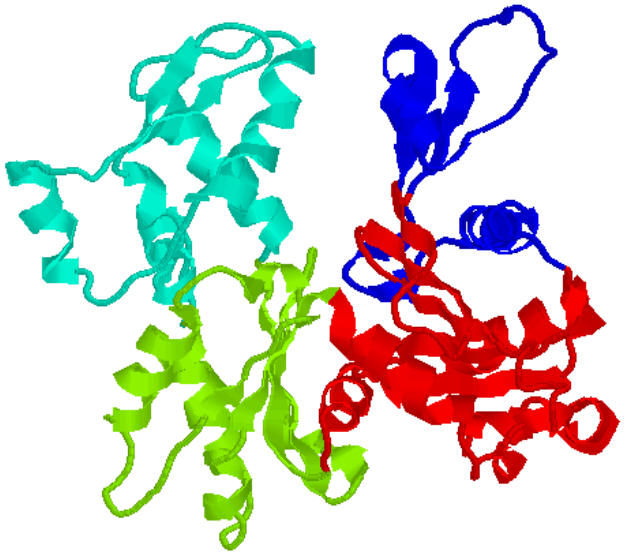
# Learning from Inconsistent and Unreliable Annotators



**by Ping Zhang and Zoran Obradovic**

**Center for Data Analytics and Biomedical Informatics, Temple University**

September, 2011

Athens, Greece

# Typical supervised classification

| Instance $\boldsymbol{x}_i \in \mathbf{R}^d$ | Label $y_i \in \mathcal{Y} = \{0, 1\}$ |
|---|---|
| $\boldsymbol{x}_1$ | 1 |
| $\boldsymbol{x}_2$ | 0 |
| $\boldsymbol{x}_3$ | 0 |
| $\boldsymbol{x}_4$ | 1 |
| . | . |
| . | . |
| $\boldsymbol{x}_N$ | 1 |

Learn a classification function

$$f : \mathbf{R}^d \rightarrow \mathcal{Y}$$

# Golden ground truth

- How to obtain the labels for training?

$$y_i \in \mathcal{Y} = \{0, 1\}$$

- Getting the actual golden ground truth can be

- Expensive
- Potentially dangerous
- Could be impossible

# Subjective ground truth from multiple annotators

- Getting golden ground truth is hard, so we use opinion from an annotator
- An annotator provides his/her subjective version of the truth
- Error prone/noisy/unreliable
- Use multiple annotators who label the same example

# Annotations from multiple annotators

Each radiologist is asked to annotate whether a lesion is malignant (1) or not (0).

| Lesion ID | Radiologist 1 | Radiologist 2 | Radiologist 3 | Radiologist 4 | Truth Unknown |
|---|---|---|---|---|---|
| 01 | 0 | 0 | 0 | 0 | x |
| 02 | 0 | 1 | 0 | 0 | x |
| 03 | 1 | 1 | 1 | 1 | x |
| 04 | 0 | 0 | 1 | 1 | x |
| 05 | 0 | 1 | 1 | 1 | x |
| 06 | 0 | 0 | 1 | 0 | x |
| 07 | 0 | 1 | 1 | 0 | x |

In practice there is a substantial amount of disagreement.

We have no knowledge of the actual golden ground truth.

Getting absolute ground truth (e.g. biopsy) can be expensive.

# We are interested in

- Building a model to answer questions
  - How to train a classifier?
  - How to evaluate annotators?
  - How to estimate the actual ground truth?

# How to judge an annotator? (1)

Sensitivity

$$\alpha^j = \Pr[y^j = 1 \mid y = 1]$$

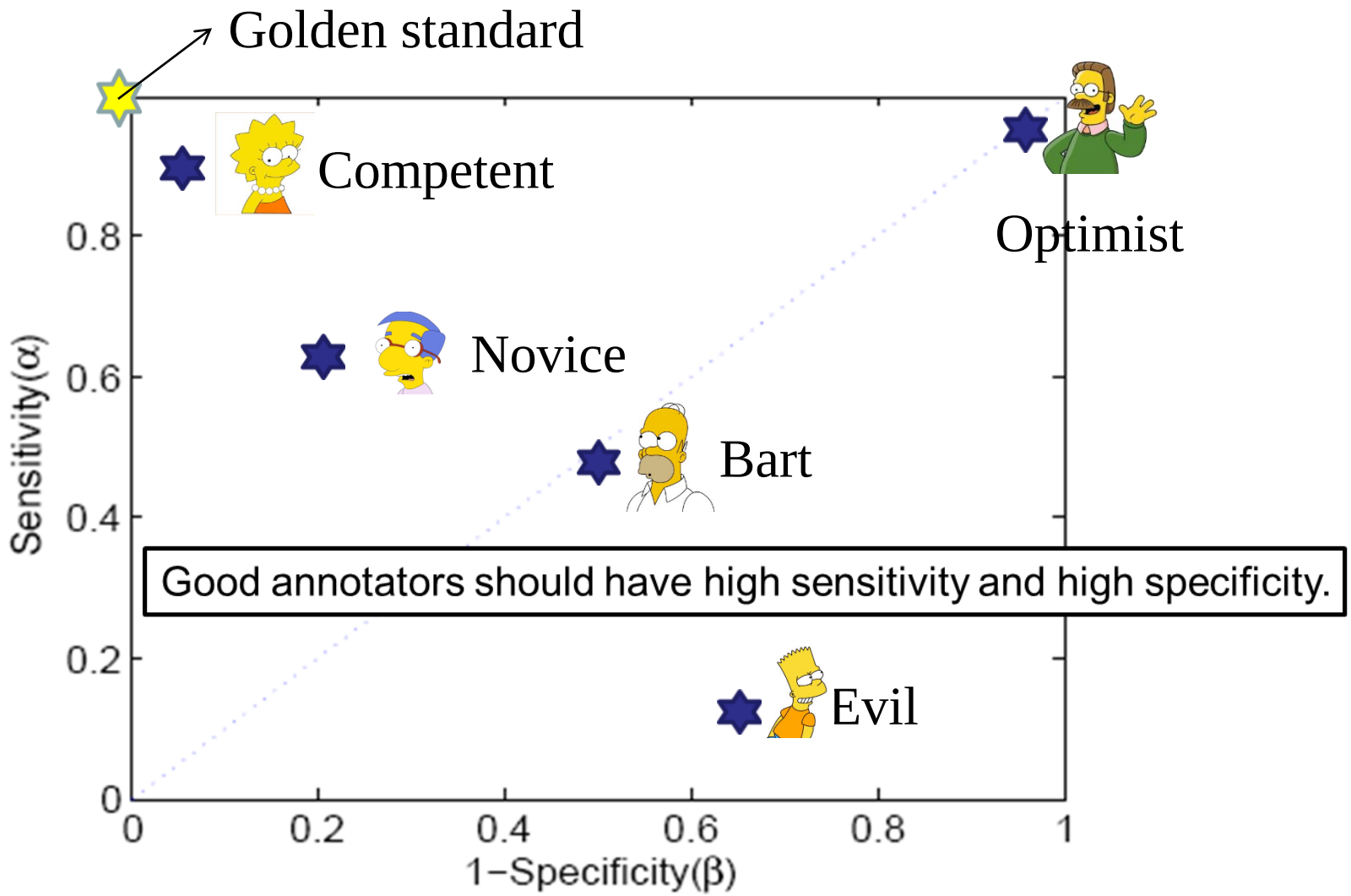
Label assigned by annotator j

True label

Specificity

$$\beta^j = \Pr[y^j = 0 \mid y = 0]$$

# How to judge an annotator? (2)



# Varying effectiveness on types of data

- In many cases annotator knowledge can fluctuate considerably depending on the groups of input instances
- Build data-dependent model based on the intuition that **inconsistent** annotators have different sensitivity and specificity for different regions of the feature space
- How to find the fittest model to approximate the distribution of the instances?

# How to approximate the distribution of the instances?

- **Gaussian mixture model (GMM)**: Linear superposition of Gaussians components
- Well-studied statistical inference techniques are available (EM algorithm)
- A "soft" group assignment is available. E-step evaluates the probability that an observation $x_i$ belongs to component k as $\tau_{ik}$
- Choose the model and the number of components by **Bayesian Information Criterion (BIC)**

# Problem Statement

- Input: Given N instances with annotations from R annotators

$$D = \{\mathbf{x}_i, y_i^{\ 1}, \ldots, y_i^{\ R}\}_{i=1}^{N}$$

- Output:
  - Sensitivities at each component
  - Specificities at each component
  - Estimates of true labels $y_1, \ldots, y_N$

# If we know the true labels

- We can learn a classifier
- To model the data-dependent behavior of annotators, we hypothesize that each annotator has its own sensitivity and specificity for each mixture component

$$\alpha_k^j = \Pr(y_i^j = 1 \mid y_i = 1, \text{ k-th Gaussian mixture component generates } \boldsymbol{x}_i)$$

$$\beta_k^j = \Pr(y_i^j = 0 \mid y_i = 0, \text{ k-th Gaussian mixture component generates } \boldsymbol{x}_i)$$

$$\alpha_k^j = \sum_{i=1}^{N} z_{ik} y_i^j \Big/ \sum_{i=1}^{N} z_{ik}$$

$$\beta_k^j = \sum_{i=1}^{N} (\tau_{ik} - z_{ik})(1 - y_i^j) \Big/ \sum_{i=1}^{N} (\tau_{ik} - z_{ik})$$

$Z_i$ is a soft label (probability that the label is 1) and $z_{ik} = z_i \tau_{ik}$

# How to find the unknown true labels (1)

- Hypothesize the behavior of annotator: Given an instance $x_i$ to label, the annotator finds the mixture component which most likely generates that instance. Then the annotators generate labels with their sensitivities and specificities at the most likely component

Again, prior probability by classifier

Again, Bayes Rule

$$z_i = \frac{\Pr[y_i^1, \ldots, y_i^R \mid y_i = 1, \phi] \cdot \Pr[y_i = 1 \mid \mathbf{x}_i, \phi]}{\Pr[y_i^1, \ldots, y_i^R \mid \phi]}$$

$$\Pr[y_i^1, \ldots, y_i^R \mid y_i = 1, \boldsymbol{\alpha}] = \Pr[y_i^1, \ldots, y_i^R \mid y_i = 1, \alpha_q^1, \ldots, \alpha_q^R]$$

where $q = \arg\max_{k=1,\ldots,K}(\tau_{ik})$

$$= \prod_{j=1}^{R} \Pr[y_i^j \mid y_i = 1, \alpha_q^j] = \prod_{j=1}^{R} [\alpha_q^j]^{y_i^j} [1 - \alpha_q^j]^{1 - y_i^j}$$

# How to find the unknown true labels (2)

- Therefore, if we know annotators' sensitivities and specificities at each component, the estimation of the hidden true label is:

$$z_i = \frac{a_i p_i}{a_i p_i + b_i (1 - p_i)}$$

where

$$p_i = \Pr[y_i = 1 \mid \boldsymbol{x}_i, \boldsymbol{w}] = \sigma(\boldsymbol{w}^T \boldsymbol{x}_i)$$

$$a_i = \prod_{j=1}^{R} [\alpha_q^j]^{y_i^j} [1 - \alpha_q^j]^{1 - y_i^j}$$

$$b_i = \prod_{j=1}^{R} [1 - \beta_q^j]^{y_i^j} [\beta_q^j]^{1 - y_i^j}$$

$$q = \arg\max_{k=1,\ldots,K} (\tau_{ik})$$

# GMM-MAPML Algorithm

Find the fittest model to approximate the distribution of the instances

GMM

If we know how good each predictor is, we can estimate the true label

$$z_i = \frac{\sigma(\boldsymbol{w}^T \boldsymbol{x}_i)\prod_{j=1}^{R}[\alpha_q^j]^{y_i^j}[1-\alpha_q^j]^{1-y_i^j}}{\sigma(\boldsymbol{w}^T \boldsymbol{x}_i)\prod_{j=1}^{R}[\alpha_q^j]^{y_i^j}[1-\alpha_q^j]^{1-y_i^j} + (1-\sigma(\boldsymbol{w}^T \boldsymbol{x}_i))\prod_{j=1}^{R}[\beta_q^j]^{1-y_i^j}[1-\beta_q^j]^{y_i^j}}$$

MAP

Iterate until convergence

Initialize using majority-voting

ML

If we know the true label we can estimate how good each predictor is at each component

$$\alpha_k^j = \sum_{i=1}^{N} z_{ik} y_i^j \Big/ \sum_{i=1}^{N} z_{ik} \qquad \beta_k^j = \sum_{i=1}^{N} (\tau_{ik} - z_{ik})(1-y_i^j) \Big/ \sum_{i=1}^{N} (\tau_{ik} - z_{ik})$$

Learn a classifier

# Analysis of the model

$$\mathrm{logit}(z_i) = \ln \frac{z_i}{1 - z_i} = \ln \frac{\Pr[y_i = 1 \mid y_i^1, \ldots, y_i^R, \boldsymbol{x}_i, \phi]}{\Pr[y_i = 0 \mid y_i^1, \ldots, y_i^R, \boldsymbol{x}_i, \phi]}$$

$$= w^T \boldsymbol{x}_i + \sum_{j=1}^{R} y_i^j \boxed{[\mathrm{logit}(\alpha_q^j) + \mathrm{logit}(\beta_q^j)]} + c$$

Observations

Annotators' labels

Constant

Consider both sensitivity and specificity as weight.

$q = \arg\max_{k=1,\ldots,K}(\tau_{ik})$ indicates data-dependent.

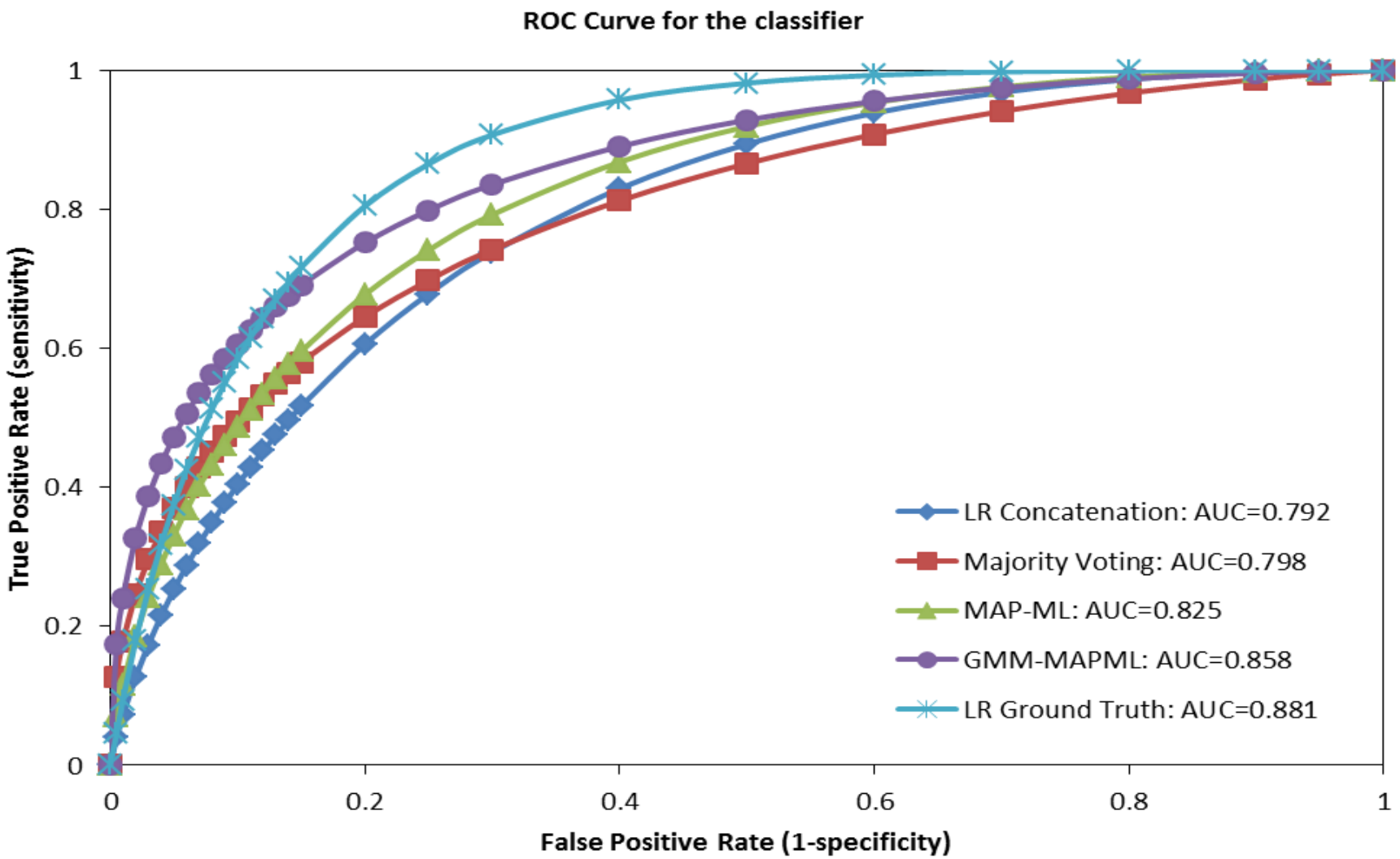
# Emotional speech classification

- Why to study emotional speech?
  - Recognition (e.g., Interface optimization in call centers)
  - Generation (e.g., TTS, games)
- Acted emotional utterance
  - Semantically neutral
  - Four acted emotions: happy , neutral , sad , angry

# Dataset: EMA database from University of Southern California

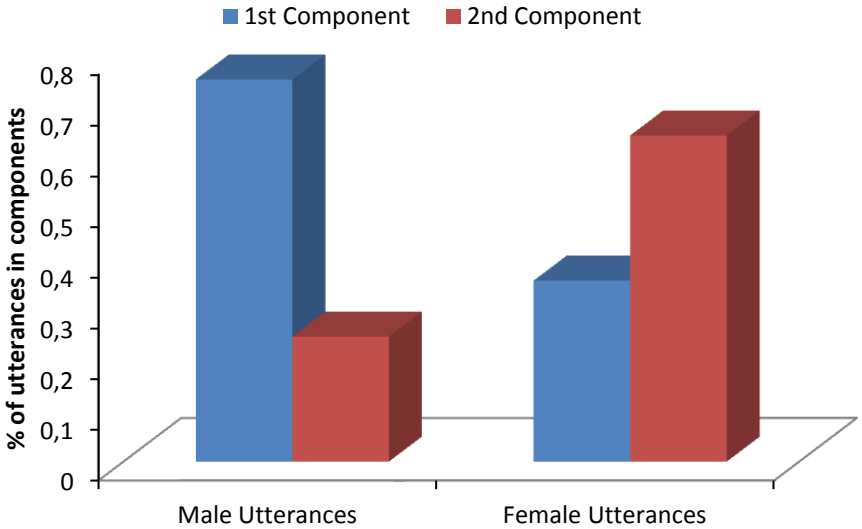
- Golden ground truth is known: 568 utterances were chosen as best emotional utterances
- 39-element feature vectors were extracted from the speech signal (WAV file) by using VOICEBOX
- Binary labels: {happy, neutral} were assigned to positive emotion (0), {sad, angry} were assigned to negative emotion (1)
- Multiple annotators: 5 annotators with different academic background. Most of them are non-native English speakers. Noisy/unreliable annotators

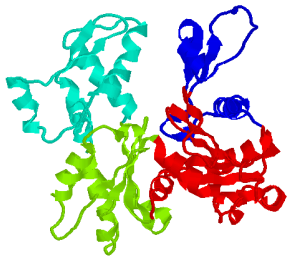
# Experiment Results: ROC comparisons



# Experiment Results: GMM-MAPML based estimates of annotators' accuracy

| | First Component | | Second Component | |
|---|---|---|---|---|
| Listeners | Estimated Sensitivity | Estimated Specificity | Estimated Sensitivity | Estimated Specificity |
| Listener 1 | 0.902 | 0.891 | 0.925 | 0.951 |
| Listener 2 | 0.843 | 0.862 | 0.814 | 0.799 |
| Listener 3 | 0.784 | 0.802 | 0.779 | 0.792 |
| Listener 4 | 0.756 | 0.744 | 0.877 | 0.861 |
| Listener 5 | 0.719 | 0.698 | 0.728 | 0.736 |





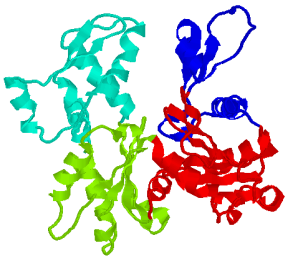
# Protein Disorder Prediction

- Lock and Key Paradigm:

  AA seq → 3-D Structure → Function
- Definition: A part of the protein or the whole protein doesn't have a fixed tertiary structure
- Importance: Involved in many important functions and in various diseases.

# CASP9 Disorder Dataset

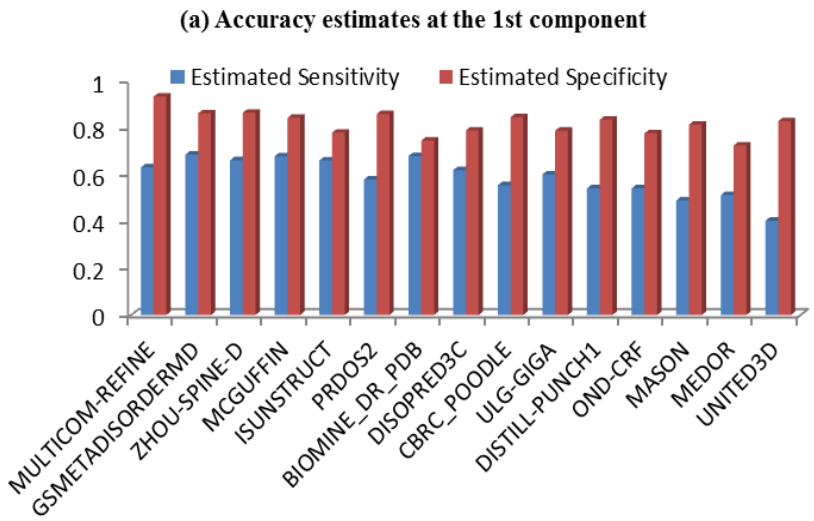
- 117 experimentally characterized targets (=26083 residues) were analyzed containing: 9.30% disordered residues and 90.70% ordered residues
- Golden ground truth is known: either X-ray or NMR experimental characterization
- 20-element feature vectors (19 amino acid composition features and 1 sequence complexity feature) were extracted from the protein sequences
- Multiple annotators: Labels by15 predictors developed **at different institutions**
- Disordered segments <4 residues were not considered

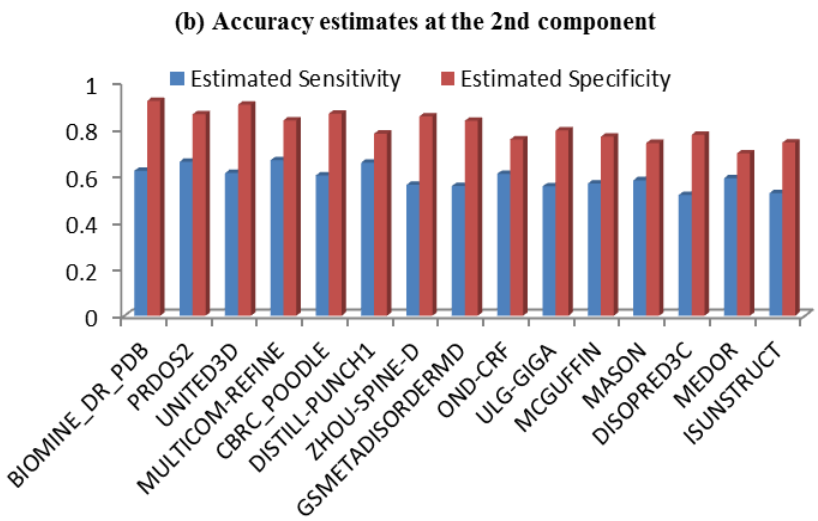


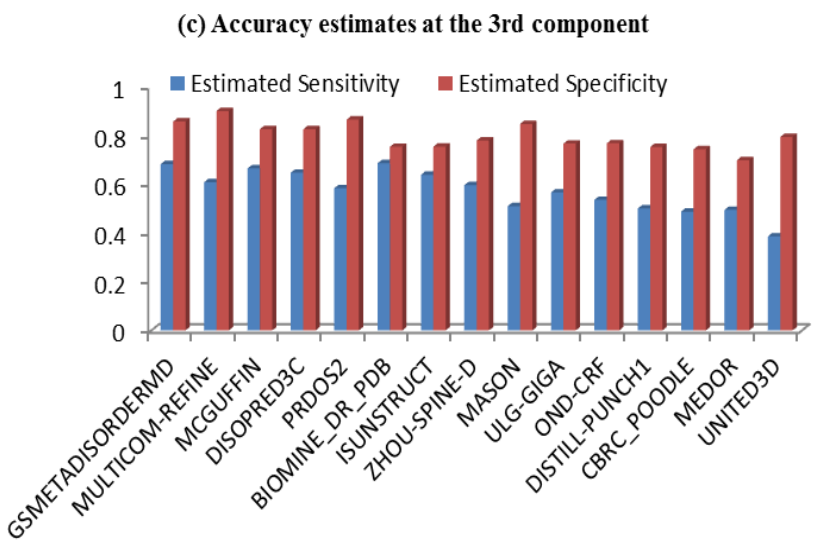
# CASP9 Assessment Scores

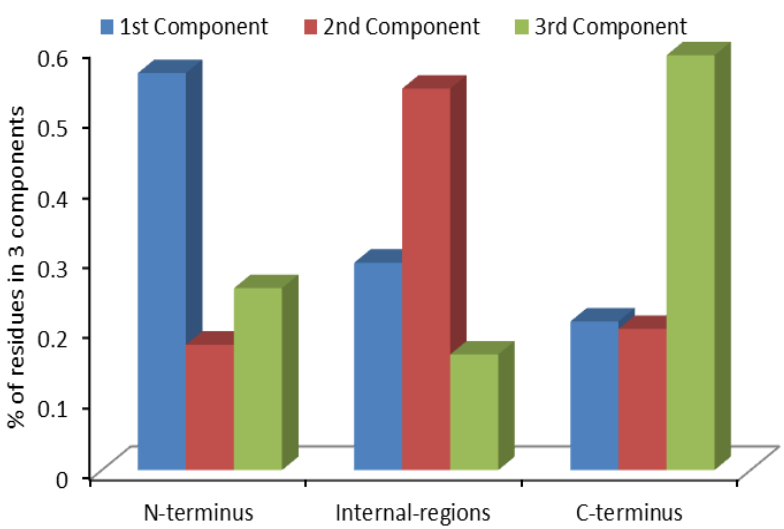
| Predictor Name | Institution | ACC | Sw | AUC |
|---|---|---|---|---|
| GMM-MAPML | | 0.785 | 0.527 | 0.874 |
| MAP-ML | | 0.764 | 0.513 | 0.859 |
| MAJORITY VOTING | | 0.735 | 0.496 | 0.776 |
| PRDOS2 | Tokyo Tech, Japan | 0.754 | 0.509 | 0.855 |
| MULTICOM-REFINE | University of Missouri, USA | 0.75 | 0.5 | 0.822 |
| BIOMINE_DR_PDB | University of Alberta, Canada | 0.741 | 0.483 | 0.821 |
| GSMETADISORDERMD | IIMCB in Warsaw, Poland | 0.738 | 0.476 | 0.816 |
| MASON | George Mason University, USA | 0.736 | 0.473 | 0.743 |
| ZHOU-SPINE-D | IU School of Medicine, USA | 0.731 | 0.462 | 0.832 |
| DISTILL-PUNCH1 | UCD Dublin, Ireland | 0.726 | 0.453 | 0.8 |
| OND-CRF | Umea University, Sweden | 0.706 | 0.412 | 0.737 |
| UNITED3D | Kitasato University, Japan | 0.704 | 0.412 | 0.781 |
| CBRC_POODLE | CBRC, Japan | 0.694 | 0.405 | 0.83 |
| MCGUFFIN | University of Reading, UK | 0.688 | 0.402 | 0.817 |
| ISUNSTRUCT | IPR RAS, Russia | 0.679 | 0.396 | 0.742 |
| DISOPRED3C | University College London, UK | 0.67 | 0.391 | 0.853 |
| ULG-GIGA | University of Liege, France | 0.585 | 0.341 | 0.726 |
| MEDOR | Aix-Marseille University, France | 0.579 | 0.338 | 0.688 |

# GMM-MAPML based estimates of CASP9 disorder predictors' accuracy



# Thank you! | Questions?

Ping Zhang: ping@temple.edu
Zoran Obradovic: zoran.obradovic@temple.edu