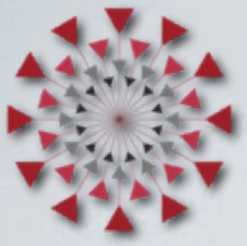


A subset of the equivalence classes are associated with observed data.

The Geometry of Inverse Sets



Find the distribution that places the most mass within the observed classes.



Monte Carlo Expectation Maximization

- We don't know which objective function is the "true" one for any given data example.

- We would like to integrate out the corresponding objective parameter θ, to maximize the likelihood.

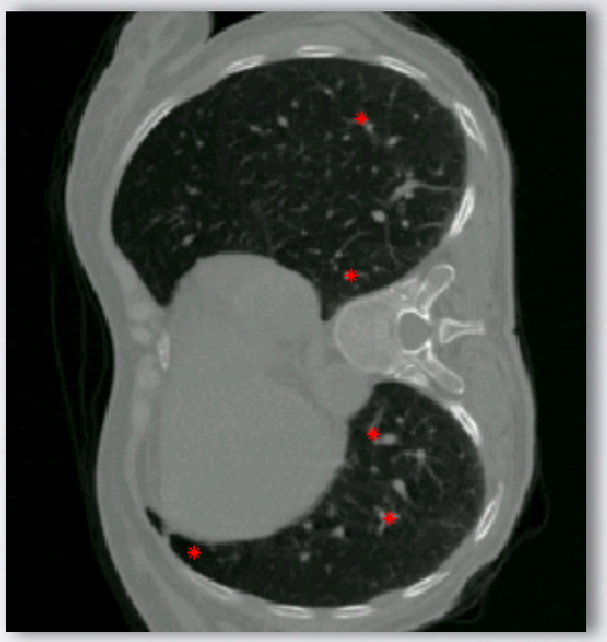
- Monte Carlo EM uses MCMC to integrate out the latent location in the data-consistent volume.

- The simple shape of the inverse set enables this to mix well.

- Dynamic variants of, e.g., graph cuts lead to efficient inside/outside oracles for slice sampling.

Lung CT Results

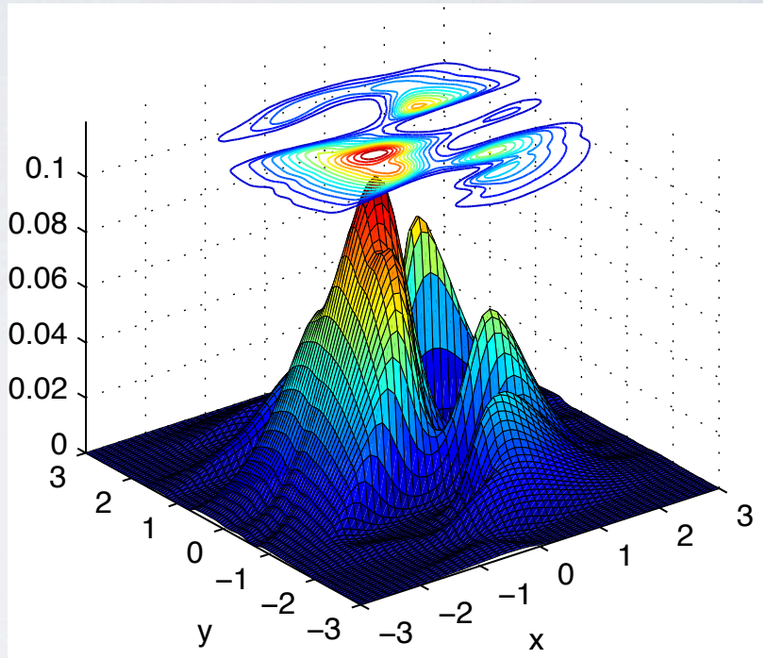
- Volumetric CT of lungs in 10 patients.
- 30 landmarks each
- Find feature-driven matches.

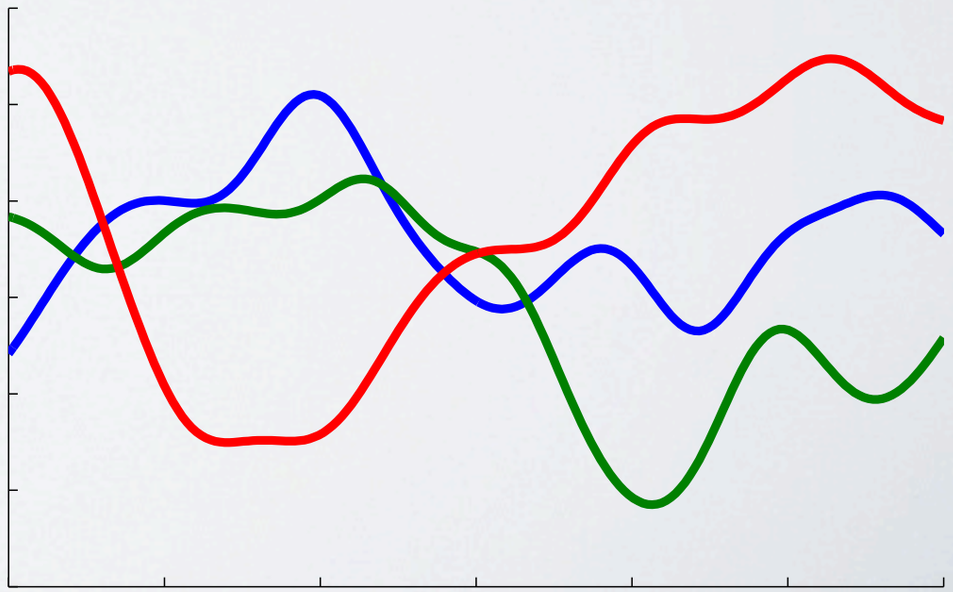


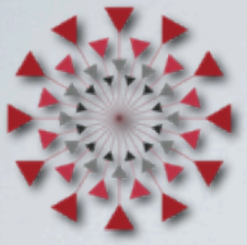
Alg	Avg. Score	Avg. Prob	Avg. Log Prob
Max-margin	91.0%	.400	0 (6)
PM(1)	95.0%	.538	-1.52 (0)
PM(10)	94.9%	.543	-1.54 (0)
MCEM(1)	98.8%	.837	-1.67 (2)
MCEM(5)	96.8%	.568	-0.62 (1)
MCEM(10)	96.7%	.544	-0.56 (1)

A New Probabilistic Modeling Tool

- We can import flexible tools for supervised and unsupervised modeling from continuous spaces and use them for structured prediction:
  - Gaussian process regression
  - Latent linear models
  - Dirichlet process mixtures
  - Latent feature models







Uncertainty in Image Segmentation

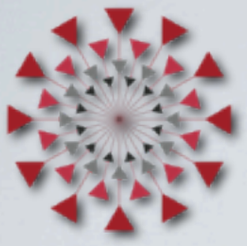
- When making predictions in low-level vision, we often want to represent our uncertainty.

- Acquiring marginals from MRFs can be expensive.

- Kohli & Torr (2008) suggested softmax min-marginals, with energy $E(\boldsymbol{y})$ and $y_d \in \{1, 2, \cdots, K\}$:

$$\Phi_d(k) = \min_{\boldsymbol{y}|y_d=k} E(\boldsymbol{y}) \qquad p(y_d = k) \approx \frac{\exp\{-\Phi_d(k)\}}{\sum_{k'=1}^{K} \exp\{-\Phi_d(k)\}}$$

- Easy to compute, but may make very poor predictions.

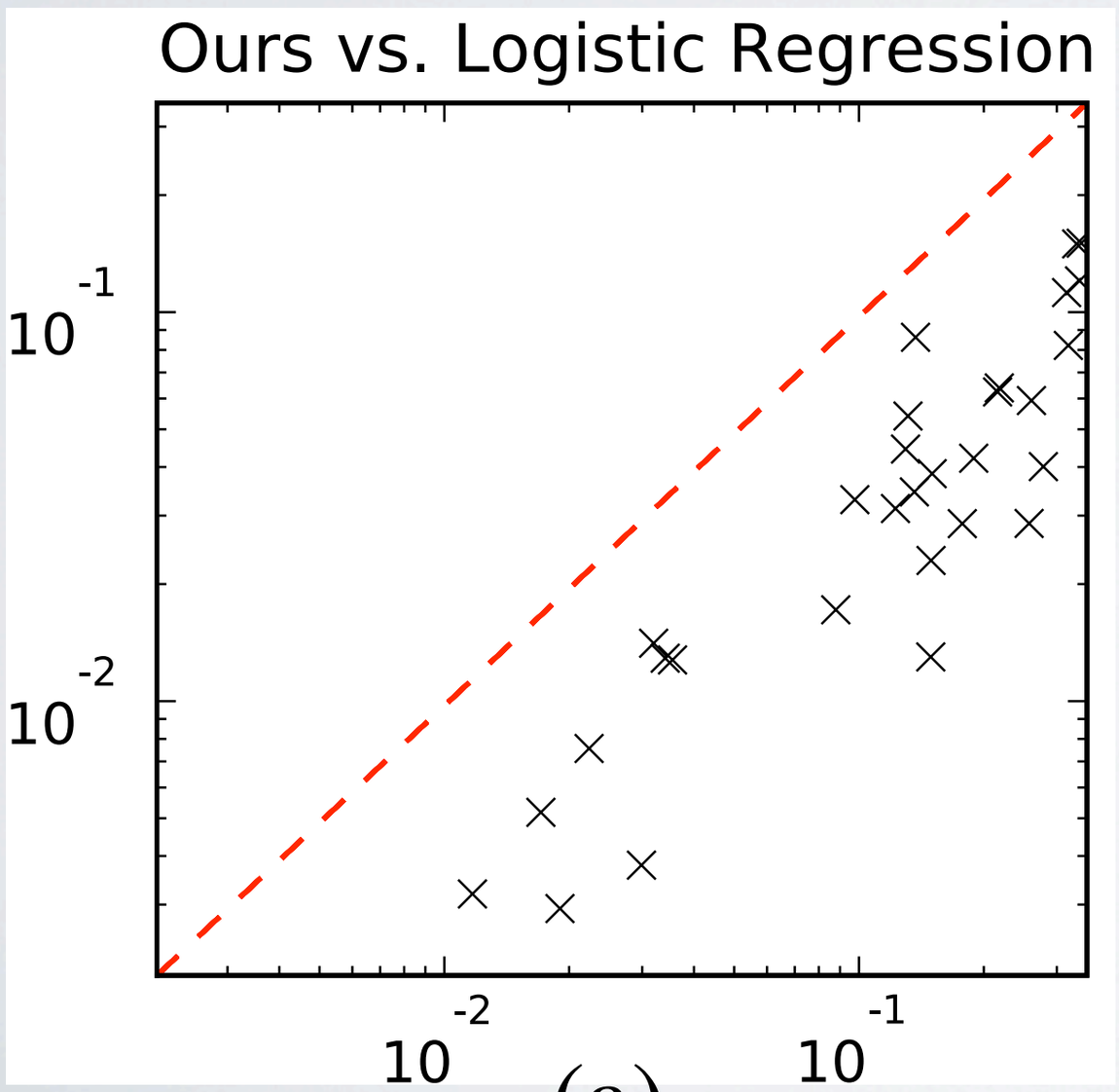


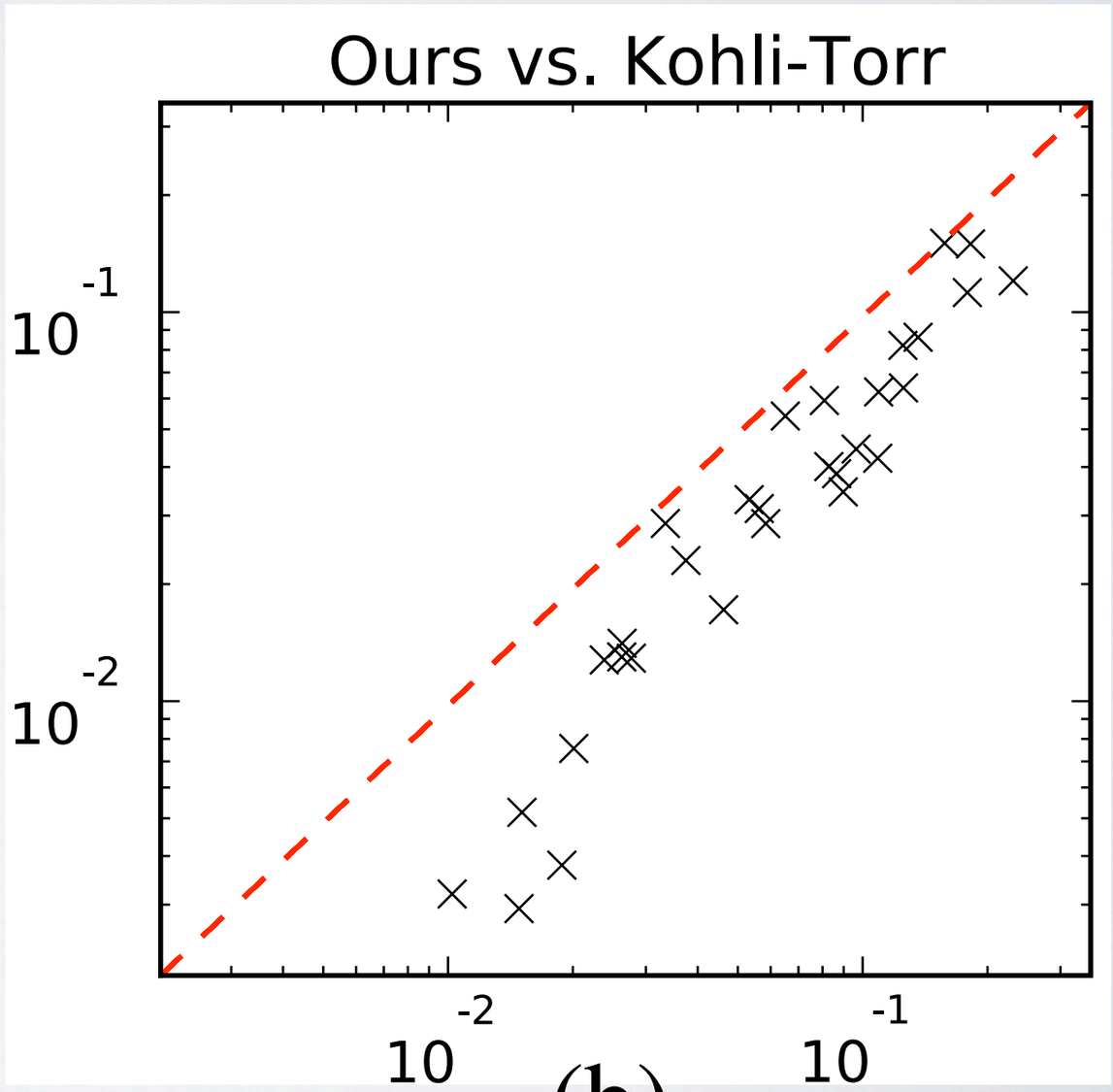
Generating Data via Min-Marginals

- As in RandOMs, why bother with the MRF at all?
- We take Kohli & Torr (2008) seriously as providing a probabilistic model for segmentation.
- **Effectively a coupled parameterization of a large collection of independent multinomials.**
- It is possible to compute the gradient of the log likelihood in terms of the underlying CRF parameters.
- The objective is to train for achieving well-calibrated test-time marginal predictions.

Evaluating Training Procedures

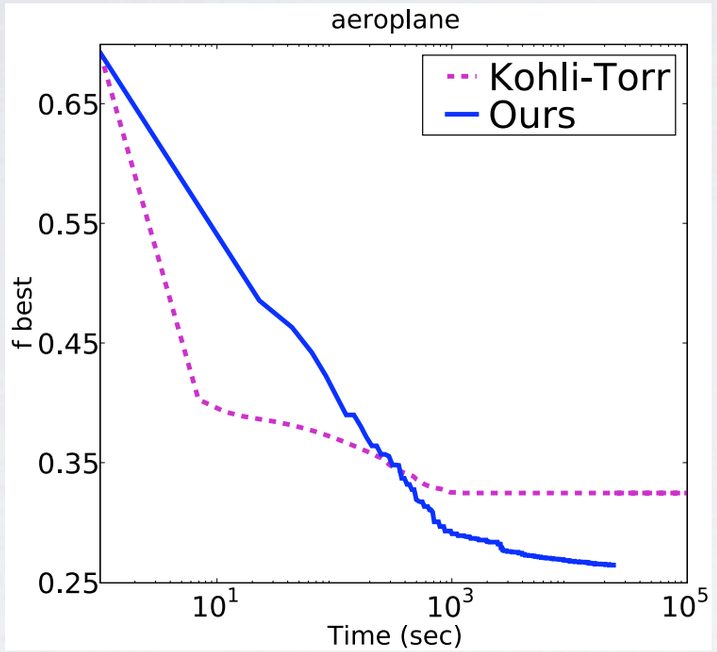
Do we get better training likelihoods by optimizing the model with the min-marginal objective?

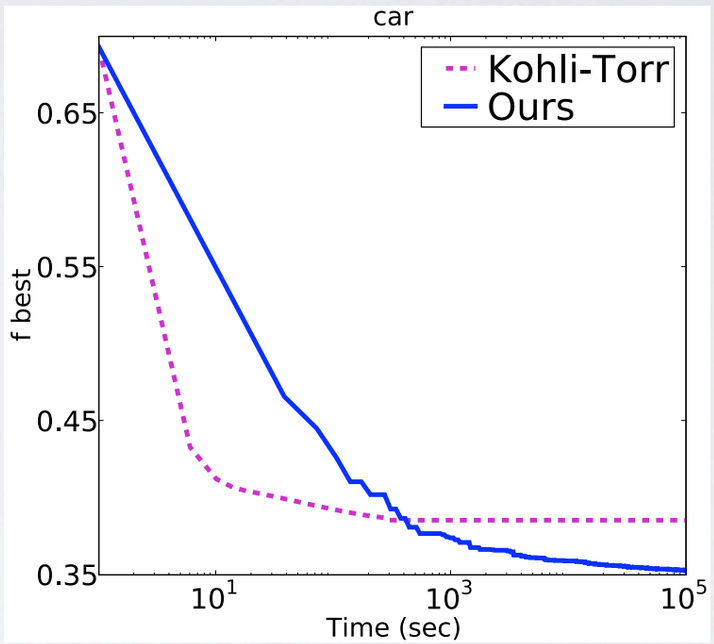


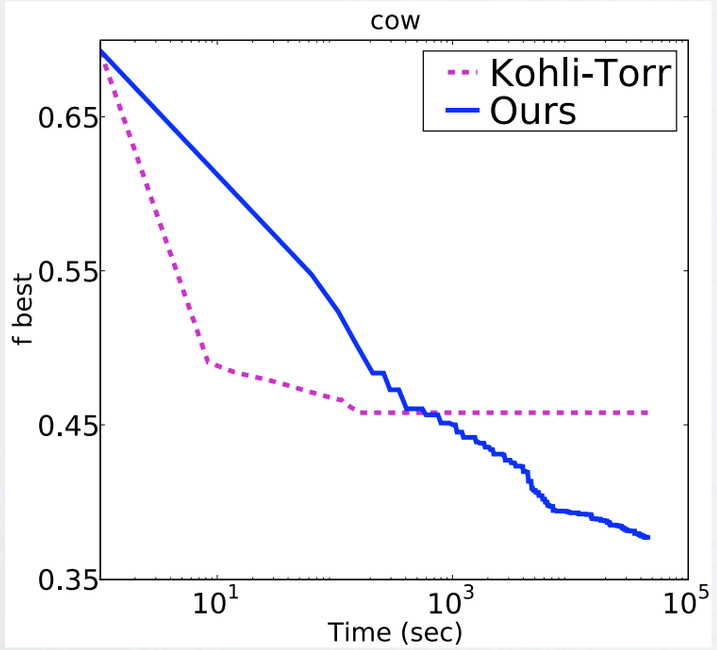


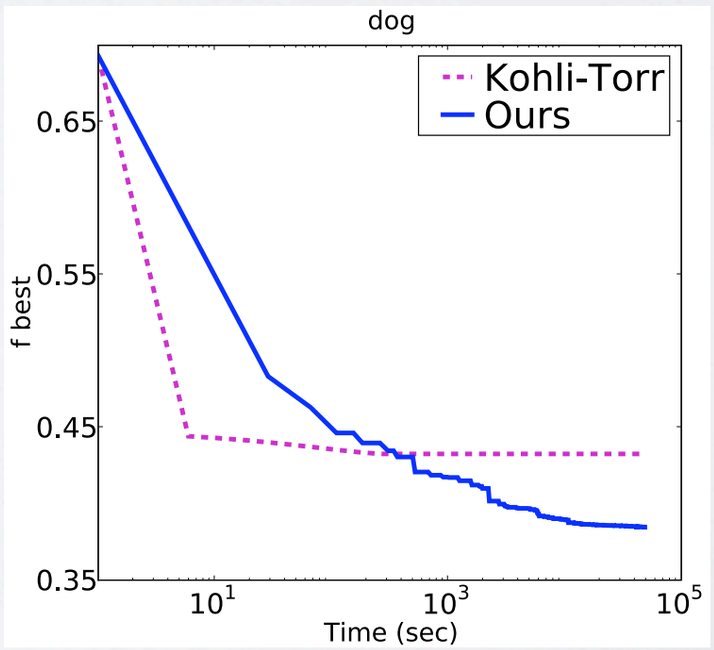
Evaluating Training Procedures

Does training time improve using the gradient of the log likelihood vs procedure of K&T?





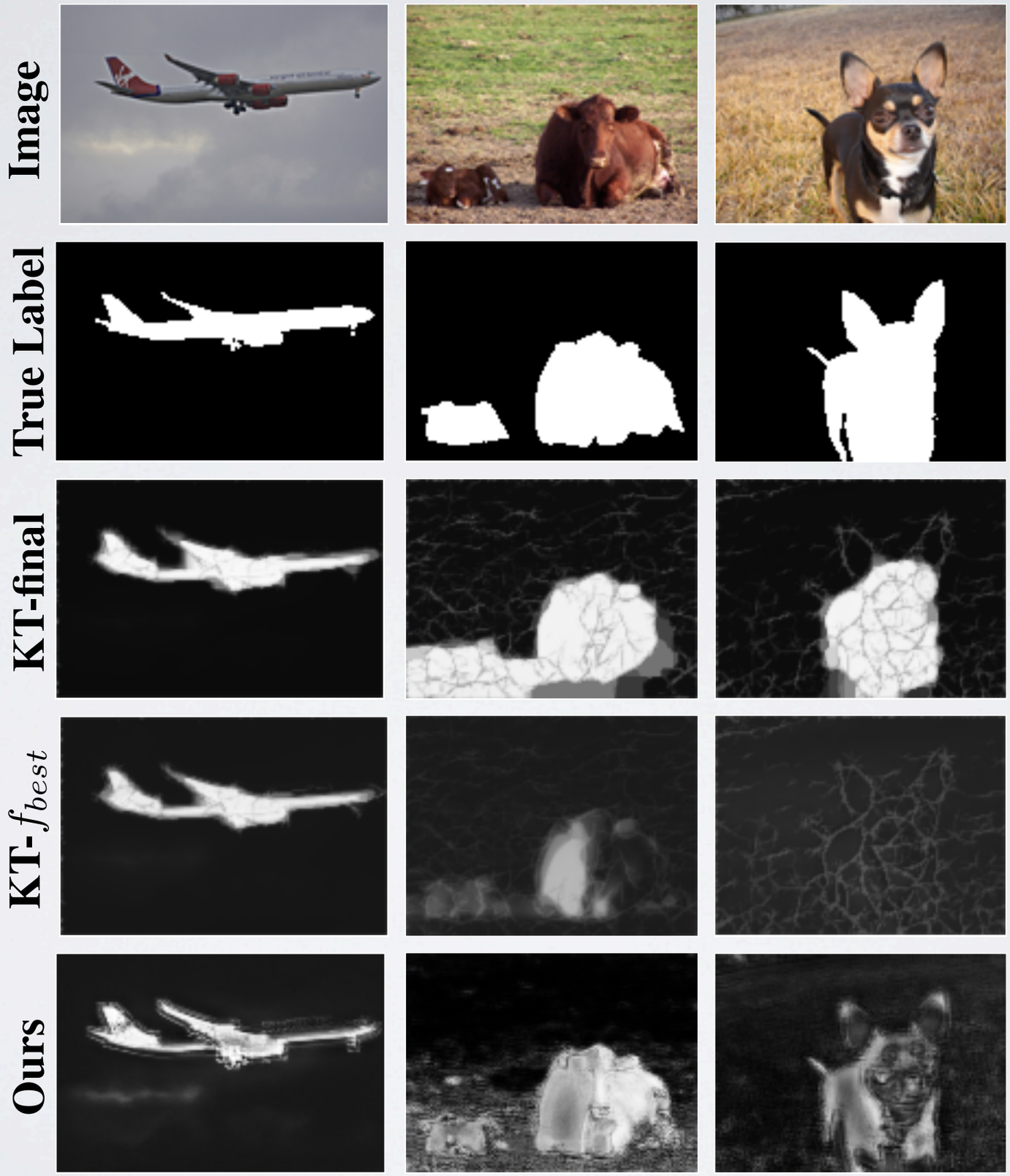




Evaluating Test-Time Performance

		Log Lik	Accuracy	AUC
Aero	KT-final	-.35 (-.29)	87.3 (89.7)	.85 (.87)
	KT-f_{best}	-.32 (-.28)	88.1 (89.9)	.85 (.87)
	Ours	**-.26** (**-.24**)	**88.9** (**90.4**)	**.90** (**.90**)
Car	KT-final	-.48 (-.65)	84.1 (78.7)	.69 (**.65**)
	KT-f_{best}	-.39 (**-.51**)	**86.2** (80.0)	.66 (.62)
	Ours	**-.35** (**-.51**)	86.1 (**81.4**)	**.76** (**.65**)
Cow	KT-final	-.54 (-.64)	79.7 (75.9)	.76 (.77)
	KT-f_{best}	-.47 (-.52)	80.9 (76.7)	.66 (.65)
	Ours	**-.38** (**-.41**)	**82.5** (**79.9**)	**.84** (**.82**)
Dog	KT-final	-.52 (-.45)	81.8 (84.7)	.64 (.66)
	KT-f_{best}	-.43 (-.38)	**84.1** (86.7)	.62 (.66)
	Ours	**-.38** (**-.34**)	84.0 (**86.8**)	**.76** (**.79**)

Evaluating Test-Time Performance

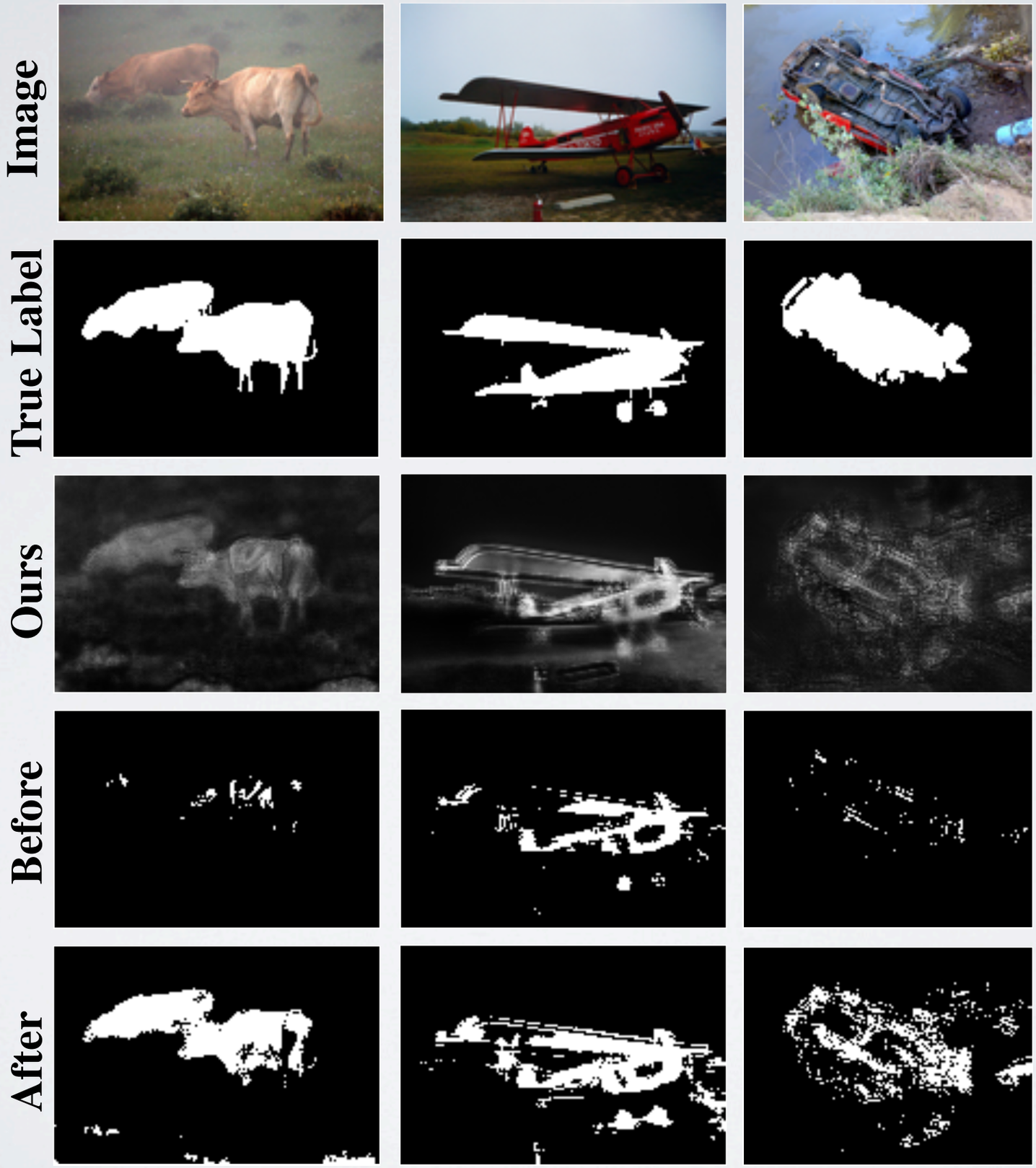


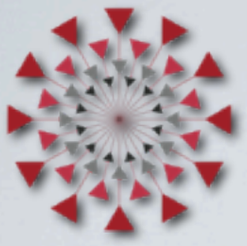
Incorporating Test-Time Losses

If we must produce **one** test-time segmentation, better-calibrated probabilities enable us to choose the segmentation that **minimizes expected loss**.

		∩/∪ **Before**	∩/∪ **After**	**Change**
Aero	KT-final	63.4	63.9	.5
	KT-f_{best}	60.7	61.8	1.1
	Ours	**64.1**	**66.6**	**2.5**
Car	KT-final	**43.5**	44.2	.7
	KT-f_{best}	40.7	43.3	2.6
	Ours	41.8	**45.9**	**4.1**
Cow	KT-final	**51.2**	51.7	.5
	KT-f_{best}	40.7	47.1	**6.4**
	Ours	46.5	**52.8**	6.3
Dog	KT-final	**52.1**	52.0	-.1
	KT-f_{best}	44.5	47.2	2.7
	Ours	48.9	**55.3**	**6.4**

INCORPORATING TEST-TIME LOSSES





Summary

- MRF doesn't stand for "Magical Random Field"

- Efficient optimization procedures can provide a useful tool in the generative modeling toolbox for structured prediction.

- Train your models to answer the questions you'll ask them at test time!

- If the full representational power of the model will never be used, there may be a significant computational gain in exactly fitting a model that is "just big enough".

THANKS

- Danny Tarlow (University of Toronto)
- Rich Zemel (University of Toronto)

- D. Tarlow, R.P. Adams, R.S. Zemel. Randomized Optimum Models for Structured Prediction. AISTATS 2012.
- D. Tarlow, R.P. Adams. Revisiting Uncertainty in Graph Cut Solutions. CVPR 2012.