Information Theoretic Evaluation of Data Processing Systems

Dr Michael B. Hurley

NIST Data Science Symposium

4-5 March 2014

LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

This work is sponsored by the Assistant Secretary of Defense for Research & Engineering under Air Force Contract #FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the United States Government.

Outline

- Performance metrics overview
- Information theoretic performance evaluations
  - Multi-target tracker evaluation
  - Classifier evaluation
- Error estimation and significance testing
- Current research focus
- Challenges
- Summary

Common Performance Measures

- **Physical measures**
  - **"Size, Weight, and Power (SWaP)"**
  - **Time / latency**
  - **Compute measures**
    - **Memory and data storage (size or space)**
    - **CPU (power)**
    - **Data rates / throughput**
    - **Scalability / extensibility**
    - **Algorithmic efficiency / computational complexity**
  - **…**

- **Utility measures**
  - **Cost / Benefit**
  - **Risk**
  - **Return on investment**
  - **…**

Information Measures

- Information theoretic measures are appropriate when data are used for critical decision making
- Quantitatively measure the uncertainty in estimates and decisions from automated decision systems (trackers, classifiers, etc.)
  - Entropy (uncertainty)
  - Mutual information (information common to a pair of data sets)
  - Conditional entropy (information unique to one of a pair of data sets)
- If truth is available, algorithms can be assessed on how well they extract information from data
  - Relative evaluations are the most meaningful
  - Can determine statistical significance of assessments with error analysis
- These measures can assess the impact of incommensurate physical measures on information content

Outline

- Performance metrics overview
- Information theoretic performance evaluations
  – Multi-target tracker evaluation
  – Classifier evaluation
- Error estimation and significance testing
- Current research focus
- Challenges
- Summary

Evaluation of Tracking Algorithms

Existing tracker metrics partially measure performance with correlations between pairs of metrics.

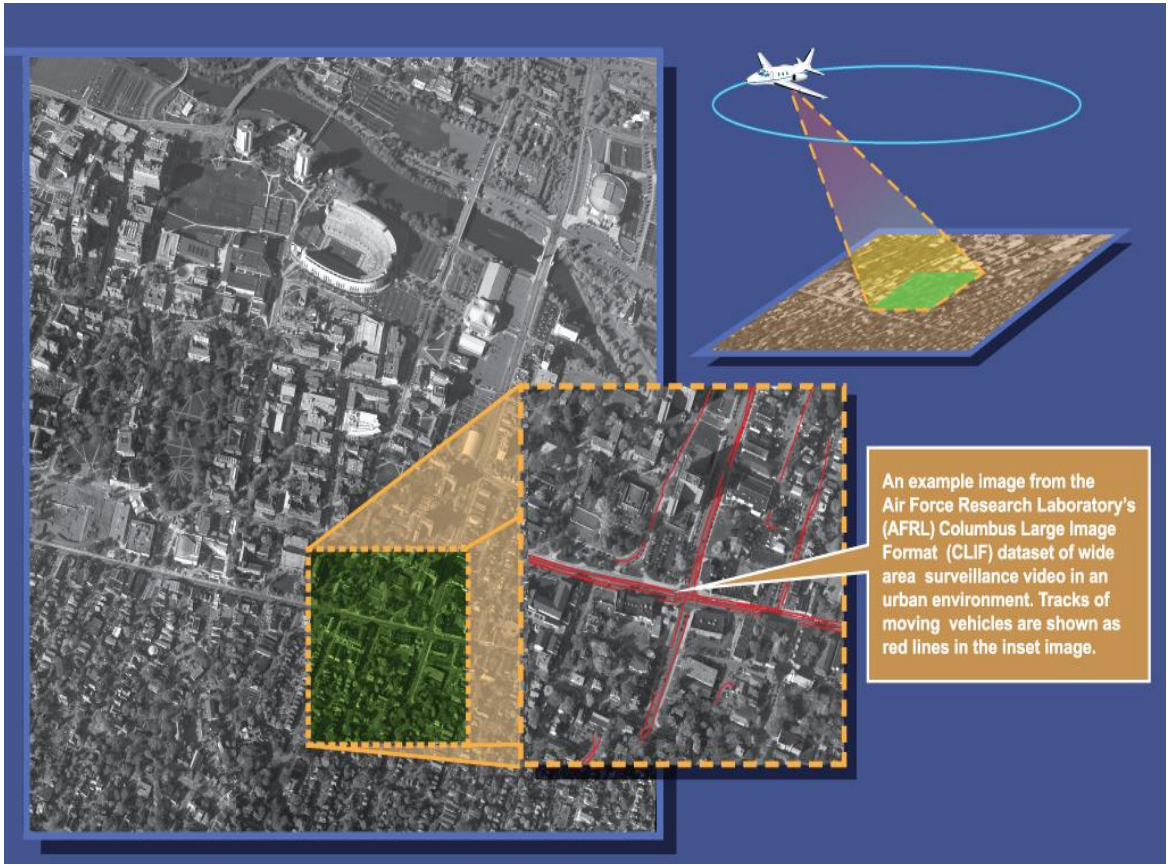
No standard method to combine measures for a holistic evaluation.

Sample List of Metrics

Most common
- Truth completeness (recall)
- Track completeness (precision)

- Truth continuity
- Track continuity
- Number of swaps
- Number of breaks
- Track lifetime

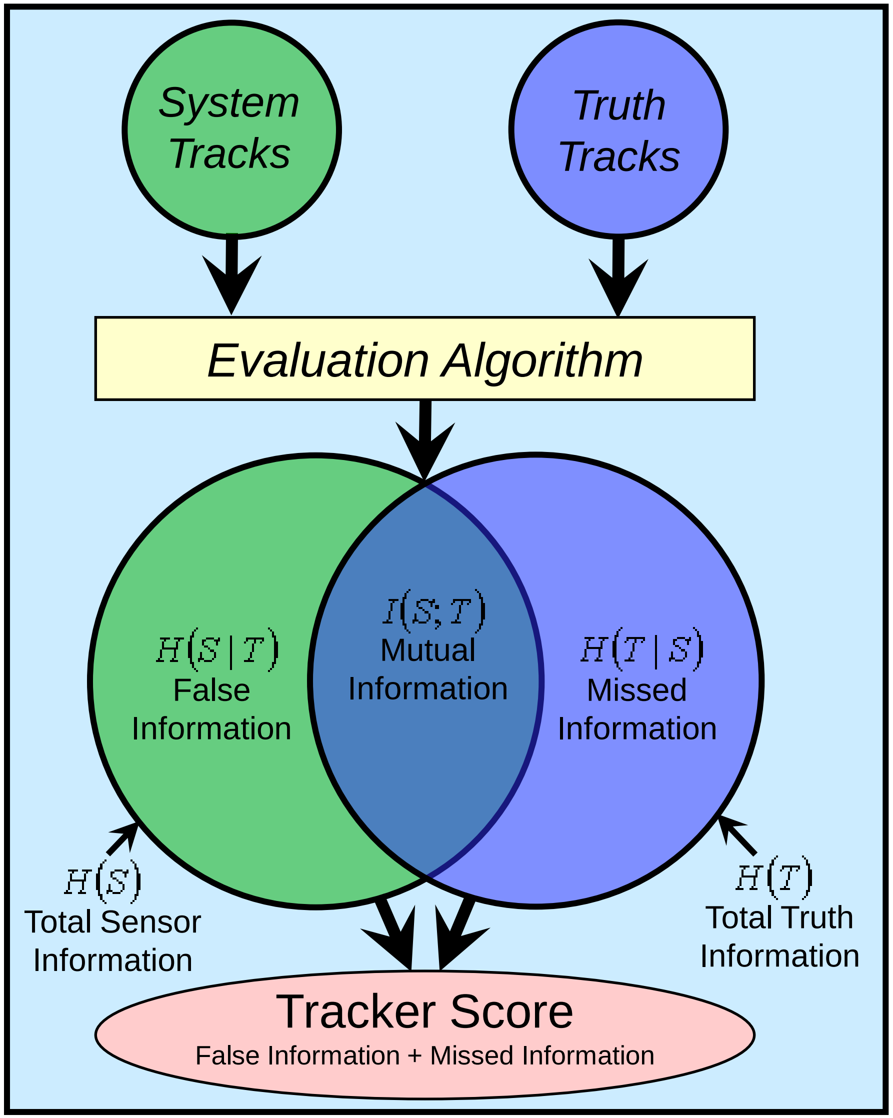
A literature review has identified 145 different tracking and classification metrics



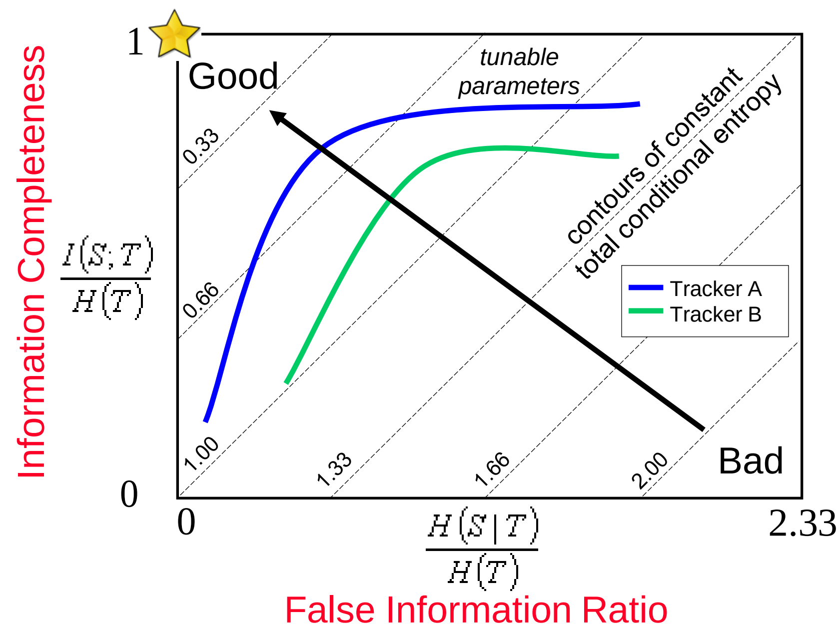
https://www.sdms.afrl.af.mil/datasets/clif2007/

Tracker Performance Evaluation

Information Theoretic Measures



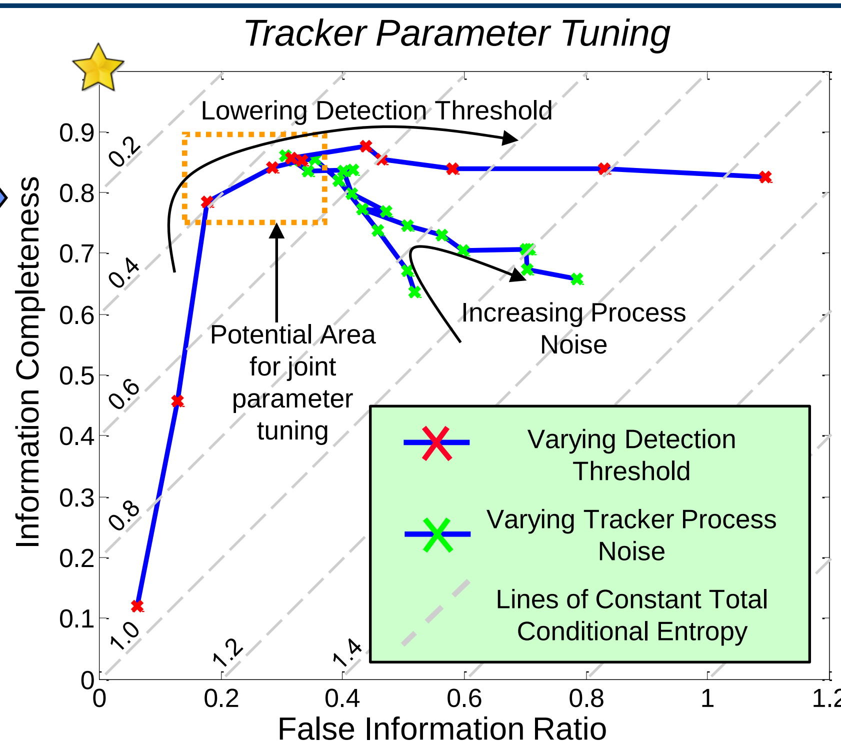
Information Coverage Plot

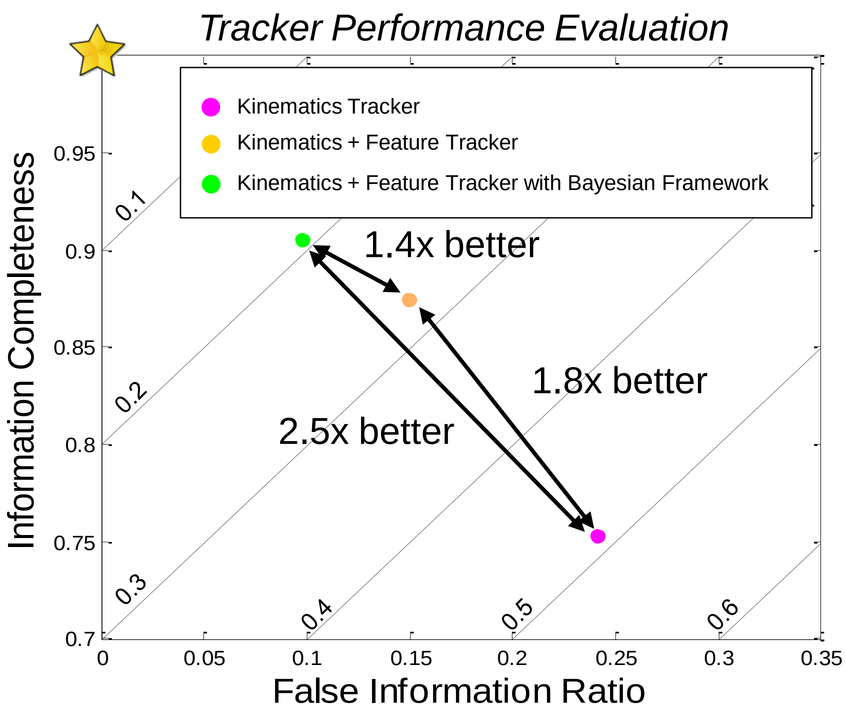


Detector theory-like ROC plots can be generated for holistic tracker performance evaluation

Parameter Tuning & Performance Evaluation

Performance of individual tracking systems can be optimized by selecting parameters that minimize erroneous information





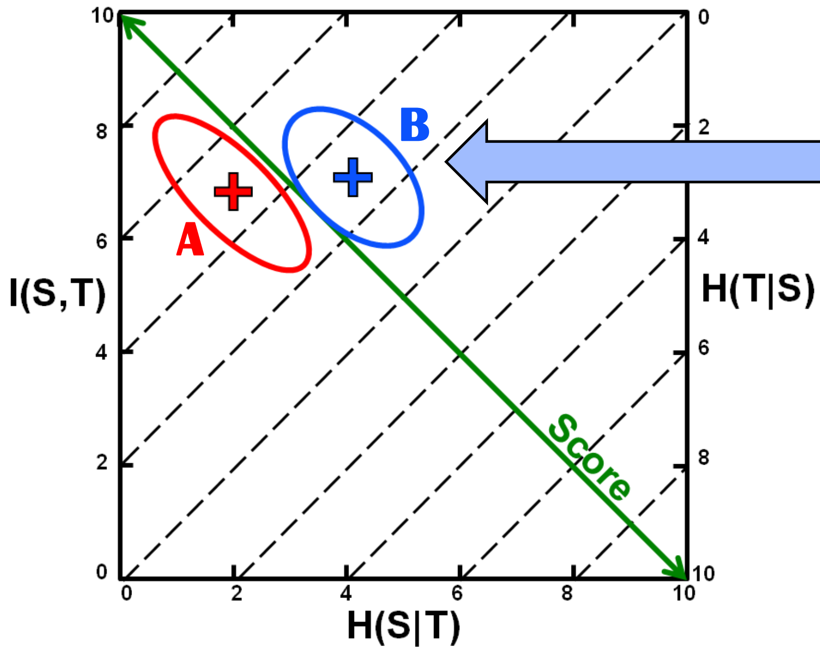
Information coverage plots provide visual evidence of the relative overall performance of different tracking systems

Outline

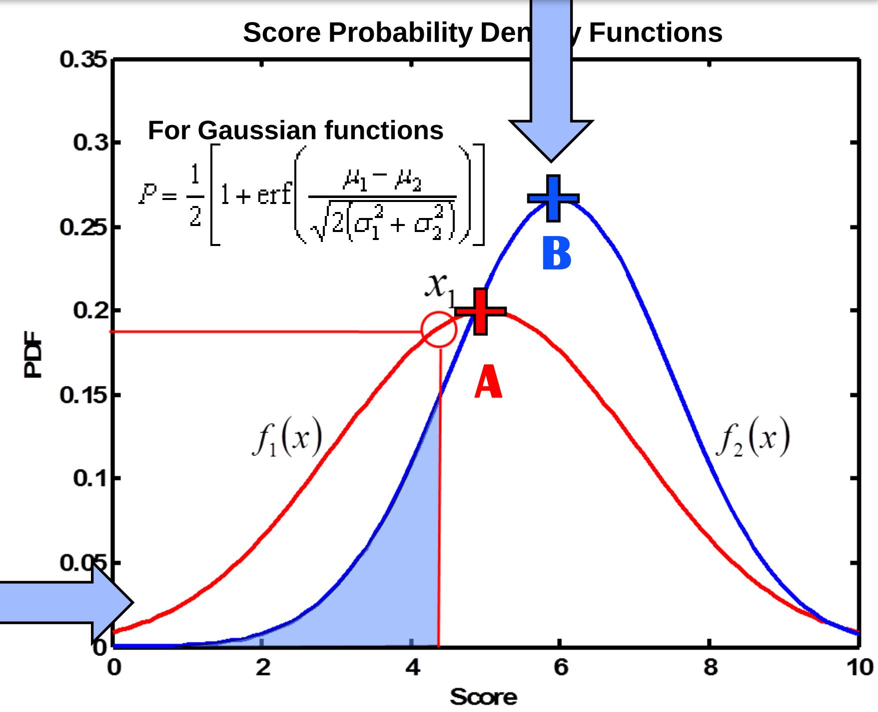
- **Performance metrics overview**
- **Information theoretic performance evaluations**
  - **Multi-target tracker evaluation**
  - **Classifier evaluation**
- **Error estimation and significance testing**
- **Current research focus**
- **Challenges**
- **Summary**

Significance Testing

Early applications of information theory to multi-target trackers and classifiers did not estimate the significance of the results

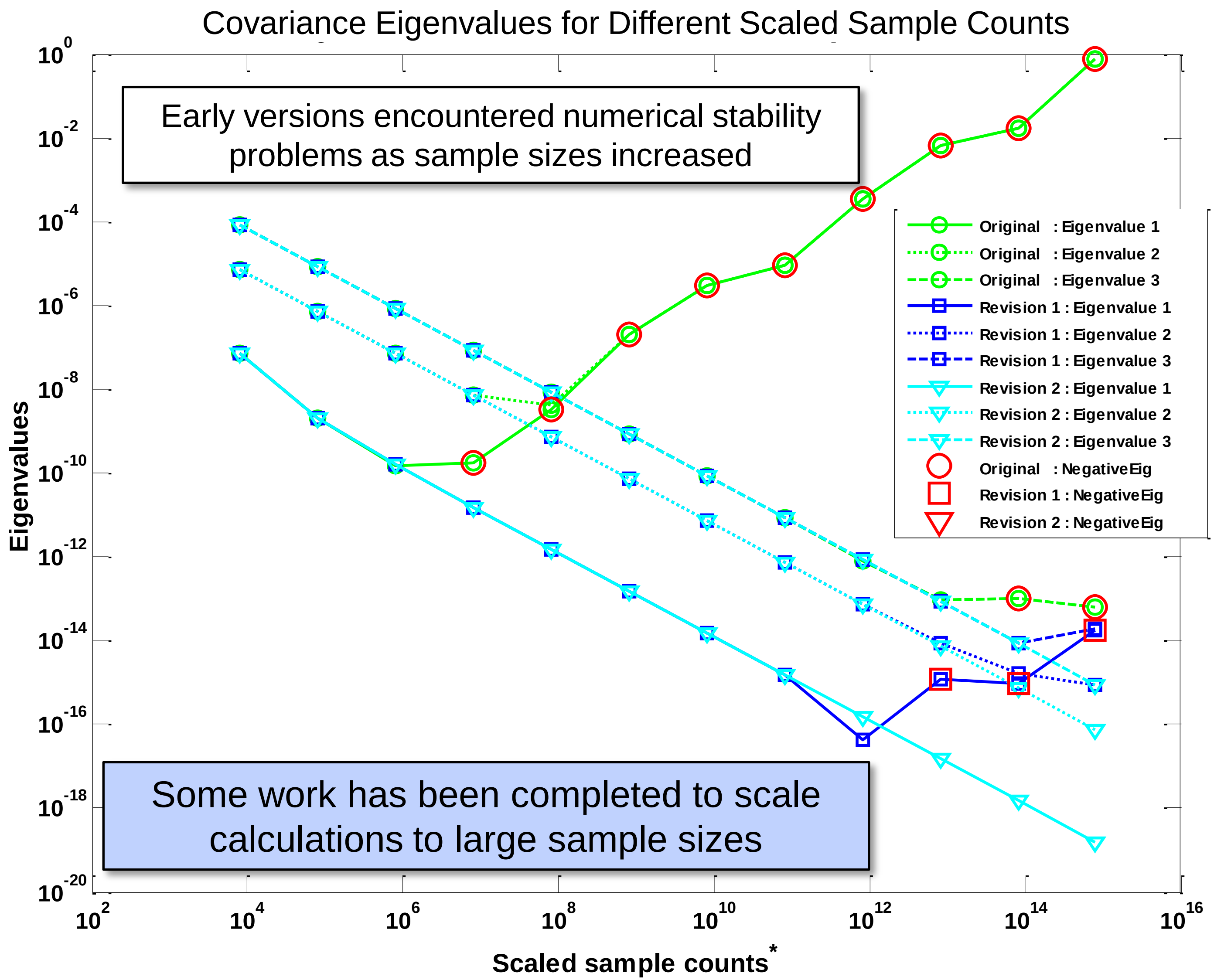


Errors on total conditional entropy are needed to determine the significance of assessment results

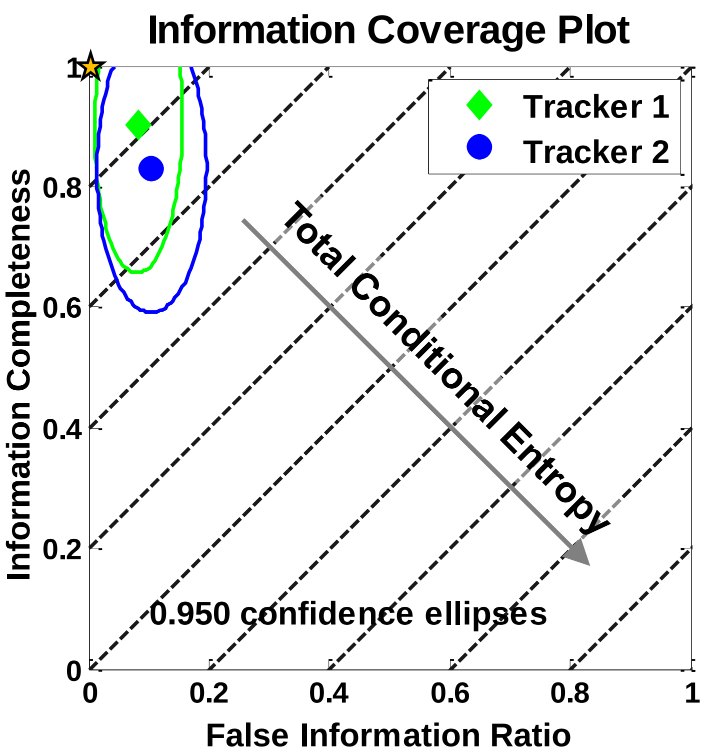


What is the probability that the true score for B is less than the true score for A?

Stability of Different Code Versions of the Wolpert and Wolf equations

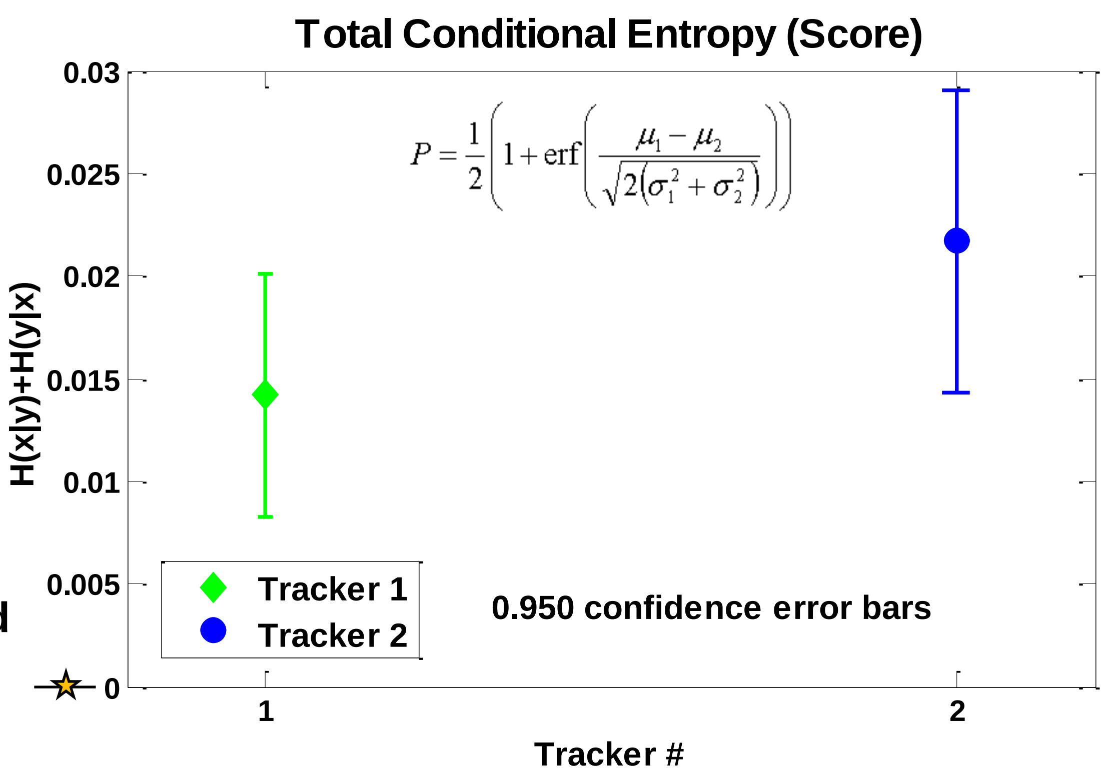


Significance of Test Results



Significance tests can now be used to estimate the chance that true performance differs from results

New software package estimates errors of information theoretic measures



Probability that tracker 2 is better than 1 is ~0.001

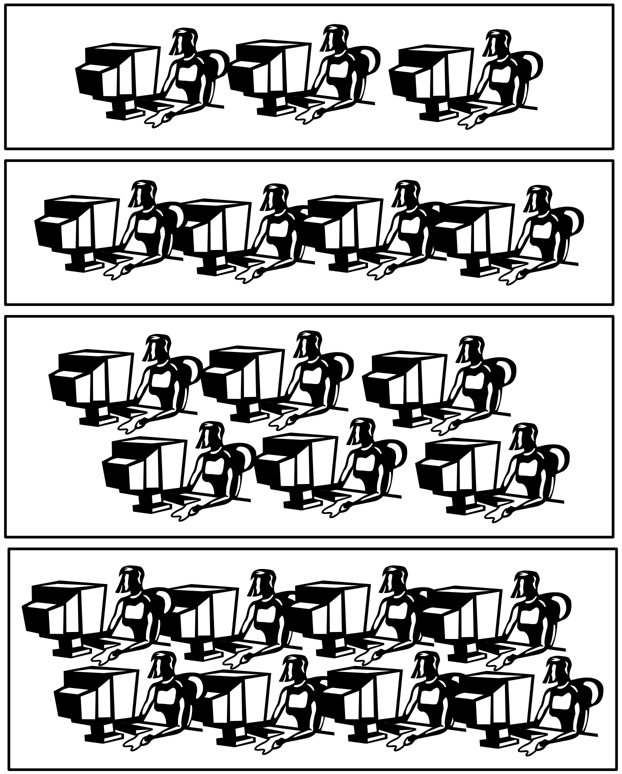
Outline

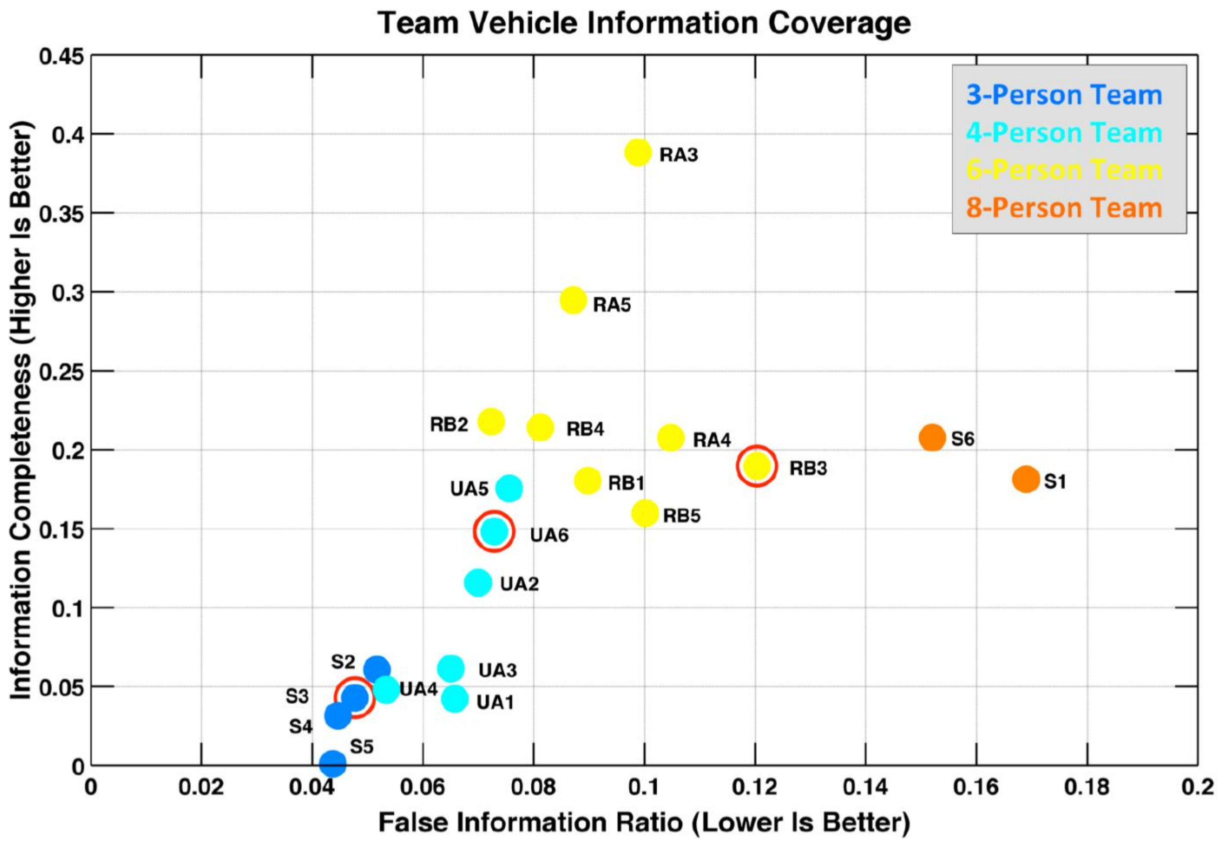
- Performance metrics overview
- Information theoretic performance evaluations
  - Multi-target tracker evaluation
  - Classifier evaluation
- Error estimation and significance testing
- Current research focus
- Challenges
- Summary

Assessment of Analyst Workflows

Red / Blue games pit teams against each other to learn how they solve decision problems

Impact of team size on discovery





Research is beginning to examine relationships between different forms of information, team behaviors, and game objectives

Big Data Science Challenges With Information Theoretic Measures

- Data processing
  - Pro: Not all data may be needed to obtain statistically significant results
  - Con: Potentially a large amount of data to process for assessments
- Information theoretic correlation measures
  - Generating accumulation matrices
  - Covariance accuracy with large-dimension accumulation matrices
- Assessment without truth data
  - Potential to use mutual information measures between systems

- Potential solutions
  - Convert software from MATLAB to a compiled language (C,C++)
  - Parallel processing
  - Optimization of hypergeometric functions and other infinite series
  - Use continued fractions for infinite series

Summary

- **Information theoretic measures provide additional performance metrics for the assessment of data processing and decision systems**
- **"Little data" may be sufficient for some evaluations of big data systems**
- **The primary challenge with information theoretic metrics for big data will be solving computation issues**

A MATLAB software package (InfoMetrics3) is available

To request, send email to hurley@ll.mit.edu

Classifier Evaluation

Ali Farhadi, Mostafa Kamali Tabrizi, Ian Endres, David Forsyth,
"A Latent Model of Discriminative Aspect,"
2009 IEEE 12th International Conference on Computer Vision (ICCV).

Eight Classes
- cell phone
- bike
- iron
- computer mouse
- shoe
- stapler
- toast
- car

T_i Truth Class S_j System Class $P(S_j \mid T_i)$ $P(T_i) = 1/8$

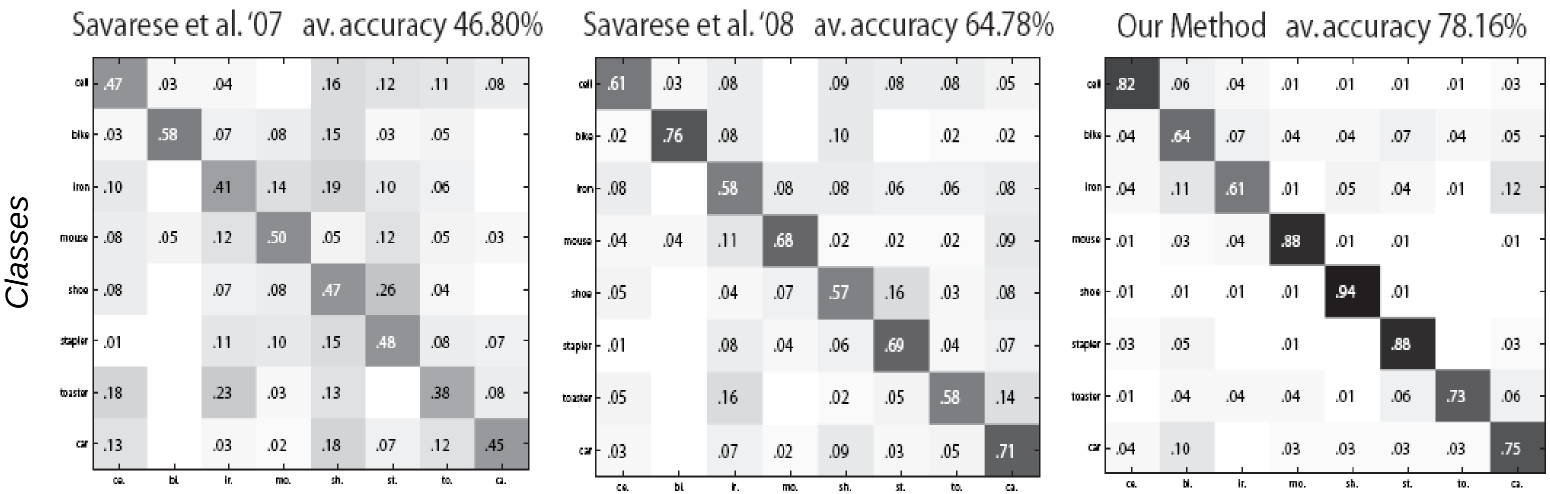
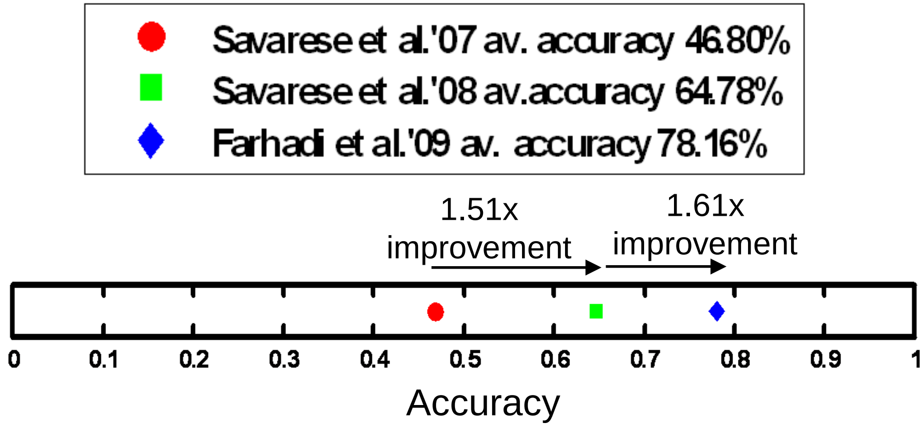
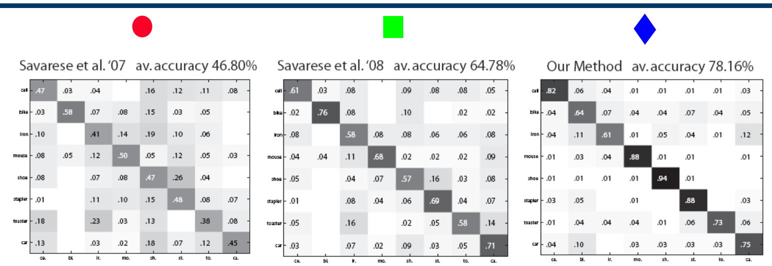


Figure 2. **A comparison** of three recognition methods for recognition in the presence of strong aspectual phenomena. On the left, the class confusion matrix for the method of Savarese et al [24], where the recognizer possesses instances of each class at each aspect. In the center, the class confusion matrix for the work of Savarese et al [23], where the recognizer possesses instances of each class at most aspects, but must interpolate models to cover some aspects. On the right, the class confusion matrix for our method, where the recognizer has no example of a test image's class at the view we want to recognize. Our model of aspect offers a substantial gain.

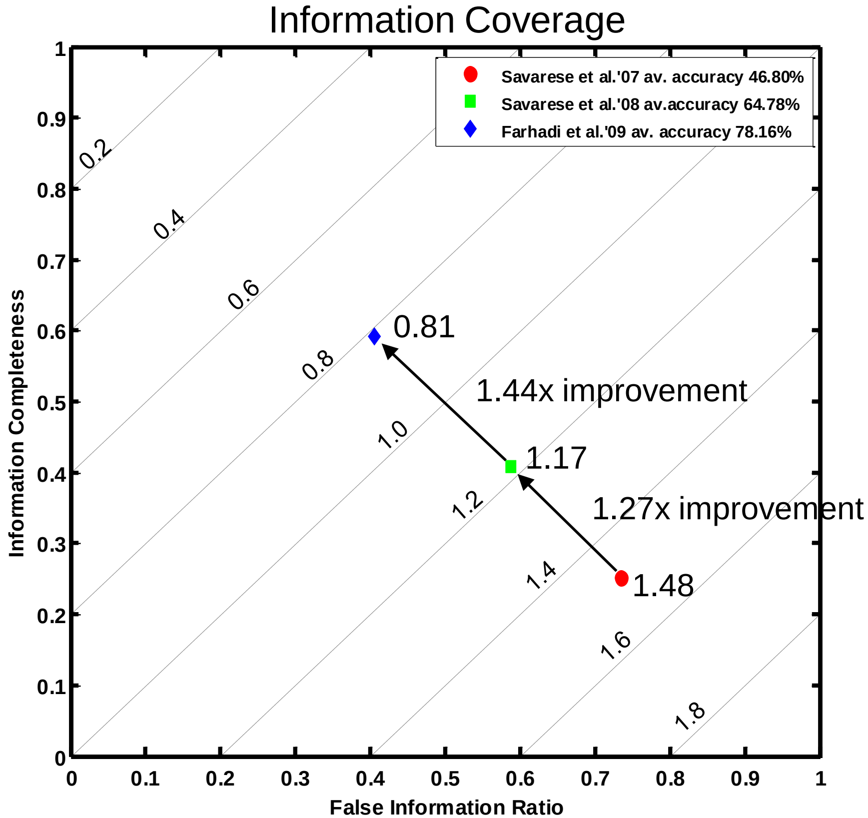
Information Theoretic Performance Measures





Performance improvement of the '09 paper is more apparent in the information theoretic metric since the spread over the misclassifications was much reduced

The classifiers are well balanced between missed information and false information

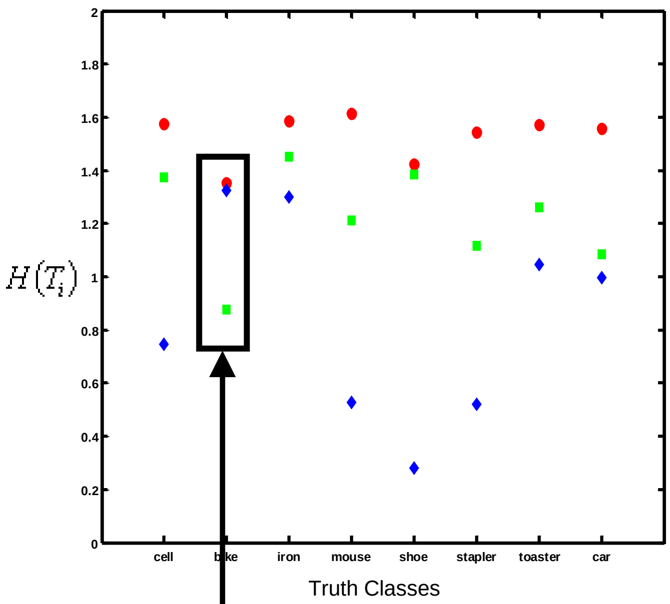


More In-depth Information Theoretic Analysis

Savarese et al.'07 av. accuracy 46.80%
Savarese et al.'08 av.accuracy 64.78%
Farhadi et al.'09 av. accuracy 78.16%

$$H(T_i) = \sum_j P(S_j \mid T_i) \ln\left(P(S_j \mid T_i)\right)$$

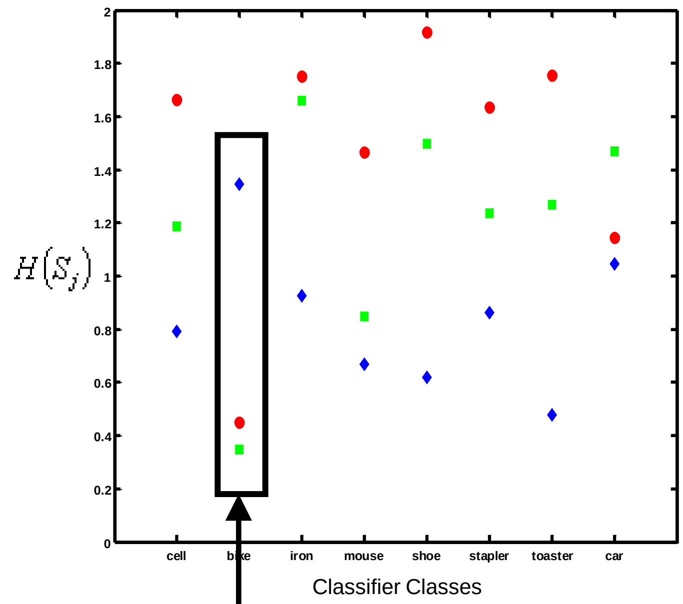
Entropy of truth classes



Second best overall algorithm is most informative when the true object is a bike

$$H(S_j) = \sum_i P(T_i \mid S_j) \ln\left(P(T_i \mid S_j)\right)$$

Entropy of decision classes



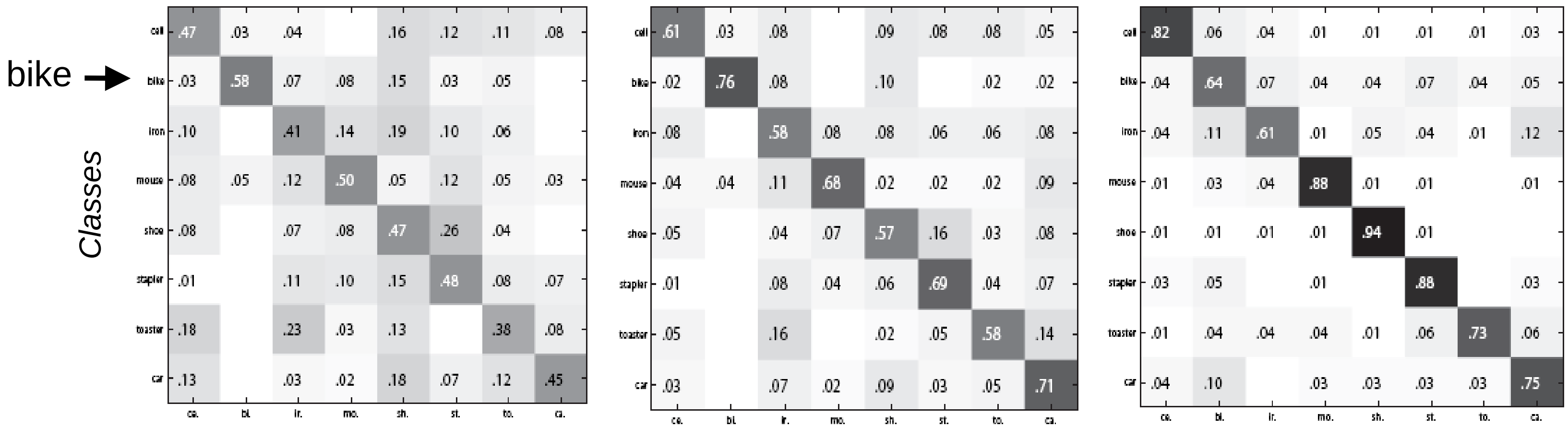
Best overall algorithm is the least informative when it reports a bike

Other information measures can be used to perform more in-depth analysis of classifier performance

Probability Comparison

$P(S_j \mid T_i)$

Savarese et al. '07 av. accuracy 46.80% | Savarese et al. '08 av. accuracy 64.78% | Farhadi av. accuracy 78.16%



$P(T_i) = 1/8$

0.46	0.36	0.06
0.88	0.92	0.62
0	0	0.11
0.76	0.05	0.03
0	0	0.01
0	0	0.05
0	0	0.04
0	0	0.10

$P(T_i \mid S_j)$

Conversion of the original confusion matrix terms to conditional probabilities of decisions given the truth shows that the best overall method is least confident when objects are classified as bikes

Entropy Equations
(Wolpert and Wolf)

Total Entropy

$$E(H(x,y)) = E\left(-\sum_{i,j} p_{ij} \ln(p_{ij}) \middle| n\right) = -\sum_{i,j} \frac{\nu_{ij}}{\nu} \Delta\Phi^{(1)}(\nu_{ij}+1, \nu+1)$$

Truth Entropy

$$E(H(x)) = E\left(\sum_{i} -p_{i\bullet} \ln(p_{i\bullet}) \middle| n\right) = -\sum_{i} \frac{\nu_{i\bullet}}{\nu} \Delta\Phi^{(1)}(\nu_{i\bullet}+1, \nu+1)$$

Algorithm Entropy

$$E(H(y)) = E\left(-\sum_{j} p_{\bullet j} \ln(p_{\bullet j}) \middle| n\right) = -\sum_{j} \frac{\nu_{\bullet j}}{\nu} \Delta\Phi^{(1)}(\nu_{\bullet j}+1, \nu+1)$$

Mutual Information

$$E(I(x,y)) = E(H(x)) + E(H(y)) - E(H(x,y))$$

Conditional Entropies

$$E(H(y \mid x)) = E(H(x,y)) - E(H(x))$$

$$E(H(x \mid y)) = E(H(x,y)) - E(H(y))$$

Counts (*n*) and priors (*r*)

$$\nu_{ij} = n_{ij} + r_{ij}$$

$$\nu = \sum_{i}\sum_{j} \nu_{ij}$$

$$\nu_{i\bullet} = \sum_{j} \nu_{ij}$$

$$\nu_{\bullet j} = \sum_{i} \nu_{ij}$$

Special functions

$$\Delta\Phi^{(n)}(z_1, z_2) = \Phi^{(n)}(z_1) - \Phi^{(n)}(z_2)$$

$$\Phi^{(n)}(z_1) = \Psi^{(n-1)}(z)$$

$$\Psi^{(n)}(z) = \partial_z^{n+1} \ln(\Gamma(z))$$

Wolpert and Wolf's Bayesian analysis of entropy measures for finite data samples provides first- and second-order moment estimates

David H. Wolpert and David R. Wolf, 'Estimating functions of probability distributions from a finite set of samples,' Phys. Rev. E, V 52, n 6, pp. 6841-6854

Entropy Covariance Equations

$$\mathcal{E}_{\overline{IJMN}}=E\left[\sum_{i,j,m,n} p_{ij}\ln(p_{ij})p_{mn}\ln(p_{mn})|\mathbf{n}\right]$$

$$=\sum_{i,j}\sum_{m,n\neq i,j}\frac{\nu_{ij}\nu_{mn}}{\nu(\nu+1)}\{\Delta\Phi^{(1)}(\nu_{ij}+1,\nu+2)\Delta\Phi^{(1)}(\nu_{mn}+1,\nu+2)-\Phi^{(2)}(\nu+2)\}$$

$$+\sum_{ij}\frac{\nu_{ij}(\nu_{ij}+1)}{\nu(\nu+1)}\{[\Delta\Phi^{(1)}(\nu_{ij}+2,\nu+2)]^2+\Delta\Phi^{(2)}(\nu_{ij}+2,\nu+2)\}\ ,$$

$$\mathcal{E}_{\overline{IM}}=E\left[\sum_{i,m} p_{i\cdot}\ln(p_{i\cdot})p_{m\cdot}\ln(p_{m\cdot}|)\mathbf{n}\right]$$

$$=\sum_{i}\sum_{m\neq i}\frac{\nu_{i\cdot}\nu_{m\cdot}}{\nu(\nu+1)}\{\Delta\Phi^{(1)}(\nu_{i\cdot}+1,\nu+2)\Delta\Phi^{(1)}(\nu_{m\cdot}+1,\nu+2)-\Phi^{(2)}(\nu+2)\}$$

$$+\sum_{i}\frac{\nu_{i\cdot}(\nu_{i\cdot}+1)}{\nu(\nu+1)}\{[\Delta\Phi^{(1)}(\nu_{i\cdot}+2,\nu+2)]^2+\Delta\Phi^{(2)}(\nu_{i\cdot}+2,\nu+2)\}$$

(to find $\mathcal{E}_{\overline{JN}}$ substitute $\nu_{\cdot i}$ for $\nu_{i\cdot}$ and $\nu_{\cdot m}$ for $\nu_{m\cdot}$ in the expression for $\mathcal{E}_{\overline{IM}}$),

$$\mathcal{E}_{\overline{IJM}}=E\left[\sum_{ij}\sum_{m} p_{ij}\ln(p_{ij})p_{m\cdot}\ln(p_{m\cdot})|\mathbf{n}\right]$$

$$=\sum_{ij}\sum_{m\neq i}\frac{\nu_{ij}\nu_{m\cdot}}{\nu(\nu+1)}\{\Delta\Phi^{(1)}(\nu_{ij}+1,\nu+2)\Delta\Phi^{(1)}(\nu_{m\cdot}+1,\nu+2)-\Phi^{(2)}(\nu+2)\}$$

$$+\sum_{i,j}\frac{\nu_{ij}(\nu_{i\cdot}+1)}{\nu(\nu+1)}\{[\Delta\Phi^{(1)}(\nu_{i\cdot}+2,\nu+2)]^2$$

$$+\Delta\Phi^{(1)}(\nu_{ij}+1,\nu_{i\cdot}+1)\Delta\Phi^{(1)}(\nu_{i\cdot}+2,\nu+2)+\Delta\Phi^{(2)}(\nu_{i\cdot}+2,\nu+2)\}$$

(to find $\mathcal{E}_{\overline{IJN}}$ substitute $\nu_{\cdot m}$ for $\nu_{m\cdot}$ in the expression for $\mathcal{E}_{\overline{IJM}}$),

$$\mathcal{E}_{\overline{IN}}=E\left[\sum_{i}\sum_{n} p_{i\cdot}\ln(p_{i\cdot})p_{\cdot n}\ln(p_{\cdot n})|\mathbf{n}\right]$$

$$=\sum_{i,n}\frac{\overline{\nu}_{in}(\overline{\nu}_{in}+1)}{\nu(\nu+1)}\Bigg\{\{[\Delta\Phi^{(1)}(\overline{\nu}_{in}+2,\nu+2)]^2+\Delta\Phi^{(2)}(\overline{\nu}_{in}+2,\nu+2)\}$$

$$\times\left[1-\frac{\nu_{i\cdot}+\nu_{\cdot n}-2\nu_{\mathrm{in}}}{\overline{\nu}_{in_n}}+\frac{(\nu_{i\cdot}-\nu_{in})(\nu_{\cdot n}-\nu_{in})}{\overline{\nu}_{in}(\overline{\nu}_{in}+1))}\right]$$

$$+\Delta\Phi^{(1)}(\overline{\nu}_{in}+2,\nu+2)\sum_{r=0}^{\infty}\frac{Q_1(r,1)}{r!}\left[\frac{(\nu_{\cdot n}-\nu_{in})_r}{(\overline{\nu}_{in})_r}\left[--\frac{\nu_{i\cdot}-\nu_{in}}{\overline{\nu}_{in}+r}\right]+\frac{(\nu_{i\cdot}-\nu_{in})_r}{(\overline{\nu}_{in})_r}\left[--\frac{\nu_{\cdot n}-\nu_{in}}{\overline{\nu}_{in}+r}\right]\right]$$

$$+\sum_{r=0}^{\infty}\sum_{s=0}^{\infty}\frac{(\nu_{i\cdot}-\nu_{in})_r(\nu_{\cdot n}-v_{in})_s}{(\overline{\nu}_{in})_{r+s}}\frac{Q_1(r,1)}{r!}\frac{Q_1(s,1)}{s!}\Bigg\}\ ,$$

Other Covariance Terms

Terms	$I(x,y)$	$H(x\mid y)$	$H(y\mid x)$	$H(x\mid y)+H(y\mid x)$
$H(x,y)$	$\mathcal{E}_{\overline{IRd}}+\mathcal{E}_{\overline{IRN}}-\mathcal{E}_{\overline{IRdN}}$	$\mathcal{E}_{\overline{IJdN}}-\mathcal{E}_{\overline{IJN}}$	$\mathcal{E}_{\overline{IRdN}}-\mathcal{E}_{\overline{IRd}}$	$2\mathcal{E}_{\overline{IRdN}}-\mathcal{E}_{\overline{IRN}}-\mathcal{E}_{\overline{IRd}}$
$H(x)$	$\mathcal{E}_{\overline{Rd}}+\mathcal{E}_{\overline{RN}}-\mathcal{E}_{\overline{IRd}}$	$\mathcal{E}_{\overline{IJd}}-\mathcal{E}_{\overline{JN}}$	$\mathcal{E}_{\overline{IRd}}-\mathcal{E}_{\overline{Rd}}$	$2\mathcal{E}_{\overline{IRd}}-\mathcal{E}_{\overline{RN}}-\mathcal{E}_{\overline{Rd}}$
$H(y)$	$\mathcal{E}_{\overline{JN}}+\mathcal{E}_{\overline{JN}}-\mathcal{E}_{\overline{IJN}}$	$\mathcal{E}_{\overline{IJN}}-\mathcal{E}_{\overline{JN}}$	$\mathcal{E}_{\overline{IJN}}-\mathcal{E}_{\overline{JN}}$	$2\mathcal{E}_{\overline{IRN}}-\mathcal{E}_{\overline{RN}}-\mathcal{E}_{\overline{RN}}$
$I(x,y)$	$\mathcal{E}_{\overline{IJdN}}+\mathcal{E}_{\overline{Jd}}+\mathcal{E}_{\overline{JN}}-2\left(\mathcal{E}_{\overline{IJd}}+\mathcal{E}_{\overline{IJN}}-\mathcal{E}_{\overline{JN}}\right)$	$-\mathcal{E}_{\overline{IJdN}}+2\mathcal{E}_{\overline{IJN}}+\mathcal{E}_{\overline{IJd}}-\mathcal{E}_{\overline{JN}}-\mathcal{E}_{\overline{JN}}$	$-\mathcal{E}_{\overline{IJdN}}+2\mathcal{E}_{\overline{IJd}}+\mathcal{E}_{\overline{IJN}}-\mathcal{E}_{\overline{JN}}-\mathcal{E}_{\overline{Jd}}$	$-2\mathcal{E}_{\overline{IRdN}}+3\mathcal{E}_{\overline{IRN}}+3\mathcal{E}_{\overline{IRd}}-2\mathcal{E}_{\overline{RN}}-\mathcal{E}_{\overline{RN}}-\mathcal{E}_{\overline{Rd}}$
$H(x\mid y)$		$\mathcal{E}_{\overline{IRdN}}+\mathcal{E}_{\overline{RN}}-2\mathcal{E}_{\overline{IRN}}$	$\mathcal{E}_{\overline{IRdN}}-\mathcal{E}_{\overline{IRd}}-\mathcal{E}_{\overline{IRN}}+\mathcal{E}_{\overline{RN}}$	$2\mathcal{E}_{\overline{IJdN}}+\mathcal{E}_{\overline{JN}}-3\mathcal{E}_{\overline{IJN}}-\mathcal{E}_{\overline{IJd}}+\mathcal{E}_{\overline{JN}}$
$H(y\mid x)$			$\mathcal{E}_{\overline{IRdN}}+\mathcal{E}_{\overline{Rd}}-2\mathcal{E}_{\overline{IRd}}$	$2\mathcal{E}_{\overline{IJdN}}+\mathcal{E}_{\overline{Jd}}-3\mathcal{E}_{\overline{IJd}}-\mathcal{E}_{\overline{IJN}}+\mathcal{E}_{\overline{JN}}$
$H(x\mid y)+H(y\mid x)$				$4\mathcal{E}_{\overline{IJdN}}-4\mathcal{E}_{\overline{IJN}}-4\mathcal{E}_{\overline{IJd}}+\mathcal{E}_{\overline{Jd}}+2\mathcal{E}_{\overline{JN}}+\mathcal{E}_{\overline{JN}}$

Formulation accounts for statistical significance of data and accepts prior probabilities

where Q_1 is given by

$$Q_1(j,\eta_1)\equiv[1-\theta(j-\eta_1-1)]\frac{(-1)^j}{(\eta_1-j)!}\sum_{r=0}^{j-1}\frac{1}{\eta_1-r}$$

$$+\theta(j-\eta_1-1)(-1)^{\eta_1+1}\Gamma(j-\eta_1)$$

David H. Wolpert and David R. Wolf, ‘Estimating functions of probability distributions from a finite set of samples,’ Phys. Rev. E, V 52, n 6, pp. 6841-6854

LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY