Mining MR Image Data by Discriminative Methods for the Diagnosis of Dementia

Ceyhun Burak Akgül
Former Marie Curie Postdoctoral Fellow @ Philips Research

www.cba-research.com
cb.akgul@gmail.com

Motivation – 1/2

Diagnose dementia (e.g., Alzheimer's disease) from MR Images

Standard medical practice:

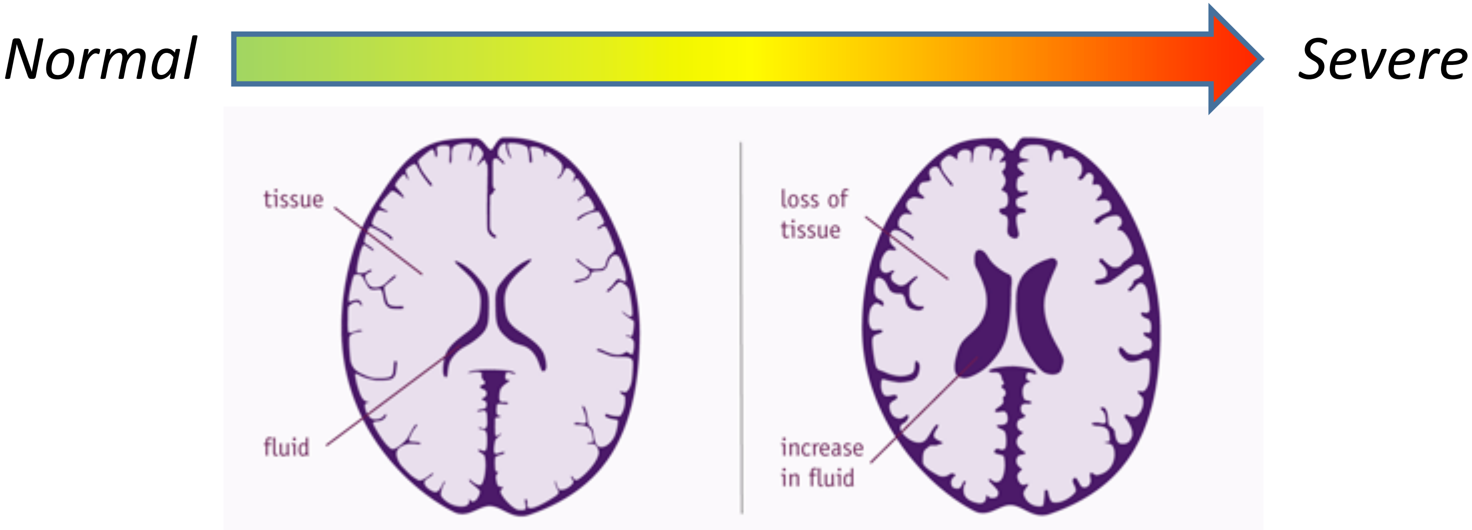
- ❑ patient history, collateral history from relatives
- ❑ clinical observations: neurological/neuropsychological features

BUT: does not often lead to <u>an early diagnosis</u>

An emerging trend: Exploit imaging data
HOW?

Motivation – 2/2

Brain Atrophy?



- requires longitudinal data: MR scans at different time stamps
- requires complex mathematical modeling and algorithms
- should quantify minute changes (that human eye can't see)

Or something else…

Data Mining Framework

Representation

learn an image representation from data: analyze images

- at each location
- at several scales
- with several patterns

Selection

discover image features using labeled data

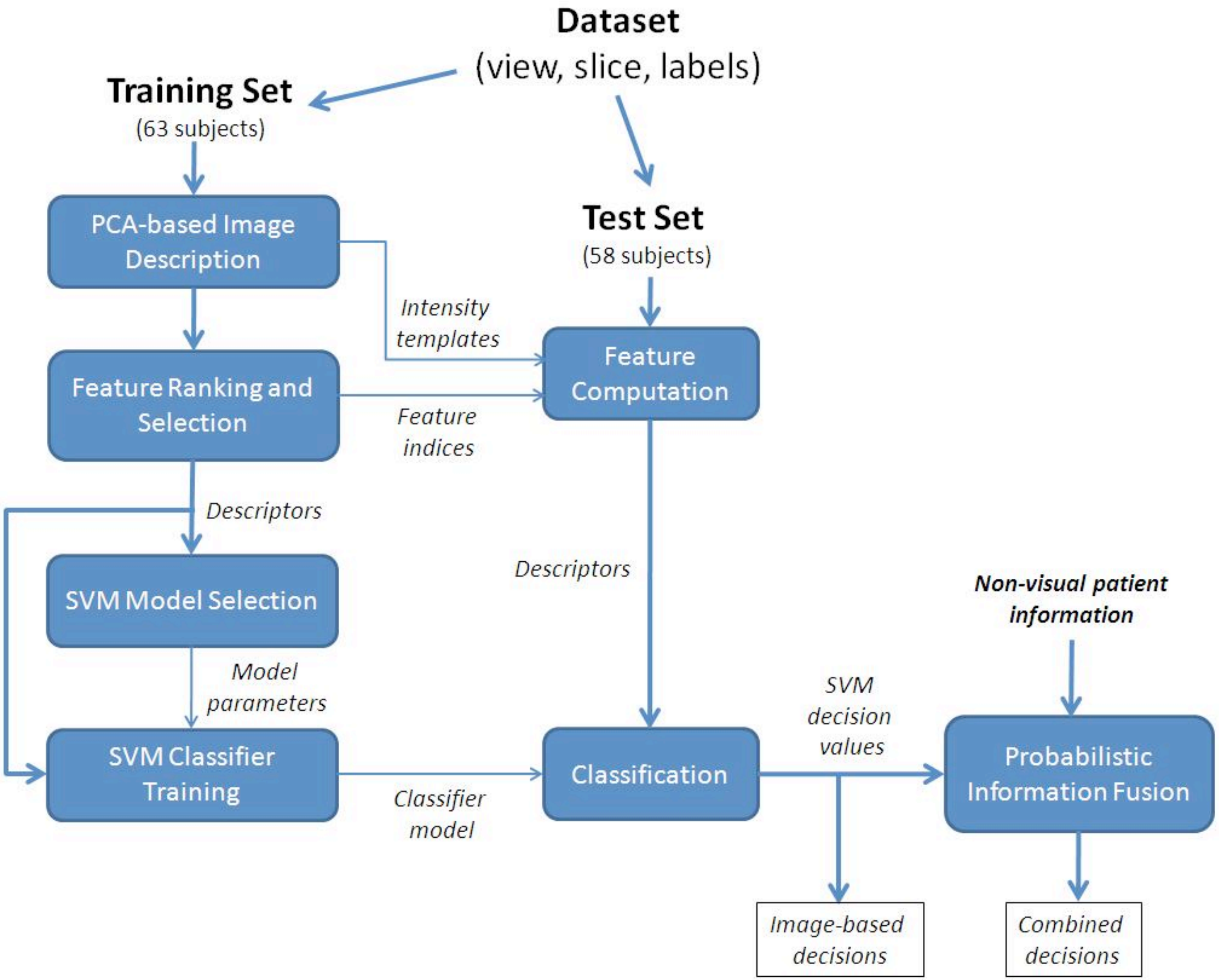
Classification

characterize patient groups discriminatively

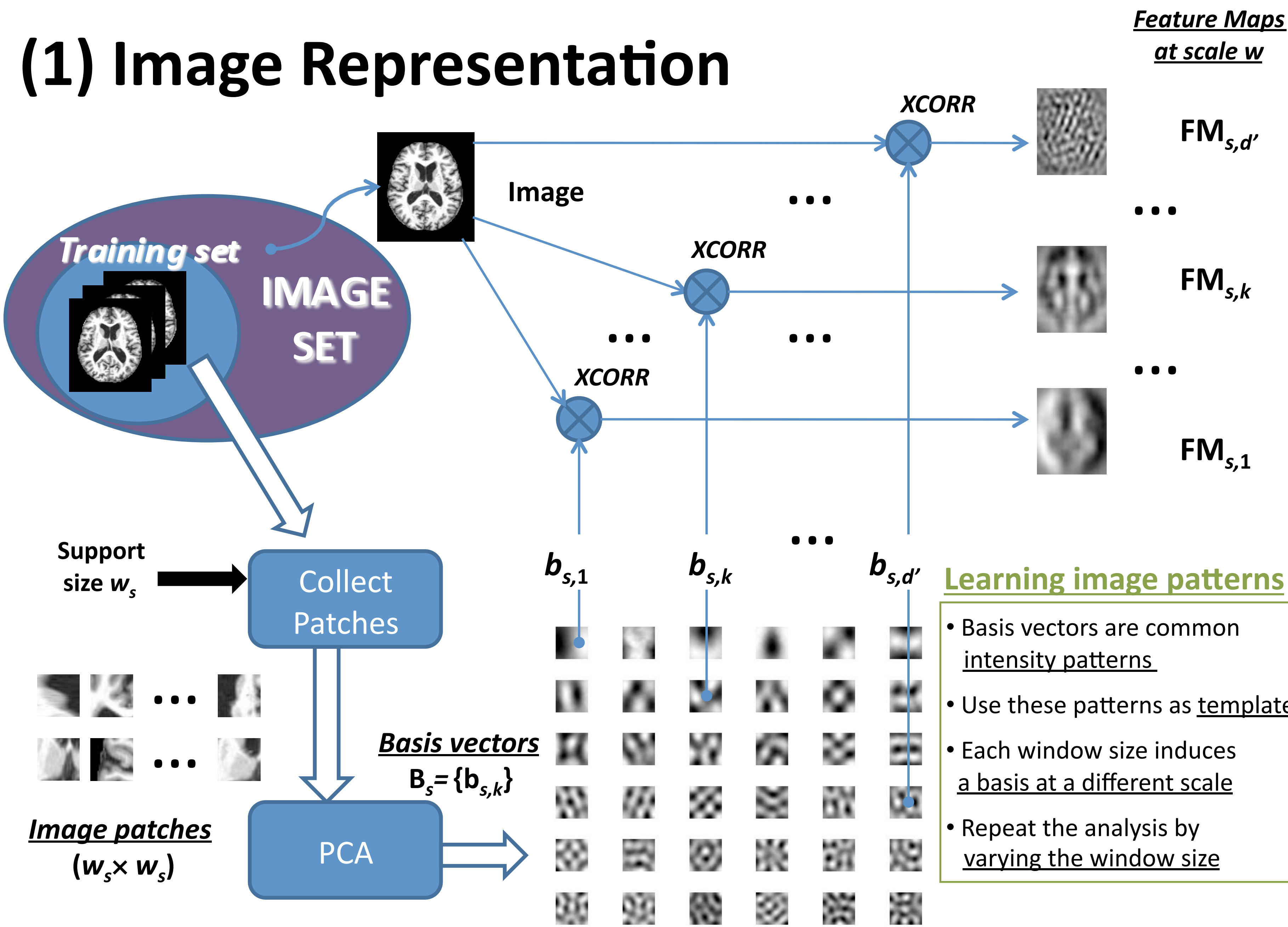
Information Fusion

combine multiple (visual or non-visual) information sources

Data Mining Framework: Overview



(1) Image Representation



Learning image patterns

- Basis vectors are common <u>intensity patterns</u>
- Use these patterns as <u>templates</u>
- Each window size induces <u>a basis at a different scale</u>
- Repeat the analysis by <u>varying the window size</u>

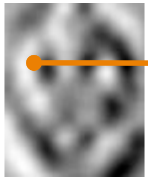
(2) Feature Selection by Ranking – 1/3

Feature Maps

***Feature value at location** (i,j)*

y = +1

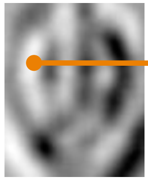
$FM^{(1)}_{s,k}$



$X^{(1)}_{s,k}(i,j)$

y = -1

$FM^{(2)}_{s,k}$

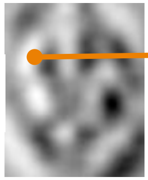


$X^{(2)}_{s,k}(i,j)$

...

y = +1

$FM^{(N)}_{s,k}$



$X^{(N)}_{s,k}(i,j)$

Scale IDs: s=1,…,S
Basis (Pattern) IDs: k=1,…,K_s
***Subject IDs*: n=1,…,N**

- Each image is described by $S \times K_s$ feature maps
- **At each pixel location, there are $S \times K_s$ feature values**
- Each feature $\leftrightarrow$ a distinct (scale, template)-pair
- At each location:
    - rank the features based on their **"usefulness"**
    - pick **the most "useful" feature** for description

"Usefulness"
$\leftrightarrow$
Mutual information
between feature and diagnostic label

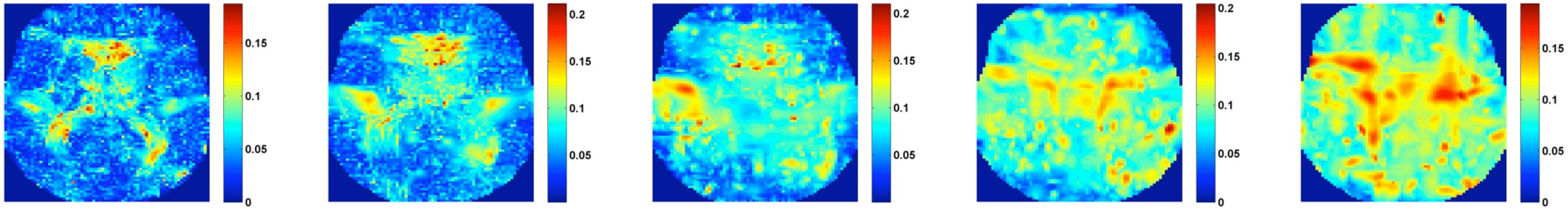
$$MI(x,y) = \sum_{y \in \{-1,+1\}} \int p(x,y) log \frac{p(x,y)}{p(x)p(y)}$$

$x \in [0,1]: normalized\ feature\ value\ at\ location\ (i,j)$
$y \in \{-1,+1\}: diagnostic\ label\ of\ the\ image$

(2) Feature Selection by Ranking – 2/3

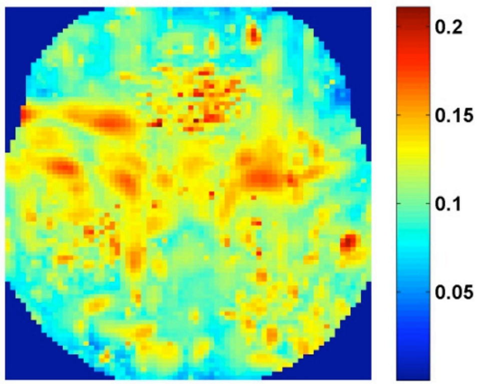
Maximum Mutual Information Maps(*) at Different Scales

Scale 1 | **Scale 2** | **Scale 3** | **Scale 4** | **Scale 5**



***w* = 5** | ***w* = 11** | ***w* = 21** | ***w* = 31** | ***w* = 41**

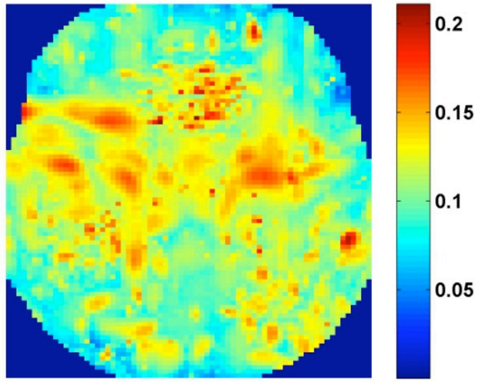
Maximum Mutual Information Map(*)
Combined over Scales



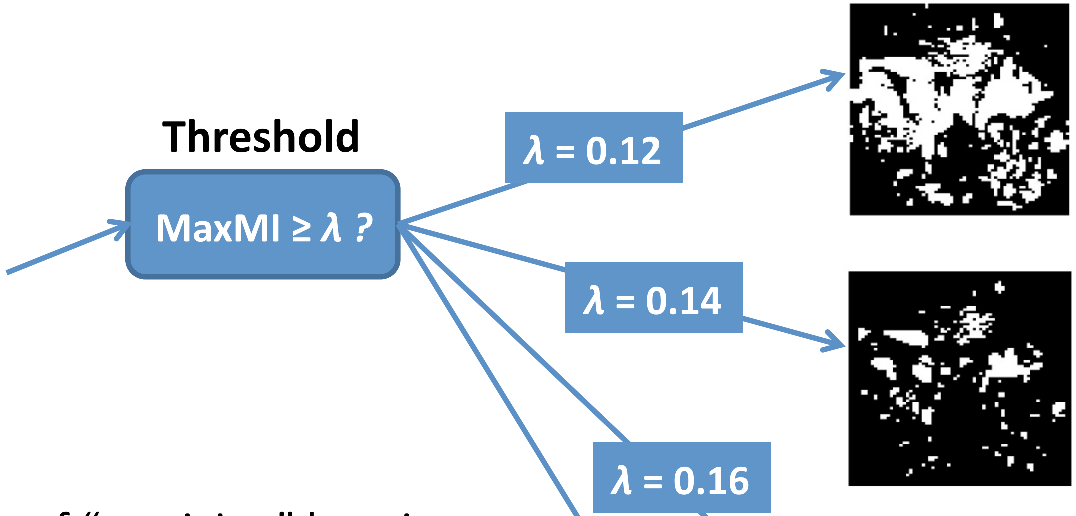
(*) Map size: 87 ×70

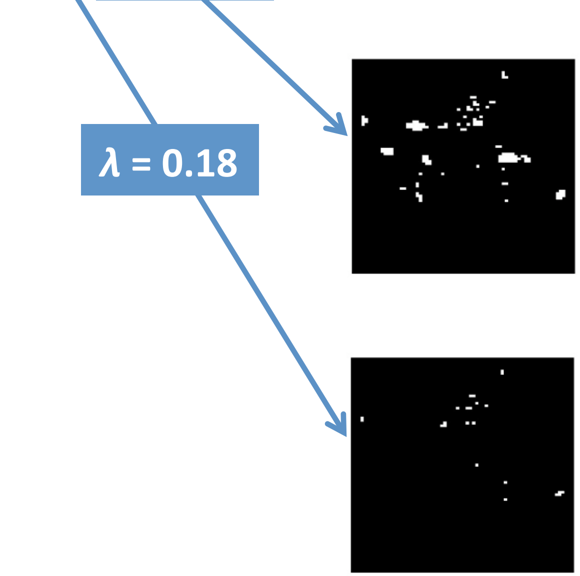
(2) Feature Selection by Ranking – 3/3

MaxMI Map



Threshold



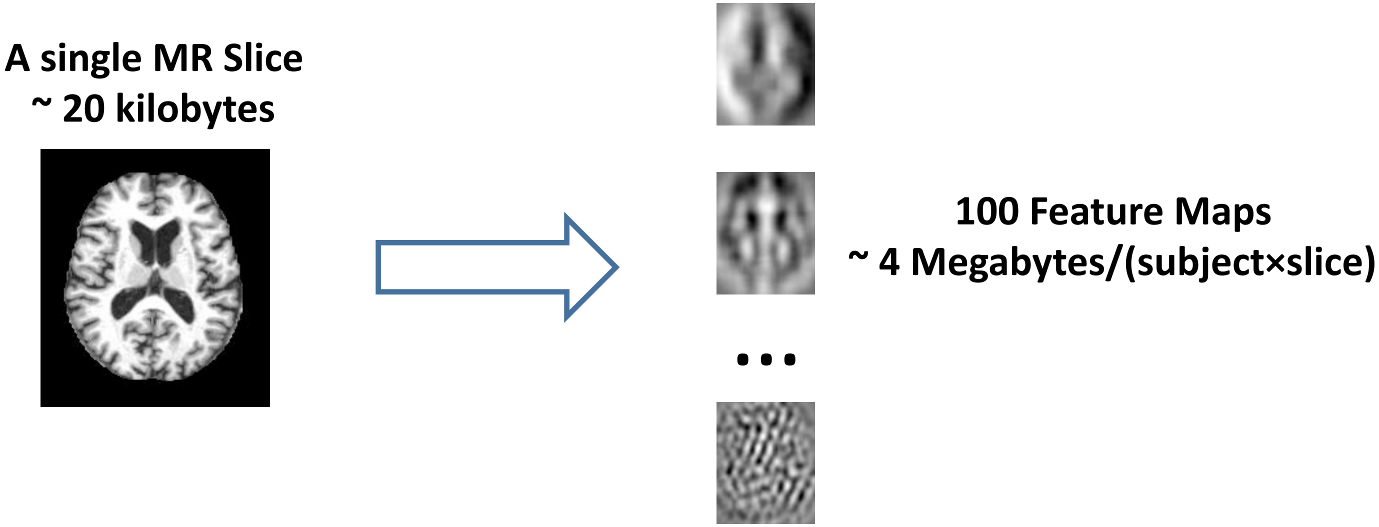


- Each λ-value induces a subset of "surviving" locations
- Remember
  - Each location is associated <u>with a (scale, basis)-pair whose feature value gives the maximum MI.</u>

→ By varying λ, one obtains <u>**nested subsets of features**</u>

- <u>**Or alternatively, retain top *k* features**</u>

Amount of Data Processed: Some Facts



- 100 slices per subject **~ 400 Megabytes/subject**
- 121 subjects **~ 50 Gigabytes** **<u>TOTAL AMOUNT OF DATA PROCESSED</u>**

- 100 informative features/(subject×slice) selected as the descriptor
  <u>< 1 kilobyte/(subject×slice)</u>

(3) Classification: SVM Basics

A non-linear SVM classifier *F* is indexed by two parameters (*C*,*γ*):

- The parameter *C* trades off training error vs. classifier complexity
- The kernel parameter *γ* determines the class of functions *F* and affects class separation

  (in some sense, it also determines the classifier complexity)

One has to specify the "best" (*C*,*γ*)-pair before testing the classifier.

A good empirical option

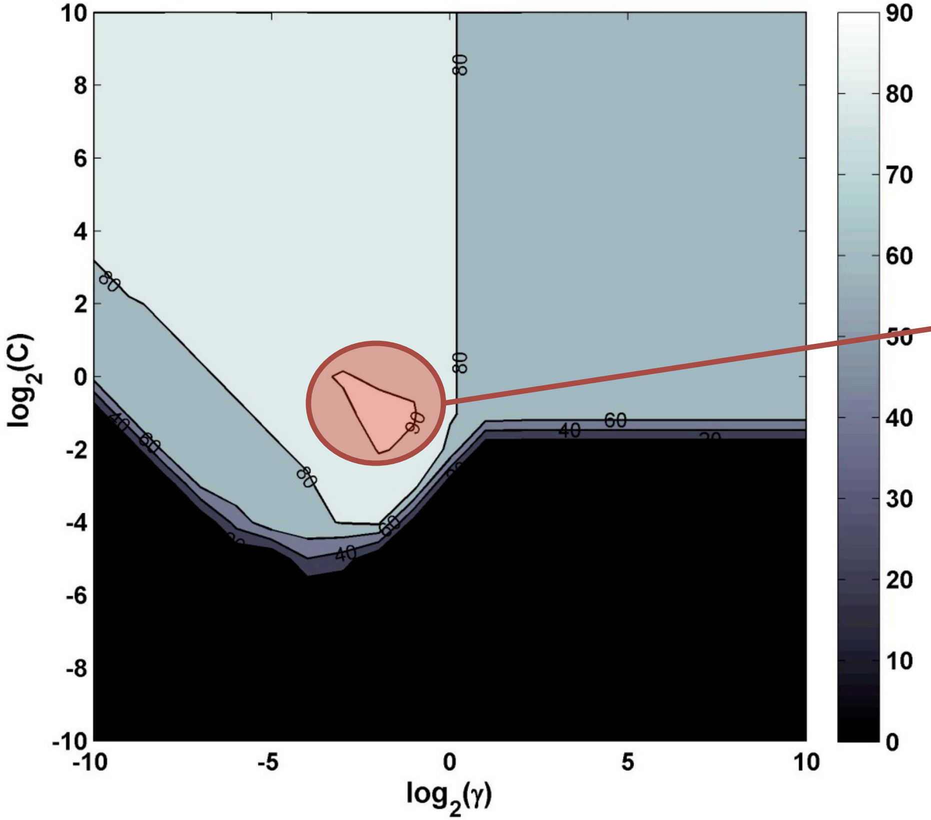
$$(C,\gamma)^* = argmin\, Err_{CV}\big(F(C,\gamma)\big)$$

$$Err_{CV}: Cross\ validation\ error$$

(3) Classification: Model Selection

- Leave-One-Out (LOO) cross-validation
- Initial search for the (C,γ)-parameters on a coarse grid

(C,γ)-Contours of LOO Accuracy (%)



{(C,γ) such that ACC>THRESH}

- Search on a finer grid
- Further heuristics – Look at:
  - Sensitivity
  - Fraction of SVs (model parsimony)
  - Specificity

(4) Probabilistic Information Fusion

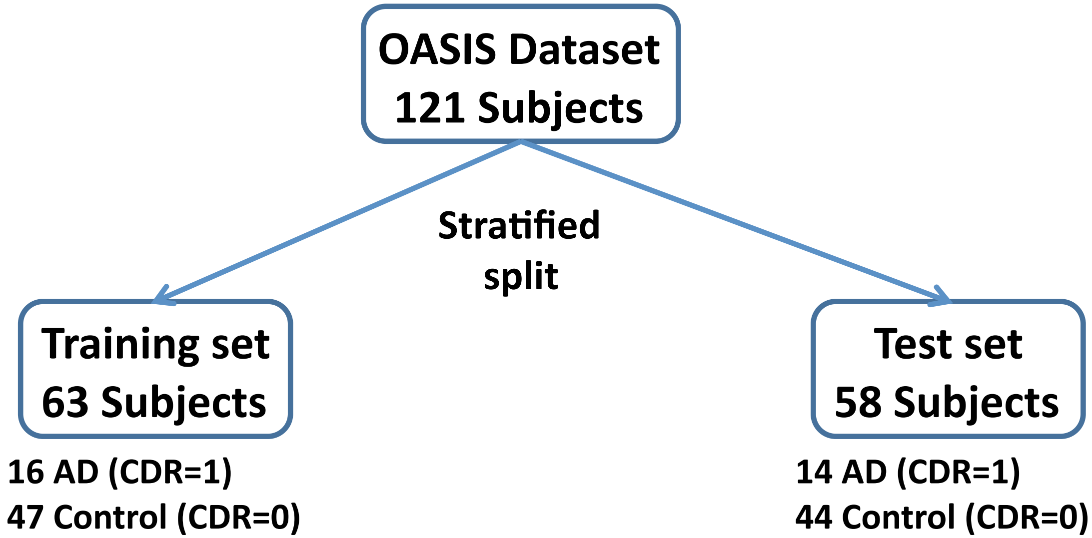
Bayesian Theory: The decision on the class label should be made on the conditional probability of the class label given all other relevant information.

Cognitive test scores, e.g., MMSE
Age
Gender
Genetics …

P(label | info) = P(label | visual, non-visual)

$$
\begin{aligned}
P(\text{label} \mid \text{info}) \;&\alpha\; P(\text{label}) \times \mathbf{P(\text{visual, non-visual} \mid \text{label})} \\
&= \mathbf{P(\text{label})} \times \mathbf{P(\text{visual} \mid \text{label})} \times \mathbf{P(\text{non-visual} \mid \text{label})} \\
&\alpha\; \underbrace{\mathbf{P(\text{label} \mid \text{visual})}}_{\text{derived from SVM outputs}} \times \underbrace{\mathbf{P(\text{non-visual} \mid \text{label})}}_{\text{class-conditional distributions estimated from training data}}
\end{aligned}
$$

Experiments: Dataset

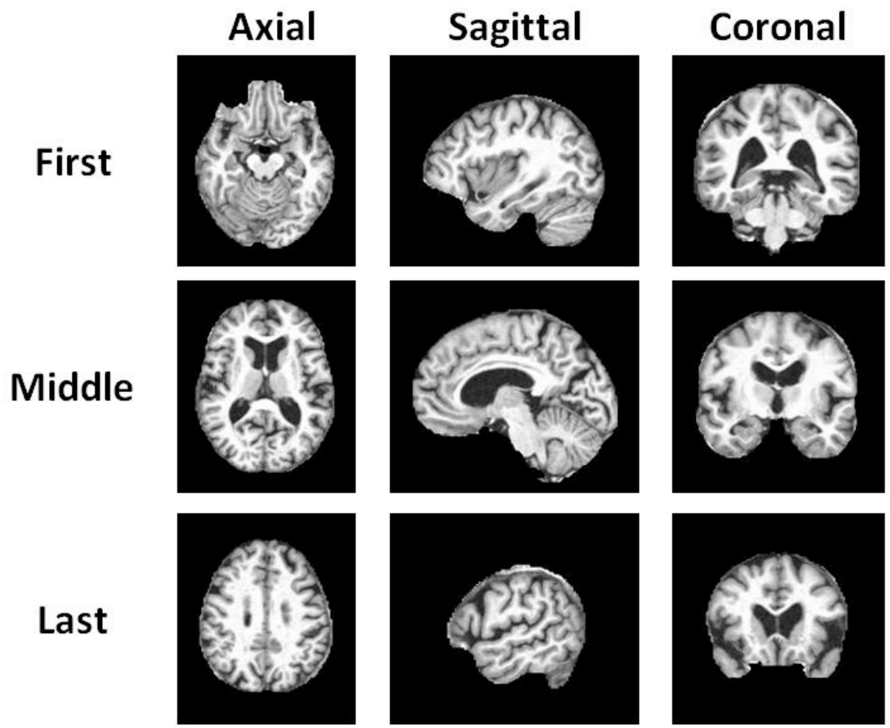


- PCA-based feature learning
- Feature ranking and filtering
- Discriminative learning (SVM)
  - SVM model selection

- Not seen during none of the training stages
- Reserved only for performance evaluation

- **CDR: Clinical Dementia Rating:** **normal → CDR = 0** **moderate dementia → CDR =1**
- **Stratified split keeps the class proportions the same in both sets (Control/AD ≈ 3)**

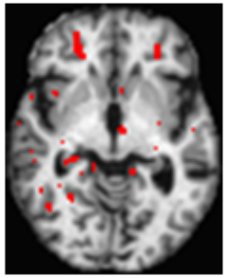
Experiments: MR Data



- 26 Axial + 46 Sagittal + 28 Coronal = <u>100 MR slices processed separately</u>
- Each slice described by <u>100 informative image features</u>

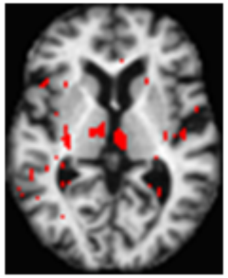
Experiments: Discriminative Slices – 1/2

Axial 6



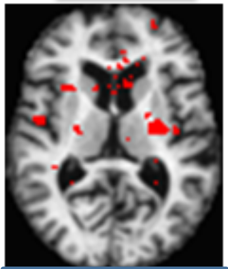
Acc = 70.7%
Sens = 64.3%
Spec = 72.7%

Axial 10



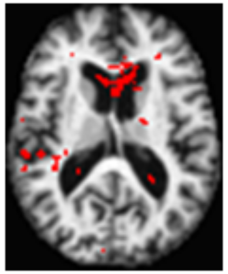
Acc = 79.9%
Sens = 71.4%
Spec = 81.8%

Axial 12



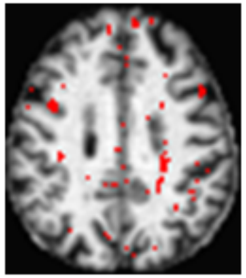
Acc = 84.5%
Sens = 78.6%
Spec = 86.4%

Axial 15



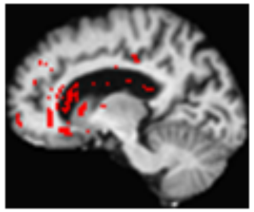
Acc = 72.4%
Sens = 64.3%
Spec = 75.0%

Axial 26



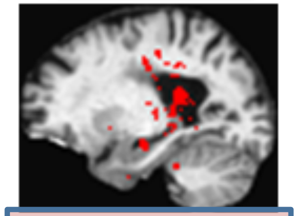
Acc = 81.0%
Sens = 64.3%
Spec = 86.4%

Sagittal 26



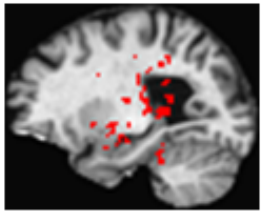
Acc = 67.2%
Sens = 57.1%
Spec = 70.5%

Sagittal 32



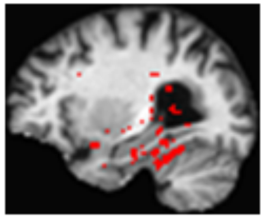
Acc = 79.3%
Sens = 71.4%
Spec = 81.8%

Sagittal 33



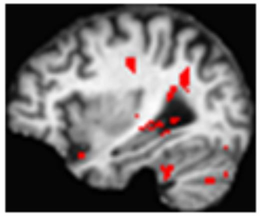
Acc = 77.6%
Sens = 64.3%
Spec = 81.8%

Sagittal 35



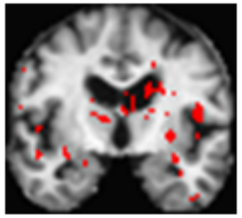
Acc = 84.5%
Sens = 64.3%
Spec = 90.9%

Sagittal 37



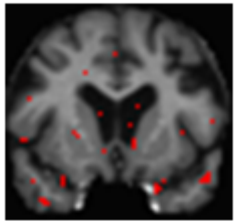
Acc = 75.9%
Sens = 57.1%
Spec = 81.8%

Coronal 15



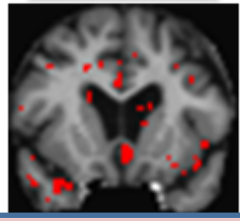
Acc = 65.5%
Sens = 57.1%
Spec = 68.2%

Coronal 25



Acc = 72.4%
Sens = 57.1%
Spec = 77.3%

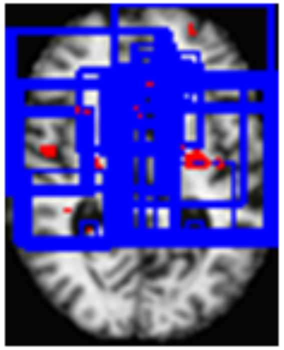
Coronal 26



Acc = 81.0%
Sens = 57.1%
Spec = 88.6%

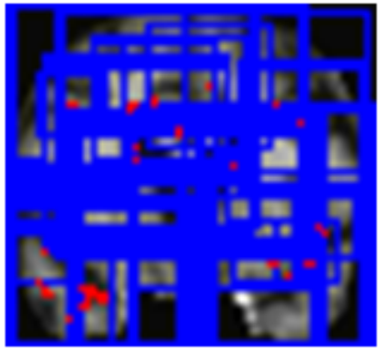
Experiments: Discriminative Slices – 2/2

Axial 12



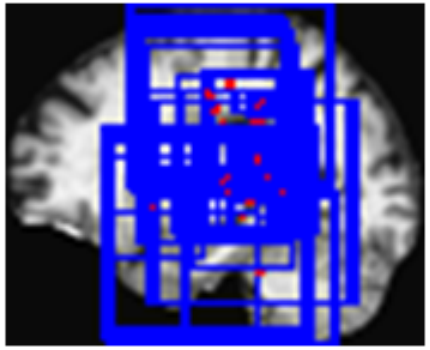
Acc = 84.5%
Sens = 78.6%
Spec = 86.4%

Coronal 26



Acc = 81.0%
Sens = 57.1%
Spec = 88.6%

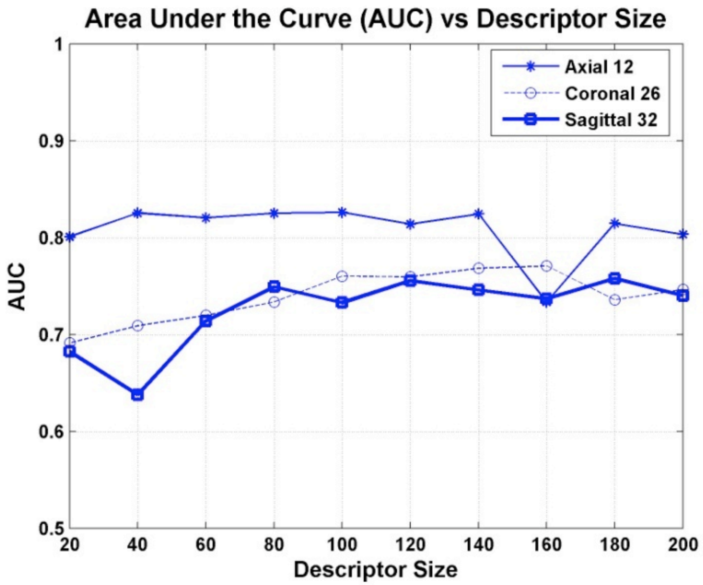
Sagittal 32

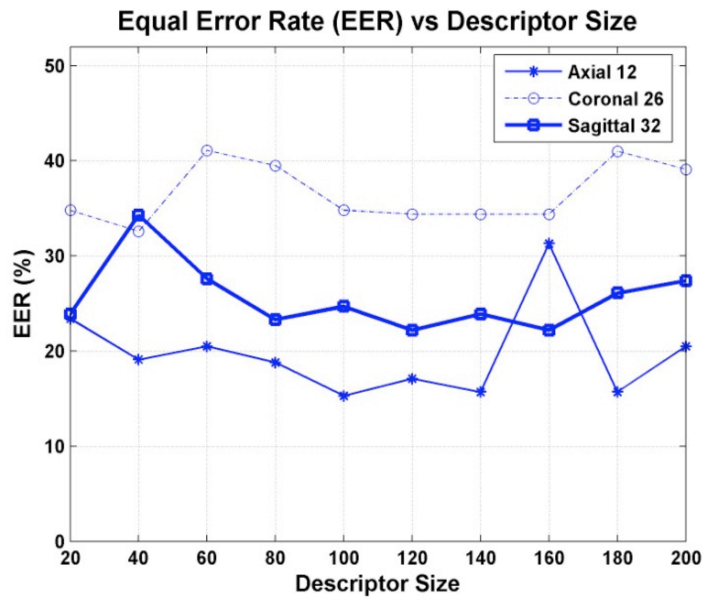


Acc = 79.3%
Sens = 71.4%
Spec = 81.8%

Axial 12 > Sagittal 32 > Coronal 26

Experiments: ROC vs. Descriptor Size

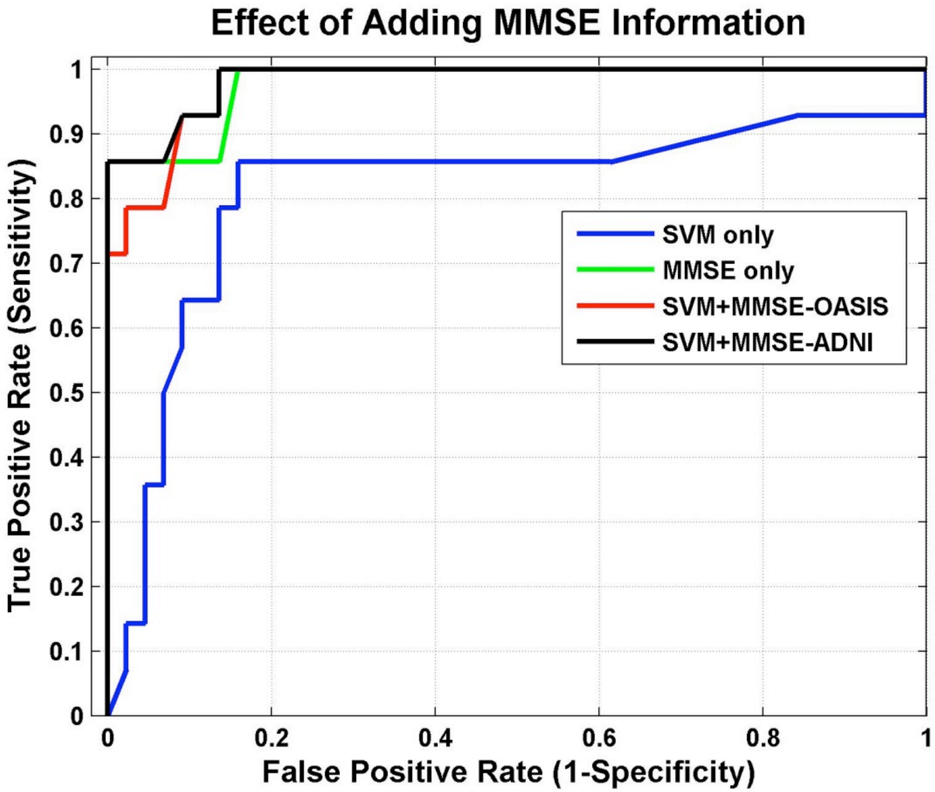




	Descriptor Size	Accuracy (%)	Sensitivity (%)	Specificity (%)
Axial 12	**100**	**84.5 (49/58)**	**85.7 (12/14)**	**84.1 (37/44)**
Sagittal 32	160	65.5 (38/58)	64.3 (9/14)	65.9 (29/44)
Coronal 25	160	77.6 (45/58)	78.6 (11/14)	77.3 (34/44)

ROC: Receiver Operating Characteristic: TPR vs. FPR
AUC: Area under the ROC curve
EER: Equal error rate (sensitivity = specificity)

Experiments: Information Fusion – 1/2



SVM-only: Image-based decisions gleaned from SVM outputs

MMSE-only: MMSE-based decisions: *<u>if MMSE<Thresh, then decide ill</u>*

SVM+MMSE-OASIS: statistics estimated from OASIS training set (63 subjects)

SVM+MMSE-ADNI: statistics estimated from ADNI dataset (322 subjects)

Experiments: Information Fusion – 2/2

ROC Summary

	AUC	EER (%)	Accuracy (%)
SVM only	0.8260	15.3	84.5
MMSE only	0.9798	13.3	86.7
SVM+MMSE-OASIS	**0.9798**	**8.7**	**91.3**
SVM+MMSE-ADNI	**0.9871**	**8.4**	**91.6**

SVM+MMSE-ADNI > SVM+MMSE-OASIS > MMSE-only > SVM-only

- Information fusion is very useful indeed
- Reliable statistics!!! ADNI (322 subjects) > OASIS (63 subjects)

229 controls **47 controls**

93 positives **16 positives**

Summary

- Data-driven image representation
  - Unsupervised learning of local image patterns via PCA
  - Localized, at several scales, with several patterns
- Feature ranking and filtering
  - Supervised: based on MI between scalar features and class labels
- Discriminative learning
  - SVM model selection via cross-validation and further heuristics
- Information fusion
  - Leverage image-only decisions by non-visual information
  - Generic: works with any kind of meta-data as long as statistics available

Proof of concept:

A promising data-driven framework for the diagnosis of dementia with high predictive performance

What's Next?

Practical

Go validate these results clinically

Do these slices, locations, scales, patterns make sense?

Acquire larger sets of labeled data

Allocate higher computational resources

Methodological

Other sparser image representations: ICA-based? NNMF-based?

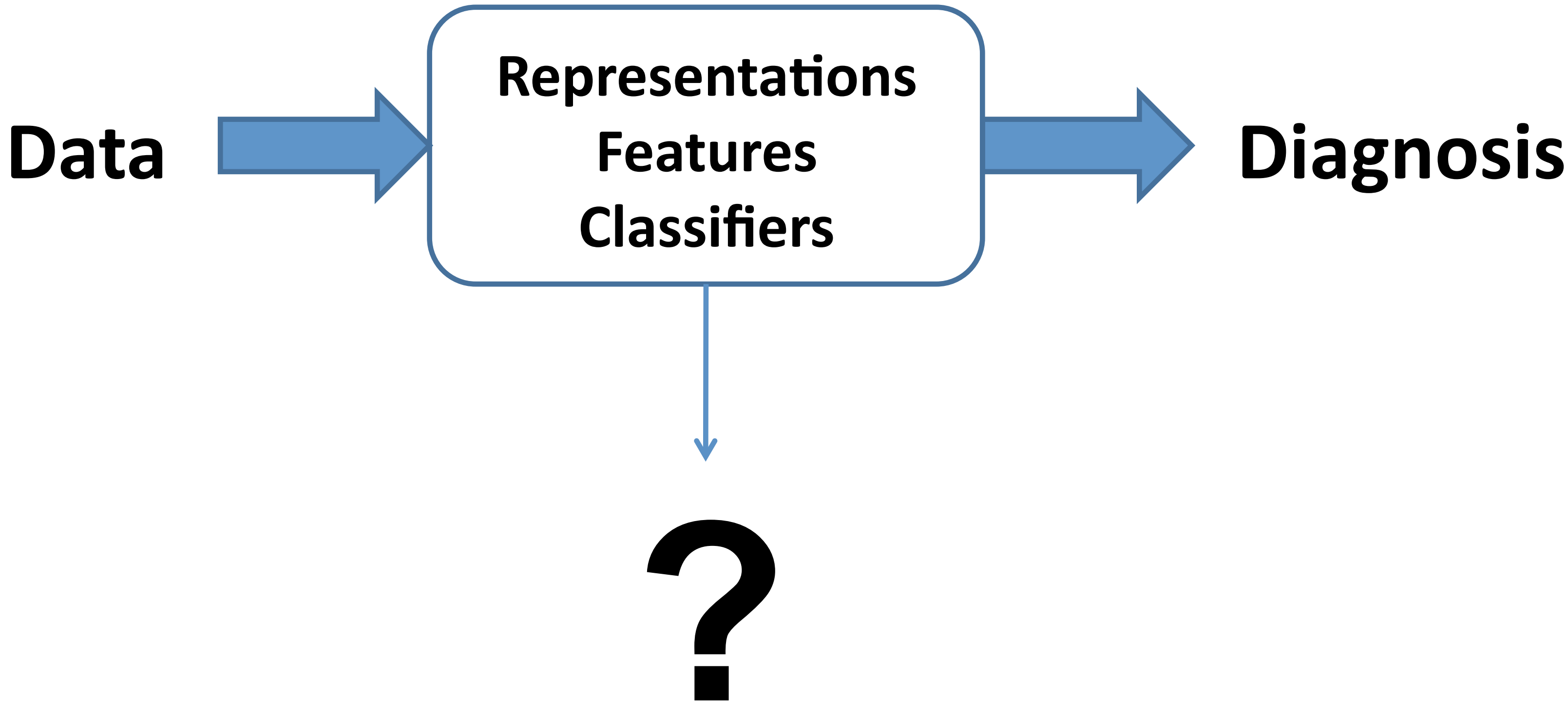
Multivariate feature selection

Model selection: Don't use one, average multiple models

Other classification schemes: AdaBoost

Theoretical …

What's Next? – *Theoretical*



convolutional networks?

To conclude…

There's nothing more practical than a good theory.

Lewin, 1952