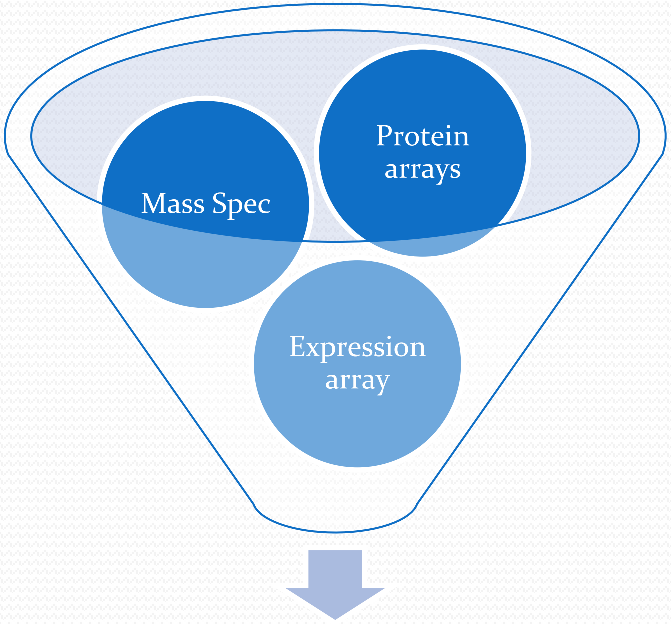
Learning predictive models for *combinations* of heterogeneous *proteomic* data *sources*

Michal Valko, Richard Pelikan, Milos Hauskrecht
University of Pittsburgh, Pittsburgh, Pennsylvania

Classifier for Pancreatic Cancer

- Measuring expression levels of protein mixtures
  - Mutliplexed protein arrays
  - Mass Spectrometry profiling
  - Expression Arrays



  - more sources » more information » better classifier

Pancreatic Cancer Dataset

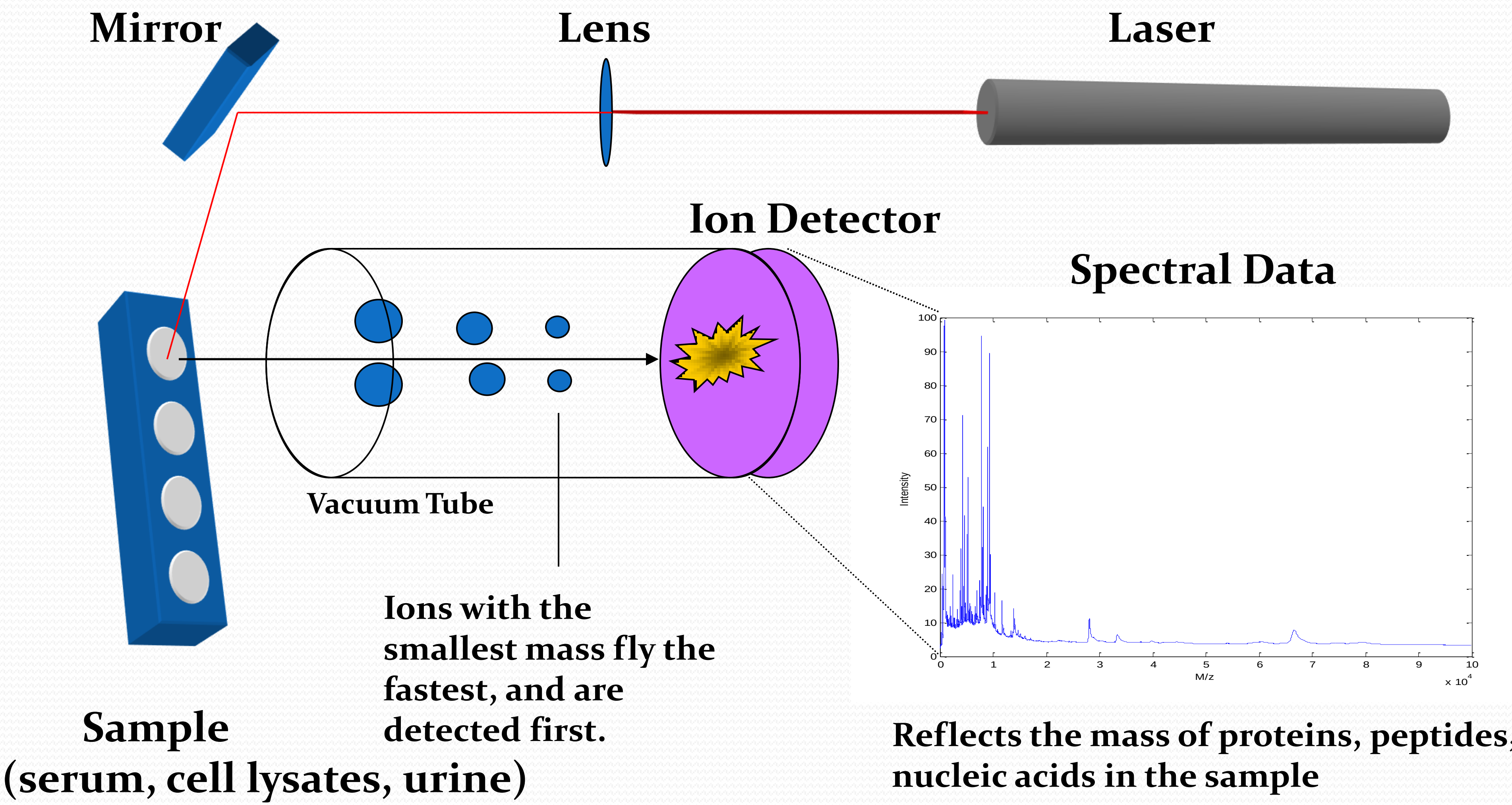
- 109 samples (from UPitt Cancer Institute)
  - 56 cases
  - 53 controls (smoking, age and gender matched to cases)

- 2 data sources
  - 1554 peaks from SELDI-TOF-Mass Spec
  - 30 measurements from Luminex xMAP ® arrays

- Several classifiers

SELDI-TOF MS

Surface Enhanced Laser Desorption/Ionization Time of Flight Mass Spectrometry



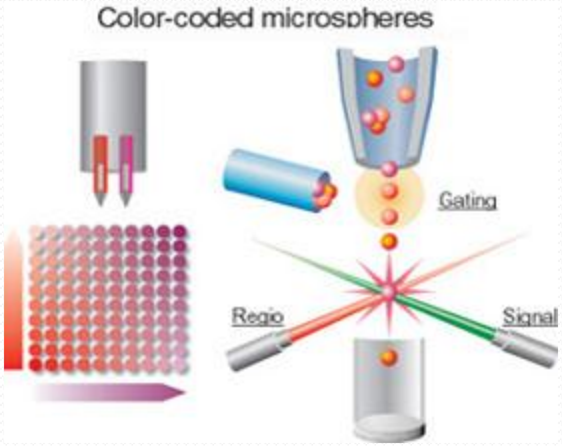
SELDI-TOF-MS preprocessing

1. variance stabilization
2. baseline correction
3. smoothing
4. intensity normalization
5. profile alignment steps

- 60264 variables from SELDI-TOF was reduced to 1554 by preprocessing

Luminex arrays

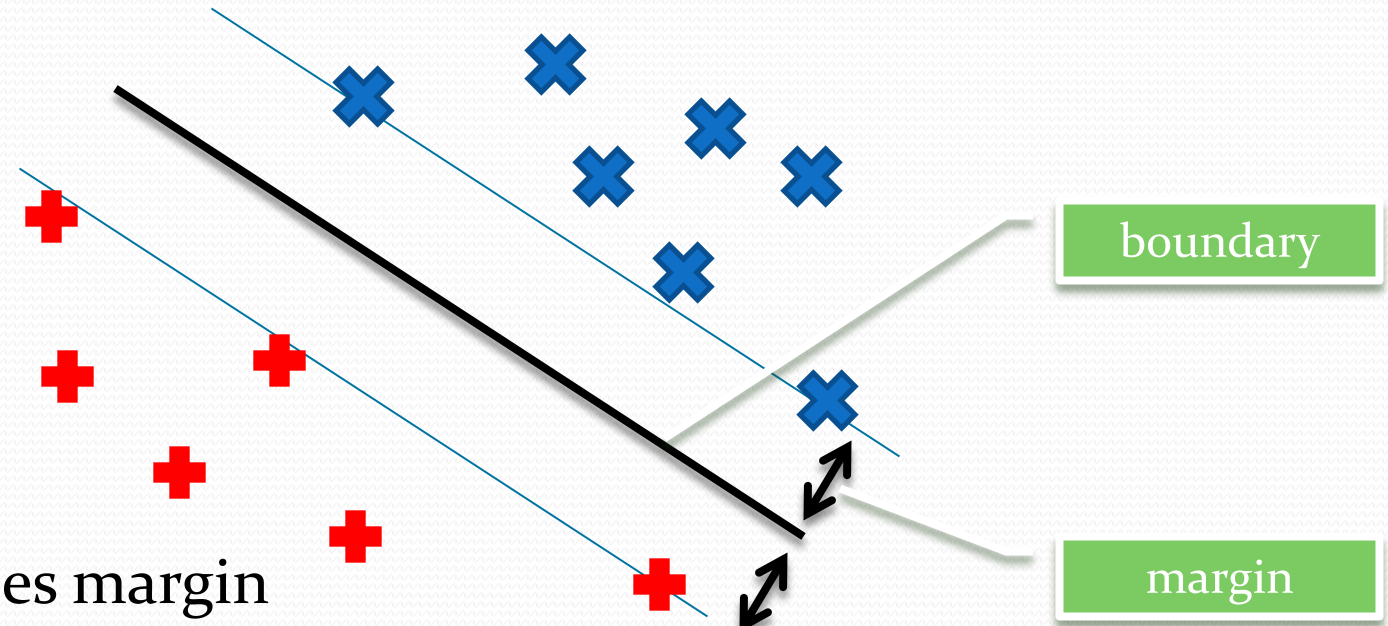
- Luminex Corporation's xMAP® technology



- Smaller number of output variables (up to 100)
- 30 variables in our data

Linear Support Vector Machine

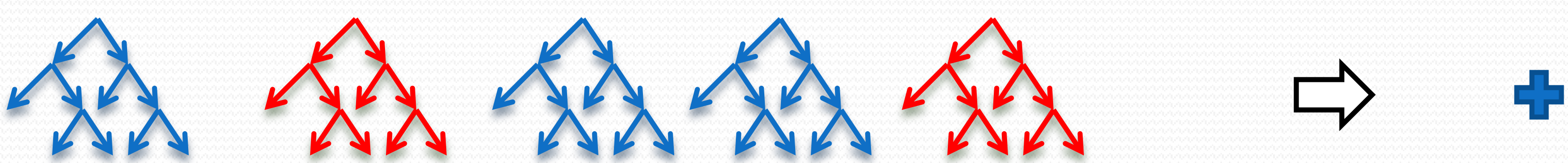
- Learn linear decision boundary
- Separates n-dimensional feature space into 2 partitions



- Maximizes margin
- Classification: which half-space new point falls in

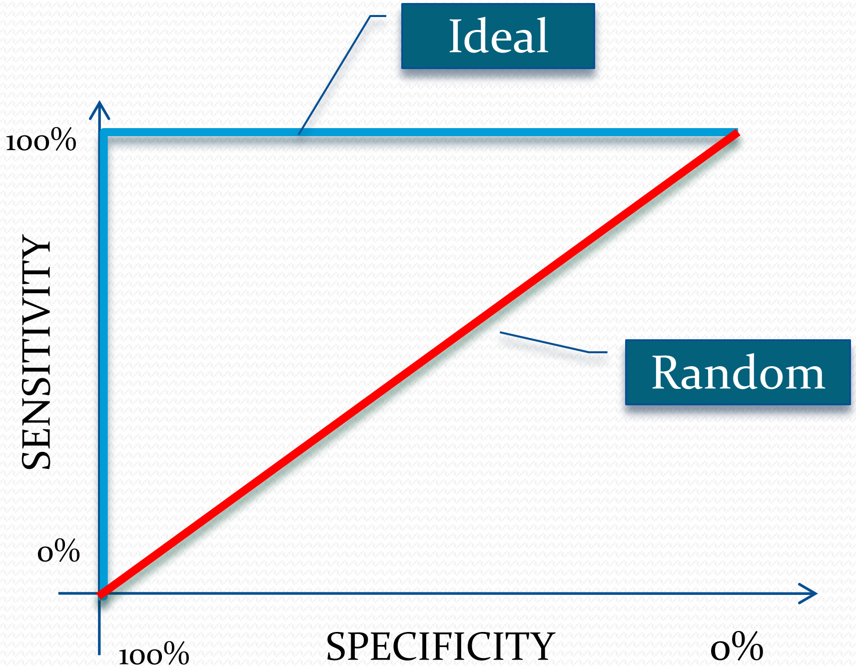
Random Forest Classifier

- Ensemble classifier :
  - Combines the result of multiple *decision trees*
  - Random Feature selection
- Construction of each tree:
  1. Sample with replacement (from training set)
  2. Randomly select subset of variables
  3. Train a tree classifier
- Class that is selected by voting



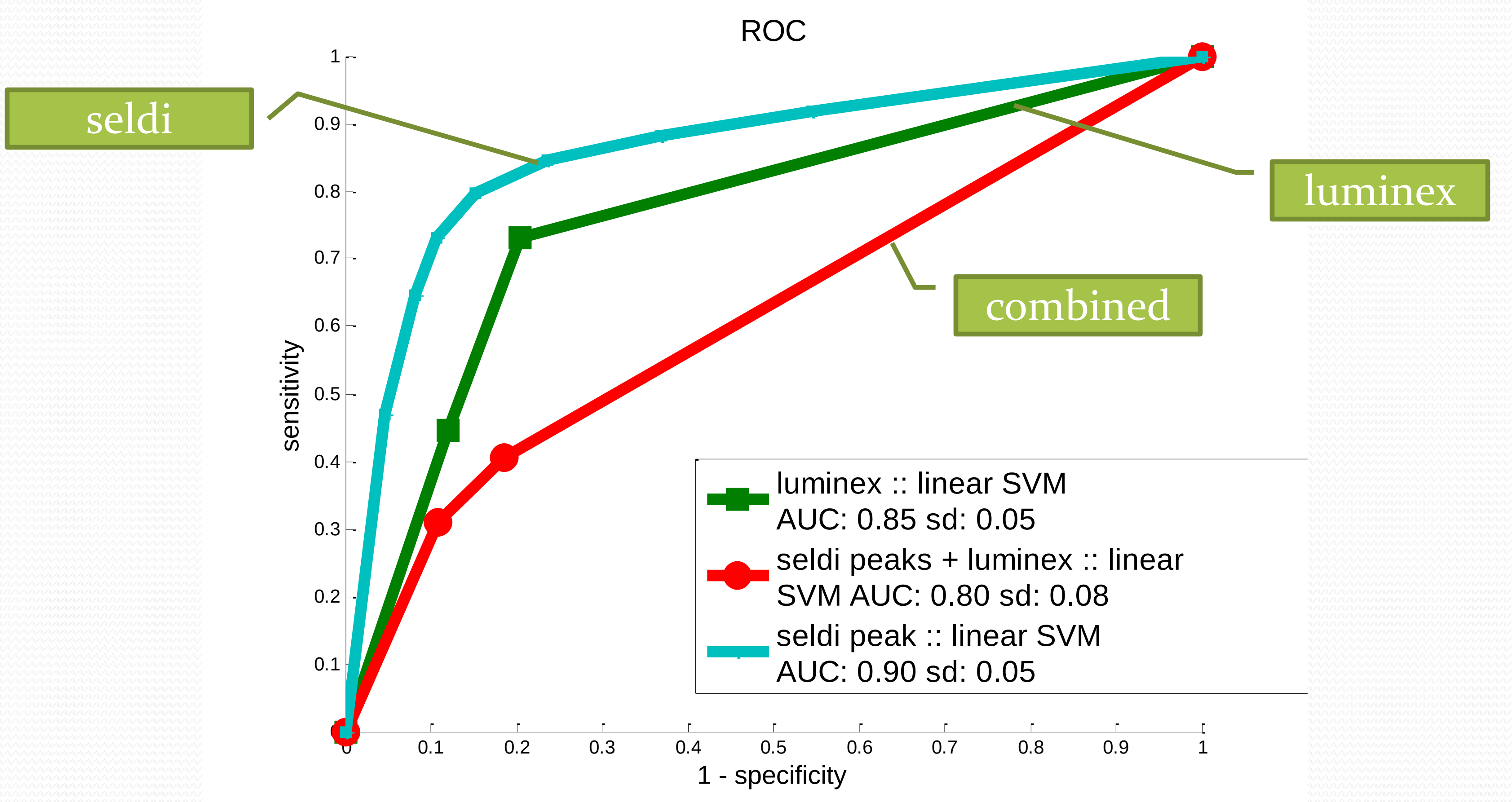
Evaluation

- Random subsampling
  - 40 splits (70% train, 30% test)
- Statistics:
  - Classification Error
  - Sensitivity
  - Specificity
    - Receiver Operating Characteristics (ROC)

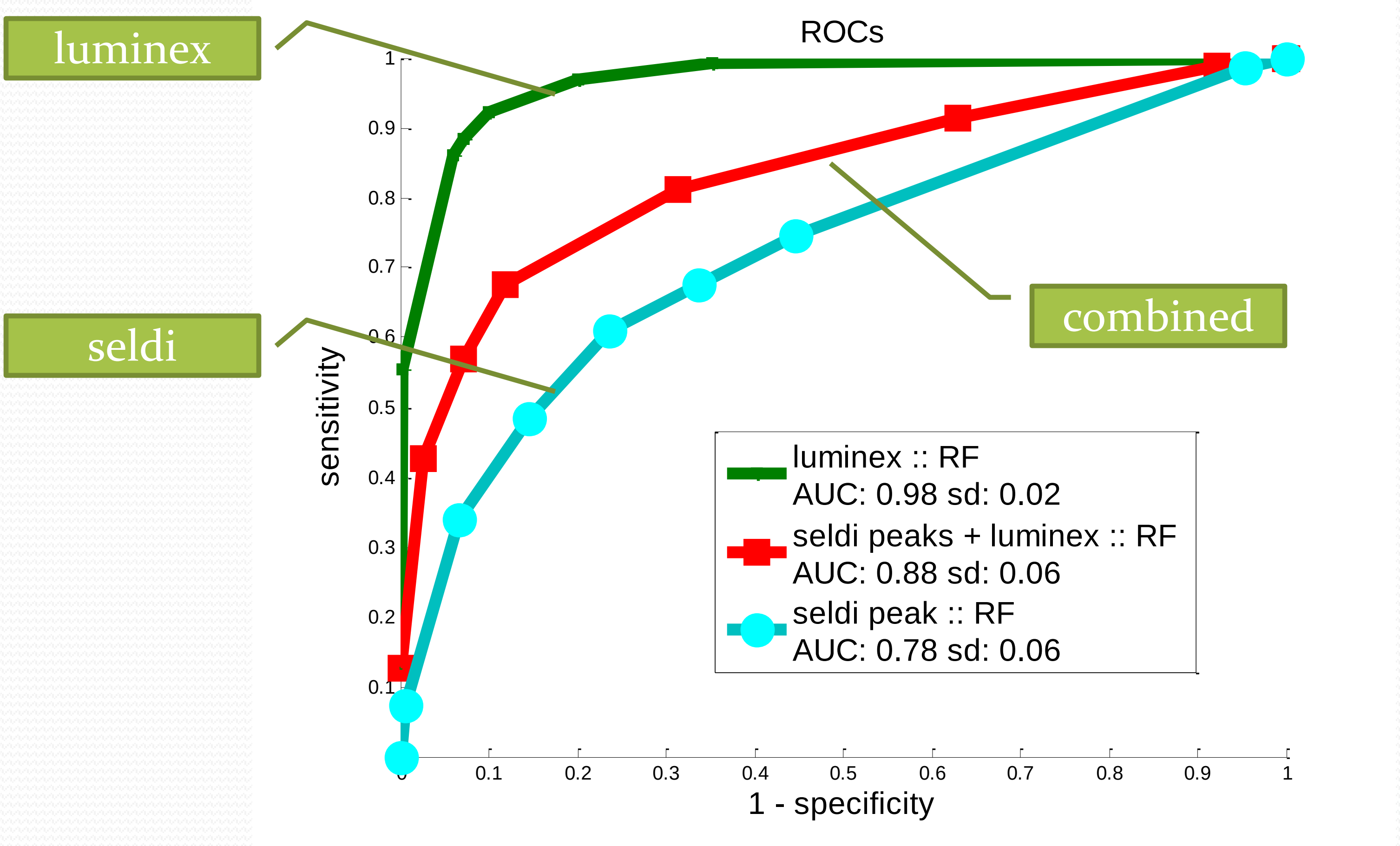


Data Fusion

Data fusion (Linear SVM)

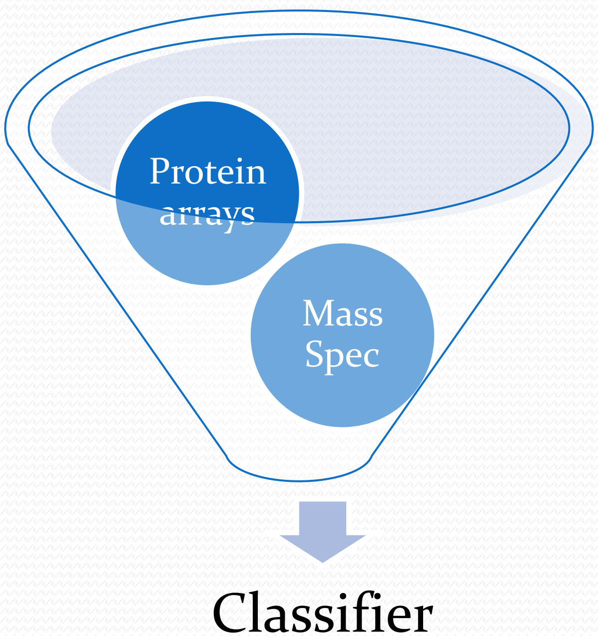


Data fusion (Random Forest)

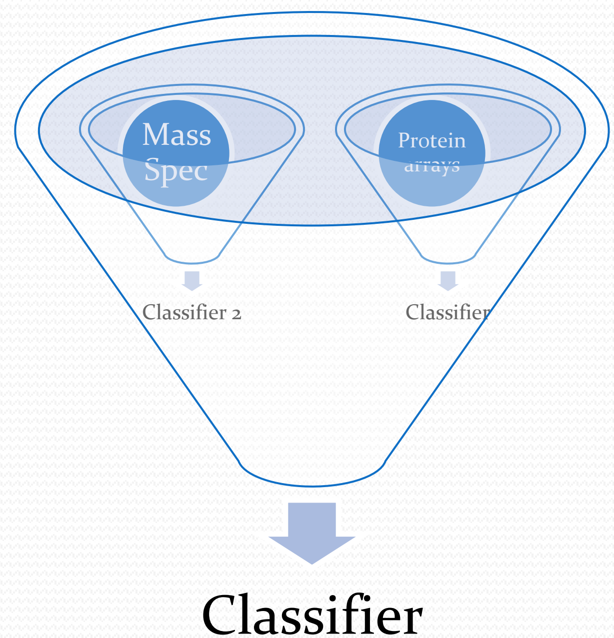


Model Fusion

- Simple data merging resulted in worse performance
- Need for classifier that combines both sources



Data Fusion

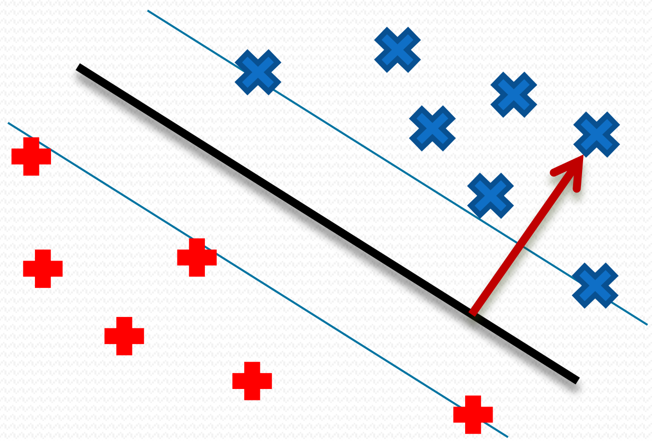


Model Fusion

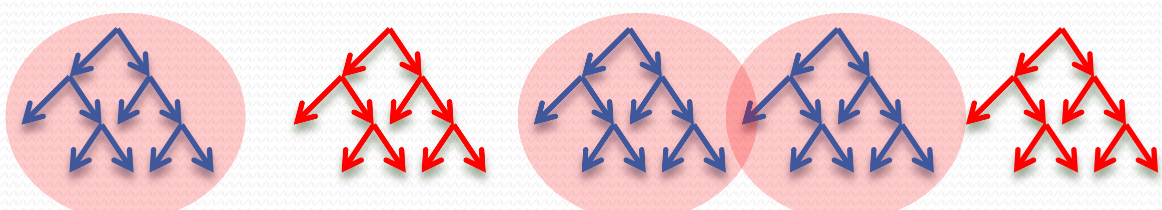
Soft Output from Classifiers

- Soft output from the best classifiers
  - SVM: distance from the separating hyperplane
  - Random Forest: Ratio of Trees that favor predicted class

Soft Output

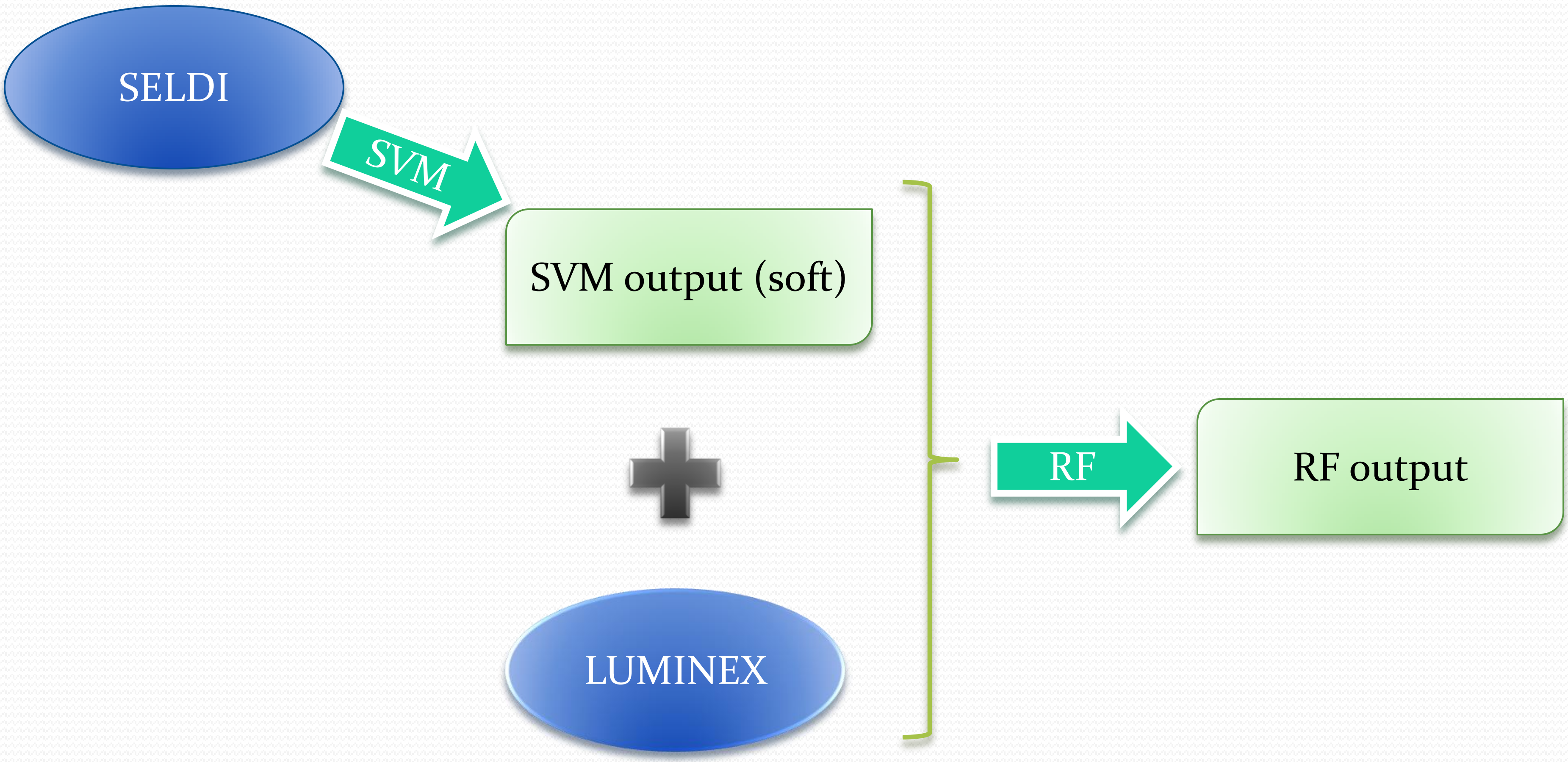


1.7

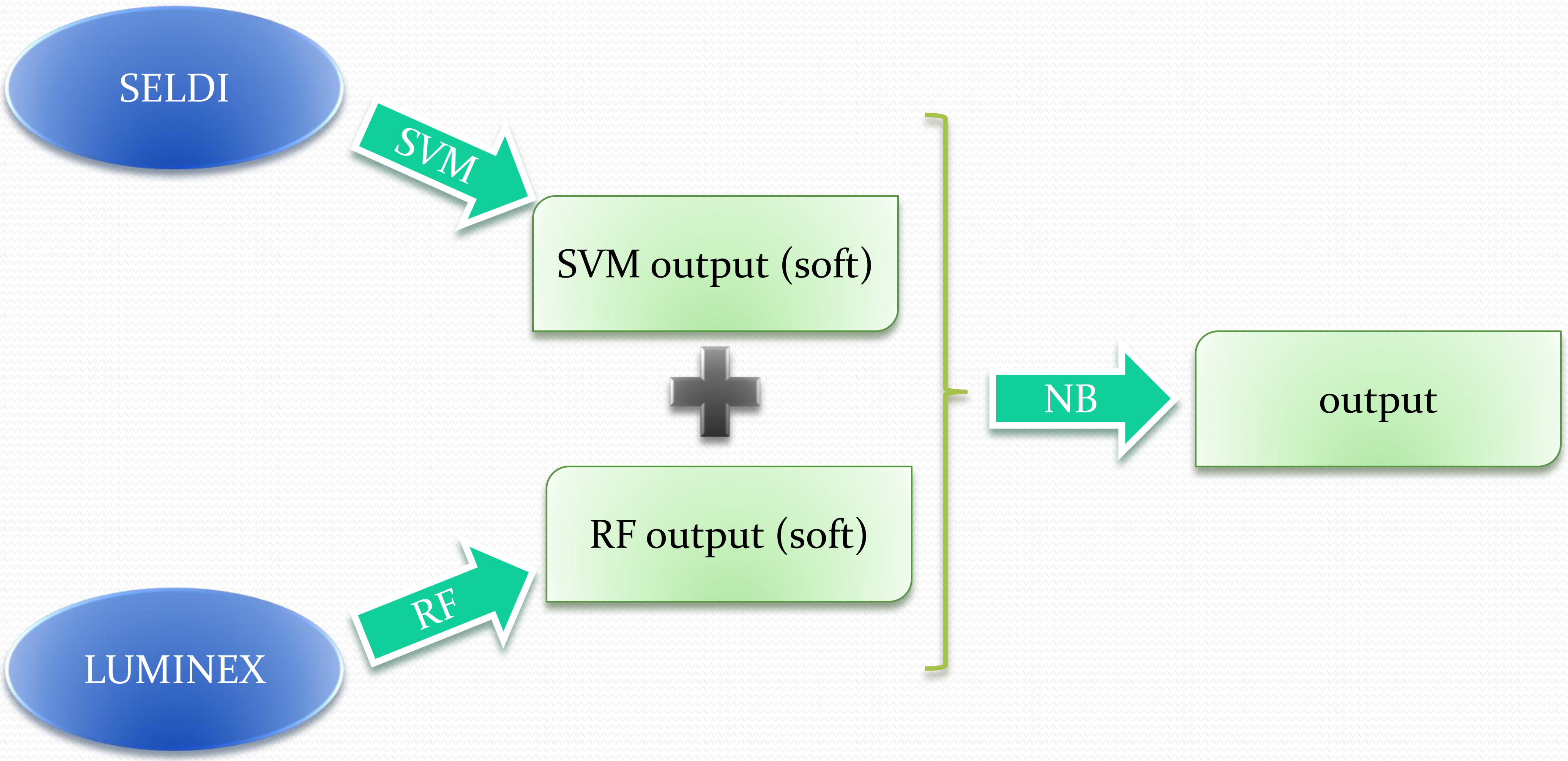


60%

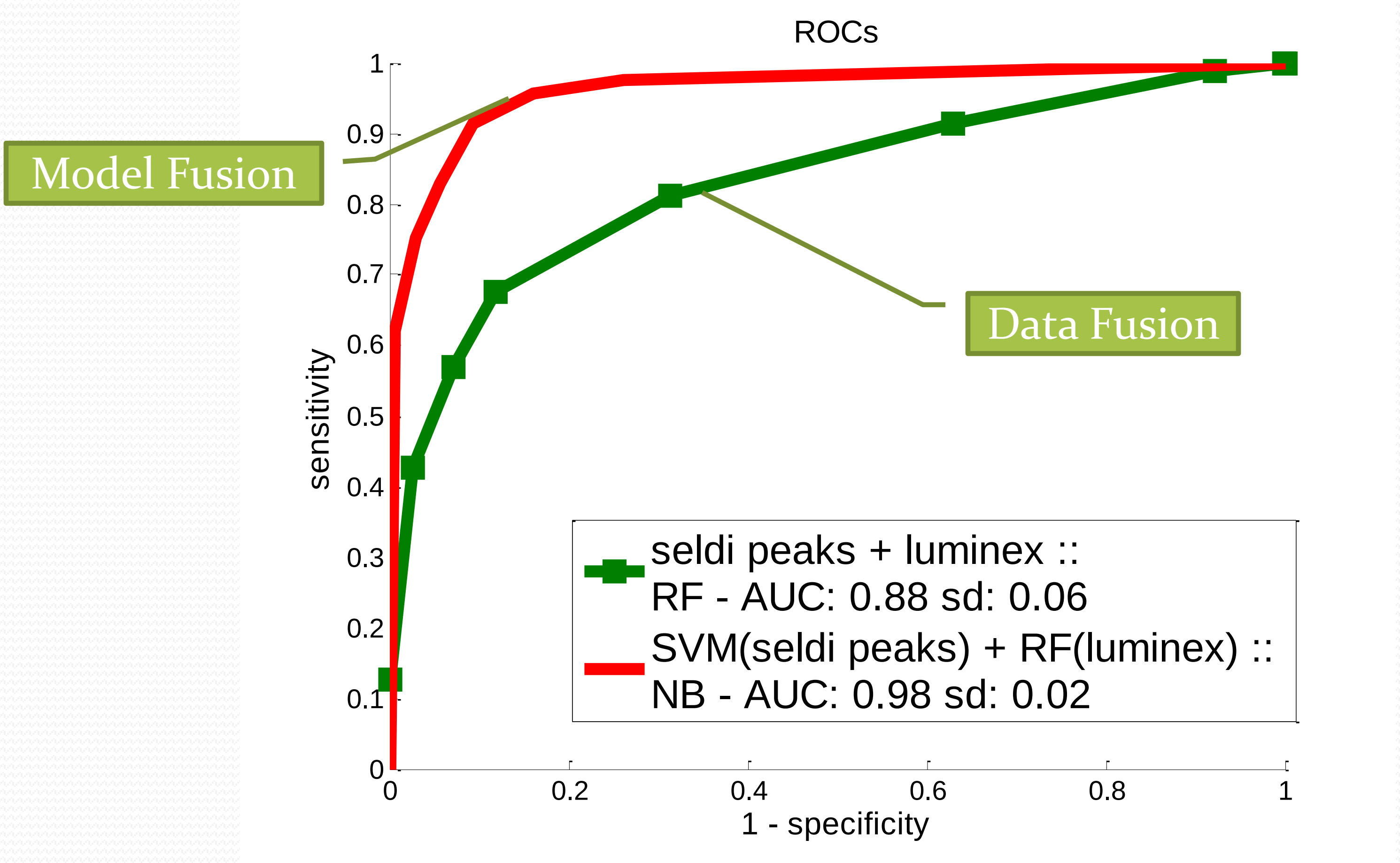
Model Inclusion



Model Composition



Model Fusion vs. Data Fusion



Data Fusion

SELDI PEAKS + LUMINEX		Error	SN	SP
	CART	22.94%	68.00%	88.87%
	std	16.07%	25.71%	22.71%
	NB	44.63%	74.21%	37.74%
	std	9.97%	26.28%	25.40%
	LogisticR	38.38%	60.72%	62.56%
	std	9.30%	12.51%	12.11%
	RF	**21.54%**	76.58%	82.37%
	std	7.98%	13.67%	13.22%
	SVM	34.49%	50.82%	79.83%
	std	12.12%	36.76%	22.42%

Standard deviation

Model Fusion

		Error	SN	SP
SVM(seldi) + RF(luminex)	NB	8.82%	91.28%	91.60%
	std	4.42%	7.90%	7.91%
SVM(seldi) + luminex	RF	9.71%	'91.29%	89.88%
	Std	4.53%	7.22%	8.40%
seldi peaks + RF(luminex)	SVM	8.46%	92.56%	91.03%
	Std	3.78%	5.98%	6.58%
T_test50(seldi) + luminex	RF	9.85%	88.83%	92.12%
	std	4.78%	9.23%	7.24%

Conclusion

- Simple data merging deteriorates the classification accuracy
- Combine classifiers that work well for certain type of data
- Using soft output from classifiers
- Model inclusion/model composition
- Significant improvement over mere data merging

Acknowledgments

- USAMRAA W81XWH-05-2-0066
- NLM training grant 5 T15 LM007059-20
- NCI grant P50 CA090440-06.
- Dr. Bigbee, Dr. Zeh, and Dr. Whitcomb for the data used in our analyses.