Generalized Agreement Statistics over Fixed Set of Experts

Mohak Shah

Accenture Technology Labs

European Conference on Machine Learning

Sep 07, 2011

Background and Settings

- Each instance in the data labeled by a fixed group of Raters
  - Expert Annotators, Opinion/Rating generators,…
- Multiple Classes (Nominal Scale)
- No ground truth labels

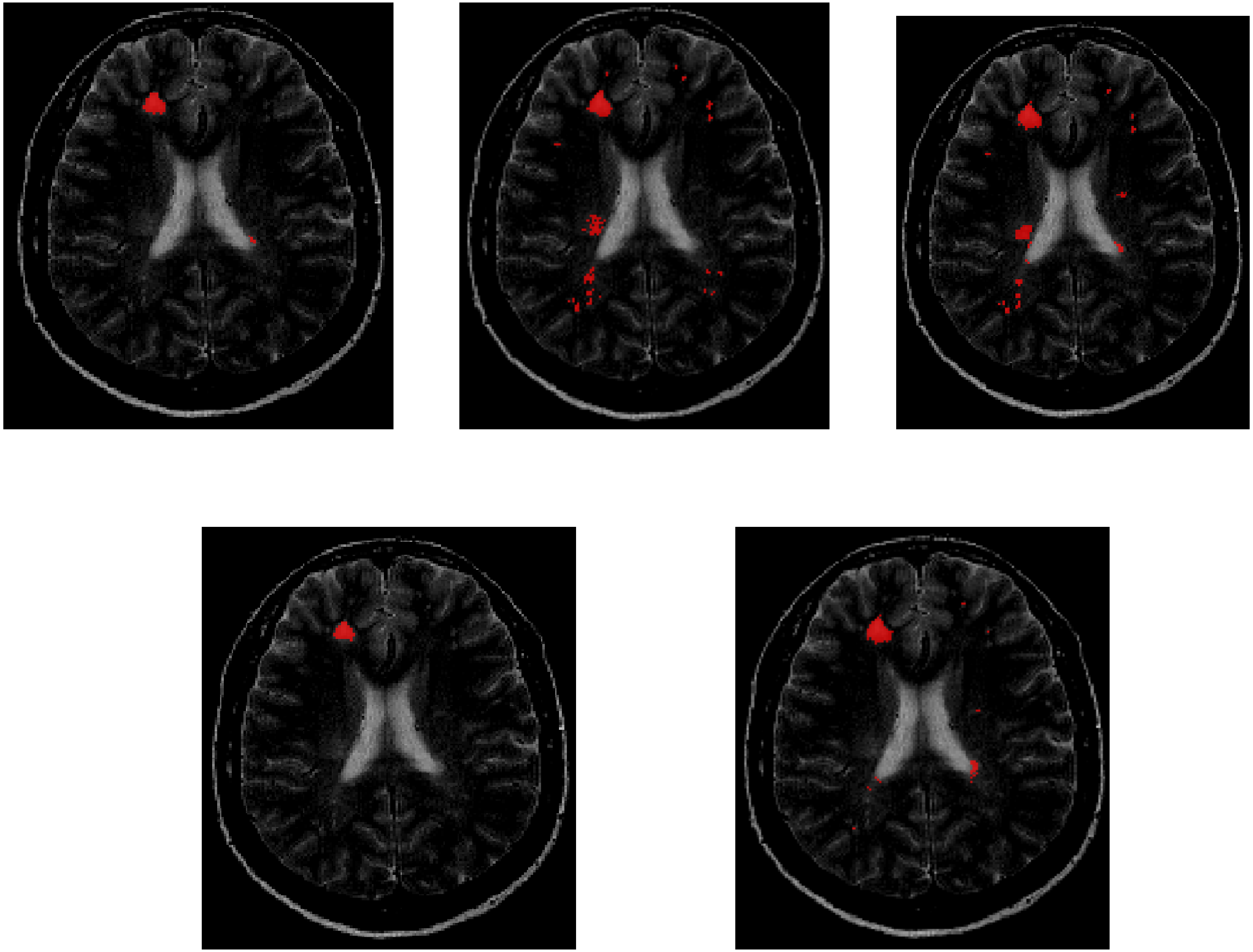
Many such scenarios

- Multiple experts' labels on multi-category examples
  - e.g., Human Intelligence Tasks (HITs)

- Medical Image Segmentation
  - e.g., Segmentation of lesion/tumor tissues from brain MRIs

- Applying ensemble methods for various tasks
  - e.g., multi-sensor radar systems for threat detection

An Example



Another Example

5 Raters suggesting positions on stocks in portfolio

Inst #	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
1	Buy	Sell	Buy	Sell	Hold
2	Buy	Metaphysical/ Epistemological	Sell	Sell	Sell
3	Buy	Buy	Buy	Buy	Buy
4	Sell	Buy	Sell	Buy	Sell
5	Hold	Hold	Buy	Hold	Sell
6	Metaphysical/ Epistemological	Metaphysical/ Epistemological	Metaphysical/ Epistemological	Sell	Sell
7	Sell	Hold	Hold	Sell	Sell

An example

Inst #	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
1	1	2	1	2	3
2	1	4	2	2	2
3	1	1	1	1	1
4	2	1	2	1	2
5	3	3	1	3	2
6	4	4	4	2	2
7	2	3	3	2	2

Two Problems

Inst #	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
1	1	2	1	2	3
2	1	4	2	2	2
3	1	1	1	1	1
4	2	1	2	1	2
5	3	3	1	3	2
6	4	4	4	2	2
7	2	3	3	2	2

Inter-expert agreement:
Overall Agreement of the group

Two Problems

Inst #	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
1	1	2	1	2	3
2	1	4	2	2	2
3	1	1	1	1	1
4	2	1	2	1	2
5	3	3	1	3	2
6	4	4	4	2	2
7	2	3	3	2	2

Inter-expert agreement: Overall Agreement of the group

Inst #	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
1	1	2	1	2	3
2	1	4	2	2	2
3	1	1	1	1	1
4	2	1	2	1	2
5	3	3	1	3	2
6	4	4	4	2	2
7	2	3	3	2	2

Classifier
1
4
1
2
3
2
2

Classifier Agreement Against the group

General Agreement Statistic

$$\kappa = \frac{\mathbf{E}_S(\mathbb{A}) - \mathbf{E}(\mathbb{A})}{\max_S(\mathbb{A}) - \mathbf{E}(\mathbb{A})}$$

General Agreement Statistic

Agreement measure

$$\kappa = \frac{\mathbf{E}_S(\mathbb{A}) - \mathbf{E}(\mathbb{A})}{\boxed{\max_S(\mathbb{A})} - \mathbf{E}(\mathbb{A})}$$

Maximum Achievable Agreement

Chance Agreement

Examples: Cohen's kappa, Fleiss Kappa, Scott's pi, ICC…

Modus operandi

- Define an agreement measure
- Derive expression for its expected value
- Define maximum achievable agreement
- Live happily ever after

Except…

this is easier said than done

Model assumptions play a big role

Problem

- To obtain general agreement measures over a **fixed set of raters** applicable in multi-class multi-rater case, accounting for coincidental concordances

- **Traditional approaches**
  - Typically applicable for 2-rater binary classification case (E.g., Cohen's kappa)
  - Generalizations assume a variable group and use marginalization argument (e.g, Fleiss kappa (Fleiss, 1971) statistic implemented in WEKA)

- **Claim:** Marginalization argument is unsuitable for the fixed experts' group case

Back to the agreement statistic

$$\kappa = \frac{\mathbf{E}_S(\mathbb{A}) - \mathbf{E}(\mathbb{A})}{\max_S(\mathbb{A}) - \mathbf{E}(\mathbb{A})}$$

Pair-wise Agreement measure

Difference arise in obtaining the expectation

Traditional Approaches: Inter-expert Agreement

The Marginalization Argument:
Consider a simple 2 rater 2 class case

Inst #	Rater 1	Rater 2
1	1	1
2	1	2
3	1	1
4	2	1
5	2	2
6	1	2
7	2	2

Agreement: 4/7

Probability of chance agreement over label 1:

$$\Pr(\text{Label}=1 | \text{ Random rater})^2 = 7/14 * 7/14 = 0.25$$

$$\text{Agreement} = \frac{4/7 - 0.5}{1 - 0.5} = 0.143$$

Traditional Approaches: Inter-expert Agreement

The Marginalization Argument:
Consider another scenario

Inst #	Rater 1	Rater 2
1	1	2
2	1	2
3	1	2
4	1	2
5	2	1
6	2	1
7	2	1

Observed Agreement: 0

Probability of chance agreement over label-1:

$$\Pr(\text{Label}=1 \mid \text{Random rater})^2 = 7/14 * 7/14 = 0.25$$

$$\text{Agreement} = \frac{0 - 0.5}{1 - 0.5} = -1$$

Traditional Approaches: Inter-expert Agreement

The Marginalization Argument:

But this holds

even when there is no evidence of a chance agreement

Inst #	Rater 1	Rater 2
1	1	2
2	1	2
3	1	2
4	1	2
5	1	2
6	1	2
7	1	2

Observed Agreement: 0

Probability of chance agreement over label-1:

$$\Pr(\text{Label}=1 \mid \text{Random rater})^2 = 7/14 * 7/14 = 0.25$$

$$\text{Agreement} = \frac{0 - 0.5}{1 - 0.5} = -1$$

Not applicable in fixed rater scenario

- Marginalization ignores rater correlation
- Ignores rater asymmetry
- Results in loose chance agreement estimates by optimistic estimation
- Hence, overly conservative agreement estimate

I: Inter-rater agreement over fixed set of raters

$$\kappa = \frac{\mathbf{E}_S(\mathbb{A}) - \mathbf{E}(\mathbb{A})}{\max_S(\mathbb{A}) - \mathbf{E}(\mathbb{A})}$$

Approach

$$A_o(\mathbf{i}_i) = \frac{1}{r(r-1)} \sum_{j=1}^{k} c_{ij}(c_{ij}-1)$$

Subscripts

i: data points
j: classes
p:raters

Pair-wise Agreement measure

$$\kappa = \frac{\mathbf{E}_S(\mathbb{A}) - \mathbf{E}(\mathbb{A})}{\max_S(\mathbb{A}) - \mathbf{E}(\mathbb{A})}$$

c_{ij}: No. of raters assigning point i to class j

Approach

$$A_o(\mathbf{i}_i) = \frac{1}{r(r-1)} \sum_{j=1}^{k} c_{ij}(c_{ij}-1)$$

Subscripts

i: data points
j: classes
p:raters

Pair-wise Agreement measure

$$\kappa = \frac{\mathbf{E}_S(\mathbb{A}) - \mathbf{E}(\mathbb{A})}{\max_S(\mathbb{A}) - \mathbf{E}(\mathbb{A})}$$

Difference arise in obtaining the expectation

$$A_e(l^j) = \frac{1}{r(r-1)} \sum_{p \in \mathcal{R}} \sum_{p' \in \mathcal{R}, p' \neq p} \left[\frac{c_{pj}}{n} \frac{c_{p'j}}{n} \right]$$

c_{ij}: No. of raters assigning point i to class j
c_{pj}: No. of **data points** that rater p assigns to class j

Approach

$$\kappa = \frac{\mathbf{E}_S(\mathbb{A}) - \mathbf{E}(\mathbb{A})}{\max_S(\mathbb{A}) - \mathbf{E}(\mathbb{A})}$$

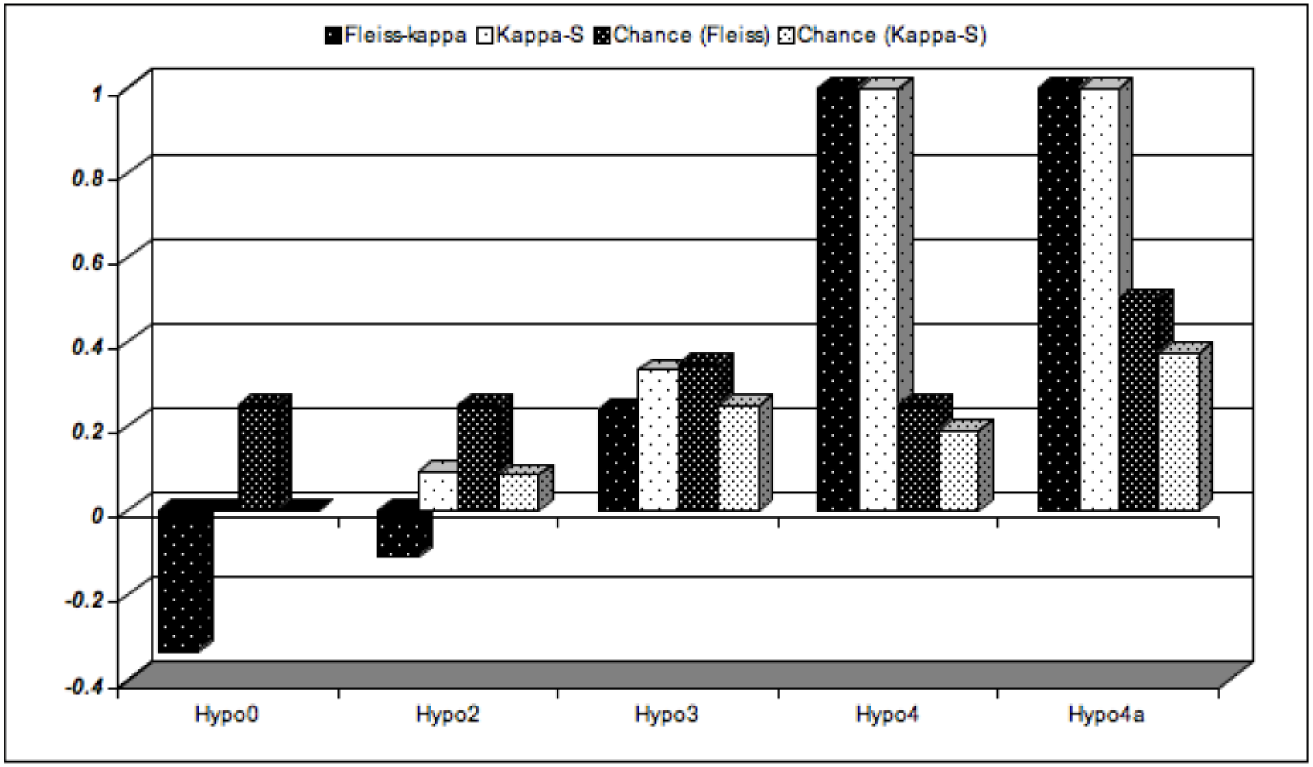
Maximum agreement possible is 1

Inter-rater Agreement: Fixed rater scenario

- Inter-rater agreement is:

$$\boxed{\kappa_S = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k} c_{ij}\cdot(c_{ij}-1) - \frac{1}{n}\sum_{j=1}^{k}\sum_{p\in\mathcal{R}}\sum_{p'\in\mathcal{R},p'\neq p}[c_{pj}c_{p'j}]}{nr(r-1)\left[1-\frac{1}{n^2r(r-1)}\sum_{j=1}^{k}\sum_{p\in\mathcal{R}}\sum_{p'\in\mathcal{R},p'\neq p}[c_{pj}c_{p'j}]\right]}}$$

Simulations on synthetic data

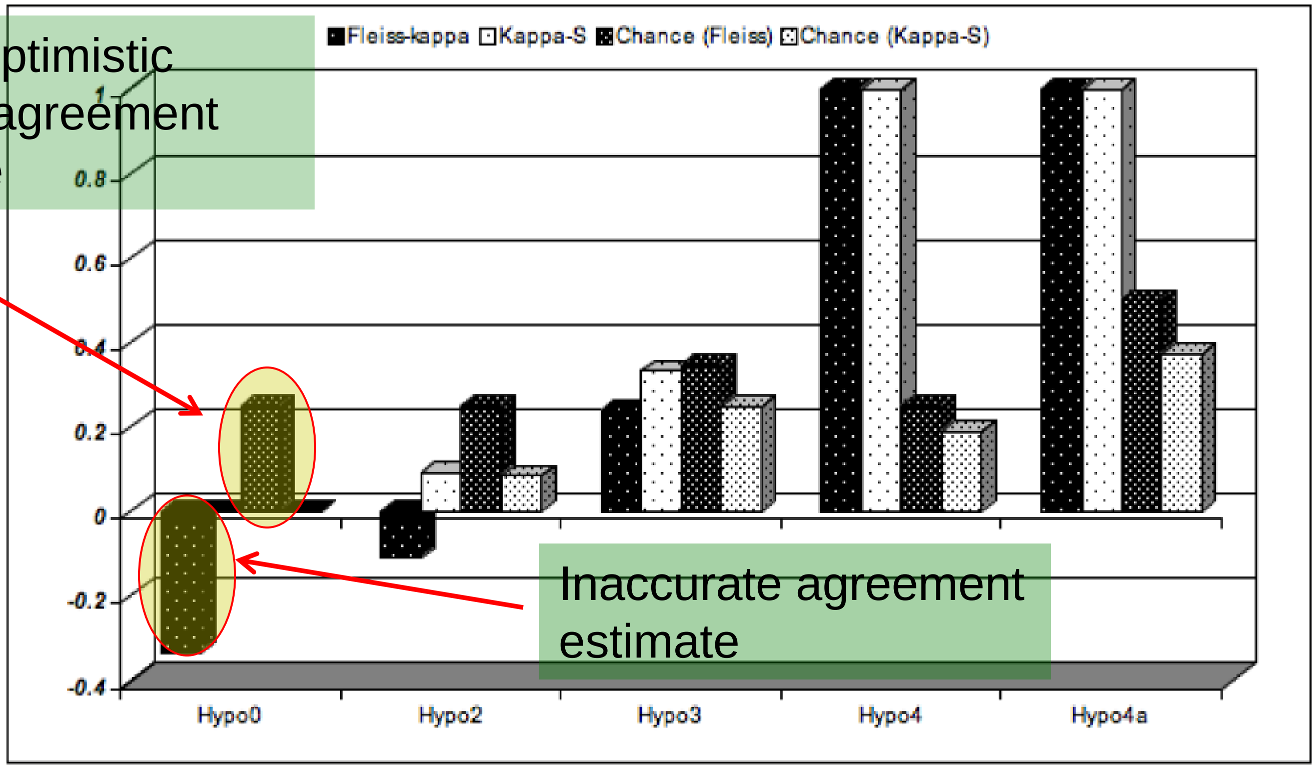


Setting: 200 data points, 4 raters, 4 classes
Hypo0: All raters disagree on all points

R1	R2	R3	R4
1	2	3	4
1	2	3	4
1	2	3	4
1	2	3	4
1	2	3	4
1	2	3	4

Simulations on synthetic data

Overly optimistic chance agreement estimate



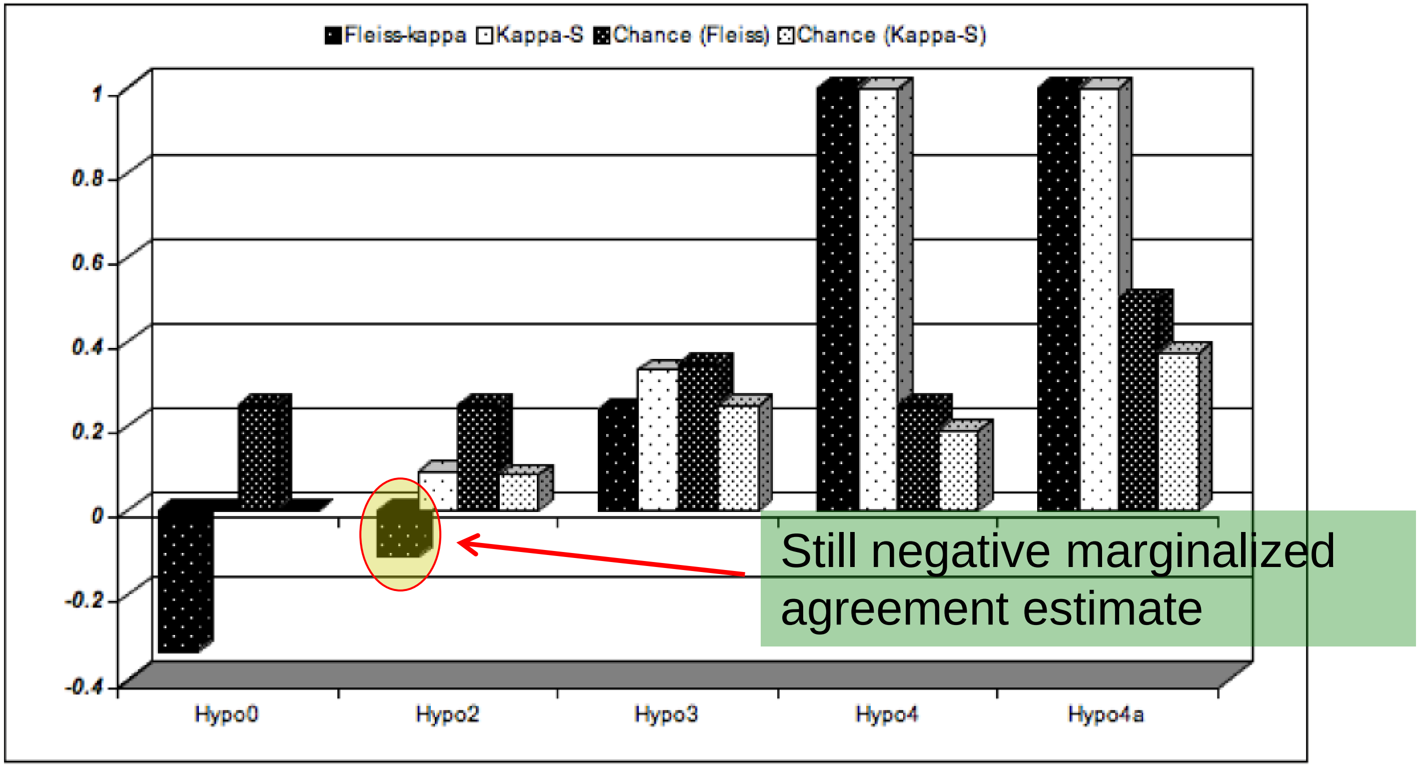
Inaccurate agreement estimate

Setting: 200 data points, 4 raters, 4 classes

Hypo0: All raters disagree on all points

R1	R2	R3	R4
1	2	3	4
1	2	3	4
1	2	3	4
1	2	3	4
1	2	3	4
1	2	3	4

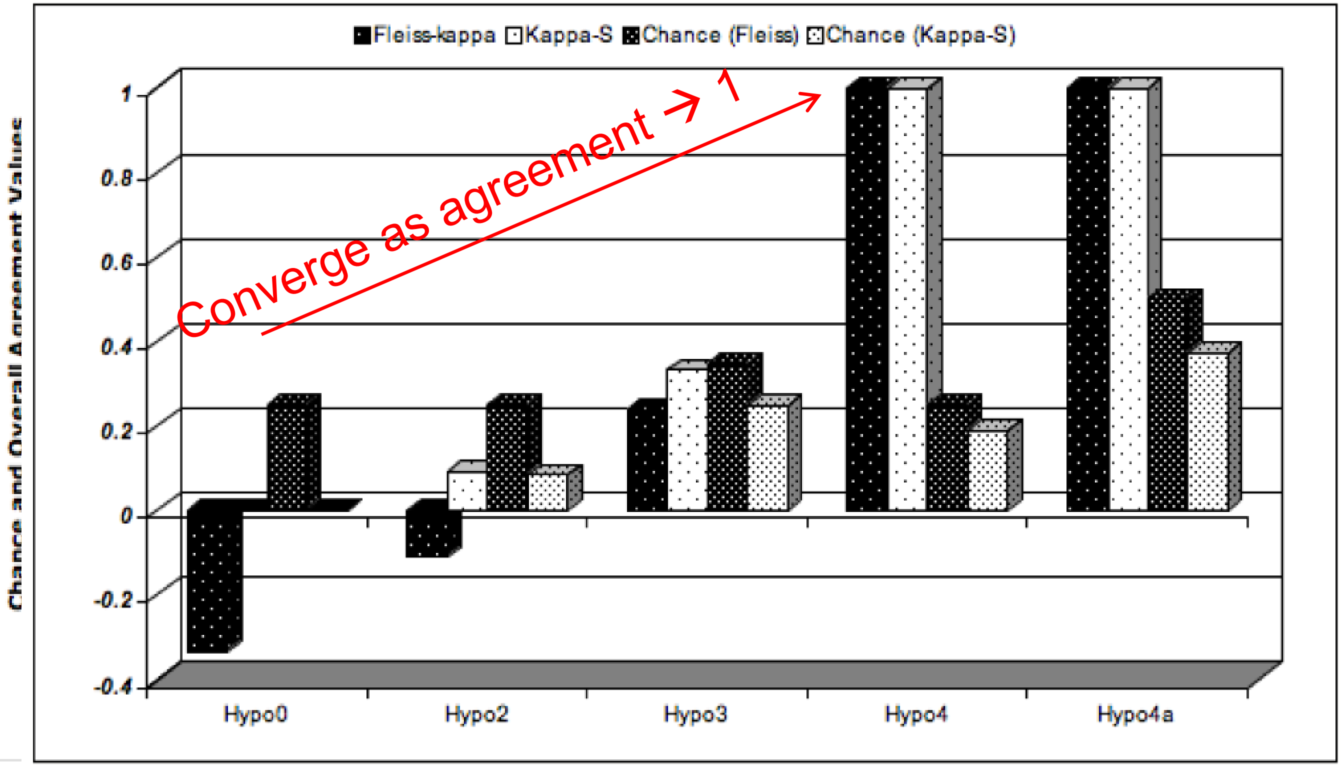
Simulations on synthetic data



Setting: 200 data points, 4 raters, 4 classes
Hypo0: All raters disagree in all points
Hypo2: 2 raters agree on all the labels

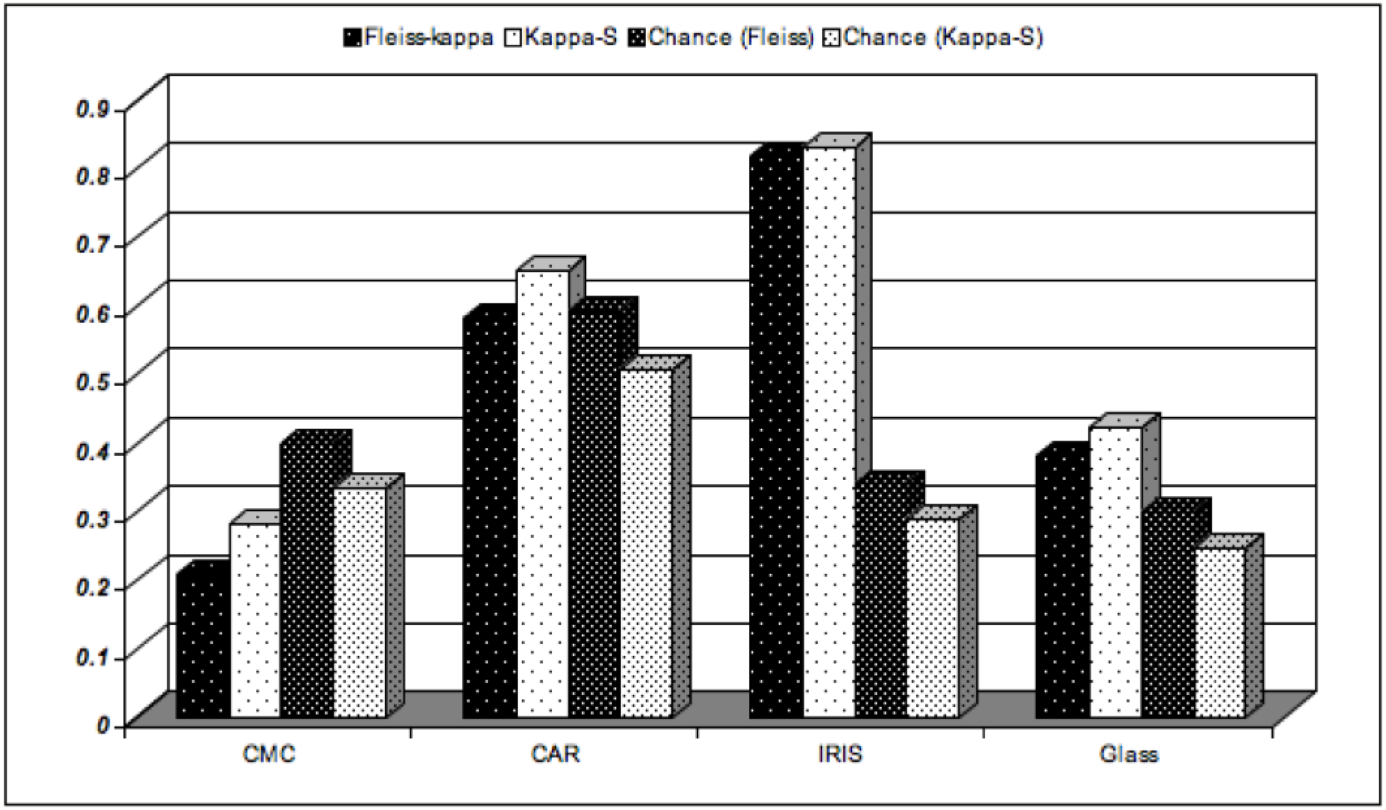
R1	R2	R3	R4
2	2	3	4
2	2	3	4
2	2	3	4
1	1	3	4
1	1	3	4
1	1	3	4

Simulations on synthetic data



Setting: 200 data points, 4 raters, 4 classes
Hypo0: All raters disagree in all points
Hypo2: 2 raters agree on all the labels
Hypo 3: 3 raters agree
Hypo4: All raters agree (50 points in each class)
Hypo4a: All raters agree (100 points each in 2 classes)

Simulations on UCI data



Setting: 7 raters (6 classifiers + 1 true label), multiple classes

Both measures converges near unity
but differs substantially on low or moderate agreement values

Inter-rater Agreement Conclusions: An upper bound on the variability

Theorem 1. *Let κ_F and κ_S denote, respectively, the agreement statistics of Fleiss (1971) and that proposed in Equation 5 computed on a population (dataset) with large sample-size n where each of the sample has been assigned one of k labels by a fixed group of r experts. If $\sigma^2(\kappa)$ denotes the variance of κ then we have that:*

$$\sigma^2(\kappa_S) \leq \sigma^2(\kappa_F)$$

with equality satisfied when the experts emulate the pool.

II. Agreement of a classifier against a group: Two Traditional Approaches

- **Extension of marginalization argument**
  - Recently appeared in Statistics literature: Vanbelle and Albert, (*stat. ner.* 2009)

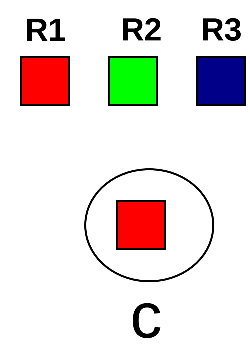
- **Consensus Based (more traditional)**
  - Almost universally used in the machine learning/data mining community
    - E.g., medical image segmentation, tissue classification, recommendation systems, expert modeling scenarios (e.g. market analyst combination)

Marginalization Approach and Issues in Fixed experts setting

- Observed agreement: Proportion of raters with which the classifier agrees
  - Ignores qualitative agreement, may even ignore group dynamics

Inst #	Rater 1	Rater 2	Rater 3
1	1	2	3
2	1	2	3
3	1	2	3
4	1	2	3
5	1	2	3
6	1	2	3
7	1	2	3

Classifier
1
3
1
2
3
2
2



Marginalization Approach and Issues in Fixed experts setting

- Observed agreement: Proportion of raters with which the classifier agrees
  - Ignores qualitative agreement, may even ignore group dynamics

Inst #	Rater 1	Rater 2	Rater 3
1	1	2	3
2	1	2	3
3	1	2	3
4	1	2	3
5	1	2	3
6	1	2	3
7	1	2	3

Classifier
1
3
1
2
3
2
2

Any random label assignment gives the same observed agreement

Marginalization Approach and Issues in Fixed experts setting

- Chance Agreement: Extend the marginalized argument
  - Not informative when the raters are fixed, ignores rater-specific correlations

Inst #	Rater 1	Rater 2	Rater 3
1	1	2	3
2	1	2	3
3	1	2	3
4	1	2	3
5	1	2	3
6	1	2	3
7	1	2	3

Classifier
1
3
1
2
3
2
2

When fixed raters never agree, chance agreement should be zero

Consensus Approach and Issues

- **Approach:**
  - Obtain a deterministic label for each instance if at least k >= r/2 raters agree
  - Treat this label set as ground truth and use dice coefficient against classifier labels
- Issues:
  - Threshold sensitive
  - Establishing threshold can be non-trivial
  - Tie breaking not clear
  - Treats estimates as deterministic
  - Ignores minority raters as well as rater correlation

Consensus approach fails in assessing classifier performance

- Dice in addition to consensus
  - No chance correction
  - Ignores agreement with minority raters
  - Dependent on consensus (and not raters' estimates)
  - Applies to two class scenario
  - Can be less sensitive, potentially even misleading, to important label changes

II. Agreement against Fixed Experts' group, derived from κ_S

$$\kappa = \frac{\mathbf{E}_S(\mathbb{A}) - \mathbf{E}(\mathbb{A})}{\max_S(\mathbb{A}) - \mathbf{E}(\mathbb{A})}$$

II. Agreement against Fixed Experts' group, derived from κ_S

No. of expert pairs agreeing with classifier assignment

$$\kappa = \frac{\mathbf{E}_S(\mathbb{A}) - \mathbf{E}(\mathbb{A})}{\max_S(\mathbb{A}) - \mathbf{E}(\mathbb{A})}$$

II. Agreement against Fixed Experts' group, derived from κ_S

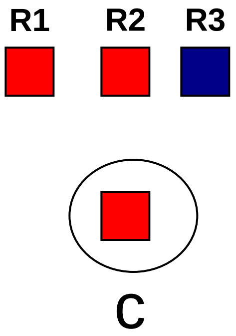
Pair-wise Agreement measure

$$\kappa = \frac{\mathbf{E}_S(\mathbb{A}) - \mathbf{E}(\mathbb{A})}{\max_S(\mathbb{A}) - \mathbf{E}(\mathbb{A})}$$

Expectation over pair-wise expert agreement and classifier assignment over all classes

II. Agreement against Fixed Experts' group, derived from κ_S

$$\kappa = \frac{\mathbf{E}_S(\mathbb{A}) - \mathbf{E}(\mathbb{A})}{\max_S(\mathbb{A}) - \mathbf{E}(\mathbb{A})}$$



Not necessarily 1, but upper bounded by the number of expert pairs agreeing

Agreement against Fixed Experts' group: The $\mathcal{S}$ measure

$$\mathcal{S} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{k}\mathfrak{t}_{ij}A_o(\mathbf{i}_i, l^j)\right] - \sum_{j=1}^{k}\mathfrak{t}_j \cdot A_e(l^j)}{\frac{1}{n}\sum_{i=1}^{n}\max_j A_o(\mathbf{i}_i, l^j) - \sum_{j=1}^{k}\mathfrak{t}_j \cdot A_e(l^j)}$$

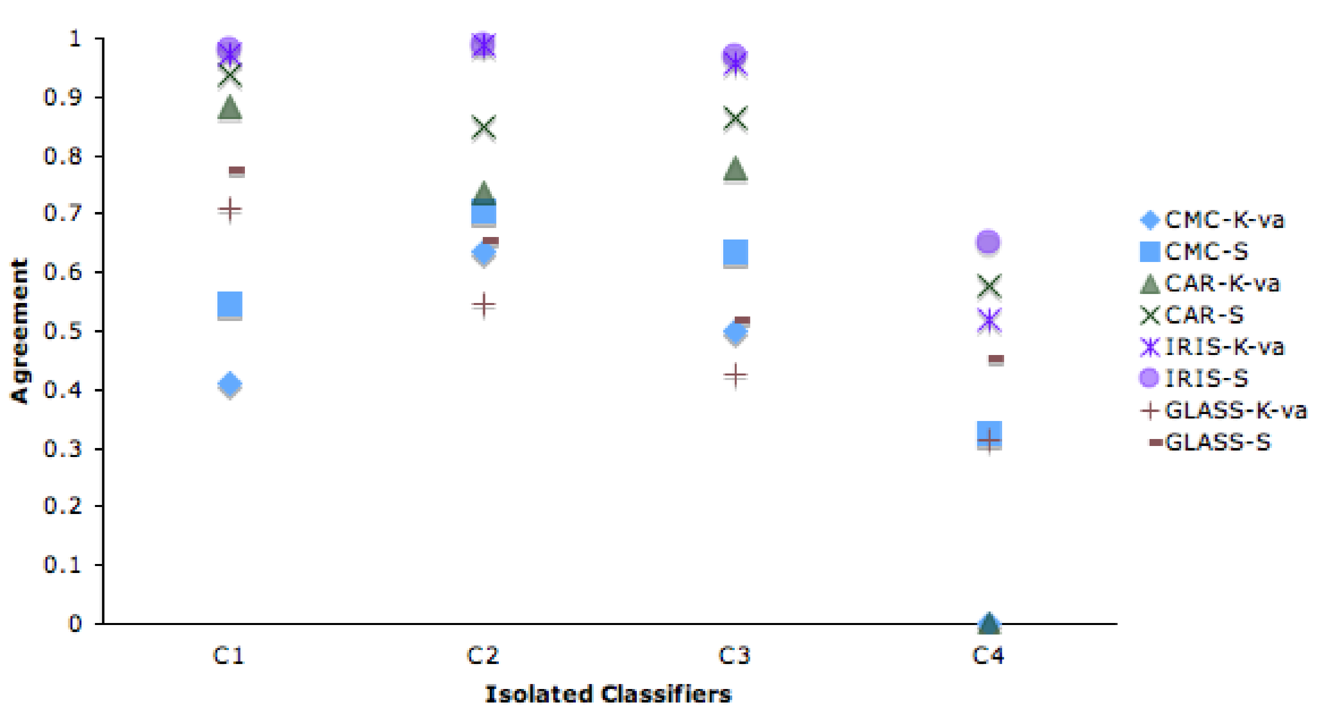
- $\mathfrak{t}$ denotes an output of learning algorithm

such that

$\mathfrak{t}_{ij}$ = 1 if the classifier assigns label j to instance i

$\mathfrak{t}_{ij}$ = 0 otherwise

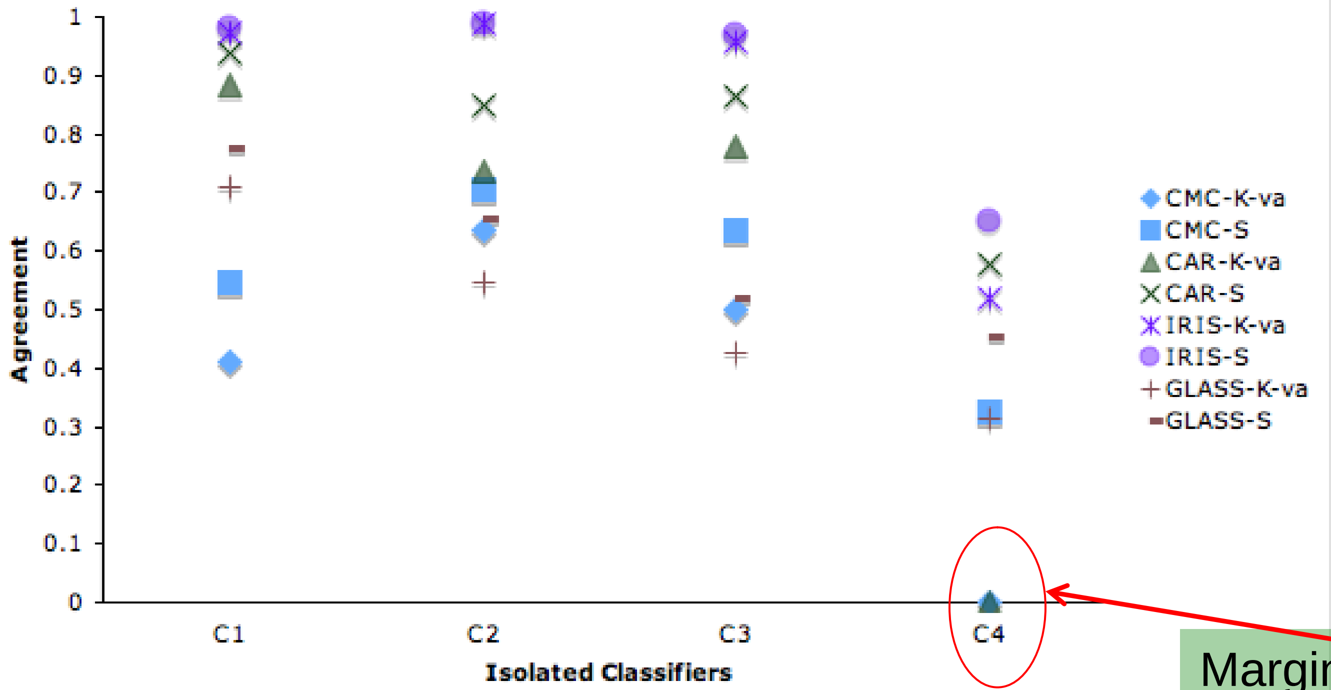
Agreement against silver standard: Illustration on UCI data



Setting:

Expert labels: True labels + 2 classifiers with highest 10-fold cv accuracy

Agreement against silver standard: Illustration on UCI data



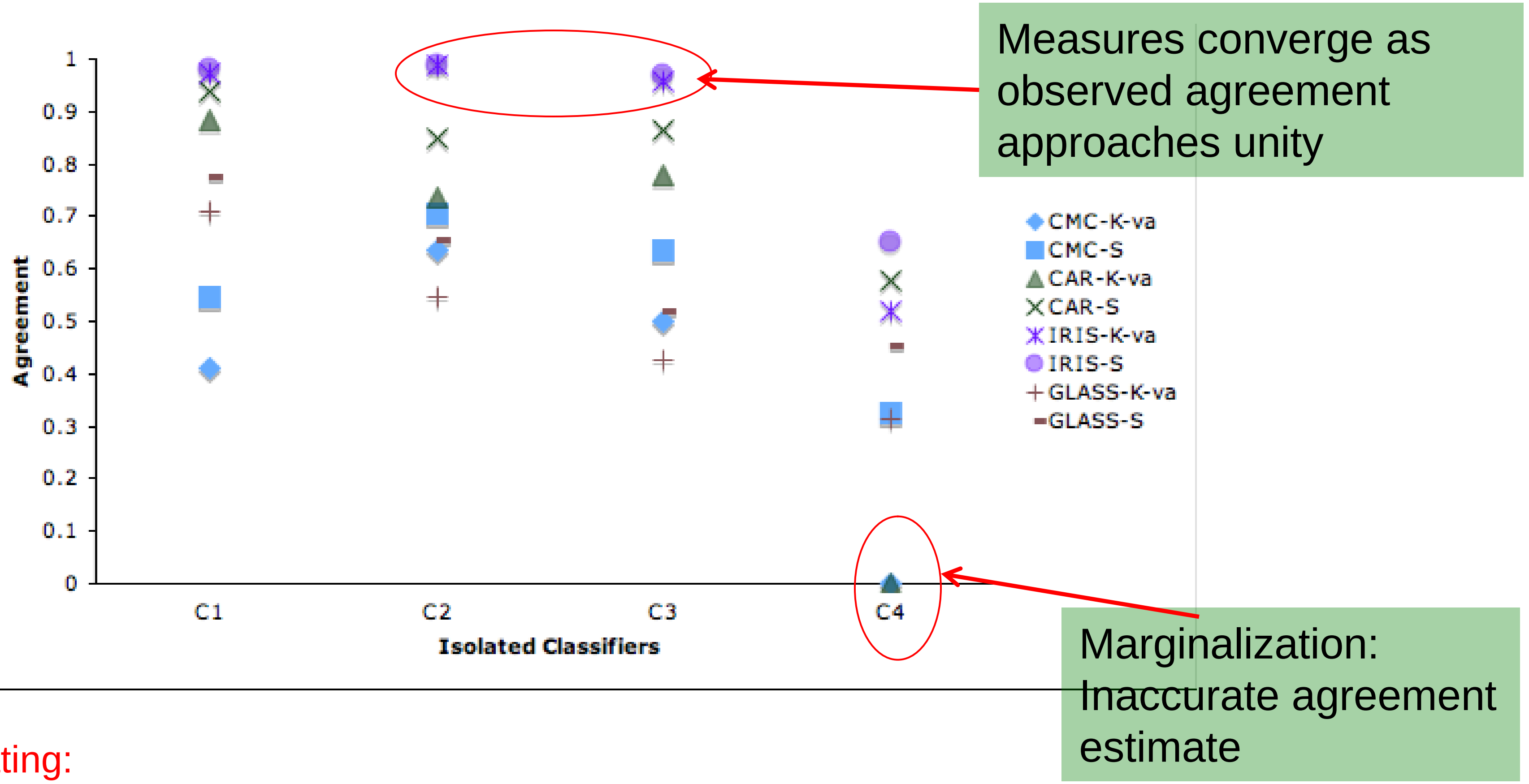
Marginalization: Inaccurate agreement estimate

Setting:

Expert labels: True labels + 2 classifiers with highest 10-fold cv accuracy

Note C4 over CAR and CMC (K-va=0)

Agreement against silver standard: Illustration on UCI data



Setting:
Expert labels: True labels + 2 classifiers with highest 10-fold cv accuracy
Note C4 over CAR and CMC (K-va=0)
Measures converge close to unity

Conclusion

- We show that the marginalization argument is unsuitable when the experts' group is fixed
- We propose generalized metrics that
  - Apply to multi-class multi-rater scenario
  - Sensitive to changing rater agreement
  - Provide more meaningful estimates
- Variance behavior can be analytically established unlike dice/consensus
- Statistical hypothesis tests can be obtained

Importance of time travel

If you'd like to discuss details or know of more results and issues, please come to my poster **yesterday!**

Thank You

Now Available: