Interface 2008

# Disparate Information Fusion

Carey E. Priebe

Department of Applied Mathematics & Statistics
Johns Hopkins University

May 21 – 24, 2008
Durham, NC

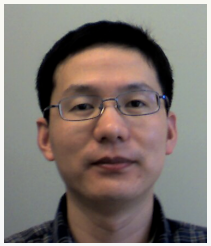
Interface 2008

# Disparate Information Fusion

Carey E. Priebe

Department of Applied Mathematics & Statistics
Johns Hopkins University

Adam Cardinal-Stakenas



Zhiliang Ma

Youngser Park

May 21 – 24, 2008
Durham, NC

# Disparate Data Types

- Text: emails, ships' logs, computer logs, $\cdots$,
- Audio: speech, cell phone, other communications, $\cdots$,
- Image: UAV, satellite, mutlispectral, hyperspectral, $\cdots$,
- Fingerprint data,
- Facial/Character recognition data,
- Other abstract data.

# Disparate Data Types

- Text: emails, ships' logs, computer logs, $\cdots$,
- Audio: speech, cell phone, other communications, $\cdots$,
- Image: UAV, satellite, mutlispectral, hyperspectral, $\cdots$,
- Fingerprint data,
- Facial/Character recognition data,
- Other abstract data.

Goal:

*To obtain superior performance in the exploitation tasks (inference, prediction, synthesis, and general mining) through the fusion of available disparate data types.*

# Dissimilarities

## Setup:

$$(X_i, Y_i) \overset{i.i.d}{\sim} F_{XY} \quad \text{with} \quad X_i : \Omega \to \Xi_1 \times \cdots \times \Xi_K, \quad \text{and}$$
$$Y_i : \Omega \to \{1, \cdots, J\}.$$

## Definition

*A dissimilarity measure* is a function $\delta : \Xi \times \Xi \to \mathbb{R}^+ \cup \{0\}$ with:

1. Positivity: $\delta(x_1, x_2) \geqslant 0$,
2. Symmetry: $\delta(x_1, x_2) = \delta(x_2, x_1)$,
3. Reflexivity: $\delta(x, x) = 0$,
4. Identifiability: $\delta(x_1, x_2) = 0 \Rightarrow x_1 = x_2$.

# Dissimilarities

## Setup:

$$(X_i, Y_i) \overset{i.i.d}{\sim} F_{XY} \quad \text{with} \quad X_i : \Omega \to \Xi_1 \times \cdots \times \Xi_K, \quad \text{and}$$
$$Y_i : \Omega \to \{1, \cdots, J\}.$$

## Definition

*A dissimilarity measure* is a function $\delta : \Xi \times \Xi \to \mathbb{R}^+ \cup \{0\}$ with:

1. Positivity: $\delta(x_1, x_2) \geqslant 0$,
2. Symmetry: $\delta(x_1, x_2) = \delta(x_2, x_1)$,
3. Reflexivity: $\delta(x, x) = 0$,
4. Identifiability: $\delta(x_1, x_2) = 0 \Rightarrow x_1 = x_2$.

*A dissimilarity representation* for a set of $n$ objects is expressed as a symmetric, nonnegative and hollow matrix $\Delta$.

# Motivation

## on the joint vs the product of marginals in disparate dissimilarity fusion

1. "The product of the marginals has less information than the joint."

- We all understand that we can derive the marginals from the joint,
- but that we cannot (necessarily) recover the joint from the marginals.
- *So . . . how shall we fuse disparate dissimilarities?*

# Motivation

## on the joint vs the product of marginals in disparate dissimilarity fusion

### 2. A Toy Example:

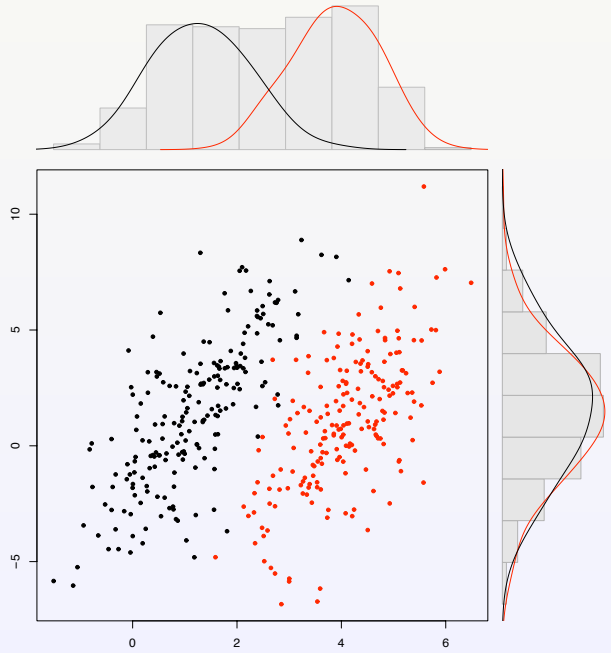
- Consider four objects: ①, ②, 3, 4.
- Consider three "marginal dissimilarities": size, ⓢhape, color.
- Each marginal dissimilarities alone produces its own clusters:
  - size → {1},{2,3,4};
  - ⓢhape → {1,2},{3,4};
  - color → {1,2,3},{4}.
- With the joint information, we can distinguish all four objects.
- Combining marginal information, in this case, also provides full information.

# Motivation

## on the joint vs the product of marginals in disparate dissimilarity fusion

### 3. Tilted Parallel Cigars:

- Neither marginal yields much discriminatory power, but the joint does.
- If the two canonical Euclidean marginal dissimilarities are considered, the joint information can be recovered.
- This is not, in general, the case.

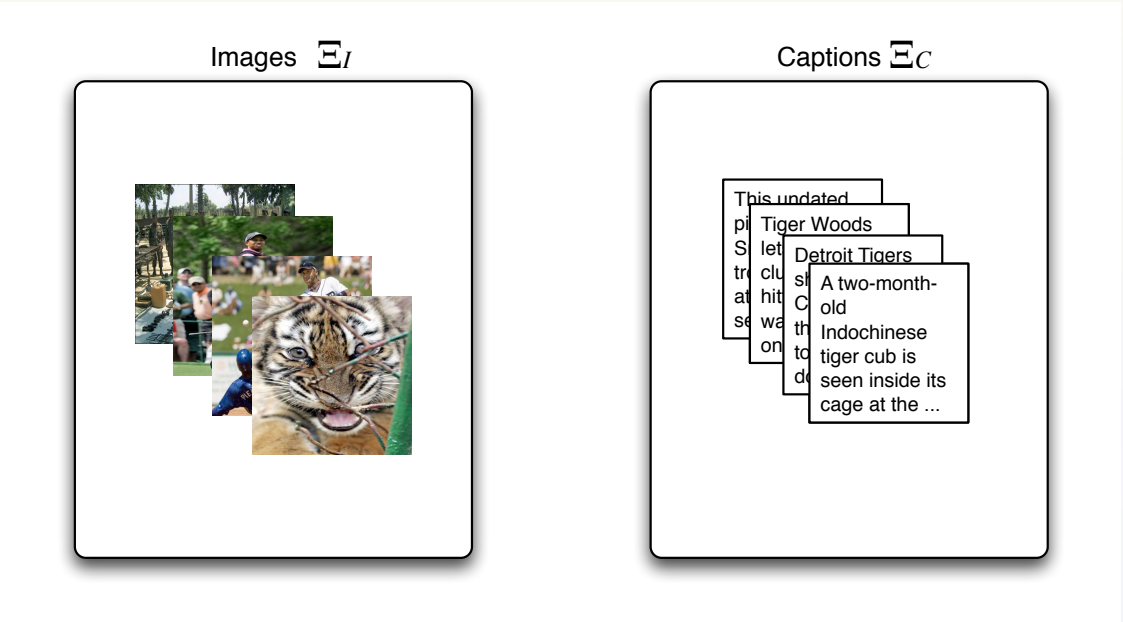


# Motivation

## on the joint vs the product of marginals in disparate dissimilarity fusion

### 4. Image/Caption Fusion:

- Images and text are most naturally considered marginally,
- Our fusion methodologies allow for subsequent "joint" analyses.

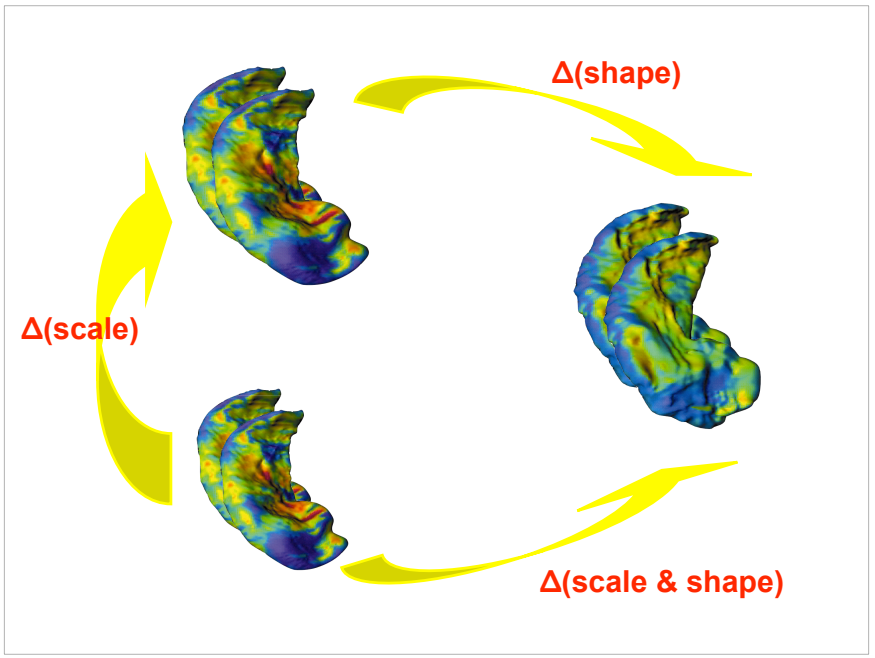


# Motivation

## on the joint vs the product of marginals in disparate dissimilarity fusion

### 5. Brain Shape Comparison:

- left & right hippocampi
- scale & shape

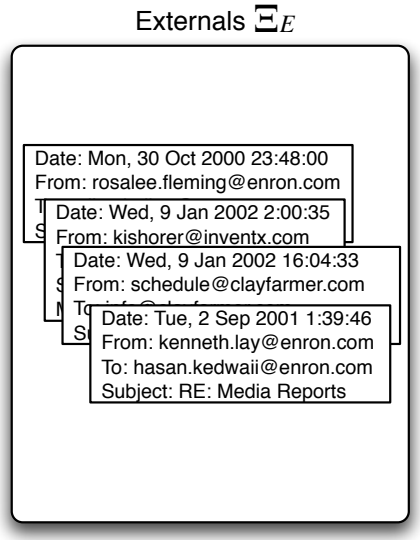


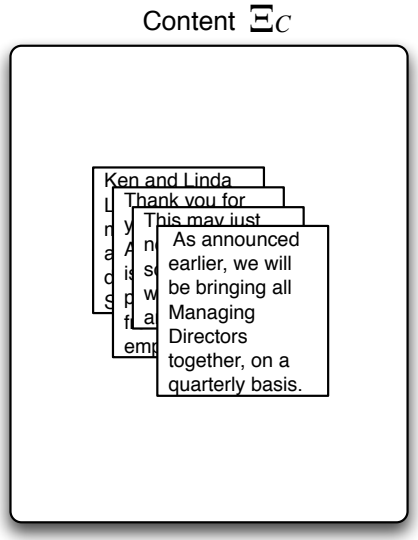
# Motivation

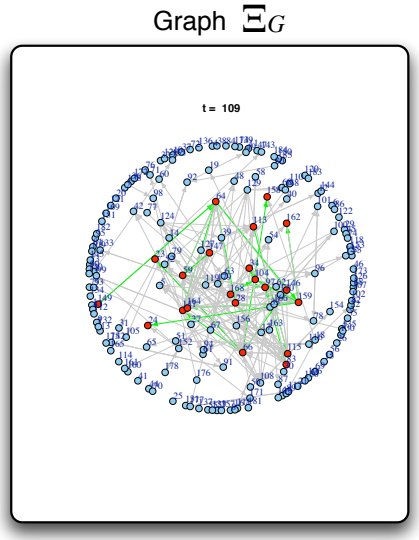
## on the joint vs the product of marginals in disparate dissimilarity fusion

## 6. Streaming Content in Context:

- Here we have $T_E$ and $T_C$, and
- $T = g(T_E, T_C)$ provides superior inference than either $T_E$ alone or $T_C$ alone.



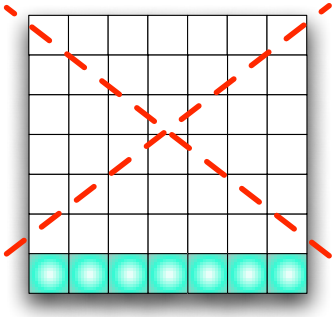



# Framework

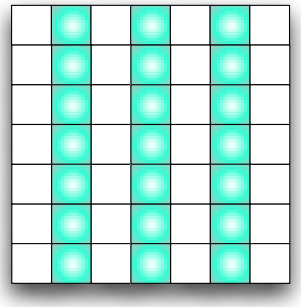
## Three dissimilarity classification approaches

Dissimilarity Matrix ($n$ x $n$)



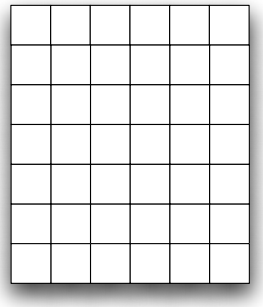
Neighborhood-based

Dissimilarity Matrix ($n$ x $n$)



Dissimilarity Space-based

Feature Matrix ($n$ x $d$)

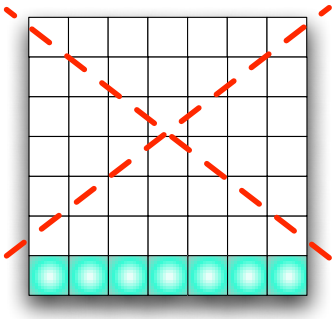


Embedding

# Framework

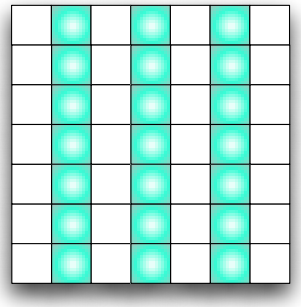
## Three dissimilarity classification approaches

Dissimilarity Matrix ($n$ x $n$)



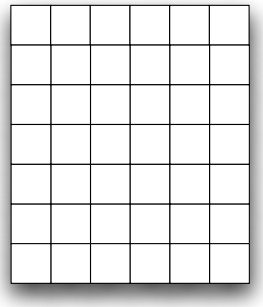
Neighborhood-based

Dissimilarity Matrix ($n$ x $n$)



Dissimilarity Space-based

Feature Matrix ($n$ x $d$)

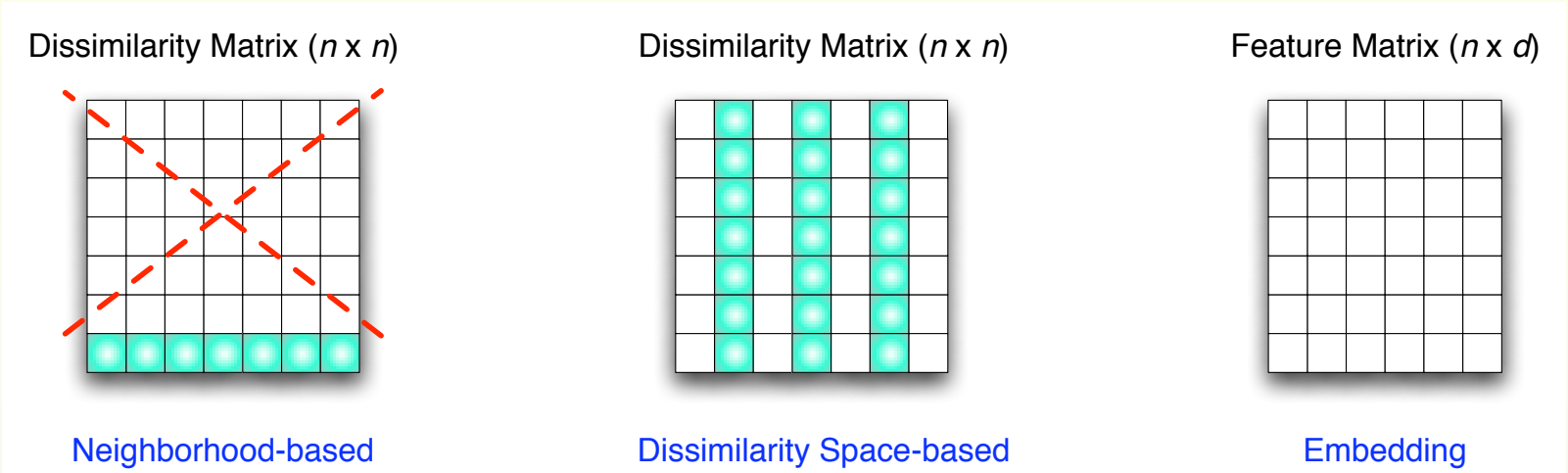


Embedding

1. The *neighborhood-based approach* interprets dissimilarities as neighborhood relations.

# Framework

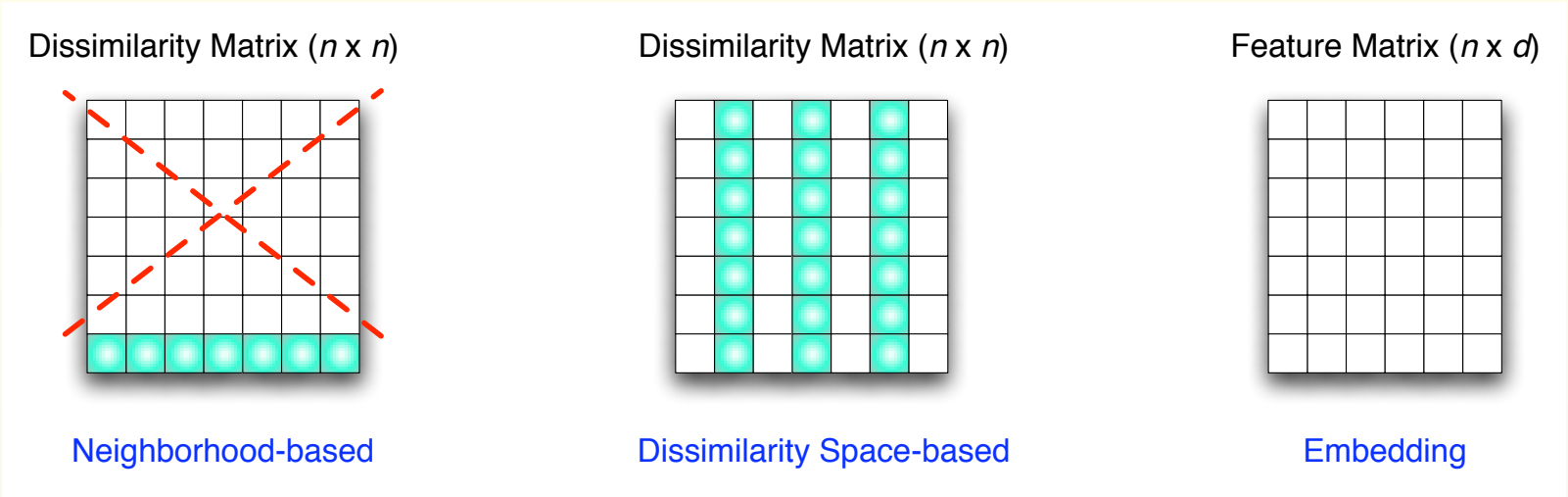
## Three dissimilarity classification approaches




1. The *neighborhood-based approach* interprets dissimilarities as neighborhood relations.
2. The *dissimilarity space approach* defines a representation set $R = \{p_1, \cdots, p_r\}$, and interprets dissimilarities from a point to each element of the representation set as features of this point.

# Framework

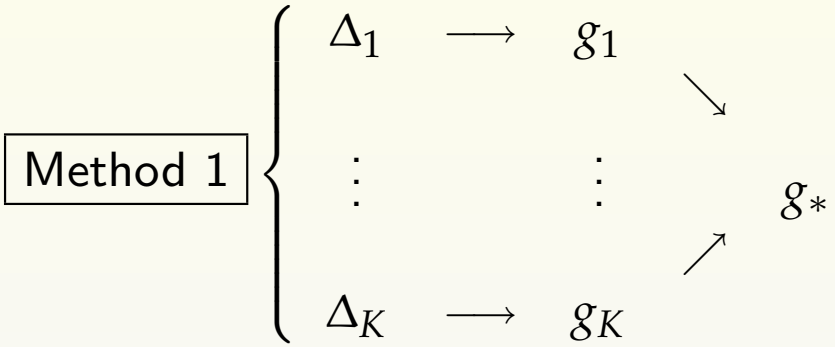
## Three dissimilarity classification approaches




1. The *neighborhood-based approach* interprets dissimilarities as neighborhood relations.
2. The *dissimilarity space approach* defines a representation set $R = \{p_1, \cdots, p_r\}$, and interprets dissimilarities from a point to each element of the representation set as features of this point.
3. The *embedding approach* embeds dissimilarities into $\mathbb{R}^d$ in such a way that the configuration's interpoint distances approximate the original dissimilarities.

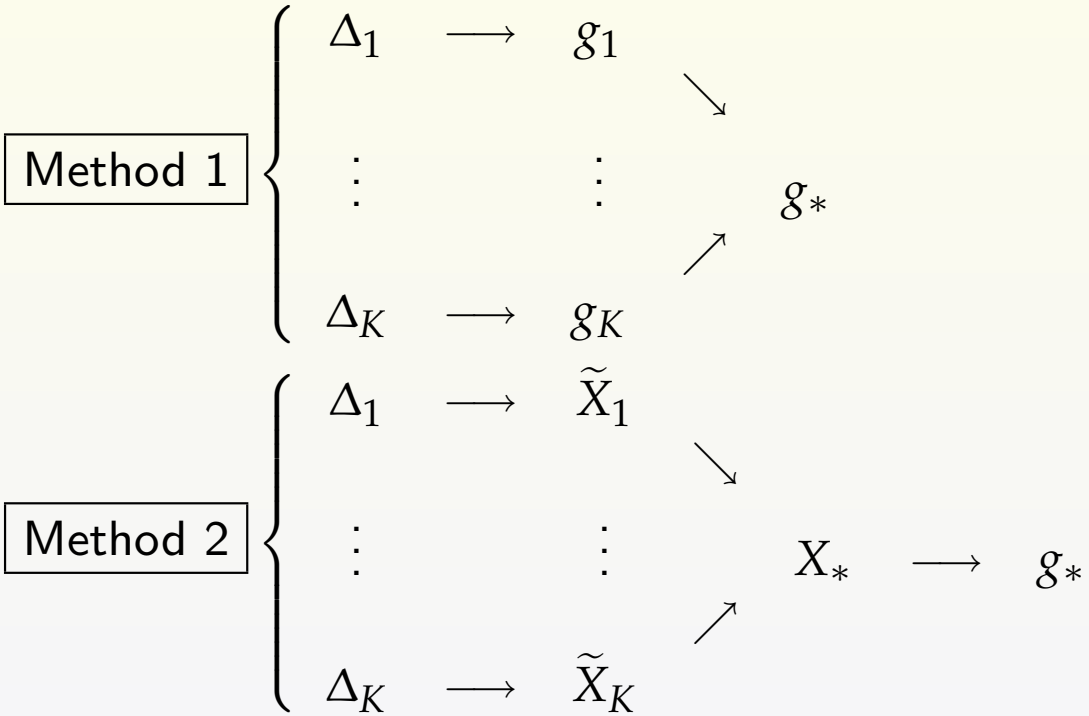
# Framework

## Three DIF approaches



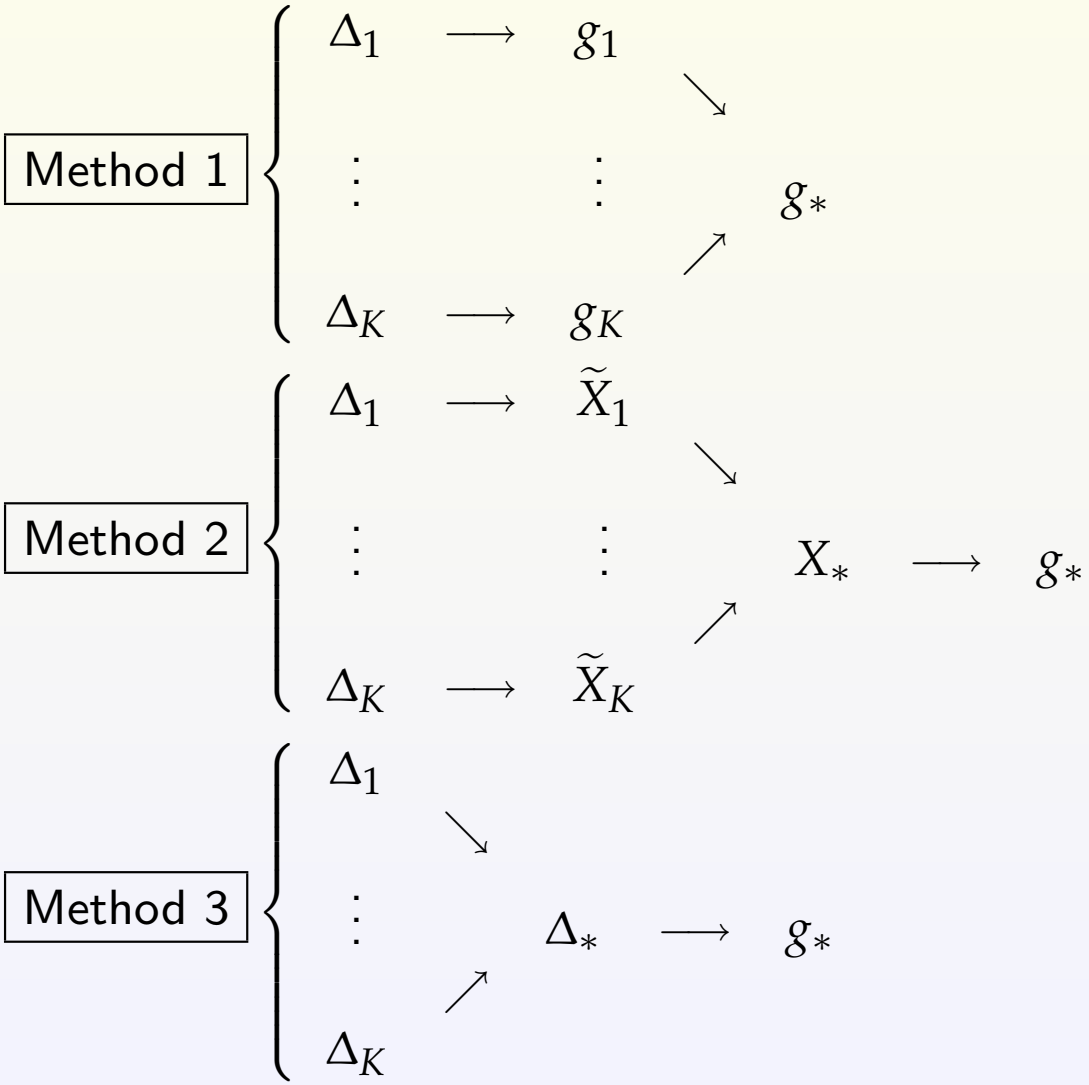
# Framework

## Three DIF approaches



# Framework

## Three DIF approaches



# Image/Caption Fusion

- 140,577 *images & captions* were collected from Yahoo! Photos website [Jeff Solka and his team].

# Image/Caption Fusion

- 140,577 *images & captions* were collected from Yahoo! Photos website [Jeff Solka and his team].
- 1,600 pairs were selected using query word "tiger" on captions.

# Image/Caption Fusion

- 140,577 *images & captions* were collected from Yahoo! Photos website [Jeff Solka and his team].
- 1,600 pairs were selected using query word "tiger" on captions.
- They were labeled manually based only on captions:

| **label** | # |
|---|---|
| animal tiger | 148 |
| Detroit Tigers baseball team | 145 |
| Tiger Woods the golfer | 897 |
| Tamil Tigers soldiers of Sri Lanka | 330 |
| Leicester Tigers rugby team | 48 |
| others | 32 |

# Image/Caption Fusion

- 140,577 *images & captions* were collected from Yahoo! Photos website [Jeff Solka and his team].
- 1,600 pairs were selected using query word "tiger" on captions.
- They were labeled manually based only on captions:

| **label** | # |
|---|---|
| animal tiger | 148 |
| Detroit Tigers baseball team | 145 |
| Tiger Woods the golfer | 897 |
| Tamil Tigers soldiers of Sri Lanka | 330 |
| Leicester Tigers rugby team | 48 |
| others | 32 |

- Two class problem: "Tiger Woods" and "Tamil Tigers".

# "Tiger" Dissimilarity Matrices

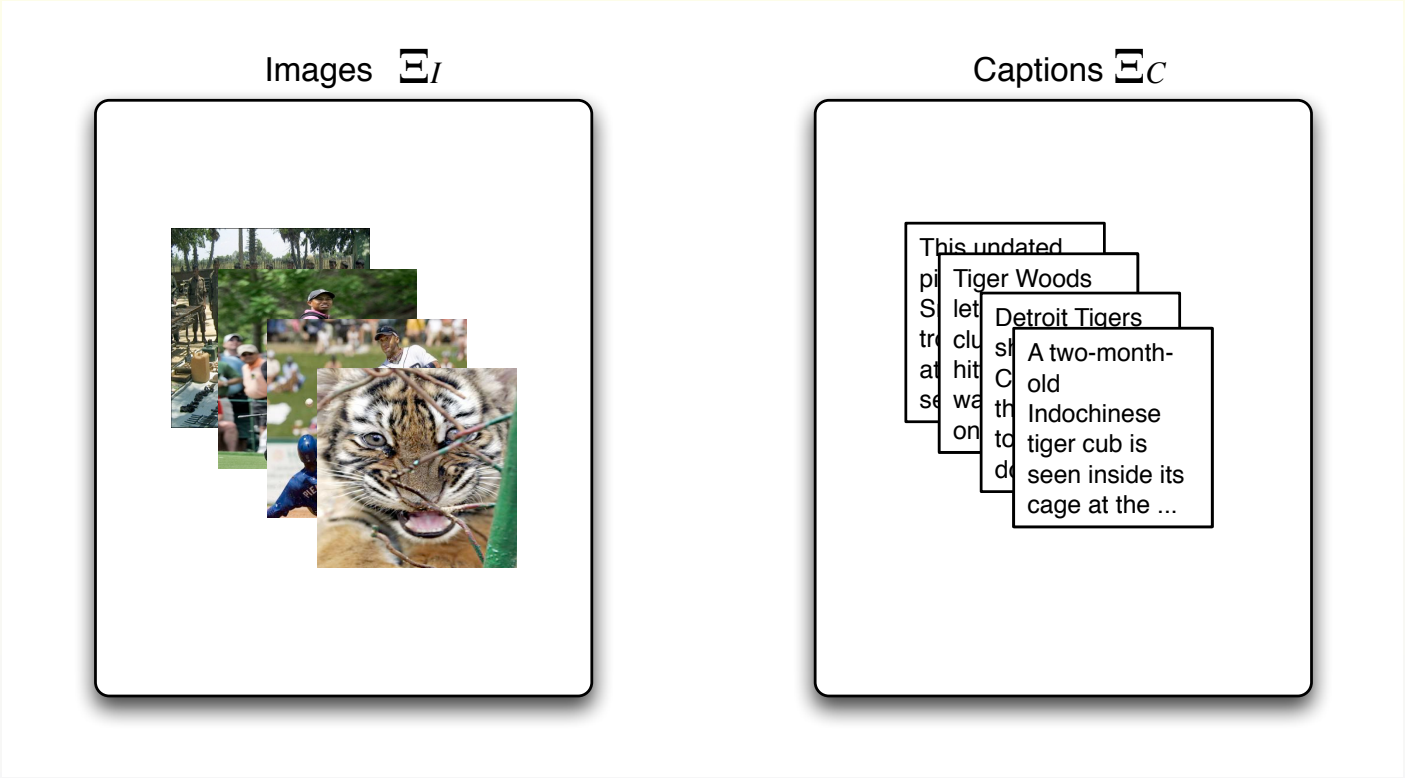
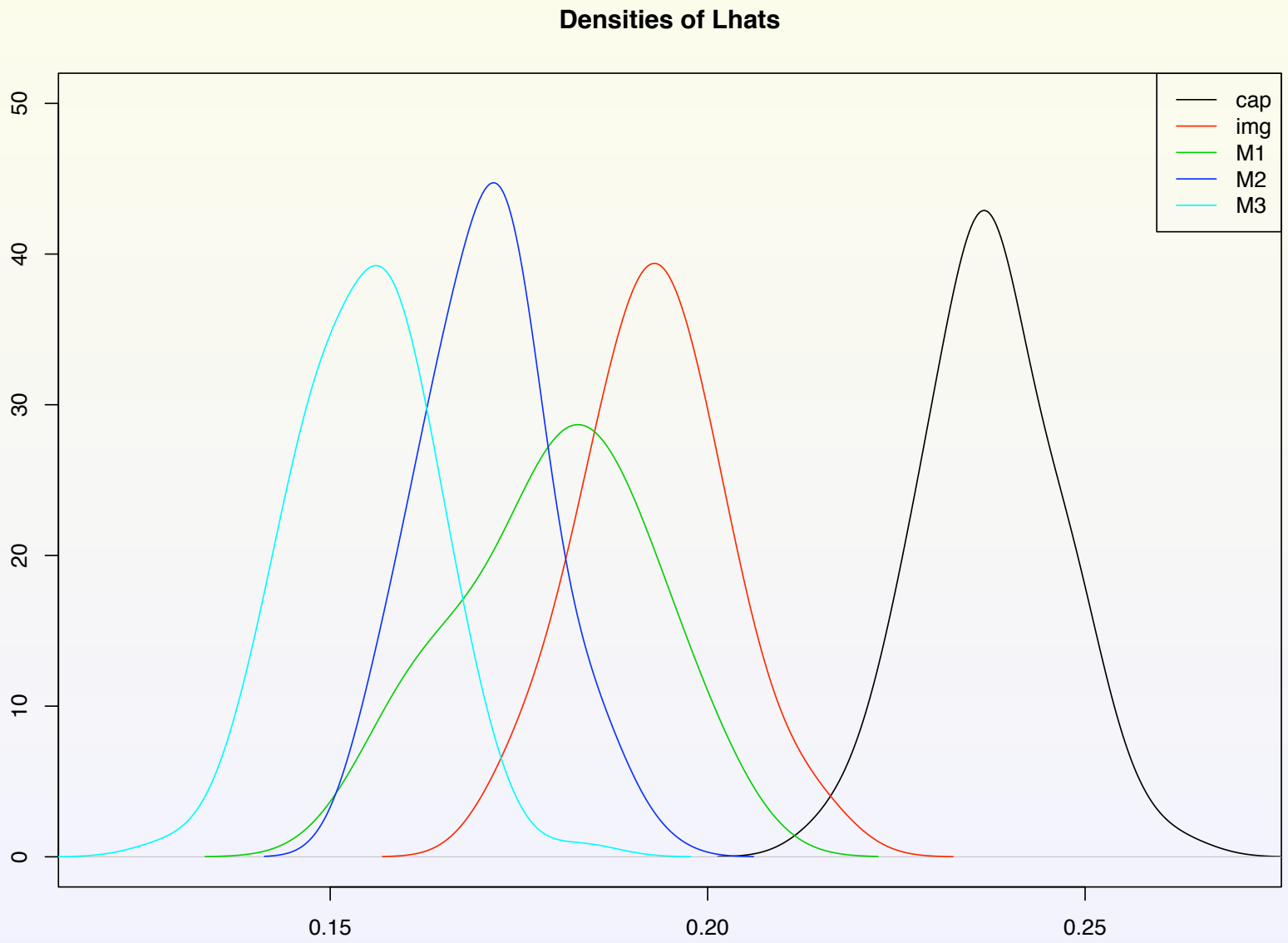



Figure: Conceptual depiction of the "tiger data set"

- *image* features: first & second order pixel derivatives [Jain].
- *caption* features: Mutual Information [Lin & Pantel].
- dissimilarities: 1 - (*Random Forest* proximity) [Breiman].

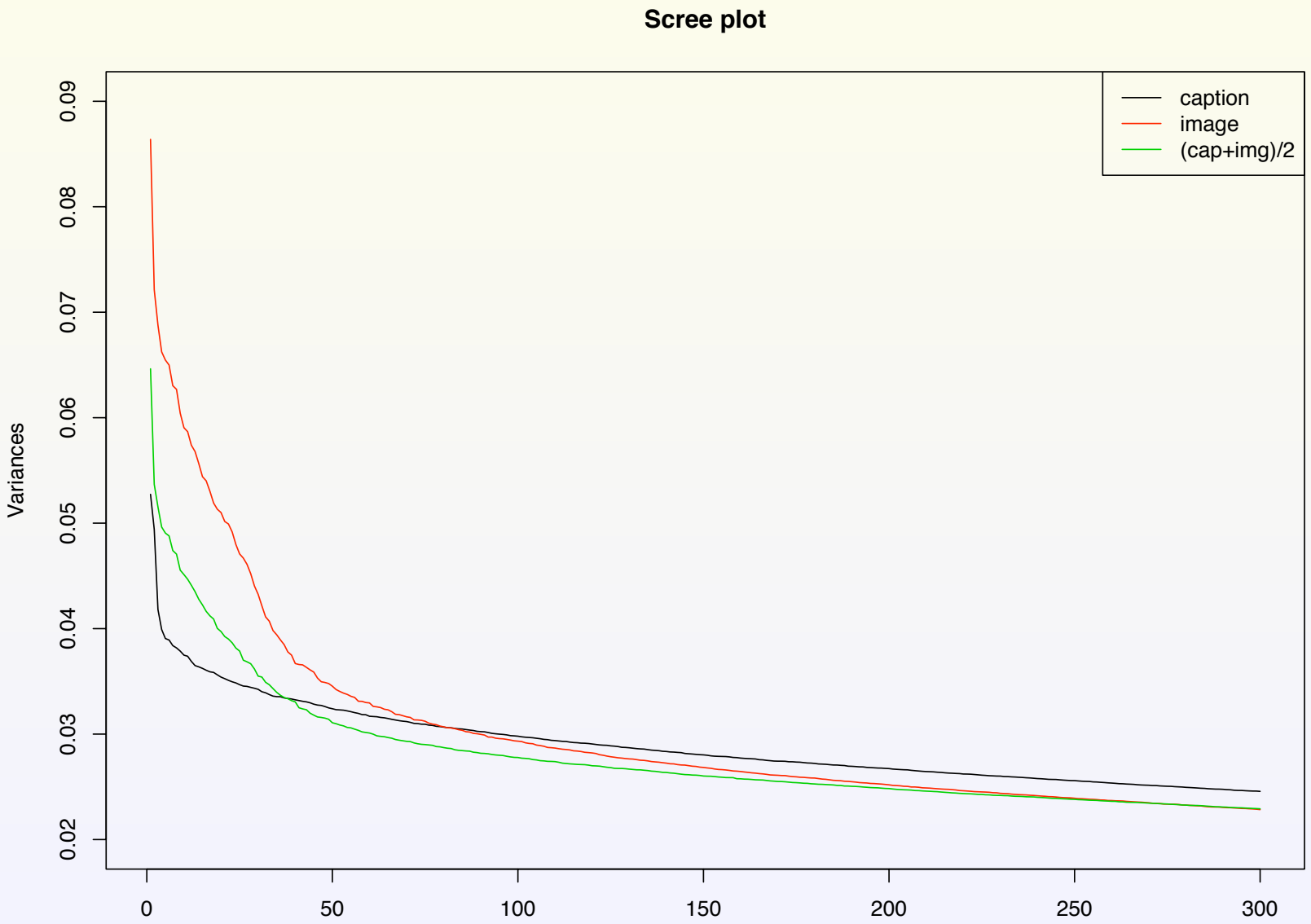
# Combining Dissimilarities

Methods 1, 2 & 3 compared with "image" only and "caption" only



# Combining Dissimilarities

## Determining the dimensions for exploitation task



# Semisupervised Learning

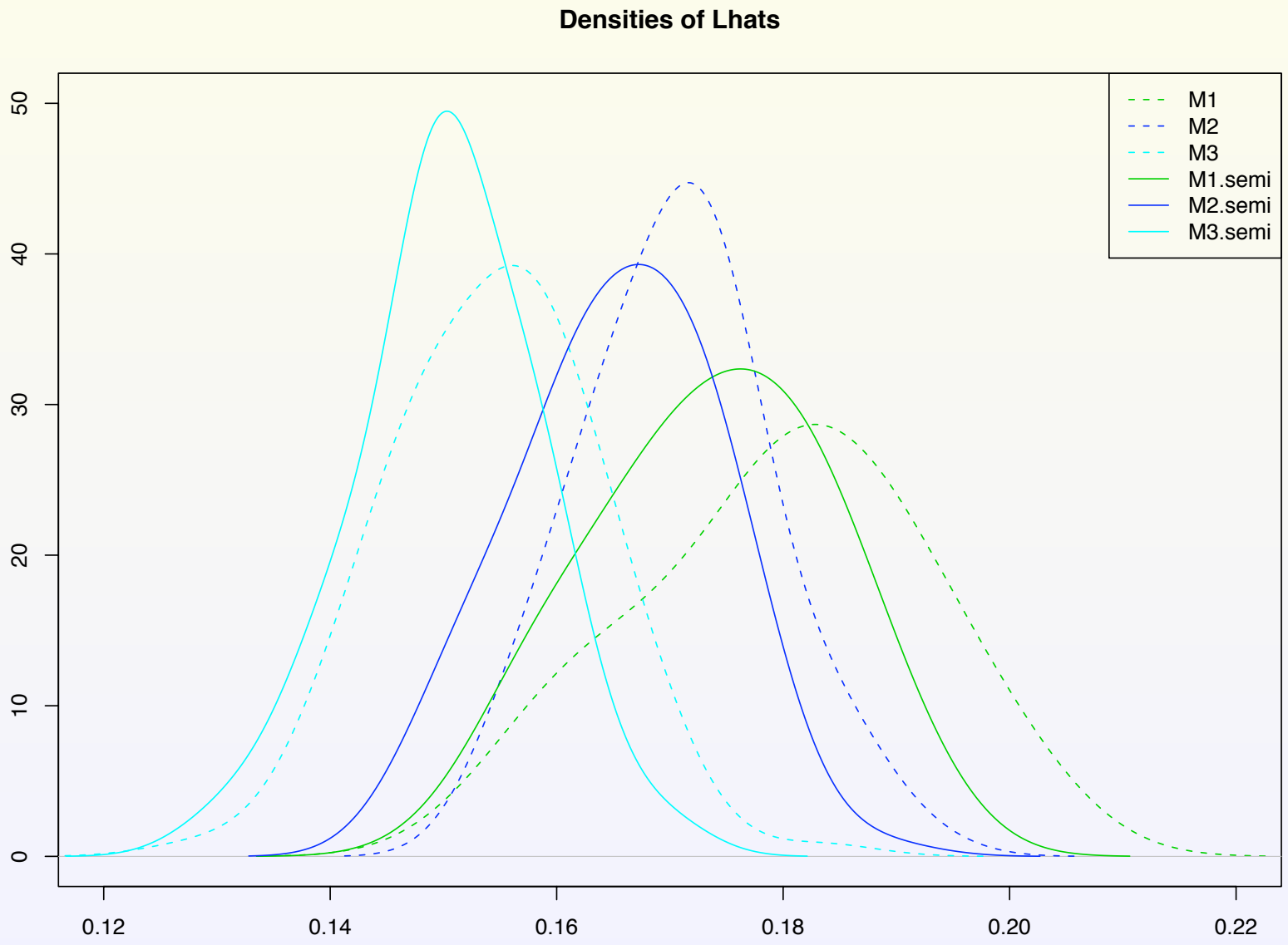
## Self-Training [Zhu2005]:

*"Self-training is a commonly used technique for semi-supervised learning. In self-training a classifier is first trained with the small amount of labeled data. The classifier is then used to classify the unlabeled data. Typically the most confident unlabeled points, together with their predicted labels, are added to the training set."*

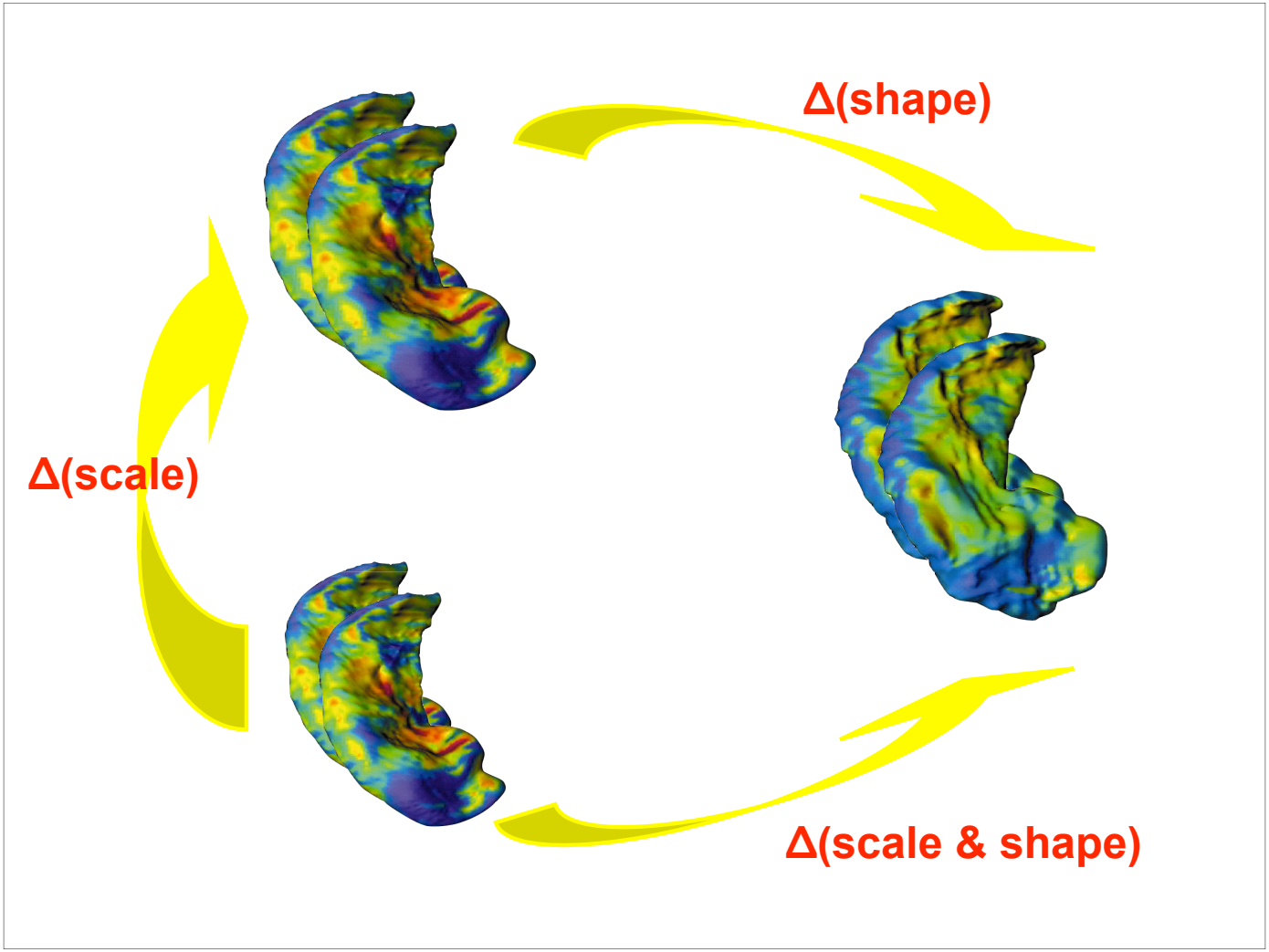

- X. Zhu. *Semi-supervised learning literature survey*. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005.

# Combining Dissimilarities

## Methods 1, 2 & 3 compared with Self-training

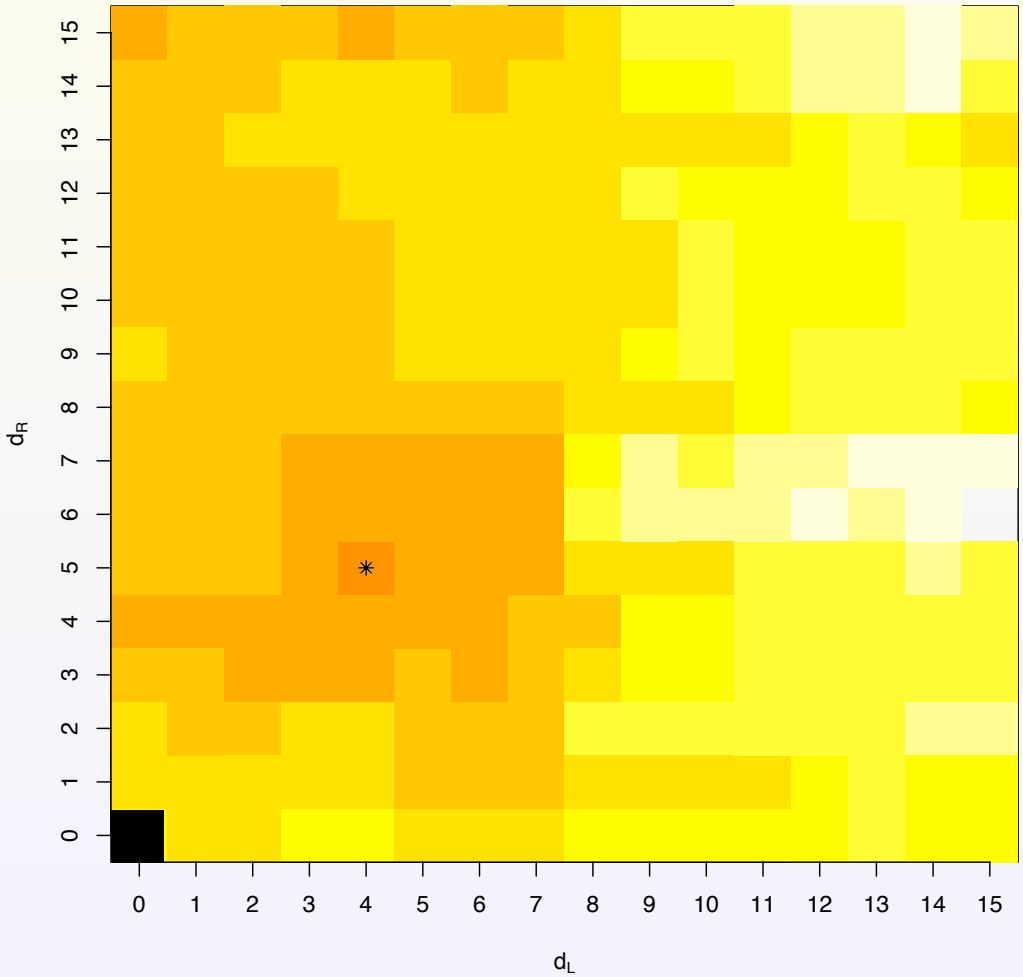


# Brain Shape Comparison

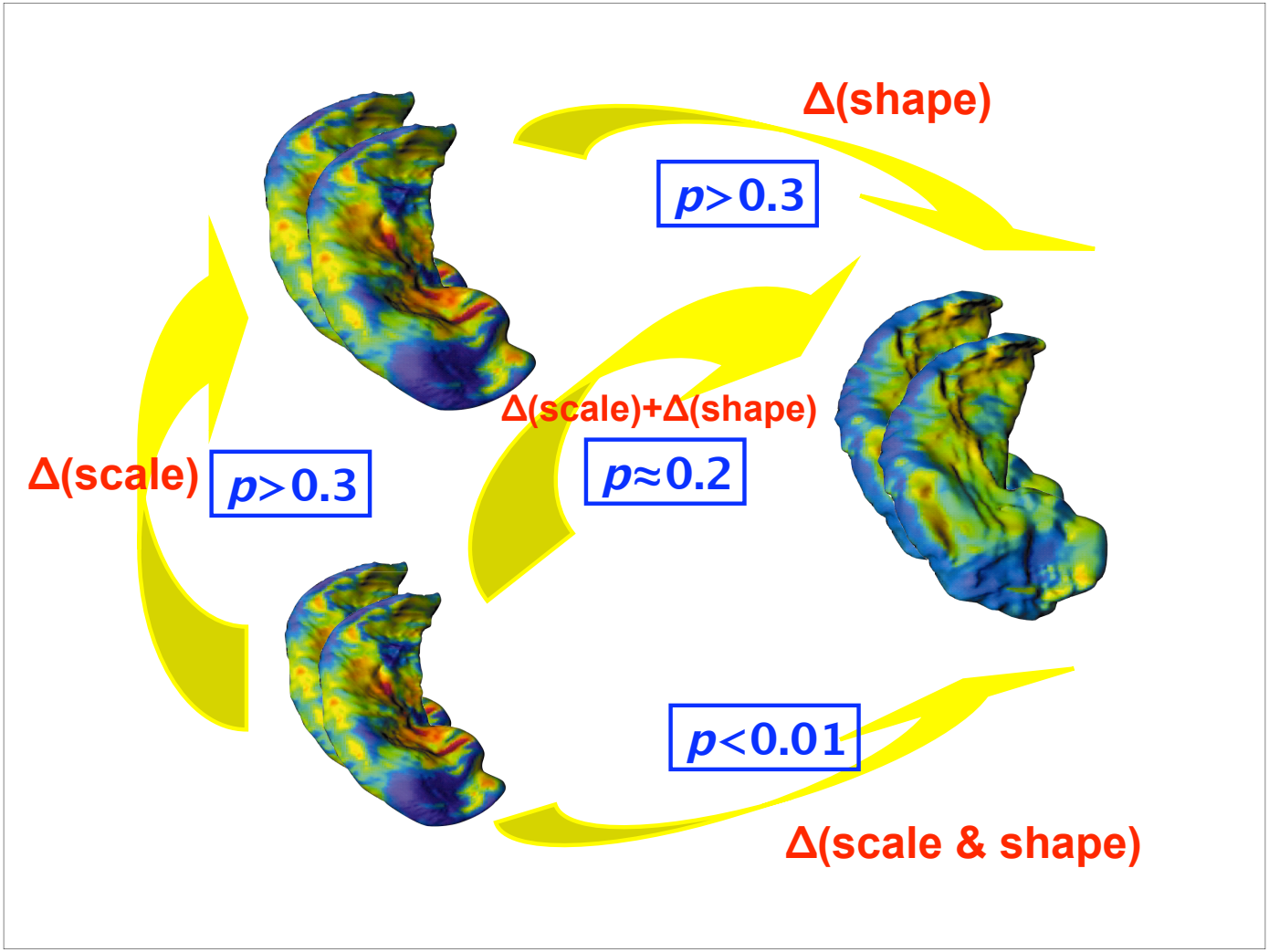


# Brain Shape Comparison

- $\widehat{L}_{d_L,d_R} = \frac{1}{m}\sum_{j=n+1}^{n+m} I\{\widehat{Y}_j \neq Y_j | (X_1, Y_1), \cdots, (X_n, Y_n)\}$
- $(d_L^*, d_R^*) \in \arg\min_{d_L,d_R} \widehat{L}_{d_L,d_R}$



# Brain Shape Comparison



# Streaming Content in Context

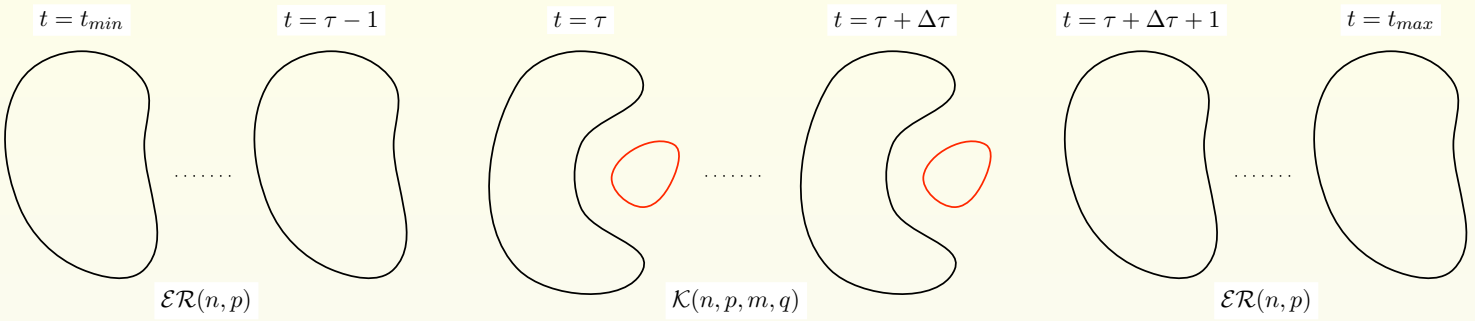



Figure: Anomaly detection in a time series of graphs

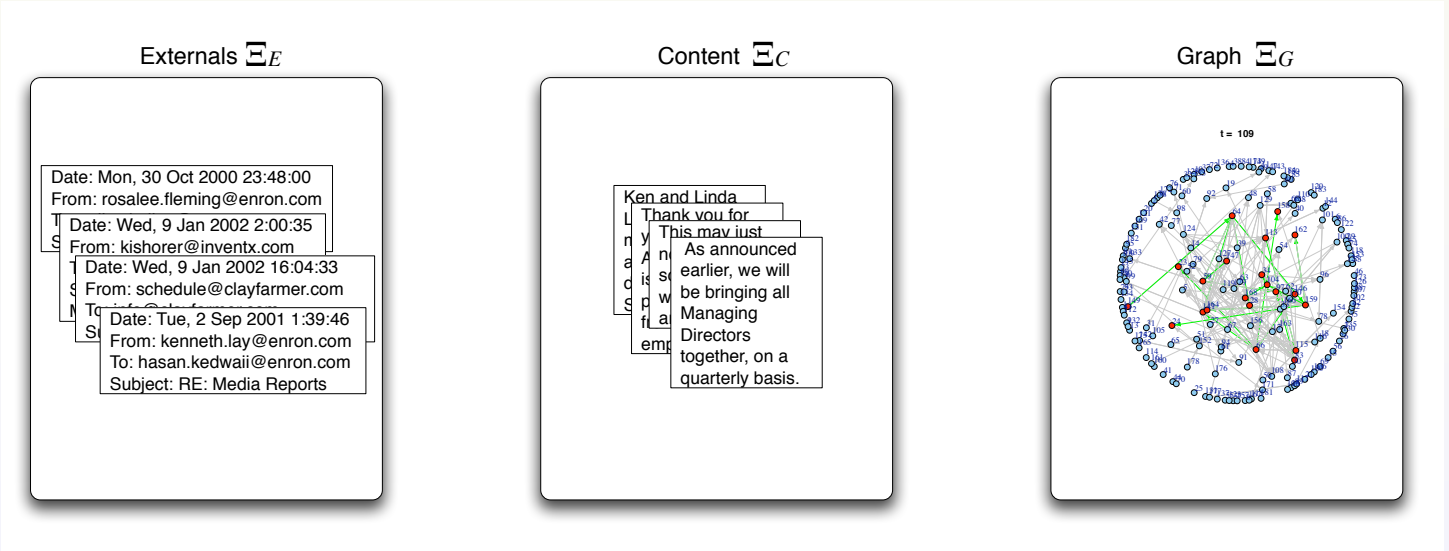



Figure: Conceptual depiction of the "Enron email data set"

# Streaming Content in Context

**Theorem**

*A test based on fusion of externals and content can be more powerful than a test based on either externals alone or content alone.*

**Proof.**

We construct an example wherein the result holds.

$H_0 : \mathcal{ER}(n, p, \theta)$

$H_A : \mathcal{K}(n, p, \theta, m, q, \theta')$

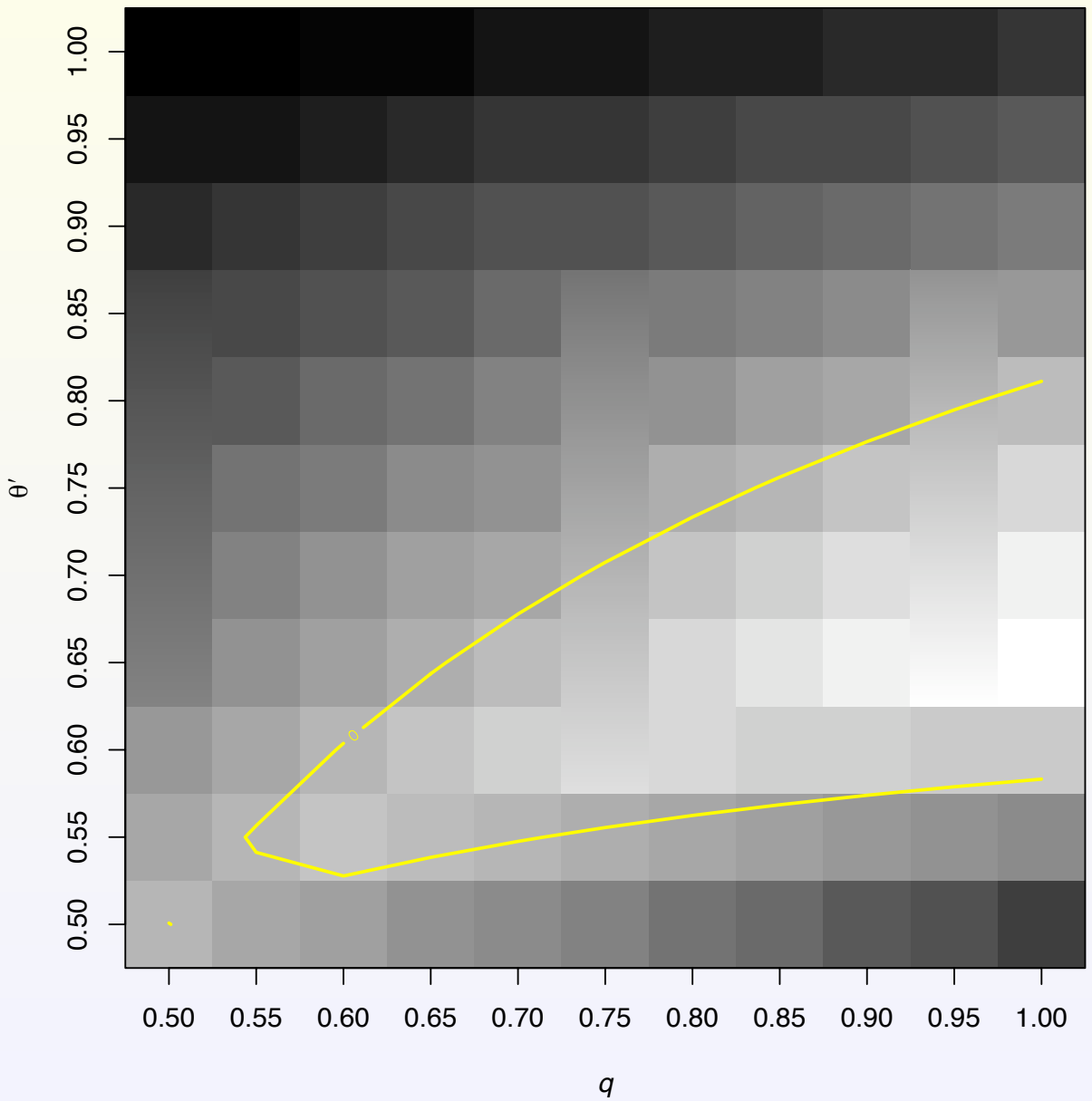
$T_1 = size(E)$: test statistic based on externals only

$T_2 = size(C)$: test statistic based on content only

$T = g(T_1, T_2)$: disparate information fusion!

We demonstrate that $\beta > \max\{\beta_1, \beta_2\}$ for appropriate choices of $n, p, \theta, \alpha, m, q, \theta'$. □

# Streaming Content in Context

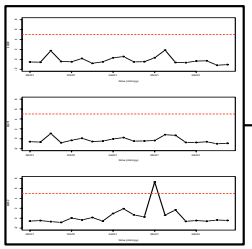


# Streaming Content in Context

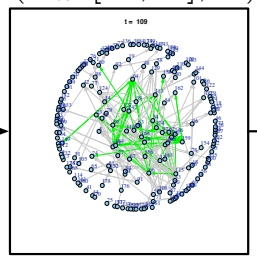
**enron content in context**

## context analysis

Detection at $k^*$, $t^*$, $v^*$

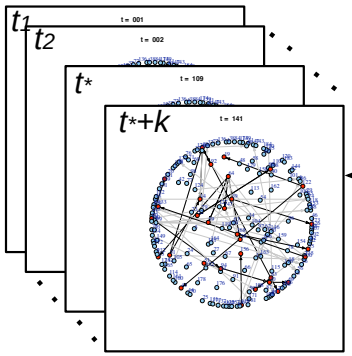


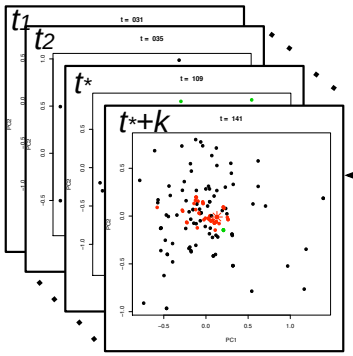
$\Omega(N_{k^*}[v^*;t^*];t^*)$




Finally, tracking vertices incident to msgs $C_t^*$, denoted $V_t^*$, forwards and backwards in time from $t^*$ provides genesis and outcome of topic of interest $C_j$.

Cluster msgs $M_t' = M_t \cup \{S_j\}$. That cluster $C_t^*$ s.t. $S_j \in C_t^*$ is the collection of msgs at time $t$ associated with $S_j$.



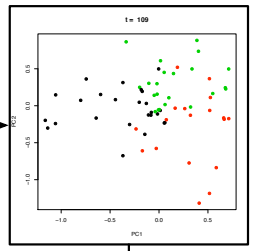




## content analysis

Cluster msgs $M_{t^*}$

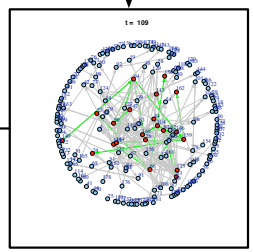
$E_{t^*} = E(\Omega(N_{k^*}[v^*;t^*];t^*))$

$M_{t^*} = \{\text{msgs } m_{vw} \text{ twixt } v\&w$ at time $t^*$ s.t. $vw \in E_{t^*}\}$




Choose cluster $C_j$, a collection of msgs of interest which make up (part of) our detection – aka, a "topic".

$S_j = \text{summary}(C_j)$

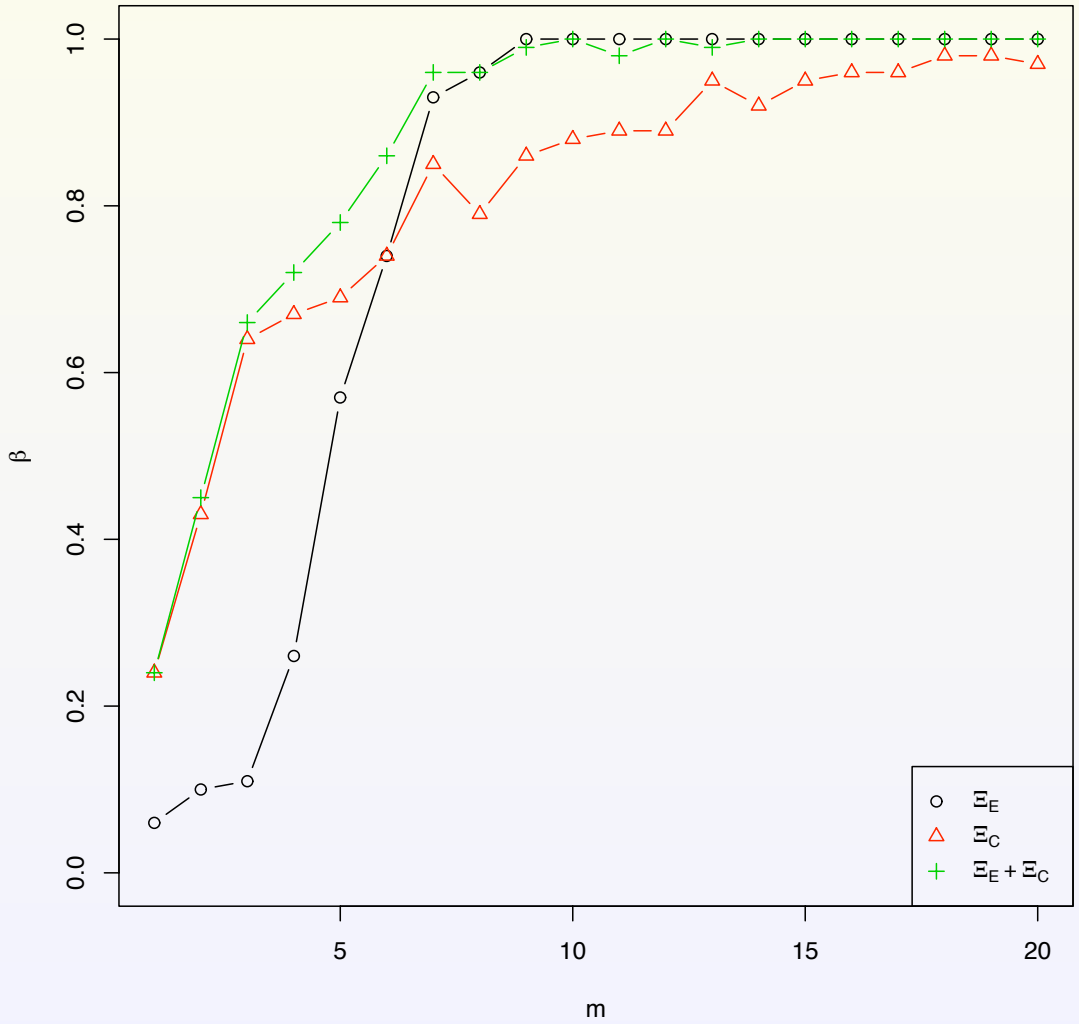



$\Omega(\{v \text{ s.t. msg } m_{vw} \in C_j\};t^*)$

# Streaming Content in Context

1000 Monte Carlo replicates at $m = 6$ yields statistical significance:

$$\widehat{\beta} \approx 0.88 > \max\{\widehat{\beta}_1 \approx 0.78, \widehat{\beta}_2 \approx 0.76\}$$



## Leopold Kronecker to Hermann von Helmholtz:

*"The wealth of your practical experience with sane and interesting problems will give to mathematics a new direction and a new impetus."*


Leopold Kronecker

Hermann von Helmholtz