



National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh

Machine Learning Overview and Applications

**School of Psychology, NUI Galway
16 Oct 2009**

Dr Michael Madden

**Machine Learning & Data Mining Group
College of Engineering and Informatics, NUI Galway**

<http://datamining.it.nuigalway.ie>

Oct 2009

NUI Galway (1)



National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh

Agenda

- Machine Learning
- Some Techniques
- Analysing Drugs and Hazardous Materials
- Biomedical Data Mining
- Reinforcement Learning
- Conclusions

(45 mins)

Oct 2009

NUI Galway (2)



What is Machine Learning?

National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh

- Various definitions:
 1. Improvement at performing **tasks** through **experience**
 2. Formulating and refining **hypotheses** that best describe **observations**
- Formal [Mitchell]:
 - A well-defined ML problem:
 - Improve over task T (e.g. **playing chess**)
 - with respect to **performance** measure P (e.g. **tournament wins**)
 - based on experience E (e.g. **games against itself**)
- Other Possible Definitions
 - Philosophical and psychological considerations ...
 - Relationship to Artificial Intelligence ...

Oct 2009

NUI Galway (3)



Data Mining: What's the Link?

National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh

- Data Mining:
 - Extract **interesting** knowledge from **large unstructured** datasets
 - **non-obvious** / **comprehensible** / **meaningful** / **useful**
- Storage Law:
(Fayyad & Uthurusamy, Comms. ACM 2002)
 - Capacity of digital storage is **doubling** every year
 - Faster than Moore's law!
 - Result: write-only "data tombs"
- Developments in ML are essential to be able to process and exploit this lost data



Oct 2009

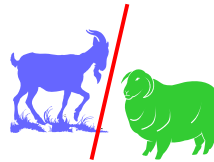
NUI Galway (4)



Major Types of Task (1)

National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh

1. **Classification**
(category prediction)



2. **Regression**
(numeric prediction)



Oct 2009

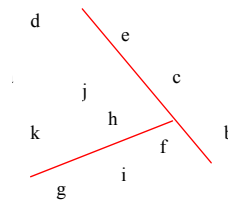
NUI Galway (5)



Major Types of Task (2)

National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh

3. **Clustering**



4. **Relationship Discovery**

beer \Leftrightarrow diapers

5. **Reinforcement Learning**



Oct 2009

NUI Galway (6)



Techniques for these Tasks

National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh

- **Classification**
Decision trees; Bayesian Classifiers
 - **Regression**
Neural nets; k-Nearest Neighbours
(good for Classification too)
 - **Clustering**
K-means
 - **Relationship Discovery**
Association Rules; Bayesian nets
 - **Reinforcement Learning**
Q-Learning
- } **Supervised**
- } **Unsupervised**
- } **Semi-supervised**

Oct 2009

NUI Galway (7)



What Do These Have in Common?

National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh

- In all cases, the machine **searches** for a **hypothesis** that **best** describes the data **presented** to it
- Choices to be made:
 - How is **search** carried out?
 - How is **hypothesis** expressed?
 - How do we measure **quality** of hypothesis?
 - **What** data and **how much**?
- Practical issues:
 - How to gather and pre-process data
 - How to express hypotheses
 - How to measure performance
 - How to interpret results

Oct 2009

NUI Galway (8)



Profile of DM&ML Group

National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh

- €1.7m in direct research funding in past 5 years
 - Enterprise Ireland
 - SFI
 - EU FP6
- Research outputs:
 - Publish in leading conferences and journals
 - 3 Patents
 - University spin-out company
- Group includes:
 - 5 MSc/PhD researchers
 - 4 post-docs
 - 1 software engineer
- Strong international collaborations:
 - University of Helsinki, University College Dublin
 - University of California: Berkeley and Irvine



Oct 2009

NUI Galway (9)



Supervised Learning

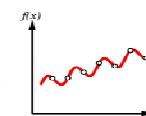
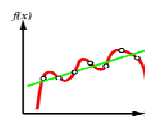
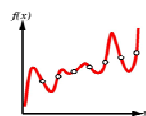
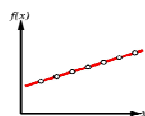
National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh

Also known as **Pure Inductive Learning**:

- Consider the *labels* (classes) to be *outputs* of some function f , the observed *attributes* are its *inputs*
- The function f is *unknown*; we want to discover it (or an approximation of it)
- All we have is a set of *examples*: inputs x and their corresponding outputs $f(x)$
- The task: **given examples, return a function h (the *hypothesis*) that approximates the 'true' function f**

■ Key issues:

- Hypothesis language
- Bias
- Overfitting



Oct 2009

NUI Galway (10)



Classification

National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh

- Is unknown sample X or O ?
- Goal
 - Find a model that correctly classifies a new sample, x_i

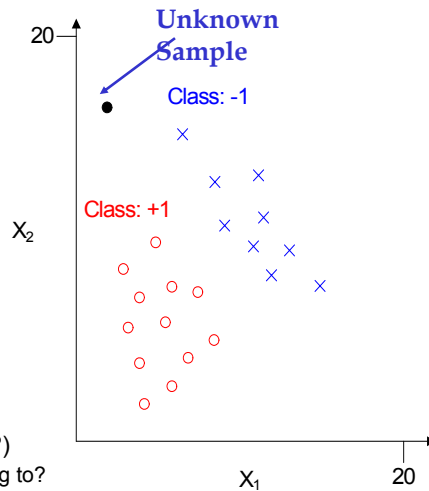
- X's are assigned : -1
- O's are assigned : +1

Training Data – $(x_i; y_i)$

- s_1 : (11, 3: +1) O
- s_2 : (3, 10: +1) O
- s_3 : (14, 17: -1) X
- s_4 : (16, 16: -1) X
-

Unknown Sample $s = (3, 15: ?)$

- What class does this belong to?



Oct 2009

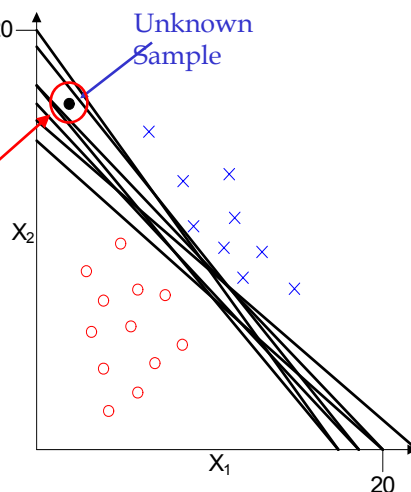
NUI Galway (11)



Linear Classification

National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh

- There are many linear classifiers to choose from
- Some linear classifiers will output a different prediction for our unknown sample
- Which should we pick?



Oct 2009

NUI Galway (12)

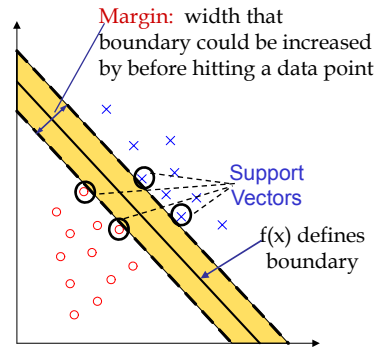


Classification: Support Vector Machines (1)

National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh

Linear Support Vector Machine:

- Finds maximum margin hyperplane that linearly separates two classes
- To classify a new sample, x :
 - $f(x) \geq 0$: sample is X
 - $f(x) < 0$: sample is O
- SVM training = finding optimum boundary (quadratic optimisation)
- Points that constrain boundary are its **support vectors**



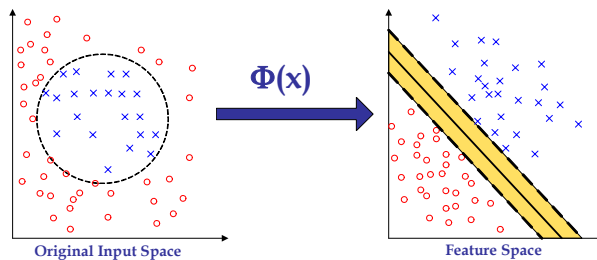
- Key point: boundary can be expressed in terms of **linear distance** between support vectors and point to be classified (dot product)

Oct 2009

NUI Galway (13)



Classification: Support Vector Machines (2)



- What do we do if data is **not linearly separable**?
- Map it on to a new feature space
- Kernel Function: $K(x, z) = \langle \Phi(x) \cdot \Phi(z) \rangle$
 - Distance between points in new feature space
 - Efficient mathematical trick: can compute boundary hyperplane in feature space **without** explicitly carrying out mapping

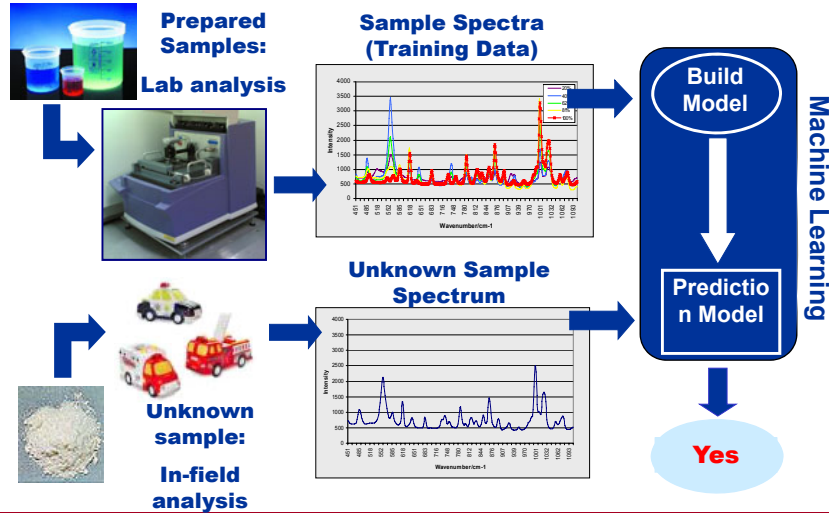
Oct 2009

NUI Galway (14)



Application: Identifying Chemical Spectra

National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh



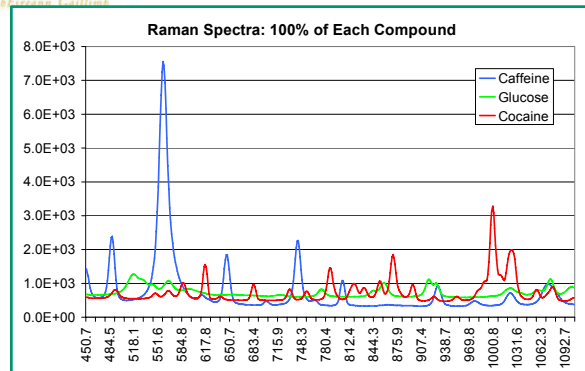
Oct 2009

NUI Galway (15)



Characteristics of Data

National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh



- Fluorescence
- Intensity variance
- Dimensionality
- Univariate
- Peak masking
- Distribution
- Noise

- **Goals:**
 - Develop new techniques that account for these characteristics
 - Focus on mixtures
 - Visualisation and added value for chemometricians
 - Ease of use for non-chemometricians

Oct 2009

NUI Galway (16)



One of Our Solutions: SVMs

National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh

- **Custom Kernel for Spectral Data**
 - Incorporates knowledge related to spectral domain
- Localised spectral regions are key to target identification
=> should compare regions of spectra
- Use pure target spectrum to assign different levels of importance to different regions

Oct 2009

NUI Galway (17)



WS Kernel: Classification

National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh

%Error 10x10 fold CV

Target	Linear Kernels			RBF Kernels			PCR
	Std.	+PCA	WSLin	Std.	+PCA	WSRBF	
Acet	2.73	3.08	1.93	2.59	3.83	0.41	4.47
Chloro	5.15	4.29	2.82	4.42	5.08	2.35	8.51
Dichloro	4.21	3.00	2.43	4.51	2.74	2.39	18.73
Trichloro	2.22	0.96	0.96	1.26	0.48	0.87	7.87
Cform	3.21	2.00	0.91	3.82	1.69	0.87	13.49

- WS Kernels outperform standard kernels and PCR
- PCA+SVM is better than standard SVM, but WS kernels are better

Oct 2009

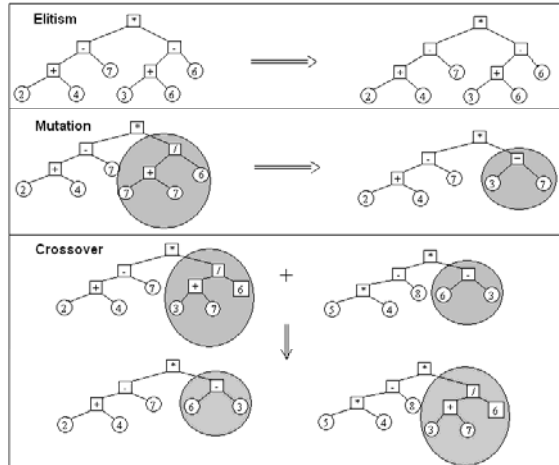
NUI Galway (18)



Genetic Programming

National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh

- Evolutionary paradigm
- Breeding of individuals, chosen based on their fitness, over a number of generations
- Represent formulae: functions and variables
 - Algebraic functions
 - Variables: spectrum points (wavelengths)
- Can be used to “grow” classification functions and logic rules



Oct 2009

NUI Galway (19)



GP for Spectroscopy Data

National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh

- Mixtures of up to 4 solvents
- Small dataset
- Overall, GP outperforms others on this data

	Prediction (Number incorrect out of 10)				
	Cyclohexane	Acetonitrile	Acetone	Toluene	Overall Error
PCR	2	0	1	0	7.5 %
PLS	1	0	1	0	5.0 %
NN	3	0	0	0	7.5 %
NB	0	2	1	4	17.5 %
Ripper	1	3	4	2	25.0 %
C4.5	0	2	4	4	25.0 %
GP-F1	4	0	0	1	12.5 %
GP-F2	0	0	0	0	0.0 %

- Not sensitive to GP parameter settings
- Certainty Factor optimisation a key mechanism

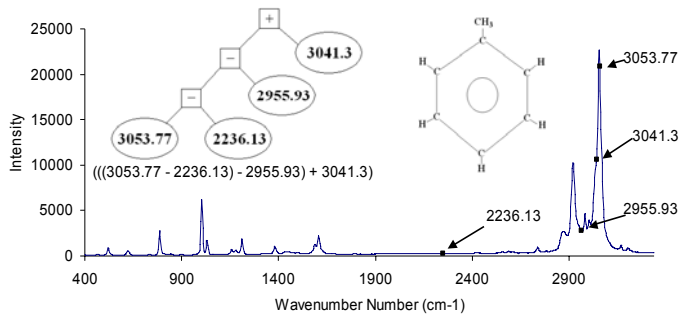
Oct 2009

NUI Galway (20)



Insight into Spectrum Classification

National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh



- Points 3053, 3041 and 2955 are in C-H bond region
 - Corresponds to domain experts' expectations
 - Intuitive visualisation of decisions

Oct 2009

NUI Galway (21)



Application: Schizophrenia Testing with fMRI

National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh

- Data collected in US NIH Project:
 - Biomedical Informatics Research Network
 - UC Irvine a partner
- 250 images over 9 sites
 - Health status unknown for 11
 - Multiple visit images for 10 more (BWH): use Visit 1
 - 115 Control + 114 Schiz. unique individuals
- 108 Controls + 105 Schiz after rejections
 - Registration failed; Poor slice prescription; Missing Data
 - Dropped one extra because scan data looked poor

Oct 2009

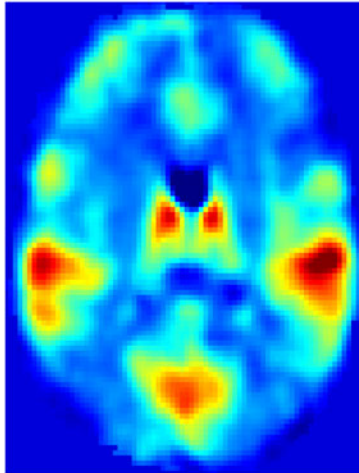
NUI Galway (22)



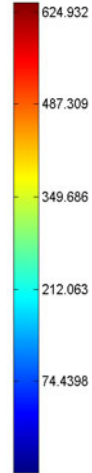
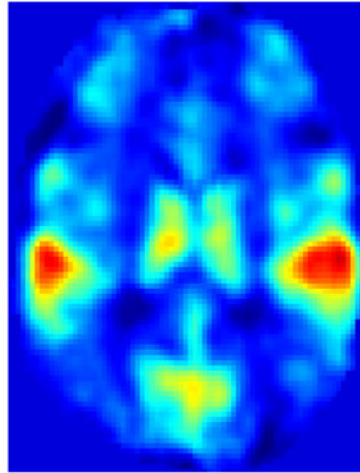
IA+MN Mean Images: Controls & Schiz.

National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh

HEALTH_IA+MN: Mean image of Controls (27 subjects)



HEALTH_IA+MN: Mean image of Schiz. (28 subjects)



Oct 2009

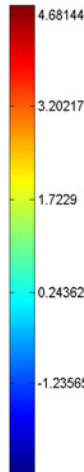
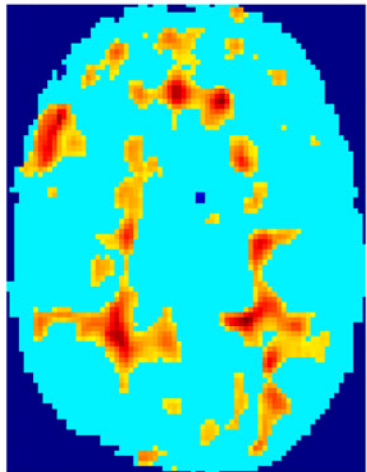
NUI Galway (23)



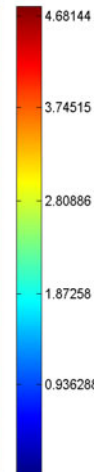
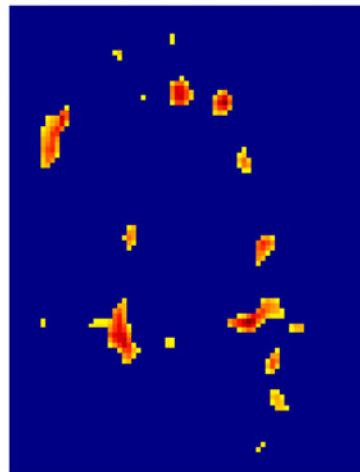
T-Tests at $\alpha = 5\%$, 0.5%

National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh

HEALTH_IA+MN: T-Test of Controls vs Schiz.: threshold at alpha=5%



HEALTH_IA+MN: T-Test of Controls vs Schiz.: threshold at alpha=0.5%



Oct 2009

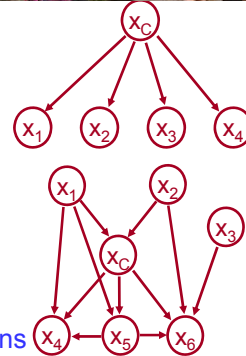
NUI Galway (24)



Bayesian Networks in One Slide

National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh

- Bayesian Networks:
 - Causal model how variables interact
 - Based on Probability
- Naive Bayes classifier:
 - Assume independence and direct causality: e.g. disease causes independent symptoms
 - Popularly used in classification (discriminative models)
- Our Research:
 - Classify using **full** Bayesian nets
 - Have found they perform better than NB
 - Have received little attention as classifiers
 - **Automatic discovery of rich model** of interactions



Oct 2009

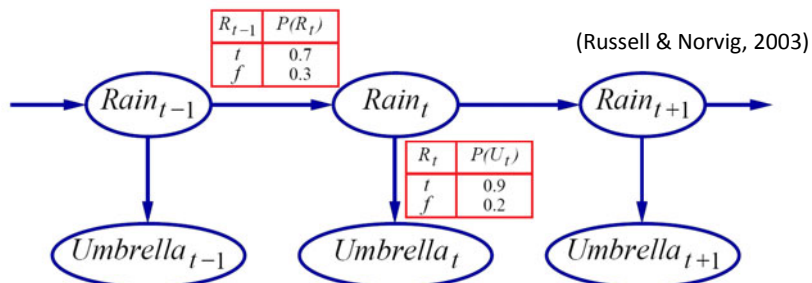
NUI Galway (25)



Dynamic BNs

National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh

- Temporal modelling
 - Tracking and prediction under uncertainty
- Key idea
 - Copy state and evidence variables at each time step
 - Allow arcs between times **t** and **t+1** (time is discretized)



Oct 2009

NUI Galway (26)



Application: ICU Data Analysis

National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh

■ Biomedical Informatics in Critical Care

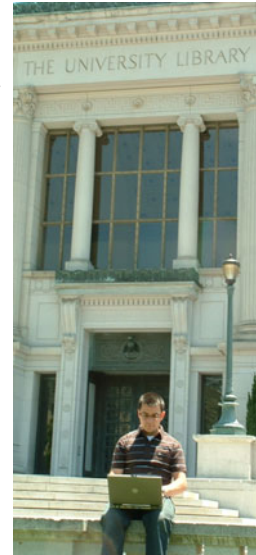
- Prof. Stuart Russell: Computer Science, Berkeley
- Prof. Geoff Manley: Neurotrauma, UCSF
- Dr Norm Aleks: PhD student, Berkeley

■ Objective: Bayesian Modelling for Clinical Decision Support

- Real-time interpretation of ICU data to determine most likely patient state
- Prediction of response to different treatments

■ Sources of Uncertainty:

- Measurement errors & artefacts
- Missing data
- Incomplete understanding



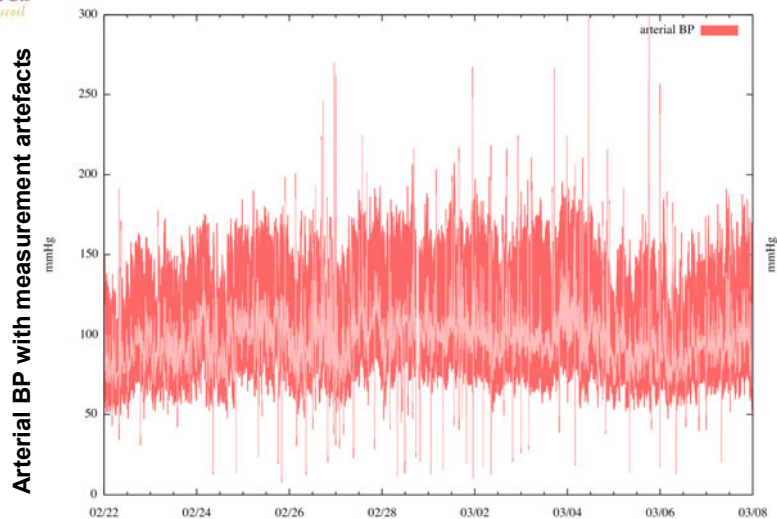
Oct 2009

NUI Galway (27)



Blood Pressure Artefacts

National University of Ireland, Galway
Ollscoil



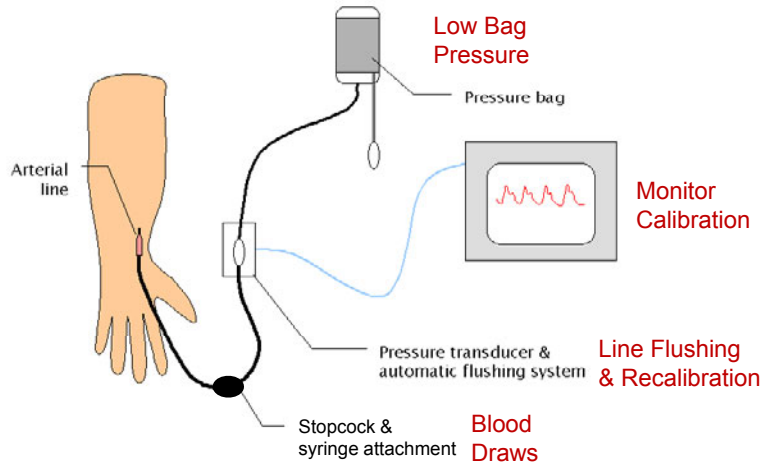
(Aleks et al., NIPS-2008)

Oct 2009

NUI Galway (28)



Sources of Artefacts

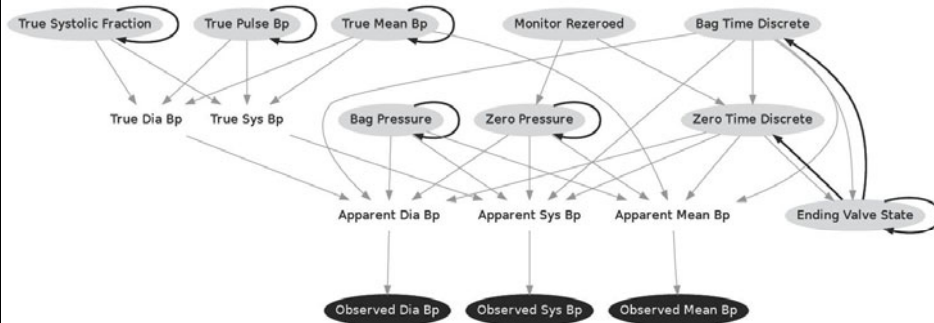


(Aleks et al., NIPS-2008)



Dynamic Bayesian Net Model of Artefacts

- Representation of domain knowledge of causes of artefacts; allows for reasoning under uncertainty

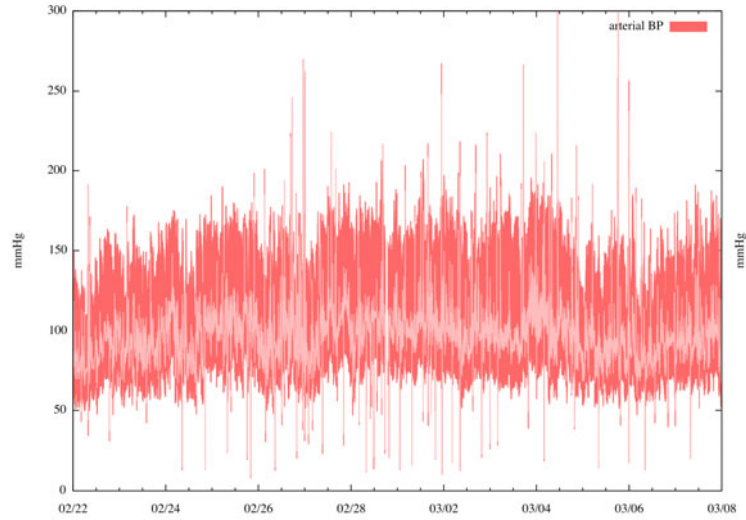


(Aleks et al., NIPS-2008)



Arterial BP With Artefacts

National University of Ireland, Galway
Ollscoil



(Aleks et al., NIPS-2008)

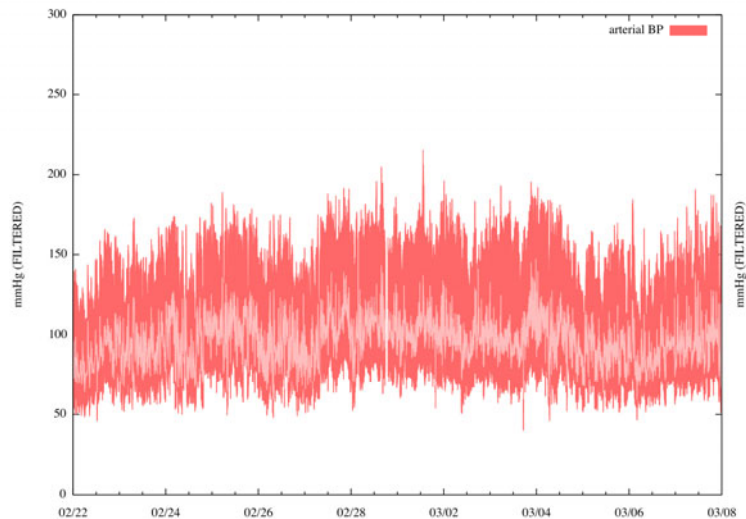
Oct 2009

NUI Galway (31)



After Filtering with DBN

National University of Ireland, Galway
Ollscoil



(Aleks et al., NIPS-2008)

Oct 2009

NUI Galway (32)



Drug Delivery Modelling

National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh

- SFI-funded project, started Oct 2008
 - M Madden; P Piiorinen; N Madden
 - J Laffey (UCHG); S Russell (Berkeley)
- Pharmacokinetics:
 - Study of how drugs are processed by the body
 - Absorption, Distribution, Metabolism, Excretion
- Given:
 - Basic PK data
 - Potentially inaccurate records (e.g. time/dosage)
 - Incomplete observations of effects
- Inference:
 - E.g. Concentration of drug in serum
- Personalised medicine
 - Individualise drug effects; simulate effects of different therapies

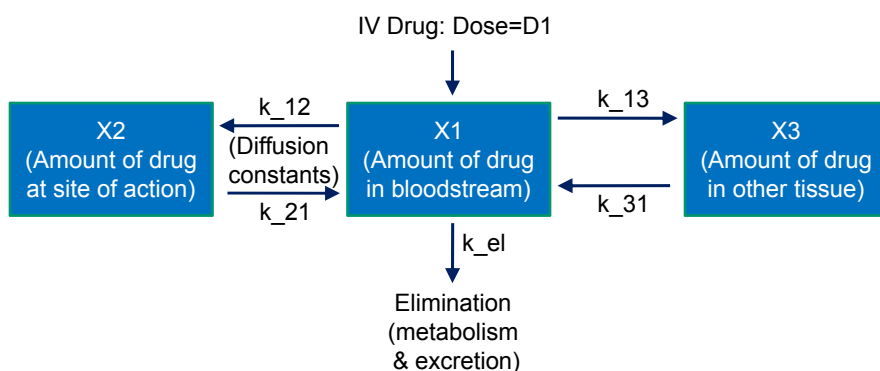
Oct 2009

NUI Galway (33)



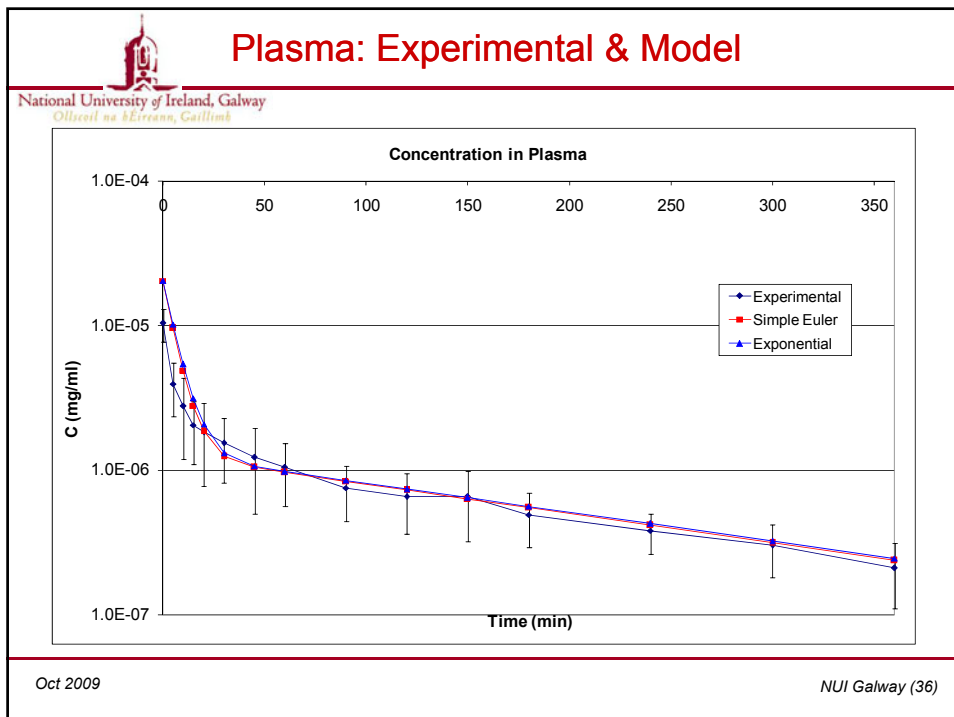
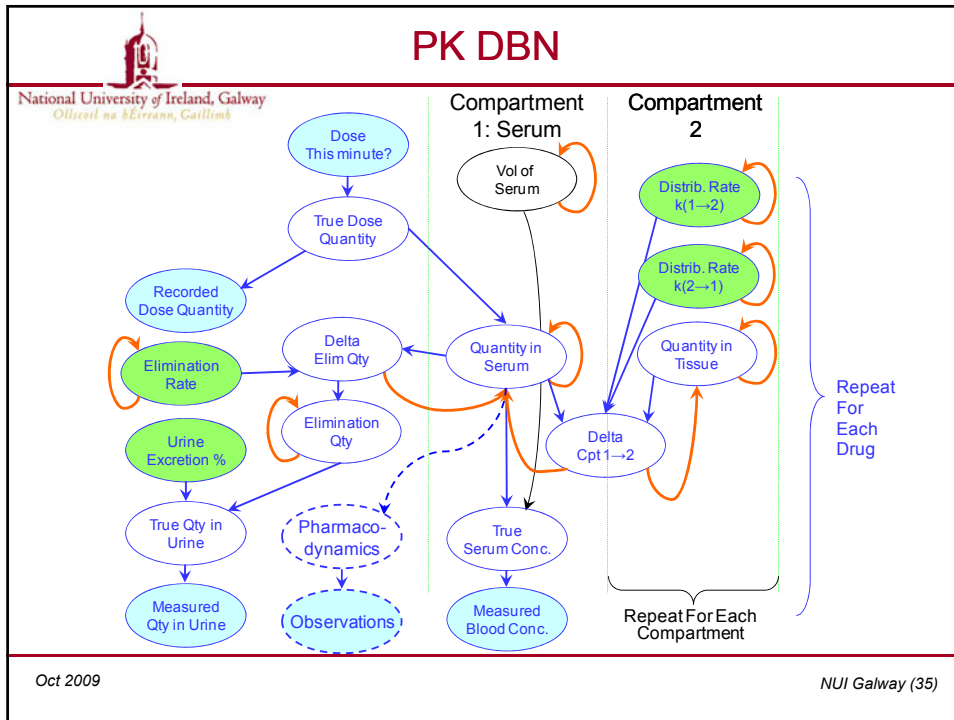
Basis: Standard Multi-Compartment Model

National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh



Oct 2009

NUI Galway (34)





Reinforcement Learning

National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh

- Learning through trial and error
 - Agent explores environment
 - Rewards for successful outcomes
 - Punishments for unsuccessful ones
- Seeks to maximise reward
 - Acts randomly at first
 - Builds map of state/actions -> rewards
 - Gradually develops an optimal strategy
- Applications
 - Poorly-defined domains
 - Don't know **how** task is done well, just **whether** it is done well

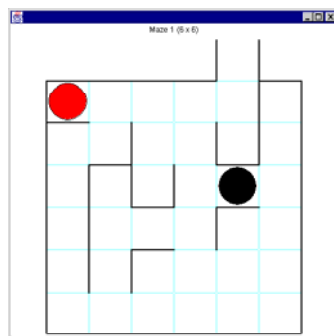
Oct 2009

NUI Galway (37)



Theseus and the Minotaur (1)

National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh



- R. Abbot:
www.logicmazes.com
- Designed for humans
- Minotaur follows fixed strategy

- RL agent initially knows **nothing** about maze
- Actions: N, S, E, W, Delay
- Reward: 1 for exit, -1 for Minotaur

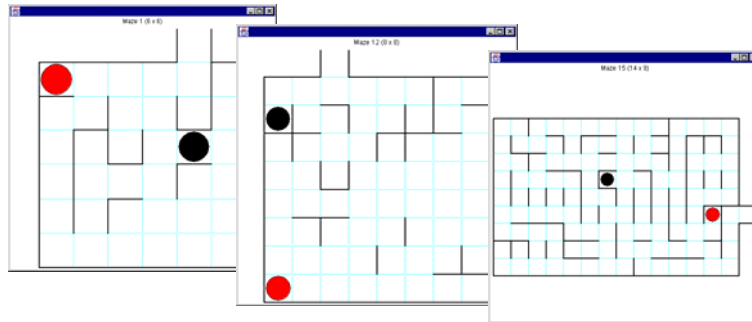
Oct 2009

NUI Galway (38)



Theseus and the Minotaur (2)

National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh



- 15 mazes: Progressive difficulty
- Standard RL: Each maze solved from scratch
 - Not ideal if learning to drive a car!
- Our goal: transfer experience gained in simpler problems to new, harder problems

Oct 2009

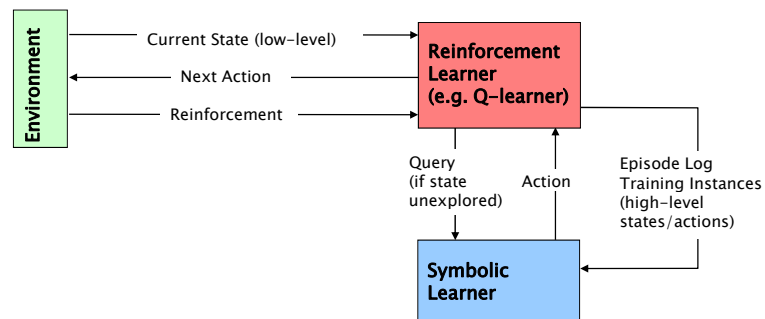
NUI Galway (39)



Progressive RL

National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh

- Alternating *experimentation* and *introspection*
- **Experimentation**: Q-Learning
- **Introspection**: Decision Tree/Naive Bayes learning



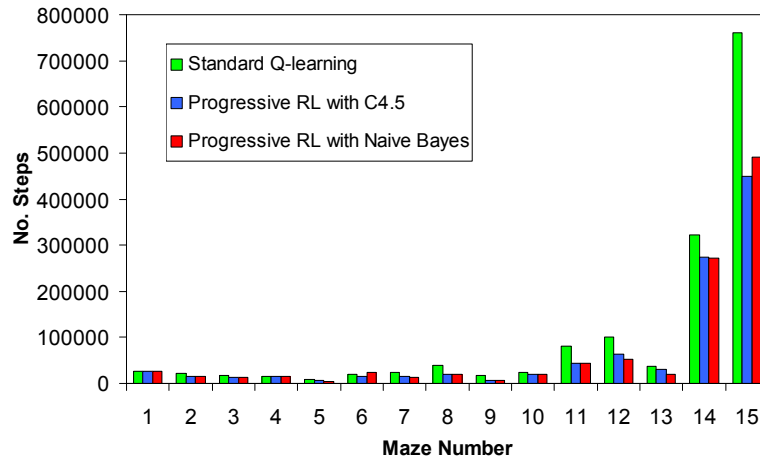
Oct 2009

NUI Galway (40)



Steps to First Win

National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh



Oct 2009

NUI Galway (41)



Conclusions & Observations

National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh

- New datasets motivates new algorithms
 - Can still have general applicability
 - Helpful to consider data characteristics

- Close collaboration with domain experts is invaluable
 - Sanity checking
 - Baseline comparisons
 - Pre-processing (normalisation, baseline correction)
 - Interpretation of results
 - Assessing value of new work

- GUIs and visualisations aid evaluation/adoption

Oct 2009

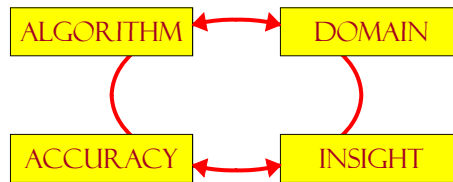
NUI Galway (42)



Conclusions & Observations

National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh

- **Insight** is essential for adoption
 - No good if accuracy is lost!
- Best to **build it in** to algorithms
 - Tailor algorithm to application
 - Based on domain considerations
 - Positive influence on accuracy



Oct 2009

NUI Galway (43)