

Online Microarray Analysis Tool using a modified support vector machine (MSVM)

Committee:

Dr. Rosemary Renaut

Dr. Huan Liu

Dr. Hongbin Guo

Student:

Wang-Juh Chen (Sting)

May 9, 2005

Outline

- ✓ Goals of internship
- ✓ Introduction and overview
- ✓ Significance
- ✓ Datasets
- ✓ Methods and Experiments
- ✓ Results and Discussion
- ✓ Conclusions
- ✓ Demo
- ✓ Future work

- ✓ **Goals of internship**
- ✓ Introduction and overview
- ✓ Significance
- ✓ Datasets
- ✓ Methods and Experiments
- ✓ Results and Discussion
- ✓ Conclusions
- ✓ Demo
- ✓ Future work

Goals of internship

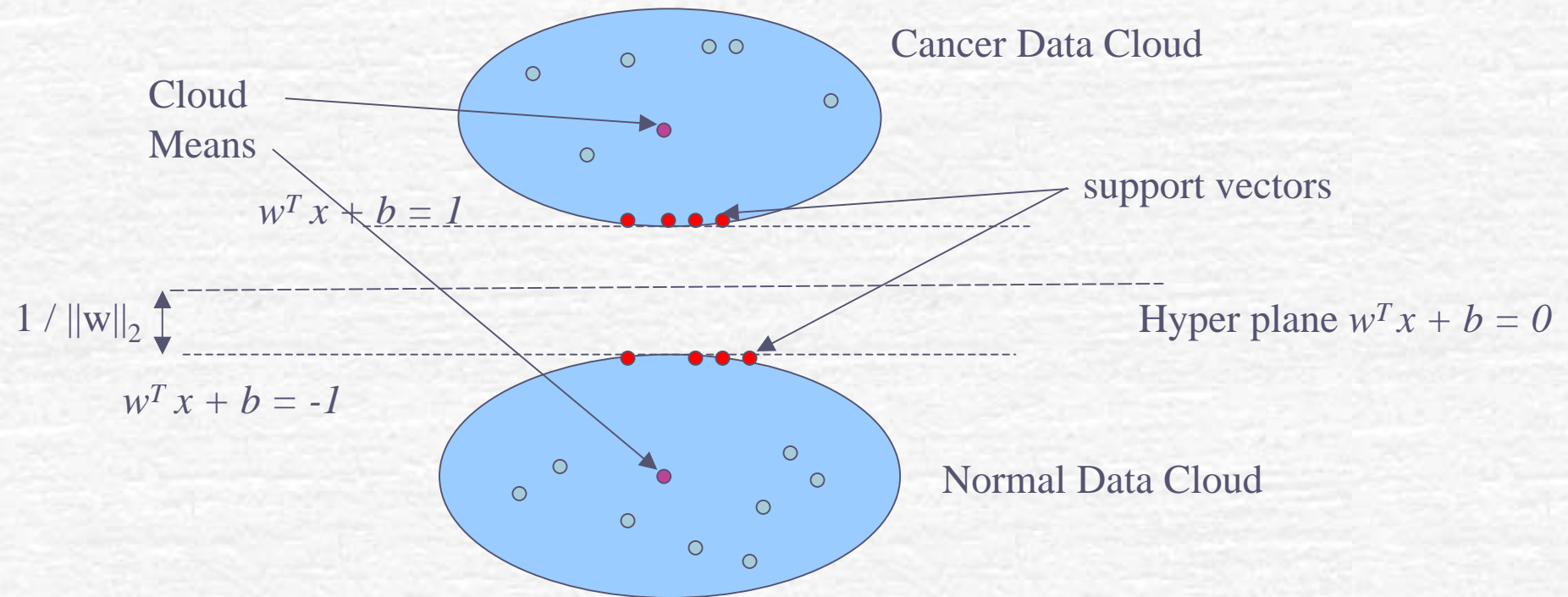
- ☛ Develop an online tool to analyze Microarray data.
- ☛ Implement a modified support vector machine (MSVM) for classifying Microarray data more accurately and apply CLUSTER platform and online submission for the service.

- ✓ Goals of internship
- ✓ **Introduction and overview**
- ✓ Significance
- ✓ Datasets
- ✓ Methods and Experiments
- ✓ Results and Discussion
- ✓ Conclusions
- ✓ Demo
- ✓ Future work

Microarray II

- ☛ The methods for analysis of Microarray
 - Statistics : T-test, FDR (False Discovery Rate)
 - Cluster – k-means
 - Classification
 - Decision Tree
 - Bayesian Network
 - Support Vector Machine

Support Vector Machine I - SVM Hyperspace



SVM II

- Standard SVM

$$\begin{aligned} \min_{w, b, \xi} J_p(w, \xi) &= \frac{1}{2} \|w\|_2^2 + \lambda \sum_{i=1}^m \xi^i, \\ \text{subject to } t^i [w^T \phi(c^i) + b] &\geq 1 - \xi^i, \\ \xi^i &\geq 0, i = 1, \dots, m, \end{aligned}$$

where ξ^i are *slack variables*

λ is a real positive constant

ϕ maps data to feature space $\phi: X \rightarrow H$

SVM III - MSVM

$$\min_{w, b, E} J_p(w, E) = \frac{\mu}{2} \|w\|_2^2 + \frac{1}{2} \|E\|_F^2,$$

subject to $(A + E)w + bt \geq e,$

where

$$t = (t^1, t^2, \dots, t^m)^T$$

$$A = \text{diag}(t) X$$

$$e = (1, 1, \dots, 1)^T$$

$\|\cdot\|_F$ represents the Frobenious norm.

E errors within the dataset

μ positive parameter

SVM IV - MSVM

$$\mathcal{L}(w, E, b, \alpha) = \frac{\mu}{2} \|w\|_2^2 + \frac{1}{2} \|E\|_F^2 - \alpha^T [(A + E)w - e + bt].$$

The dual problem is
given by

$$\max_{\alpha \geq 0} \min_{w, E, b} \mathcal{L}(w, \varepsilon, b, \alpha).$$

$$\nabla_{\varepsilon} \mathcal{L} = E - \alpha w^T = 0,$$

$$\nabla_w \mathcal{L} = \mu w - (A + E)^T \alpha = 0,$$

$$\frac{\partial \mathcal{L}}{\partial b} = -\alpha^T t = 0.$$

Introduce $\zeta = \frac{1}{\mu - \|\alpha\|_2^2}$

Final dual problem

$$\begin{aligned} \min & \frac{\zeta}{2} \|A^T \alpha\|_2^2 - \sum \alpha_i, \\ \text{subject to} & \alpha^T t = 0, \quad \text{and} \quad \alpha \geq 0. \end{aligned}$$

Computation platform – CLUSTER I

☞ The need for CLUSTER

- Reduce the time during computation
- LOO C.V on lymphoma dataset: 280 sec -> 26 sec

☞ The architecture of CLUSTER platform

● Frontend

- Server.
- Open to public.
- Users login, submission of jobs, code compilation etc..

● Compute nodes

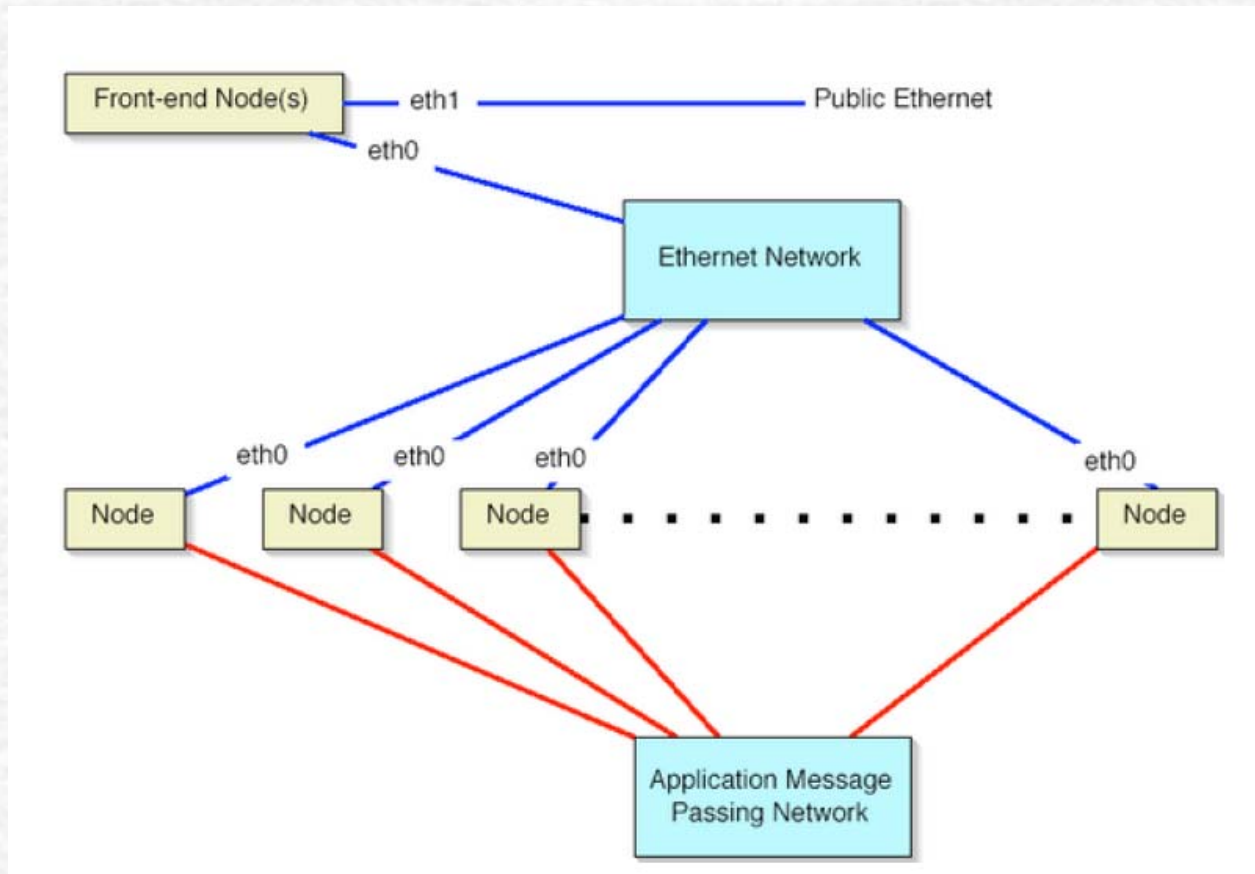
- Slaves or Workes.
- Could not be seen on the public.
- Connect with high-performance network

● Ethernet Network

● Application Message Passing Network

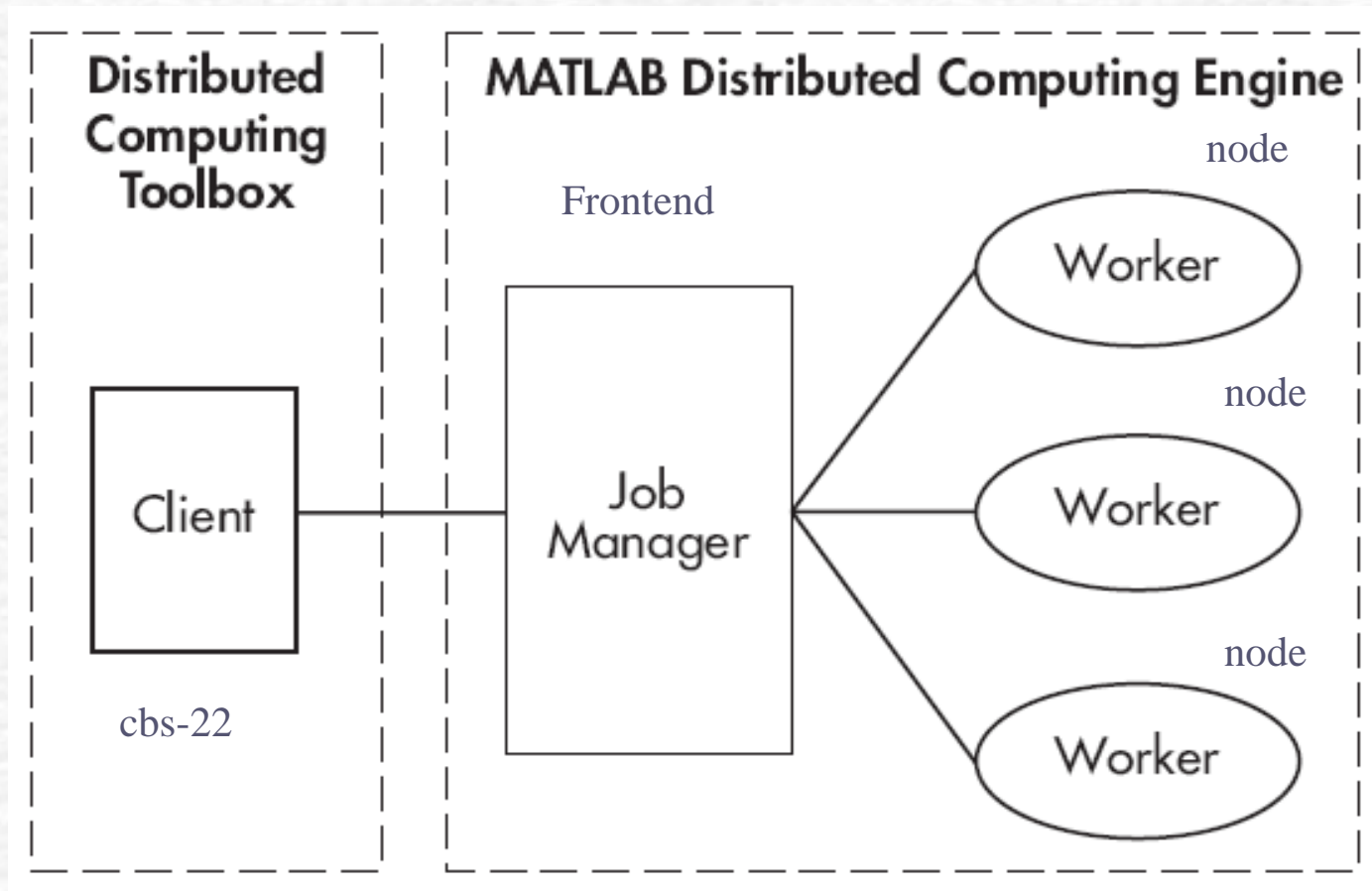
- MPI (MatLab Distributed Computing Engine)

Computation platform – CLUSTER II



The architecture of CLUSTER platform(ROCKS)
Resource: <http://www.rocksCLUSTERs.org/Rocks/>

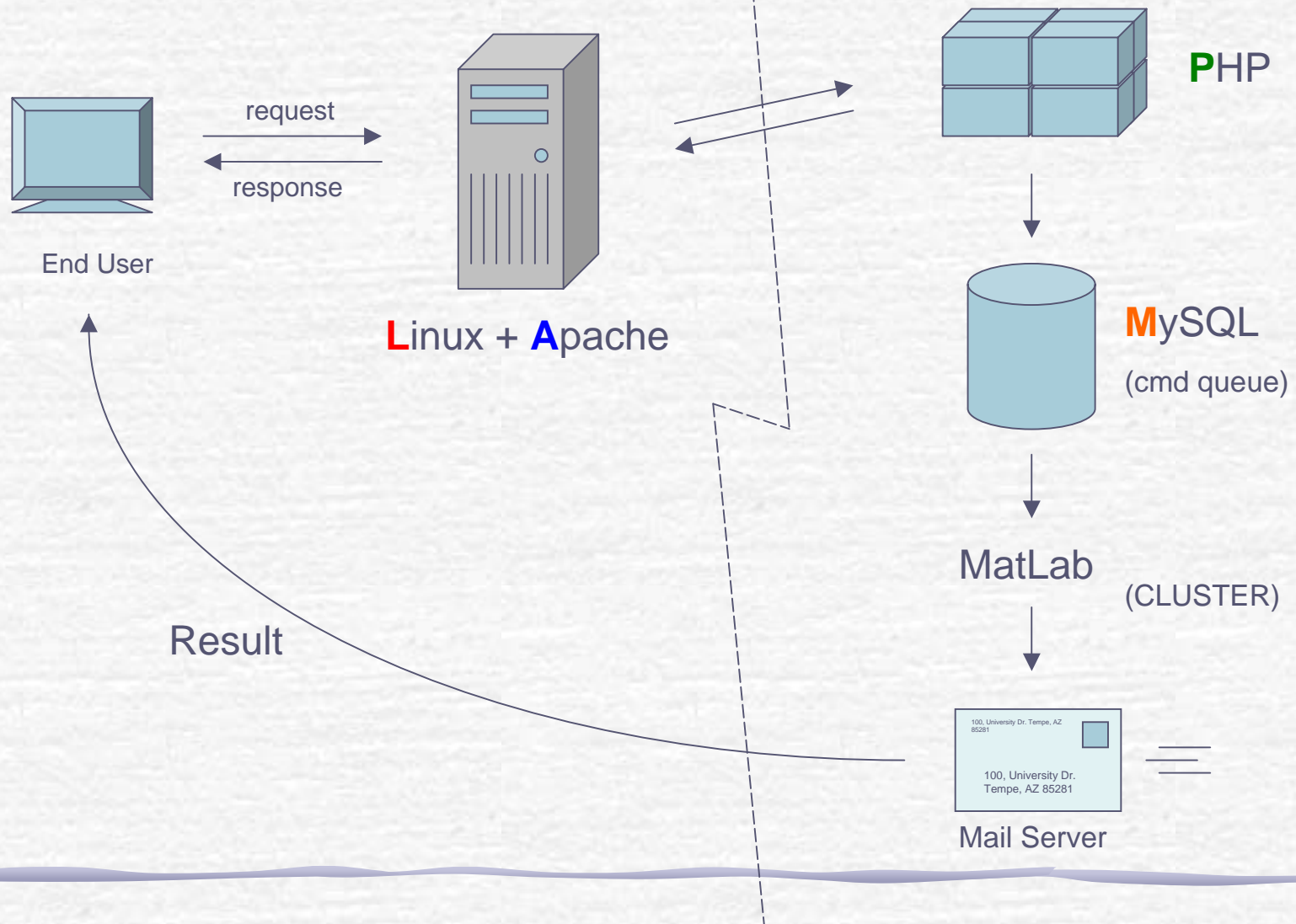
MATLAB Distributed Computing Toolbox and Engine



Basic Distributed Computing Configuration

Resource: © 1994-2005 The MathWorks, Inc

LAMP - Linux-Apache-Mysql-Php



- ✓ Goals of internship
- ✓ Introduction and overview
- ✓ **Significance**
- ✓ Datasets
- ✓ Methods and Experiments
- ✓ Results and Discussion
- ✓ Conclusions
- ✓ Demo
- ✓ Future work

Significance

MSVM

- Assumes noise in data measurement (feature space), and provides another way to extract significant genes (feature selection).

CLUSTER

- Reduces time during the validation part. Easy to implement.

LAMP

- Cross-platform and User-Friendly.

- ✓ Goals of internship
- ✓ Introduction and overview
- ✓ Significance
- ✓ **Datasets**
- ✓ Methods and Experiments
- ✓ Results and Discussion
- ✓ Conclusions
- ✓ Demo
- ✓ Future work

Datasets

	Gene Number	Cancer	Normal	Sample Size
lymphoma Training	7129	11	27	38
lymphoma Testing	7129	14	20	34
ovarian	87558	14	17	31
myeloma	7129	74	31	105

- ☞ Goals of internship
- ☞ Introduction and overview
- ☞ Significance
- ☞ Datasets
- ☞ **Methods and Experiments**
- ☞ Results and Discussion
- ☞ Conclusions
- ☞ Demo
- ☞ Future work

Data preprocessing

Why data preprocessing?

- Remove some genes bases on the knowledge.
- Remove outlier to increase the accuracy.
- Normalization - reduce the differences across datasets and eliminate artifacts.

$$X_j = \frac{X_j}{\bar{X}}$$

Error in the dataset

$$X = (1 + \epsilon * \text{randn}) X.$$

ϵ

perturbed error %

e.g. : 0%, 5%, and 50%

randn :

normal distribution with mean zero

variance one and standard deviation one.

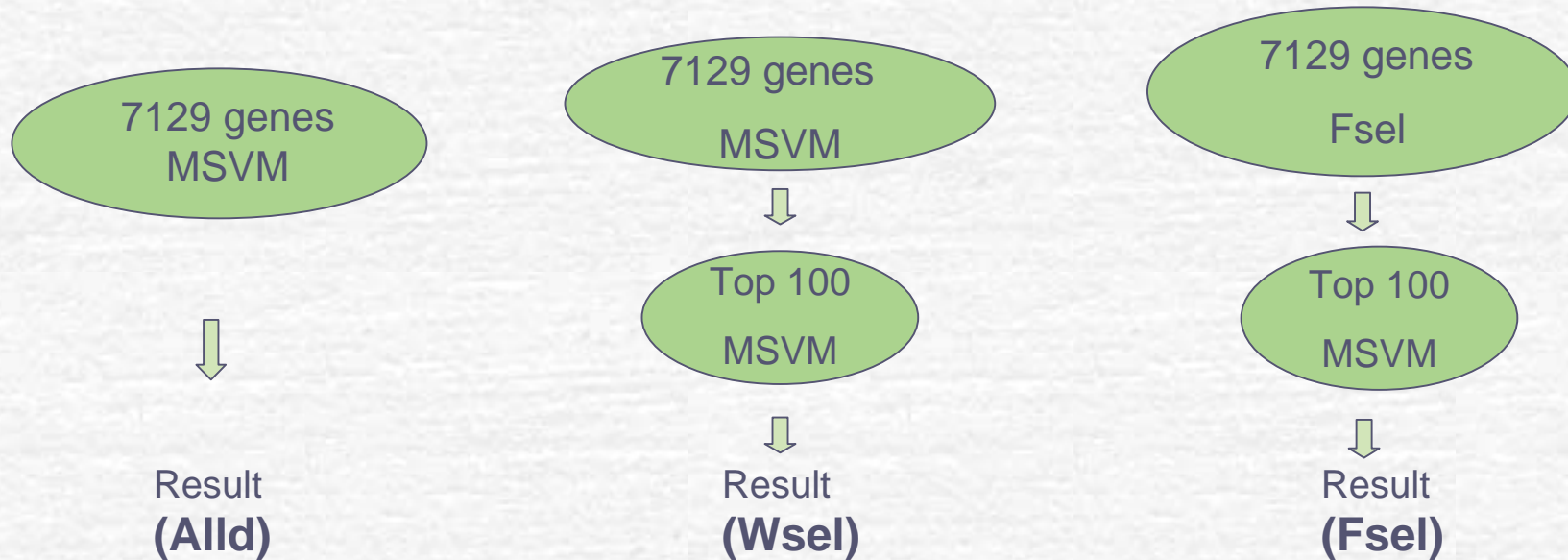
Gene selection - Alll, Wsel, and Fsel

7129 Genes

Top 100 genes

Alll	use all genes to analyze
Wsel	Use Alll to get top 100 genes, then use these for analysis
Fsel	Use F-score to get top 100 genes, then use these for analysis

$$F(x_j) = \left| \frac{\mu_j^+ - \mu_j^-}{\sigma_j^- + \sigma_j^+} \right|$$



Regularization

- ☛ To show the advantage of our proposed algorithm, we compare with the SVM written Steve Gunn (SVM).
- ☛ In order to prevent the problems when the Hessian is badly conditioned
- ☛ SVM : 10^{-10}
- ☛ MSVM :

$$\tau = 10^{-5} \times \|H\|_2,$$

$$H = H + \tau \times I$$

Choosing Support Vectors

- ☛ Use for representing the whole dataset for the future analysis.
- ☛ SVM : 10^{-5}
- ☛ MSVM :

Support Vectors $\in \{ \alpha, |\alpha| \leq \delta \},$

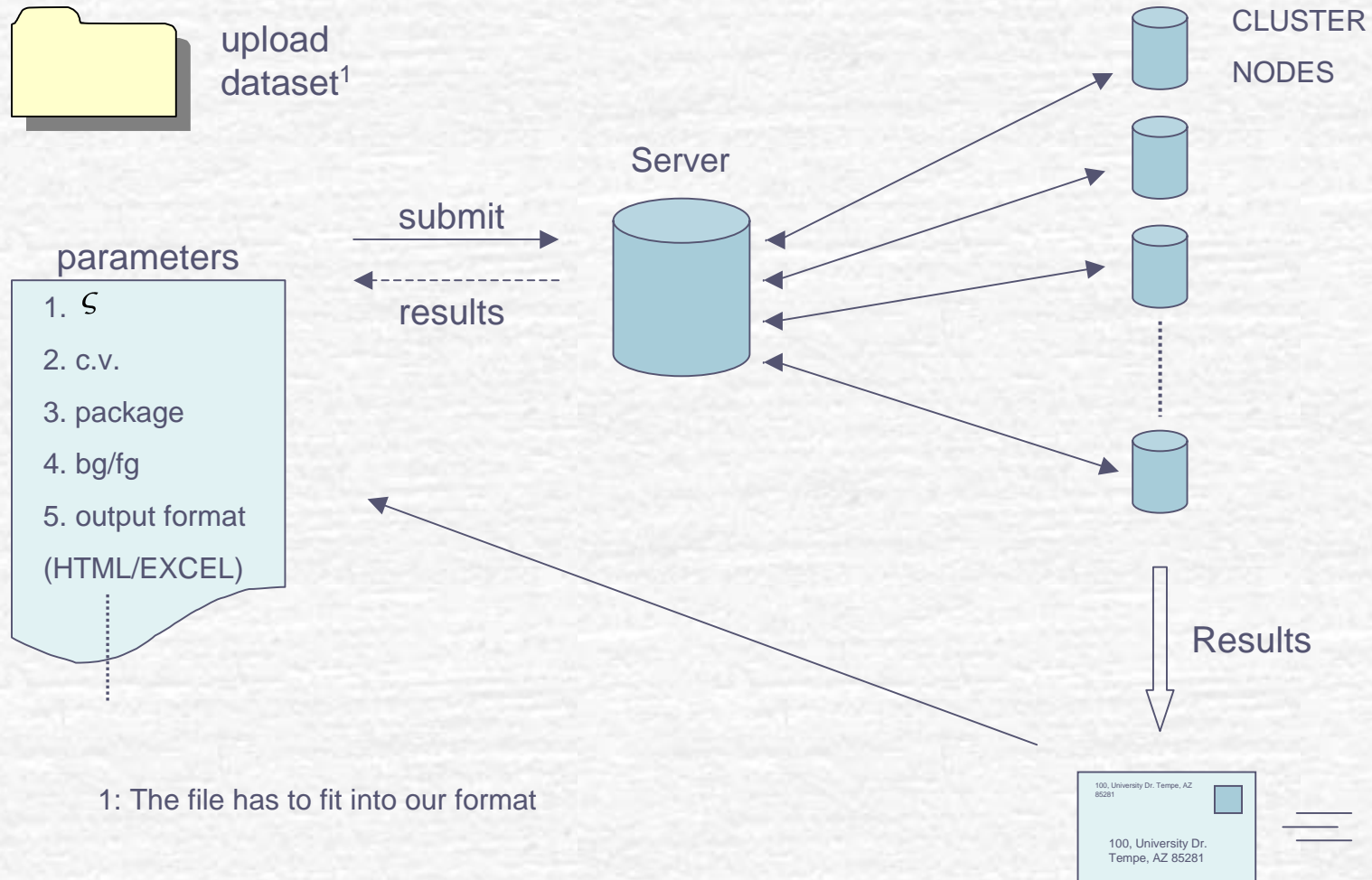
where $\delta = 0.01 \times \max(|\alpha|).$

Outlier and mislabeled data

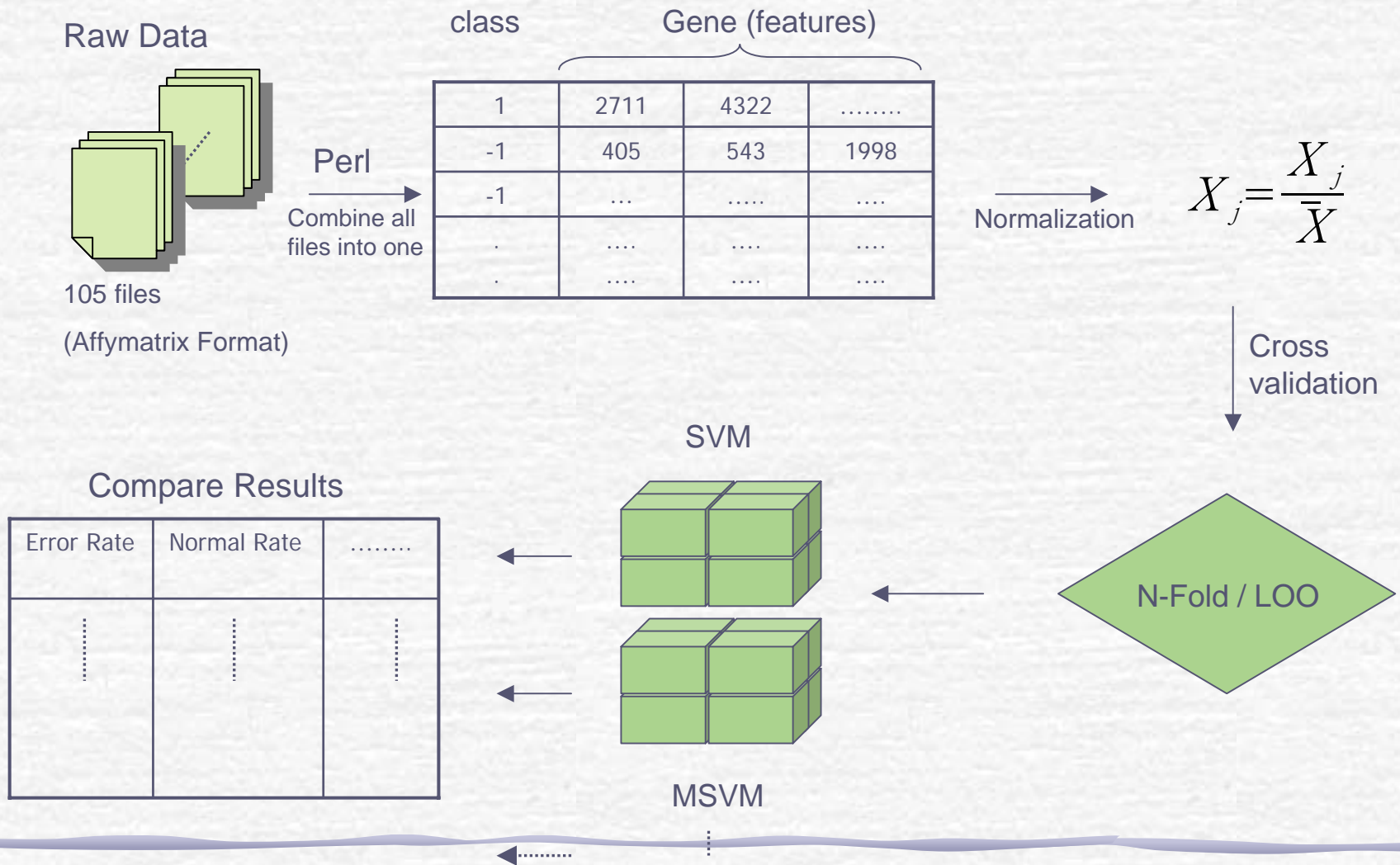
- ξ *slack variables* which represents distances of $f(x^i)$ to the margins $f(x)=1$ and $f(x)=-1$ for misclassified samples ([SVM](#))

	Sample Size	Cancer	Normal
Correct Mislabeled + Remove outlier	30	15	15
Correct Mislabeled (with outlier)	31	15	16
Original (mislabeled + outlier)	31	14	17

Flow Chart of The Application



Flow Chart of The Whole Process



- ☞ Goals of internship
- ☞ Introduction and overview
- ☞ Significance
- ☞ Datasets
- ☞ Methods and Experiments
- ☞ **Results and Discussion**
- ☞ Conclusions
- ☞ Demo
- ☞ Future work

Choosing the Support Vectors

The error rate of ovarian dataset (LOO C.V.)

Method	GenNo	ERate	SGERate	normW	normE
Alld	87558	29.00%	64.50%	3.13E-002	6.28E-005
Fsel	1000	16.10%	16.10%	1.62E-001	1.43E-003
Wsel	1000	32.30%	32.30%	7.68E-002	1.22E-004

- The worst 54.84% (17/31)
- The best 26% [19] due to the nature of dataset (not linear-separable)

$$0.641 \times 10^{-4} \leq \alpha \leq 0.25 \times 10^{-5}$$

	SVM	MSVM
Threshold	10^{-5}	0.641×10^{-6}
Support Vectors	31 -> 27	31 -> 31

Regularization

The error rate of Ionosphere dataset with Alll (LOO C.V.)

FeatureNo	ERate	SGERate	normW	normE
35	31.90%	33.00%	1.81E-001	1.12E-002

- 351 instances (225 class I and 126 class II)
- $0 \leq H_{ij} \leq 2.99 \times 10^6$, $\|H\|_2 = 5.70 \times 10^6$

	SVM	MSVM
Regularization	10^{-10}	$10^{-5} \times \ H\ _2 = 57.04$

Error in datasets

The error rate of lymphoma dataset with Alld (LOO C.V.)

Epsilon	GenNo	ERate	SGERate	normW	normE
0.00%	7129	2.80%	9.70%	3.93E-002	9.83E-006
5.00%	7129	2.80%	11.10%	3.93E-002	9.79E-006
50.00%	7129	8.30%	22.20%	3.58E-002	7.14E-006

PS:

9.7% → 11.1% = 1 misclassification

2.8% → 8.30% = 4 misclassifications

11.1% → 22.2% = 8 misclassifications

Outliers and mislabeled data

The class proportion of ovarian dataset in different cases and the number of times which MSVM outperforms SVM

	Sample Size	Cancer	Normal	Alld	Wsel (100 genes)
Correct Mislabeled + Remove outlier	30	15	15	9/9	2/9
Correct Mislabeled (with outlier)	31	15	16	9/9	0/9
Original (mislabeled + outlier)	31	14	17	9/9	1/9

Overall performance

The overall performance

Total cases	MSVM outperformance	SVM outperformance	Equality
288	103	10	175

Gene selection

- ☞ It doesn't work well enough during the experiments.
- ☞ The main reason is not the way we define the significant genes but how to determine the number.
- ☞ The number of significant genes used for these experiments is based on the experience rather than theory.

CLUSTER

Performance with and without CLUSTER

Dataset	Sample Size	File Size (MB)	Time without CLUSTER (sec)	Time with CLUSTER (sec)	Speedup
Lymphoma	72	9.9	279.190	25.950	10.76
Ovarian	31	45.1	173.880	84.850	2.05
Myeloma	105	11.7	274.420	43.750	6.27

P.S.:

1. PC Specs:

- AMD Opteron(TM) processor model 250 dual CPUs
- 2.40 Ghz Processor internal clock speed

2. 7 nodes for the CLUSTER

- ✓ Goals of internship
- ✓ Introduction and overview
- ✓ Significance
- ✓ Datasets
- ✓ Methods and Experiments
- ✓ Results and Discussion
- ✓ **Conclusions**
- ✓ Demo
- ✓ Future work

Conclusions

- ☞ We demonstrate the advantages on choosing Support Vectors, doing the regularization, and handling the errors in dataset.
- ☞ However....how to determine the significant gene number?
- ☞ How?
 - FCBF, Fast Correlation-Based Filter solution, from Huan Liu and Lei Yu[26].
 - Within the sorted weights, a method is used to filter out the most significant genes by searching the critical weight which separates the weights into significance and non-significance.

- ✓ Goals of internship
- ✓ Introduction and overview
- ✓ Significance
- ✓ Datasets
- ✓ Methods and Experiments
- ✓ Results and Discussion
- ✓ Conclusions
- ✓ **Demo**
- ✓ Future work

Demo

Matlab Form - Mozilla Firefox
檔案(F) 編輯(E) 檢視(V) 瀏覽(G) 書籤(B) 工具(T) 說明(H)
http://sols-cbs-20.la.asu.edu/~ta/cgi/Matlab/Sting/runml.html

SVM

The diagram illustrates a Support Vector Machine (SVM) model. It shows two clusters of data points: a top cluster labeled 'Cancer Data Cloud' and a bottom cluster labeled 'Normal Data Cloud'. Each cluster has a 'Cloud Means' indicated by a red dot. A horizontal dashed line represents the 'Hyper plane $w^T x + b = 0$ '. Two points on the hyperplane are labeled 'support vectors'. The distance from the hyperplane to the Cancer cloud is labeled $w^T x + b - 1$, and the distance to the Normal cloud is labeled $w^T x + b - 1$. A vertical arrow labeled $\|w\|$ indicates the perpendicular distance from the hyperplane to the origin.

Please select the conditions

Choose the Gene Select Method: All	Choose a dataset: Lym
Choose the way of splitting the dataset: LOO	Choose Gene Number for feature selection: 50
Choose VARSIGMA: 0.01	Choose EPSILON: 0
Execute Mode: ForeGround	Output Format: HTML
<input checked="" type="checkbox"/> RTLS <input checked="" type="checkbox"/> SteveGunn	Upload DataSet
submit	

完成

- ✓ Goals of internship
- ✓ Introduction and overview
- ✓ Significance
- ✓ Datasets
- ✓ Methods and Experiments
- ✓ Results and Discussion
- ✓ Conclusions
- ✓ Demo
- ✓ Future work

Future work

- ☛ Focus on how to determine the significant gene number.
- ☛ Publish the application to make this service available on the website so that we can get more comments from others and improve our algorithm.

Acknowledge

- ☞ Dr. Renaut
- ☞ Dr. Guo
- ☞ Dr. Liu
- ☞ Dr. Mittelmann
- ☞ Renate Mittelmann
- ☞ My wife - Connie

Q ?

No..

Thank you

謝謝