10-418 / 10-618 Machine Learning for Structured Data

Machine Learning Department
School of Computer Science
Carnegie Mellon University

Structured SVM

Matt Gormley
Lecture 15
Oct. 16, 2019

Reminders

- **Midterm Exam**
  - Thu, Oct. 17 at 6:30pm – 8:00pm
- **Homework 3: Structured SVM**
  - Out: Fri, Oct. 18
  - Due: Fri, Nov. 1 at 11:59pm

aka. Max-Margin Markov Networks (M^3Ns)

STRUCTURED SVM

Structured SVM

Whiteboard

- Warmup: Binary SVM
- Warmup: Binary SVM Hinge Loss
- Structured Large Margin
- Structured Hinge Loss
- Gradient of Structured Hinge Loss
- SGD for Structured SVM
- Loss Augmented MAP Inference

Max vs "Soft-Max" Margin

- SVMs:

$$\min_{\mathbf{w}} k||\mathbf{w}||^2 - \sum_i \Bigg(\underbrace{\mathbf{w}^\top \mathbf{f}_i(\mathbf{y}^i) - \max_{\mathbf{y}} \left(\mathbf{w}^\top \mathbf{f}_i(\mathbf{y}) + \ell_i(y) \right)}_{\text{Hard (Penalized) Margin}} \Bigg)$$

- Maxent:

$$\min_{\mathbf{w}} \ k||w||^2 - \sum_i \Bigg(\underbrace{\mathbf{w}^\top \mathbf{f}_i(\mathbf{y}^i) - \log \sum_{\mathbf{y}} \exp \left(\mathbf{w}^\top \mathbf{f}_i(\mathbf{y}) \right)}_{\text{Soft Margin}} \Bigg)$$

- Very similar! Both try to make the true score better than a function of the other scores.
  - The SVM tries to beat the augmented runner-up
  - The maxent classifier tries to beat the "soft-max"

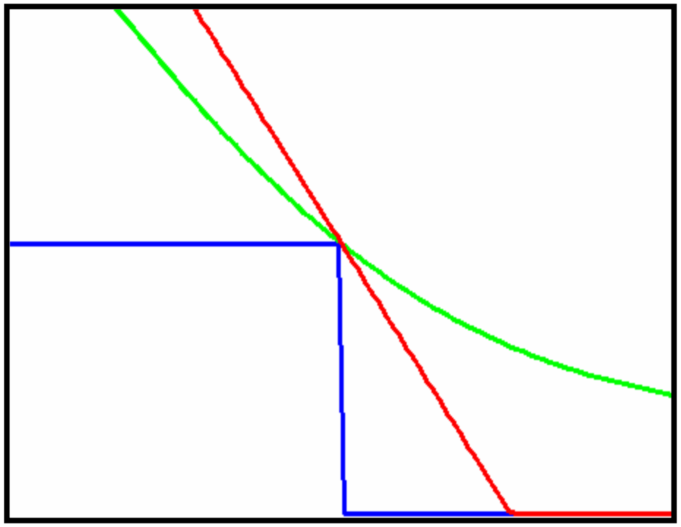
Slide from Klein & Taskar (ACL 2005 Tutorial)

Hinge Loss

- Consider the per-instance SVM objective:

$$\min_{\mathbf{w}} k||\mathbf{w}||^2 - \sum_i \left(\mathbf{w}^\top \mathbf{f}_i(\mathbf{y}^i) - \max_{\mathbf{y}} \left[\mathbf{w}^\top \mathbf{f}_i(\mathbf{y}) + \ell_i(y) \right] \right)$$

- This is called the "hinge loss"
  - Upper bounds zero-one loss
  - Unlike maxent / log loss, you stop gaining objective once the true label wins by enough
  - You can start from here and derive the SVM objective



$$\mathbf{w}^\top \mathbf{f}_i(\mathbf{y}^i) - \max_{\mathbf{y} \neq \mathbf{y}^i} \mathbf{w}^\top \mathbf{f}_i(\mathbf{y})$$

Slide from Klein & Taskar (ACL 2005 Tutorial)

Max (Conditional) Likelihood

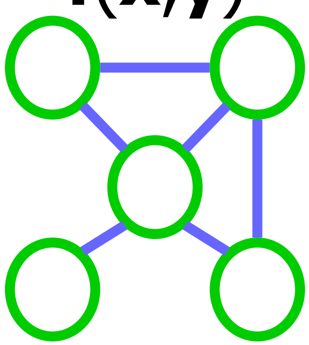
D

$\mathbf{x}^1, \mathbf{t}(\mathbf{x}^1)$

…

$\mathbf{x}^m, \mathbf{t}(\mathbf{x}^m)$

$\mathbf{f}(\mathbf{x},\mathbf{y})$



Estimation

$$\text{maximize}_{\mathbf{w}} \sum_{\mathbf{x}\in D} \log P_{\mathbf{w}}(\mathbf{t}(\mathbf{x}) \mid \mathbf{x})$$

Classification

$$\arg\max_{\mathbf{y}} \mathbf{w}^{\top}\mathbf{f}(\mathbf{x},\mathbf{y})$$

$$\log P_{\mathbf{w}}(\mathbf{y} \mid \mathbf{x}) = \mathbf{w}^{\top}\mathbf{f}(\mathbf{x},\mathbf{y}) - \log Z_{\mathbf{w}}(\mathbf{x})$$

Don't need to learn entire distribution!

Results: Handwriting Recognition

Length: ~8 chars

Letter: 16x8 pixels

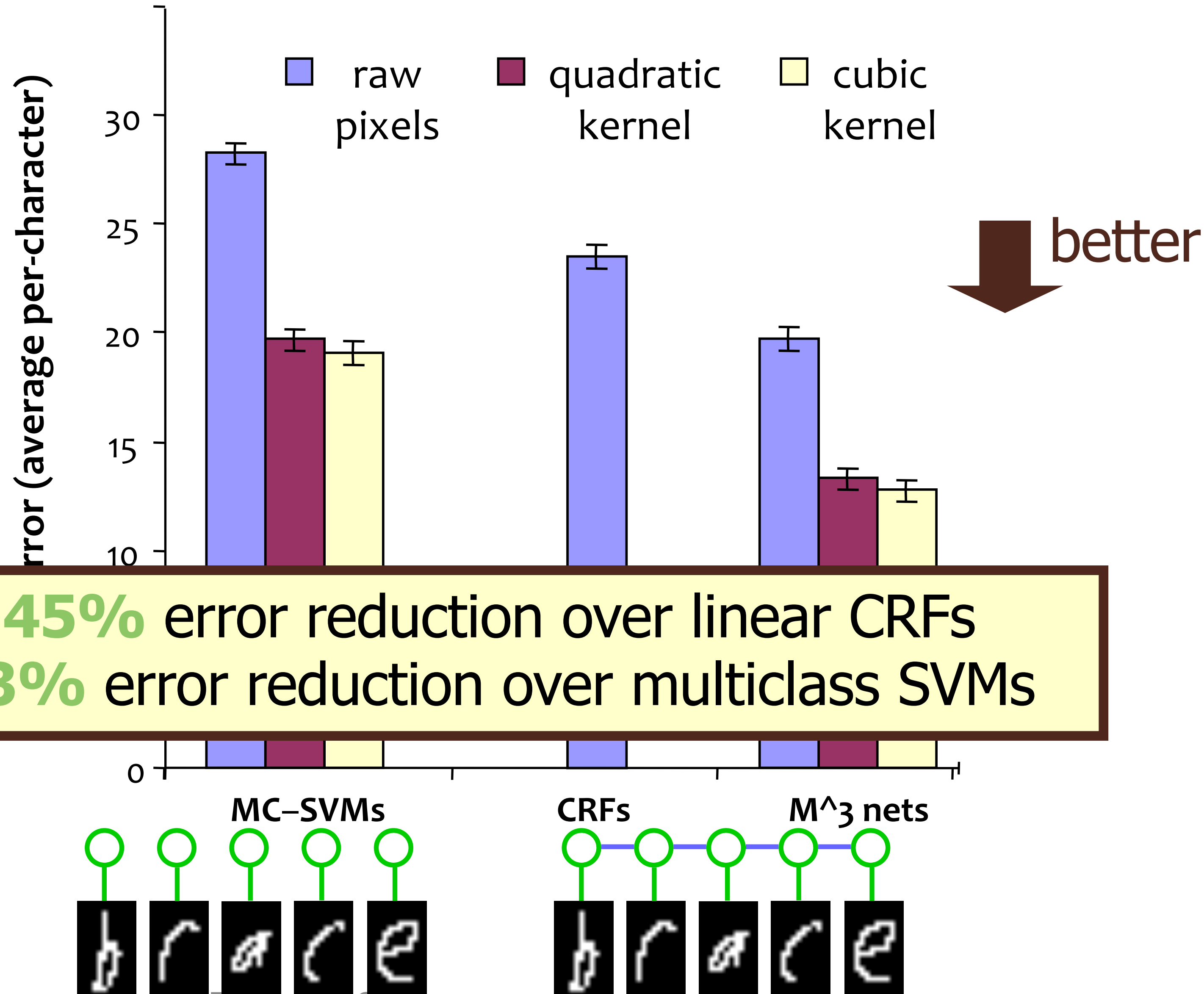
10-fold Train/Test

5000/50000 letters

600/6000 words

Models:

- Multiclass-SVMs*
- CRFs
- M^3 nets



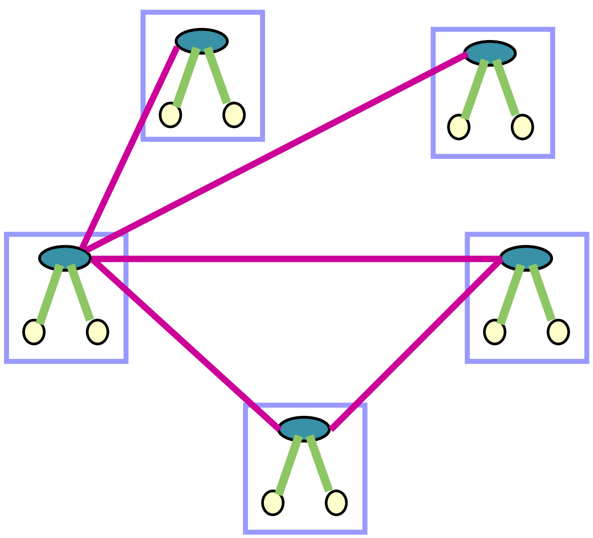
45% error reduction over linear CRFs

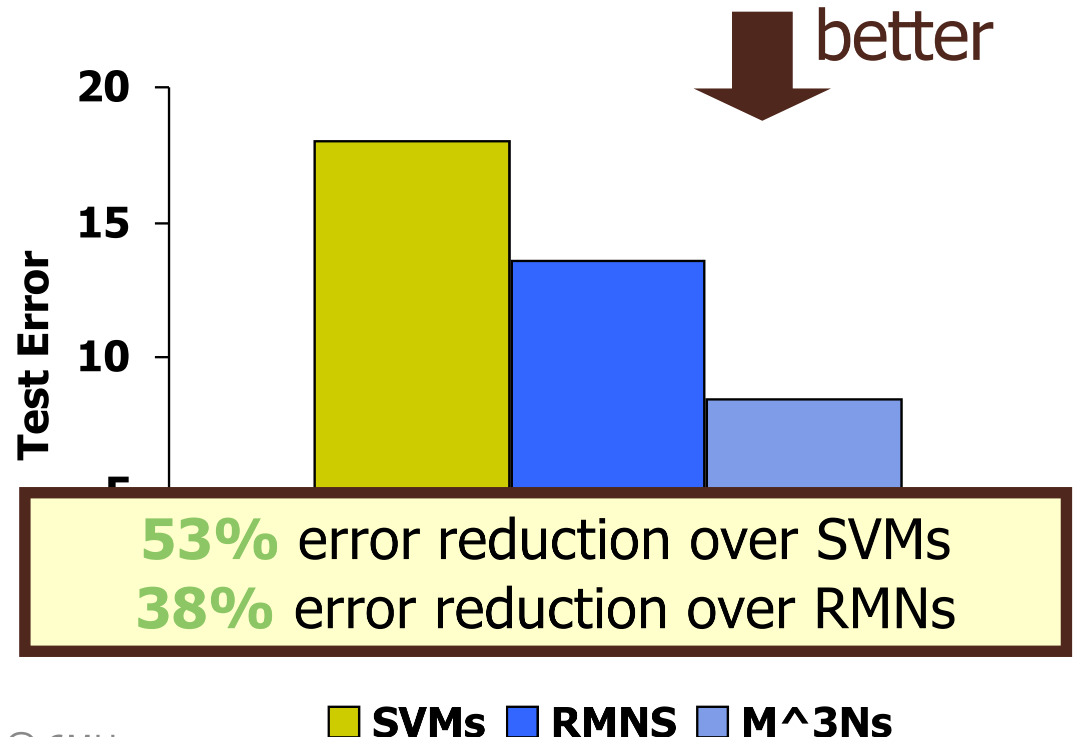
33% error reduction over multiclass SVMs

*Crammer & Singer 01

Results: Hypertext Classification

- WebKB dataset
  - Four CS department websites: 1300 pages/3500 links
  - Classify each page: faculty, course, student, project, other
  - Train on three universities/test on fourth
- Inference: loopy belief propagation
- Learning: relaxed dual

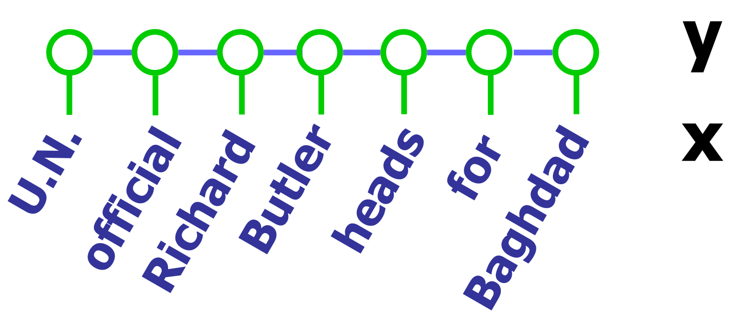




*Taskar et al 02

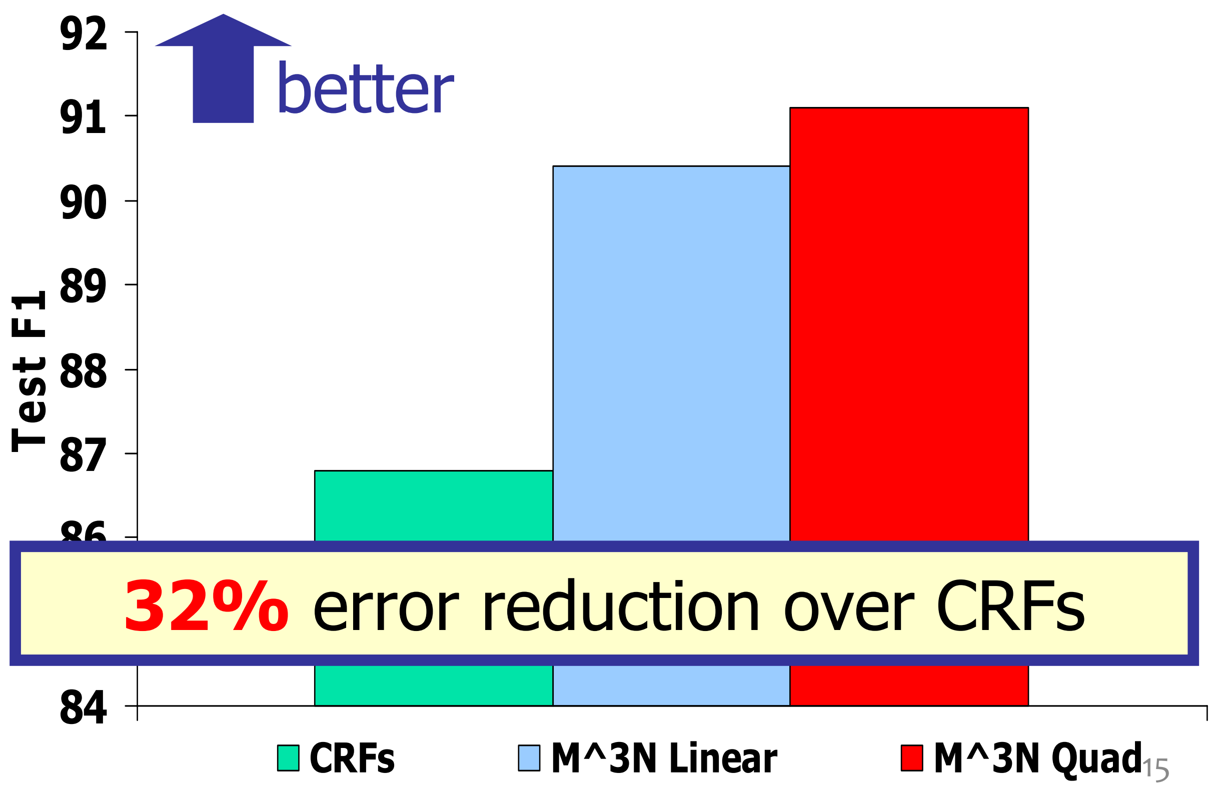
Named Entity Recognition

- Locate and classify named entities in sentences:
  - 4 categories: organization, person, location, misc.
  - e.g. "U.N. official Richard Butler heads for Baghdad".
- CoNLL 03 data set (200K words train, 50K words test)



y_i = org/per/loc/misc/none

$\mathbf{f}(y_i, x) = [\ldots,$
$I(y_i\text{=org}, x_i\text{=“U.N.”}),$
$I(y_i\text{=per}, x_i\text{=capitalized}),$
$I(y_i\text{=loc}, x_i\text{=known city}),$
$\ldots,]$



Slide from Guestrin, 10-701, 2005

Associative Markov networks

$$P(\mathbf{y} \mid \mathbf{x}) \;\propto\; \underbrace{\prod_i \phi_i(y_i, \mathbf{x}_i)}_{\text{Point features}} \prod_{ij} \underbrace{\phi_{ij}(y_i, y_j, \mathbf{x}_{ij})}_{\text{Edge features}} = \exp\{\mathbf{w}^\top \mathbf{f}(\mathbf{x}, \mathbf{y})\}$$

Point features: spin-images, point height

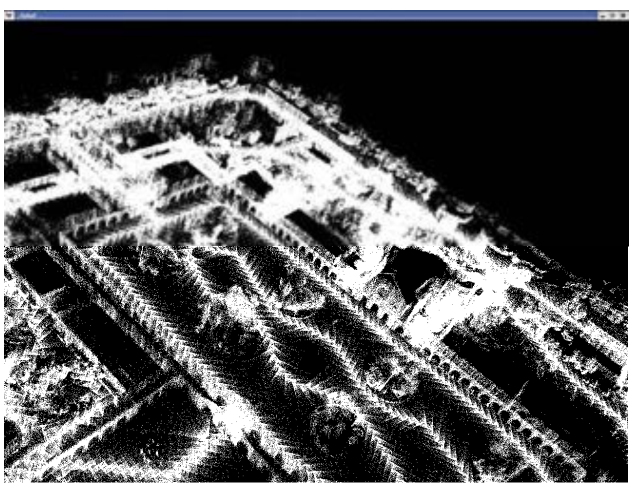
Edge features: length of edge, edge orientation

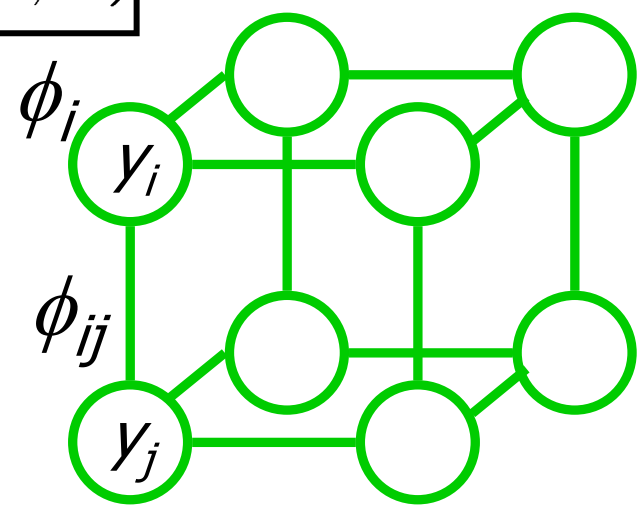
"associative" restriction

$$\phi_{ij}(y_i, y_j) \;=\; \boxed{\begin{matrix} \phi_{ij}(1,1) & & 1 \\ & \ddots & \\ 1 & & \phi_{ij}(K,K) \end{matrix}}$$

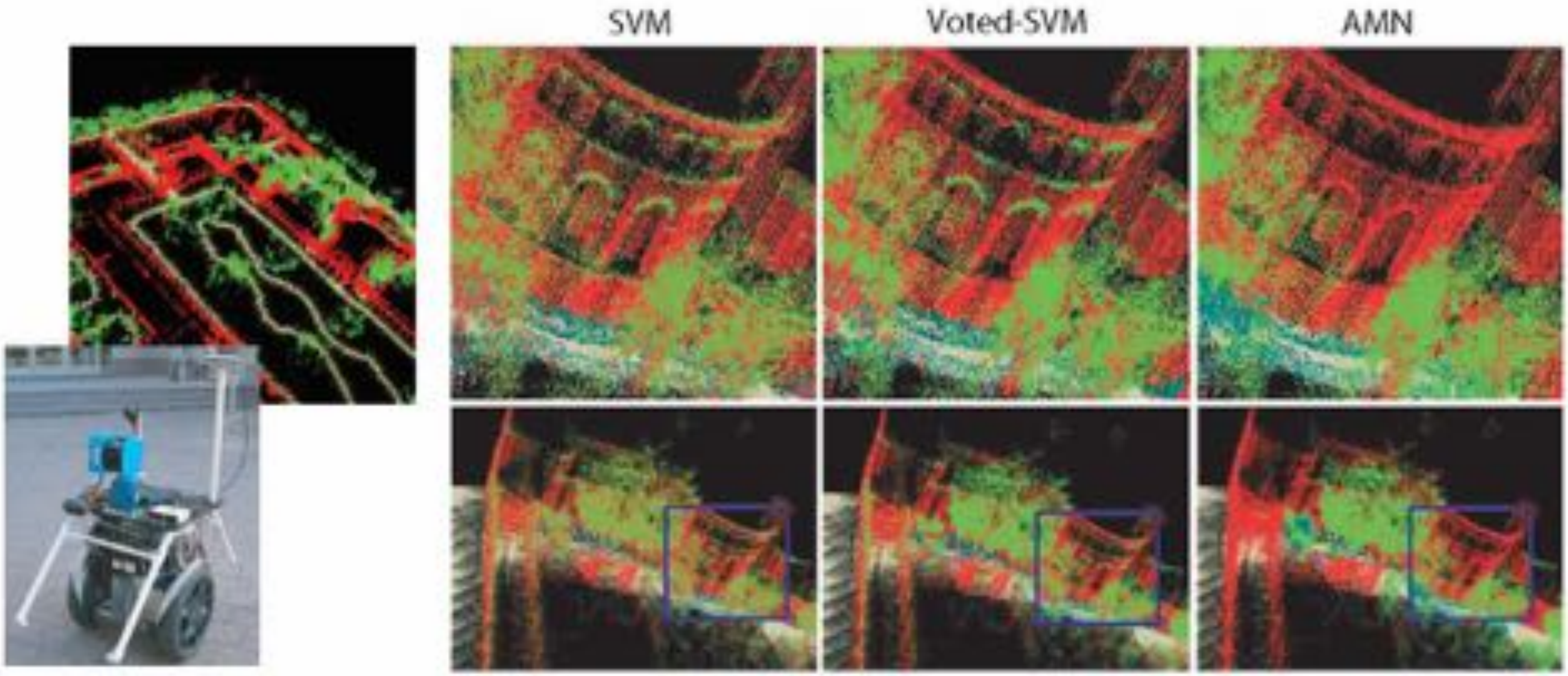
bonus

$$\phi_{ij}(k,k) \geq 1$$



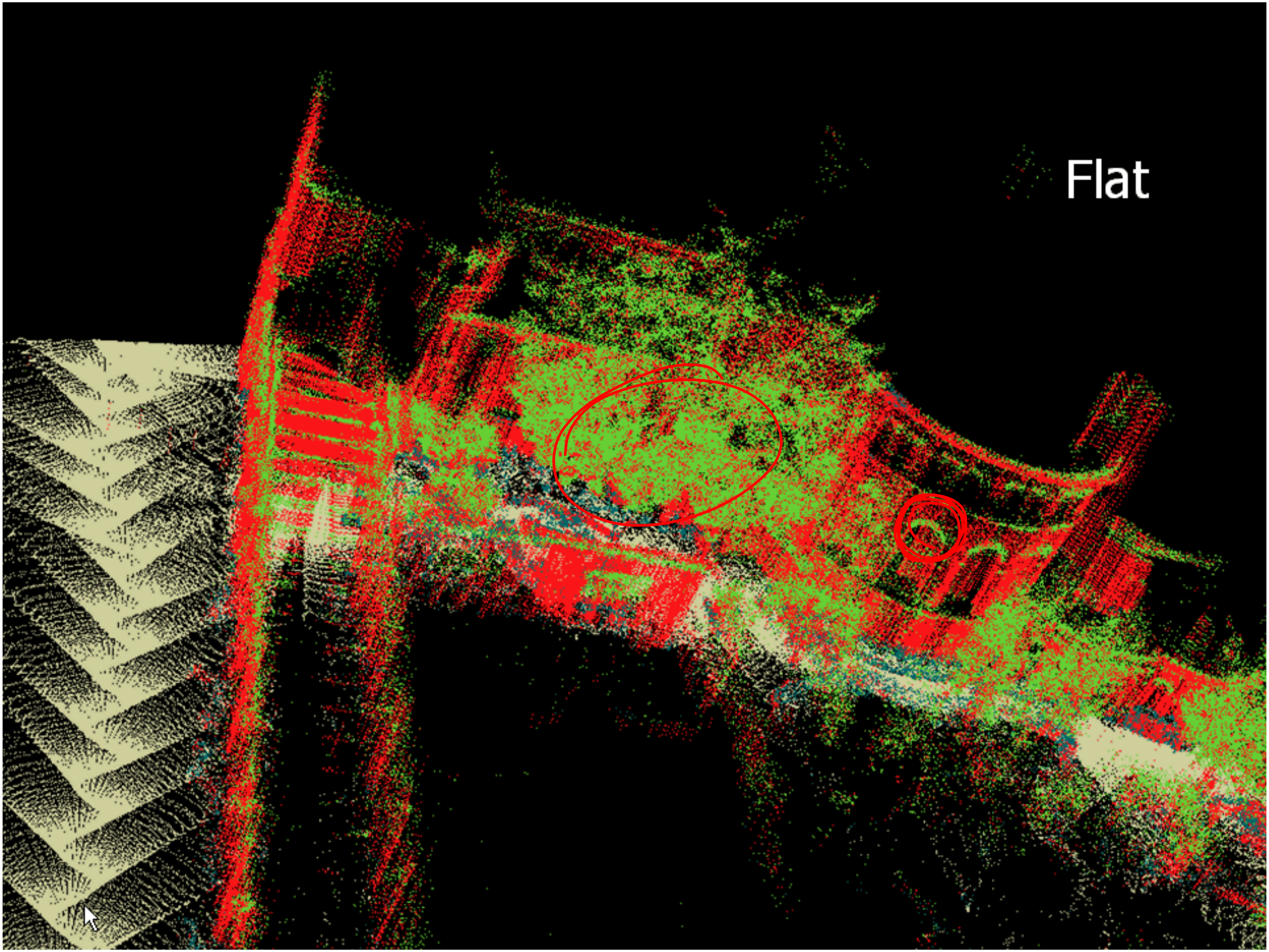


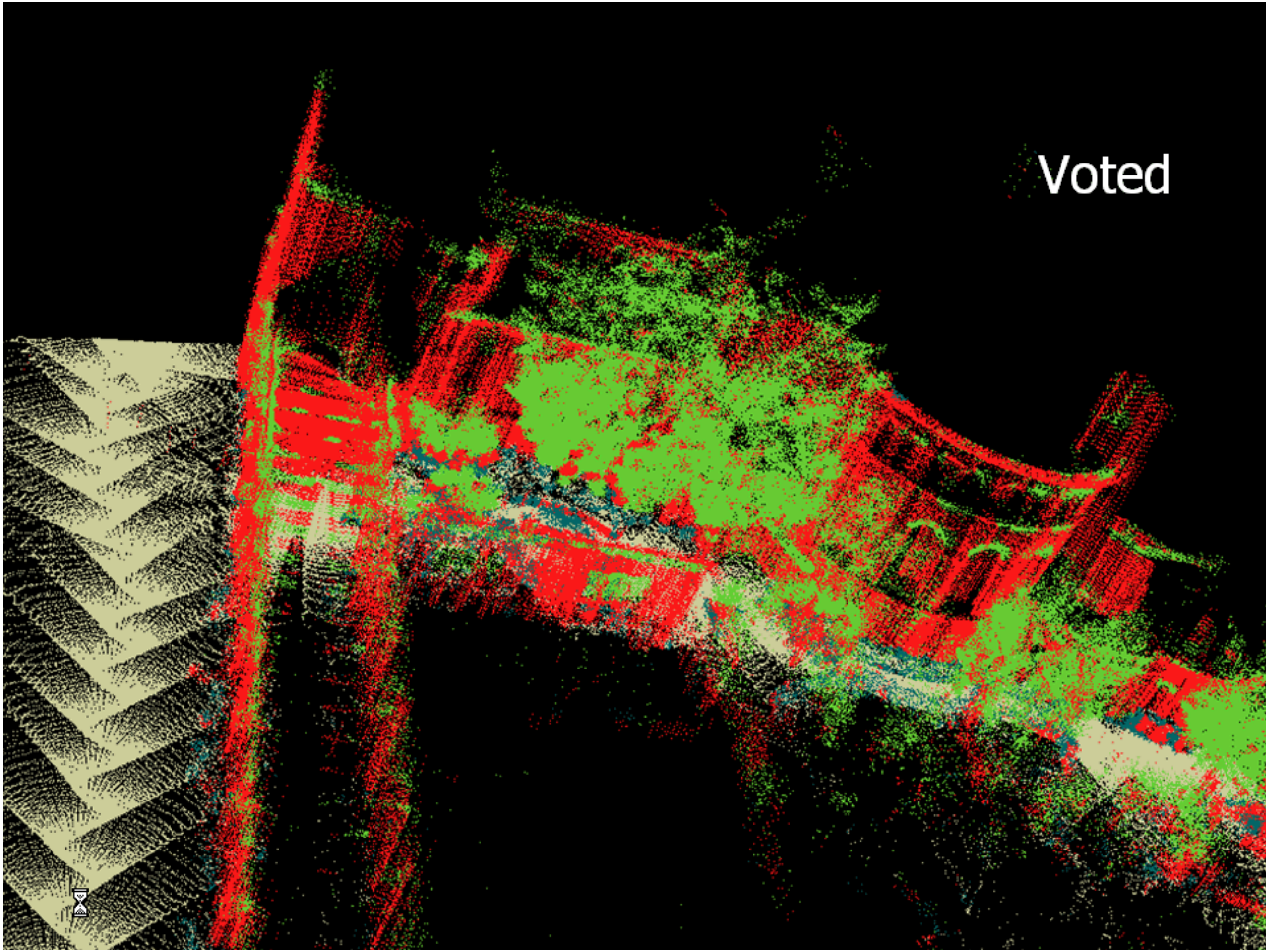
Max-margin AMNs results



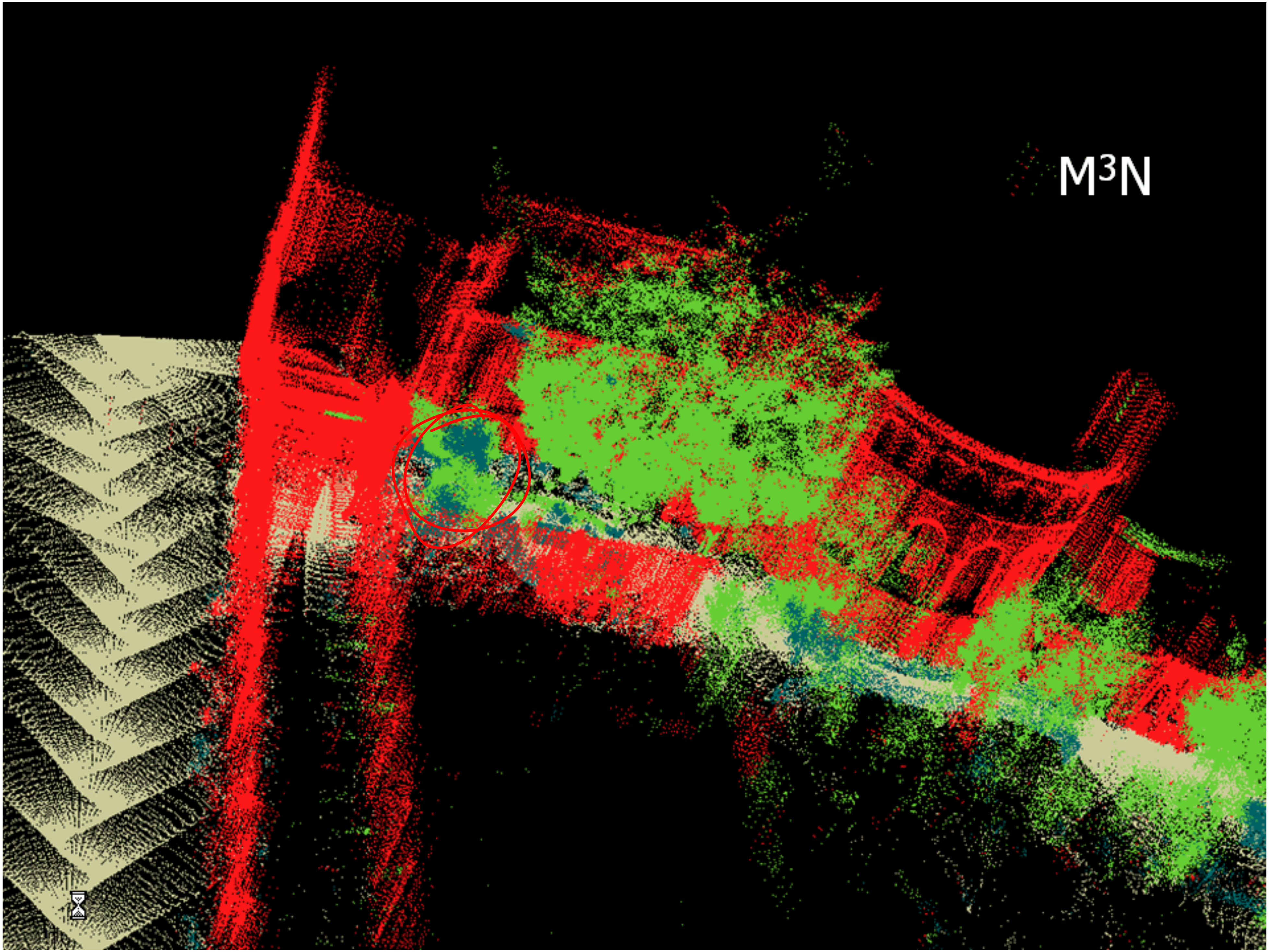
Label: ground, building, tree, shrub

Training: 30 thousand points   Testing: 3 million points


Flat


Voted

Slide from Guestrin, 10-701, 2005



Segmentation results

Hand labeled 180K test points

Model	Accuracy
SVM	68%
V-SVM	73%
M^3N	93%