

**SMA 5303: Statistical Learning and Data Mining in Bioinformatics**

(also listed as 15.077: Statistical Learning and Data Mining (MIT))

Spring Term (Feb – May 2010)

MIT Faculty: Professor Roy Welsch

Singapore Faculty: Professor Jagath Rajapakse

Graduate Student Tutors (Singapore): Fransiskus Xaverius Ivan, Iti Chaturvedi

Accurate as of Jan 27, 2010

	MIT Day & Time	S'pore Day & Time	Lecture/Recitation Topic	MIT Lecturer	NTU Lecturer	References
1 (3 way)	Wed 03 Feb 7:00-8:30 PM 3-370	Thurs 04 Feb 8:00-9:30 AM NTU: Smart classroom NUS: CIT Auditorium	SMA 5303 – L1 Sampling and statistical distributions  <b>Homework # 1 handed out</b>	Roy Welsch	Jagath Rajapakse	R 6, 7.1-7.3, 7.5-7.6
1 (3 way)	Mon 08 Feb 7:00-8:30 PM 3-370	Tues 09 Feb 8:00-9:30 AM NTU: Smart classroom NUS: CIT Auditorium	SMA 5303 – L2 Estimation, confidence intervals, and the bootstrap	Roy Welsch	Jagath Rajapakse	R 8.1, 8.3-8.5, 8.7, 8.9, 10.4.6
1 (2 way)	Wed 10 Feb 7:00-8:30 PM 3-370	Thurs 11 Feb 8:00-9:30 AM NTU: Smart classroom	SMA 5303 – L3 Hypothesis testing , likelihood ratios, goodness-of-fit tests, approximate methods	Roy Welsch	Jagath Rajapakse	R 9.1-9.6, 8.2, 4.6
7	Tues 9 Feb	Thurs 11 Feb 10:30-11:30 AM BIRC*	<u>Rec.</u> : Computing: graphics and the bootstrap			R 9.8, 10.2.3, 10.3, 10.6-10.8
			<b>NO CLASS ---- Singapore Holiday – Chinese New Year from 14 Feb to 16 Feb 2010</b>	<b>NO CLASS - MIT holiday – Presidents Day on 15 Feb 2010</b>		
3 (3 way)	Tues 16 Feb 7:00-8:30 PM 3-370	Wed 17 Feb 8:00-9:30 AM NTU: Smart classroom NUS: CIT Auditorium	SMA 5303 – L4 Bayesian inference, Molecular biology fundamentals	Roy Welsch (MIT Monday schedule of classes to be held)	<b>Jagath Rajapakse</b>	R 3.5.2 (94 ,95), 8.6, BB 2.1-2.3, CB 1.0-1.7
3 (2 way)	Wed 17 Feb 7:00-8:30 PM 3-370	Thur 18 Feb 8:00-9:30 AM NTU: Smart classroom	SMA 5303 – L5 Die models of sequences, Markov models, pairwise sequence alignment <b>Tentative homework # 1 due dates</b> <b>Homework # 2 handed out</b>	Roy Welsch	<b>Jagath Rajapakse</b>	BB 3.1; EG 4.5-4.10, 5.2-5.4, 6.1-6.4
7	Thurs 18 Feb	Thurs 18 Feb 10:30-11:30 AM BIRC	<u>Rec.</u> : Testing and Bayesian Inference			

\*: 1. Live video-casting from MIT; 2. Taped lecture from MIT; 3. Live video-casting from S'pore; 4. Taped lecture from S'pore; 5. Classroom lecture in S'pore; 6. MIT faculty teaches in S'pore; 7. Recitation by MIT faculty to students at MIT and by NTU faculty to students at Singapore; 8. Other-Please specify

**SMA 5303: Statistical Learning and Data Mining in Bioinformatics**  
 (also listed as 15.077: Statistical Learning and Data Mining (MIT))  
 Spring Term (Feb – May 2010)  
 MIT Faculty: Professor Roy Welsch  
 Singapore Faculty: Professor Jagath Rajapakse  
 Graduate Student Tutors (Singapore): Fransiskus Xaverius Ivan, Iti Chaturvedi

Accurate as of Jan 27, 2010

	MIT Day & Time	S'pore Day & Time	Lecture/Recitation Topic	MIT Lecturer	NTU Lecturer	References
3 (3 way)	Mon 22 Feb 7:00-8:30 PM 3-370	Tues 23 Feb 8:00-9:30 AM NTU: Smart classroom NUS: CIT Auditorium	SMA 5303 – L6 Substitution matrices, multiple sequence alignment, Markov chain Monte Carlo, simulated annealing, Gibbs sampling, BLAST	Roy Welsch	Jagath Rajapakse	EG 6.5-6.6, 10.1-10.5, 11.1-11.7 Brooks paper
3 (2 way)	Wed 24 Feb 7:00-8:30 PM 3-370	Thurs 25 Feb 8:00-9:30 AM NTU: Smart classroom	SMA 5303 – L7 Hidden Markov models: gene structure prediction, profile HMM, Expectation-Maximization algorithm	Roy Welsch	Jagath Rajapakse	BB 7.1-7.5 EG 12.1-12.3
7	Tues 23 Feb	Thurs 25 Feb 10:30-11:30 AM BIRC	<u>Rec.</u> : Die models; Markov modeling			BB 3.1 EG 4.5-4.10, 11.1-11.4
1 (3 way)	Mon 01 Mar 7:00-8:30 PM 3-370	Tues 02 Mar 8:00-9:30 AM NTU: Smart classroom NUS: CIT Auditorium	SMA 5303 – L8 Linear regression and smoothing	Roy Welsch	Jagath Rajapakse	R4.4.2, 14.1-14.5, 14.7
1 (2 way)	Wed 03 Mar 7:00-8:30PM 3-370	Thurs 04 Mar 8:00-9:30 AM NTU: Smart classroom	SMA 5303 – L9 Regression diagnostics, collinearity, and robust regression <b>Tentative homework # 2 due dates</b> <b>Homework # 3 handed out</b>	Roy Welsch	Jagath Rajapakse	Notes, R10.4.2-10.4.5, 14.8
7	Tues 02 Mar	Thurs 04 Mar 10:30-11:30 AM BIRC	<u>Rec.</u> : Computing: regression and Gene structure prediction with HMM			
1 (3 way)	Mon 08 Mar 7:00-8:30PM 3-370	Tues 09 Mar 8:00-9:30 AM NTU: Smart classroom NUS: CIT Auditorium	SMA 5303 – L10 Comparing two samples; non-parametric methods and experimental design	Roy Welsch	Jagath Rajapakse	R11.1-11.5
1 (2 way)	Wed 10 Mar 7:00-8:30PM 3-370	Thurs 11 Mar 8:00-9:30 AM NTU: Smart classroom	SMA 5303 – L11 Analysis of categorical data	Roy Welsch	Jagath Rajapakse	R13.1,13.3-13.4
7	Tues 09 Mar	Thurs 11 Mar 10:30-11:30 AM BIRC	<u>Rec.</u> : Computing: diagnostics and two-sample			

\*: 1. Live video-casting from MIT; 2. Taped lecture from MIT; 3. Live video-casting from S'pore; 4. Taped lecture from S'pore; 5. Classroom lecture in S'pore; 6. MIT faculty teaches in S'pore; 7. Recitation by MIT faculty to students at MIT and by NTU faculty to students at Singapore; 8. Other-Please specify

**SMA 5303: Statistical Learning and Data Mining in Bioinformatics**  
 (also listed as 15.077: Statistical Learning and Data Mining (MIT))  
**Spring Term (Feb – May 2010)**  
**MIT Faculty: Professor Roy Welsch**  
**Singapore Faculty: Professor Jagath Rajapakse**  
**Graduate Student Tutors (Singapore): Fransiskus Xaverius Ivan, Iti Chaturvedi**

Accurate as of Jan 27, 2010

	MIT Day & Time	S'pore Day & Time	Lecture/Recitation Topic	MIT Lecturer	NTU Lecturer	References
<i>Daylight Savings Time Start Note – Time Change at NTU and MIT (starts from 14 March 09 - 12 hours difference)</i>						
1 (2 way)	Mon 15 Mar 8:00-9:30PM 3-370	Tues 16 Mar 8:00-9:30 AM NTU: Smart classroom	SMA 5303 – L12 Analysis of variance <b>Tentative homework # 3 due dates</b>	Roy Welsch	Jagath Rajapakse	R12.1-12.4
7		Tues 16 Mar 10:30-11:30 AM BIRC	<u>Rec.</u> : Categorical Data and ANOVA			
8 (no beaming)	Wed 17 Mar 8:00-9:30PM 3-370	Thurs 18 Mar 8:00-9:30 AM NTU: Smart classroom	<b>Midterm Examination (in-class)</b>	Roy Welsch	Jagath Rajapakse	
<b>Spring Vacation 22-28 March, MIT (Mon – Sun)</b>						
1 (3 way)	Mon 29 Mar 8:00-9:30PM 3-370	Tues 30 Mar 8:00-9:30 AM NTU: Smart classroom NUS: CIT Auditorium	SMA 5303 – L14 Learning from data <b>Homework # 4 handed out</b>	Roy Welsch	Jagath Rajapakse	H1, 2
1 (2 way)	Wed 31 Mar 8:00-9:30PM 3-370	Thurs 01 Apr 8:00-9:30 AM NTU: Smart classroom	SMA 5303 – L15 Model Assessment	Roy Welsch	Jagath Rajapakse	H7.1-7.7, 7.10-7.11
7	Tues 30 Mar	Thurs 01 Apr 10:30-11:30 AM BIRC	<u>Rec.</u> : Insightful Miner Basics			
1 (3 way)	Mon 05 Apr 8:00-9:30PM 3-370	Tues 06 Apr 8:00-9:30 AM NTU: Smart classroom NUS: CIT Auditorium	SMA 5303 – L16 Regression Selection: Ridge, PCR, PLS, LAR	Roy Welsch	Jagath Rajapakse	H3.1-3.6,3.9
3 (2 way)	Wed 07 Apr 8:00-9:30PM 3-370	Thurs 08 Apr 8:00-9:30 AM NTU: Smart classroom	SMA 5303 – L17 Discriminant Analysis; Logistic Regression <b>Tentative homework # 4 due dates</b> <b>Homework # 5 handed out</b>	Roy Welsch	Jagath Rajapakse	H4.1-4.4

\*: 1. Live video-casting from MIT; 2. Taped lecture from MIT; 3. Live video-casting from S'pore; 4. Taped lecture from S'pore; 5. Classroom lecture in S'pore; 6. MIT faculty teaches in S'pore; 7. Recitation by MIT faculty to students at MIT and by NTU faculty to students at Singapore; 8. Other-Please specify

**SMA 5303: Statistical Learning and Data Mining in Bioinformatics**  
 (also listed as 15.077: Statistical Learning and Data Mining (MIT))  
**Spring Term (Feb – May 2010)**  
**MIT Faculty: Professor Roy Welsch**  
**Singapore Faculty: Professor Jagath Rajapakse**  
**Graduate Student Tutors (Singapore): Fransiskus Xaverius Ivan, Iti Chaturvedi**

**Accurate as of Jan 27, 2010**

	<b>MIT Day &amp; Time</b>	<b>S'pore Day &amp; Time</b>	<b>Lecture/Recitation Topic</b>	<b>MIT Lecturer</b>	<b>NTU Lecturer</b>	<b>References</b>
7	Tues 06 Apr	Thurs 08 Apr 10:30-11:30 AM BIRC	<u>Rec.</u> : Regression Selection			
3 (3 way)	Mon 12 Apr 8:00-9:30PM 3-370	Tues 13 Apr 8:00-9:30 AM NTU: Smart classroom NUS: CIT Auditorium	SMA 5303 – L18 Generalized Additive Models and Trees: CART	<b>Roy Welsch</b>	Jagath Rajapakse	H9.1-9.2
3 (2 way)	Wed 14 Apr 8:00-9:30PM 3-370	Thurs 15 Apr 8:00-9:30 AM NTU: Smart classroom	SMA 5303 – L19 Support Vector Machines; Support Vector Regression; Prediction of protein features: secondary structures, solvent accessibility, and accessibility area	Roy Welsch	<b>Jagath Rajapakse</b>	H4.5, 12.1-12.3, (omit 12.3.3, 12.3.5) Nguyen & Rajapakse 2005, 2006, 2007
7	Tues 13 Apr	Thurs 15 Apr 10:30-11:30 AM BIRC	<u>Rec.</u> : Classification, logistics Reg. , and SVM			
<b>MIT holiday Patriots Day, 19 - 20 Apr (Mon , Tues) No Class</b>						
3 (3 way)	Wed 21 Apr 8:00-9:30PM 3-370	Thurs 22 Apr 8:00-9:30 AM NTU: Smart classroom	SMA 5303 – L20 Neural Networks, prediction of signal sites in genomic sequences	Roy Welsch	<b>Jagath Rajapakse</b>	H1.1,11.3- 11.10 Rajapakse & Ho 2005
1 (3 way)	Mon 26 Apr 8:00-9:30PM 3-370	Tues 27 Apr 8:00-9:30 AM NTU: Smart classroom NUS: CIT Auditorium	SMA 5303 – L21 Cluster analysis, k-means, hierarchical clustering, clustering of gene expressions, biclustering  <b>Tentative homework # 5 due dates Homework # 6 handed out</b>	Roy Welsch	<b>Jagath Rajapakse</b>	H13.1-13.2, 14.3 (omit 14.3.9)
3 (2 way)	Wed 28 Apr 8:00-9:30PM 3-370	Thurs 29 Apr 8:00-9:30 AM NTU: Smart classroom NUS: CIT Auditorium	SMA 5303 – L22 Bagging and Boosting, AdaBoost, Random Forests	<b>Roy Welsch</b>	Jagath Rajapakse	H8.7-8.9, 10.1-10.14, 15.1- 15.3

\*: 1. Live video-casting from MIT; 2. Taped lecture from MIT; 3. Live video-casting from S'pore; 4. Taped lecture from S'pore; 5. Classroom lecture in S'pore; 6. MIT faculty teaches in S'pore; 7. Recitation by MIT faculty to students at MIT and by NTU faculty to students at Singapore; 8. Other-Please specify

**SMA 5303: Statistical Learning and Data Mining in Bioinformatics**  
 (also listed as 15.077: Statistical Learning and Data Mining (MIT))  
 Spring Term (Feb – May 2010)  
 MIT Faculty: Professor Roy Welsch  
 Singapore Faculty: Professor Jagath Rajapakse  
 Graduate Student Tutors (Singapore): Fransiskus Xaverius Ivan, Iti Chaturvedi

Accurate as of Jan 27, 2010

	MIT Day & Time	S'pore Day & Time	Lecture/Recitation Topic	MIT Lecturer	NTU Lecturer	References
7	Tue 27 April	Thurs 29 Apr 10:30 -11:30AM BIRC	Rec.: Neural Nets, CART, Bagging and Boosting			
7	Mon 03 May	Tues 04 May 9:00 – 11:00 AM BIRC	SMA 5303 – L23 Project consultation	Roy Welsch	Jagath Rajapakse	
7	Wed 05 May	Thurs 06 May 9:00 – 11:00 AM BIRC	SMA 5303 – L24 Project consultation <b>Tentative homework # 6 due dates</b>	Roy Welsch	Jagath Rajapakse	
7	Tues 04 May	Thurs 06 May 11:00AM -12:00 PM BIRC	Rec.: Clustering and Neural Nets			
7	Mon 10 May	Tues 11May 9:00-11:00 AM BIRC	SMA 5303 – L25 Project presentations	Roy Welsch	Jagath Rajapakse	
7	Tues 11 May	Tues 11May 11:00 AM -12:00PM BIRC	Rec.: Project help			
7	Wed 12 May	Thurs 13 May 9:00 – 11:00 AM BIRC	SMA 5303 – L26 Project presentations  <b>Project report due</b>	Roy Welsch	Jagath Rajapakse	

Texts:

1. John A. Rice, *Mathematical Statistics and Data Analysis* (Third Edition, 2007) [R]

An alternative to Rice, especially if you are interested in bioinformatics might be:

2. Warren J. Ewens, Gregory R. Grant, *Statistical Methods in Bioinformatics: An Introduction*, Second Edition [EG]

3. Hastie, Tibshirani, and Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* [H]

On reserve or portions handed out:

\*: 1. Live video-casting from MIT; 2. Taped lecture from MIT; 3. Live video-casting from S'pore; 4. Taped lecture from S'pore; 5. Classroom lecture in S'pore; 6. MIT faculty teaches in S'pore; 7. Recitation by MIT faculty to students at MIT and by NTU faculty to students at Singapore; 8. Other-Please specify

**SMA 5303: Statistical Learning and Data Mining in Bioinformatics**  
(also listed as 15.077: Statistical Learning and Data Mining (MIT))  
**Spring Term (Feb – May 2010)**  
**MIT Faculty: Professor Roy Welsch**  
**Singapore Faculty: Professor Jagath Rajapakse**  
**Graduate Student Tutors (Singapore): Fransiskus Xavier Ivan, Iti Chaturvedi**

**Accurate as of Jan 27, 2010**

4. P. Baldi and S. Brunak, *Bioinformatics: The Machine Learning Approach*, Second Edition [BB]
5. P. Clote and R. Backofen, *Computational Molecular Biology: An Introduction* [CB]
6. R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acid* [DEKM]

\*Note that BIRC lab sessions are held at SCE Multi-Media Lab at N4-01A-02

\*: 1. Live video-casting from MIT; 2. Taped lecture from MIT; 3. Live video-casting from S'pore; 4. Taped lecture from S'pore; 5. Classroom lecture in S'pore; 6. MIT faculty teaches in S'pore; 7. Recitation by MIT faculty to students at MIT and by NTU faculty to students at Singapore; 8. Other-Please specify