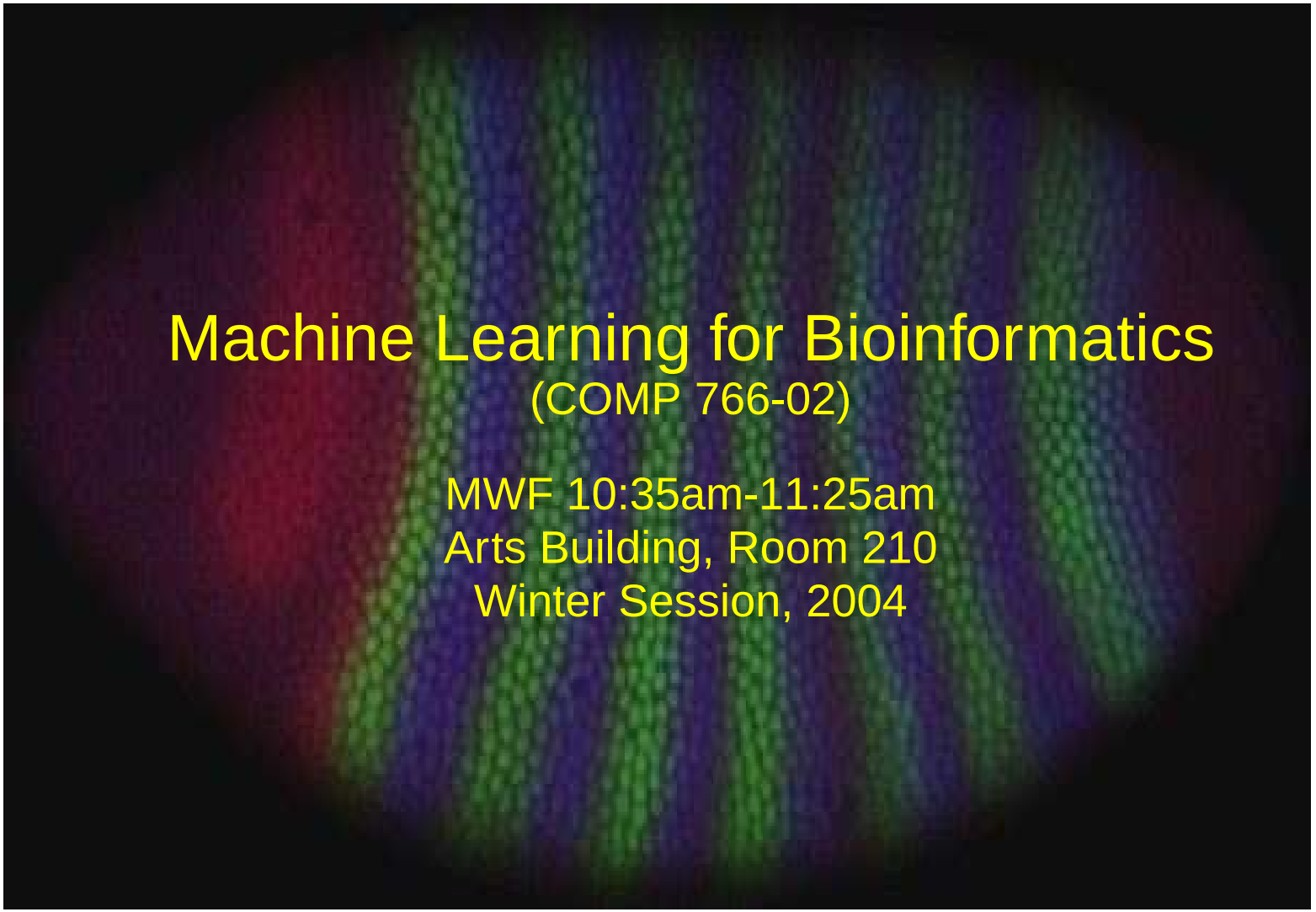

Machine Learning for Bioinformatics
(COMP 766-02)
MWF 10:35am-11:25am
Arts Building, Room 210
Winter Session, 2004

What is machine learning?

(or data mining, pattern recognition, knowledge discovery, signal processing, system identification. . . ?)

From “Data Mining: Practical Machine Learning Tools and Techniques with JAVA Implementations” by Ian H. Witten and Eibe Frank:

> *If* data *is characterized as recorded facts, then* information *is the set of patterns, or expectations that underlie the data. . . information that is potentially important but has not yet been discovered or articulated. Our mission is to bring it forth.*

ML in a bioinformatics context. . .

- . . . is computer-aided *discovery science.*

Exploration	Visualization
Summarization	Generalization
Prediction	Estimation
Modeling	Hypothesis generation

- It's usually *not* about testing a specific hypothesis, as is common is statistics.
  (Though some ML borrows heavily from statistics.)

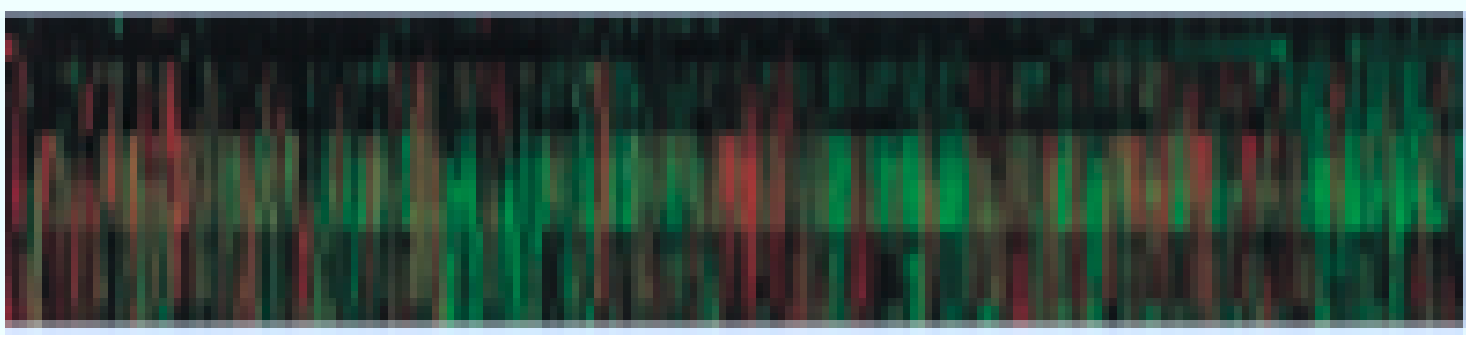
We'll study four problem types:

1. Summarization
2. Prediction
3. Probabilistic modeling
4. Modeling dynamical systems

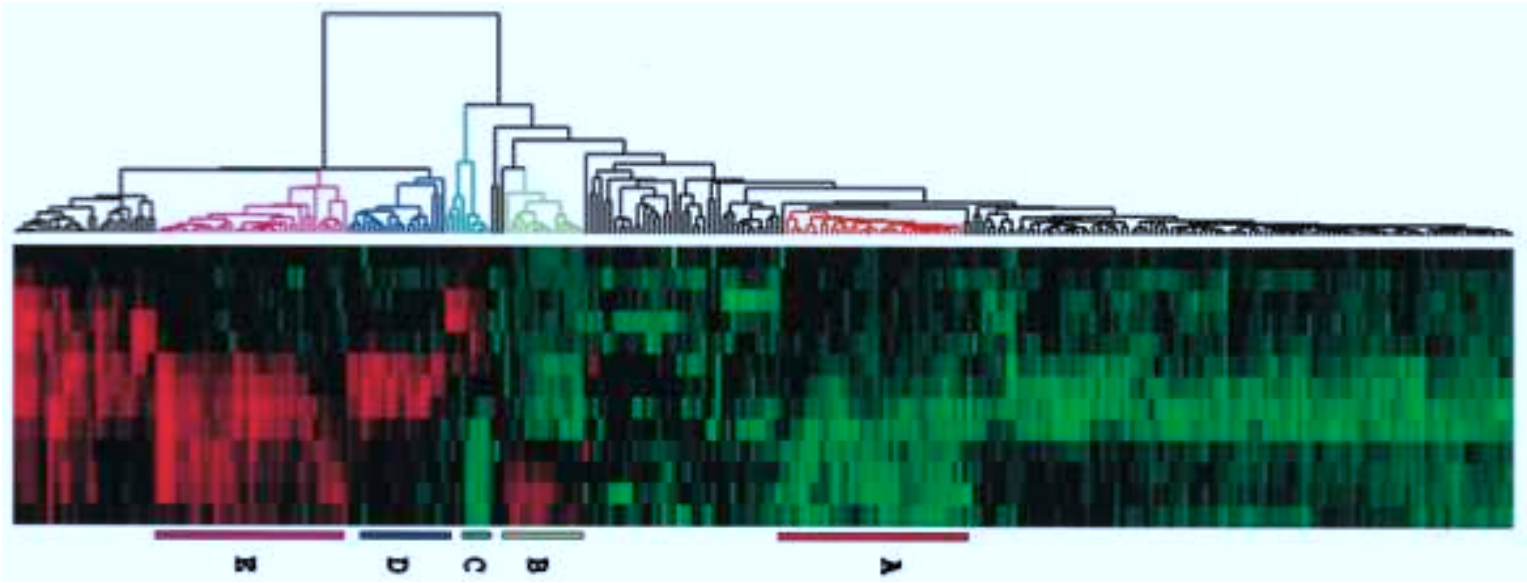
We'll study four problem types:

1. Summarization, e.g.:
   - Which genes express similarly?
2. Prediction
3. Probabilistic modeling
4. Modeling dynamical systems

Consider a time-series of microarrays (columns are genes, rows are time points):



Hierarchically clustering the genes gives a clearer picture.



from Eisen *et al.*, PNAS vol. 95, pp. 14863–14868, 1998.

We'll study four problem types:

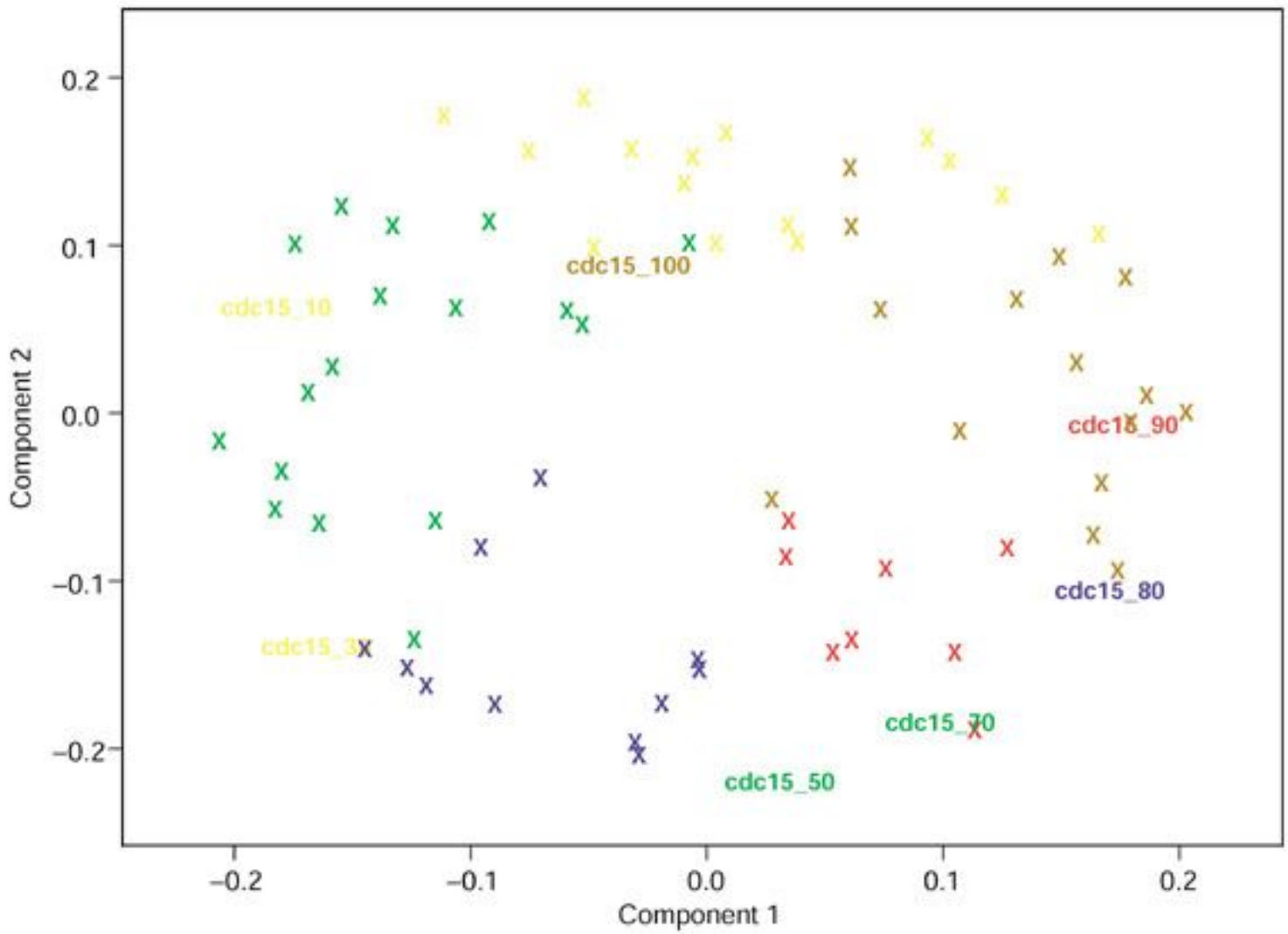
1. Summarization, e.g.:
   - Which genes express similarly?
2. Prediction
3. Probabilistic modeling
4. Modeling dynamical systems

We'll study four problem types:

1. Summarization, e.g.:
   - Which genes express similarly?
   - Which cancers are similar? Are there subtypes?
2. Prediction
3. Probabilistic modeling
4. Modeling dynamical systems

We'll study four problem types:

1. Summarization, e.g.:
   - Which genes express similarly?
   - Which cancers are similar? Are there subtypes?
   - How can we visualize high dimensional data?
2. Prediction
3. Probabilistic modeling
4. Modeling dynamical systems



From Landgrebe, Wurst, Welzl. Genome Biology Vol. 3 Iss. 4 (2002)

We'll study four problem types:

1. Summarization, e.g.:
   - Which genes express similarly?
   - Which cancers are similar? Are there subtypes?
   - How can we visualize high dimensional data?
2. Prediction
3. Probabilistic modeling
4. Modeling dynamical systems

We'll study four problem types:

1. Summarization, e.g.:
   - Which genes express similarly?
   - Which cancers are similar? Are there subtypes?
   - How can we visualize high dimensional data?
   - How many "degrees of freedom" in the human genome? Proteome?
2. Prediction
3. Probabilistic modeling
4. Modeling dynamical systems

We'll study four problem types:

1. Summarization
2. Prediction, e.g.
3. Probabilistic modeling
4. Modeling dynamical systems

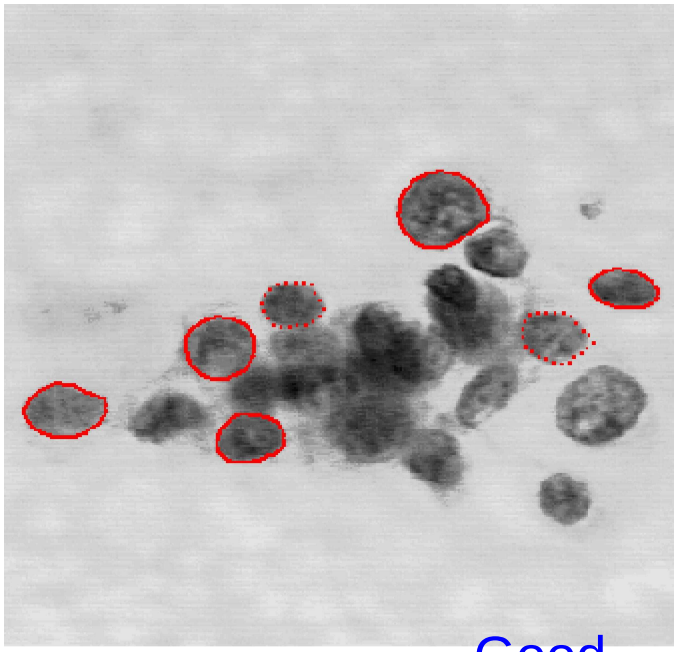
We'll study four problem types:

1. Summarization
2. Prediction, e.g.
   - Given measurements X, is the tumor benign or malignant?
3. Probabilistic modeling
4. Modeling dynamical systems

We'll study four problem types:

1. Summarization
2. Prediction, e.g.
   - Given measurements X, is the tumor benign or malignant?
   - Given medical test results X, how long does the patient live?
3. Probabilistic modeling
4. Modeling dynamical systems

From http://www.cs.wisc.edu/∼olvi



⇒ Features such as tumor size (from surgery), and cell area, perimeter, texture (from image).

	Good	no chemo recommended
⇒	Intermediate	chemo likely to prolong survival
	Poor	chemo may or may not enhance survival

We'll study four problem types:

1. Summarization
2. Prediction, e.g.
   - Given measurements X, is the tumor benign or malignant?
   - Given medical test results X, how long does the patient live?
3. Probabilistic modeling
4. Modeling dynamical systems

We'll study four problem types:

1. Summarization
2. Prediction, e.g.
   - Given measurements X, is the tumor benign or malignant?
   - Given medical test results X, how long does the patient live?
   - Is DNA sequence X a transcription factor binding site?
3. Probabilistic modeling
4. Modeling dynamical systems

We'll study four problem types:

1. Summarization
2. Prediction, e.g.
   - Given measurements X, is the tumor benign or malignant?
   - Given medical test results X, how long does the patient live?
   - Is DNA sequence X a transcription factor binding site?
   - Does amino acid sequence X fold into α-helix, β-sheet, ...
3. Probabilistic modeling
4. Modeling dynamical systems

We'll study four problem types:

1. Summarization
2. Prediction, e.g.
   - Given measurements X, is the tumor benign or malignant?
   - Given medical test results X, how long does the patient live?
   - Is DNA sequence X a transcription factor binding site?
   - Does amino acid sequence X fold into α-helix, β-sheet, ...
   - Do proteins X and Y interact?
3. Probabilistic modeling
4. Modeling dynamical systems

We'll study four problem types:

1. Summarization
2. Prediction
3. Probabilistic modeling, e.g.
4. Modeling dynamical systems

We'll study four problem types:

1. Summarization
2. Prediction
3. Probabilistic modeling, e.g.
   - If patient has nausea and fever, what is chance of influenza?
4. Modeling dynamical systems

We'll study four problem types:

1. Summarization
2. Prediction
3. Probabilistic modeling, e.g.
   - If patient has nausea and fever, what is chance of influenza?
   - If patient has influenza, what is chance of nausea and dizziness?
4. Modeling dynamical systems

We'll study four problem types:

1. Summarization
2. Prediction
3. Probabilistic modeling, e.g.
   - If patient has nausea and fever, what is chance of influenza?
   - If patient has influenza, what is chance of nausea and dizziness?
   - What happens to genes X and Y if Z and W are knocked out?
4. Modeling dynamical systems

We'll study four problem types:

1. Summarization
2. Prediction
3. Probabilistic modeling, e.g.
   - If patient has nausea and fever, what is chance of influenza?
   - If patient has influenza, what is chance of nausea and dizziness?
   - What happens to genes X and Y if Z and W are knocked out?
   - (How) are bases in a TF binding site related?
4. Modeling dynamical systems

We'll study four problem types:

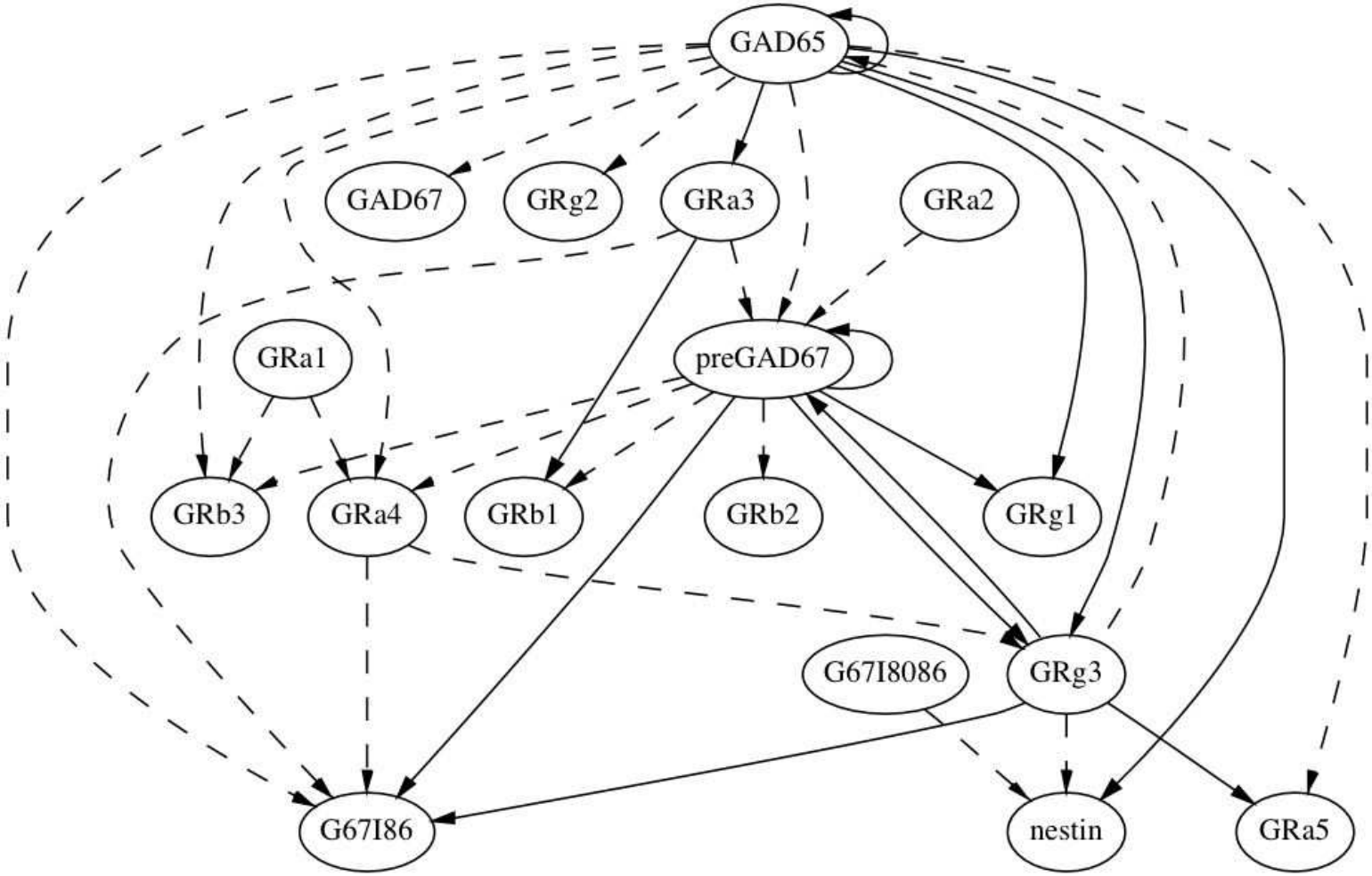
1. Summarization
2. Prediction
3. Probabilistic modeling
4. Modeling dynamical systems, e.g.

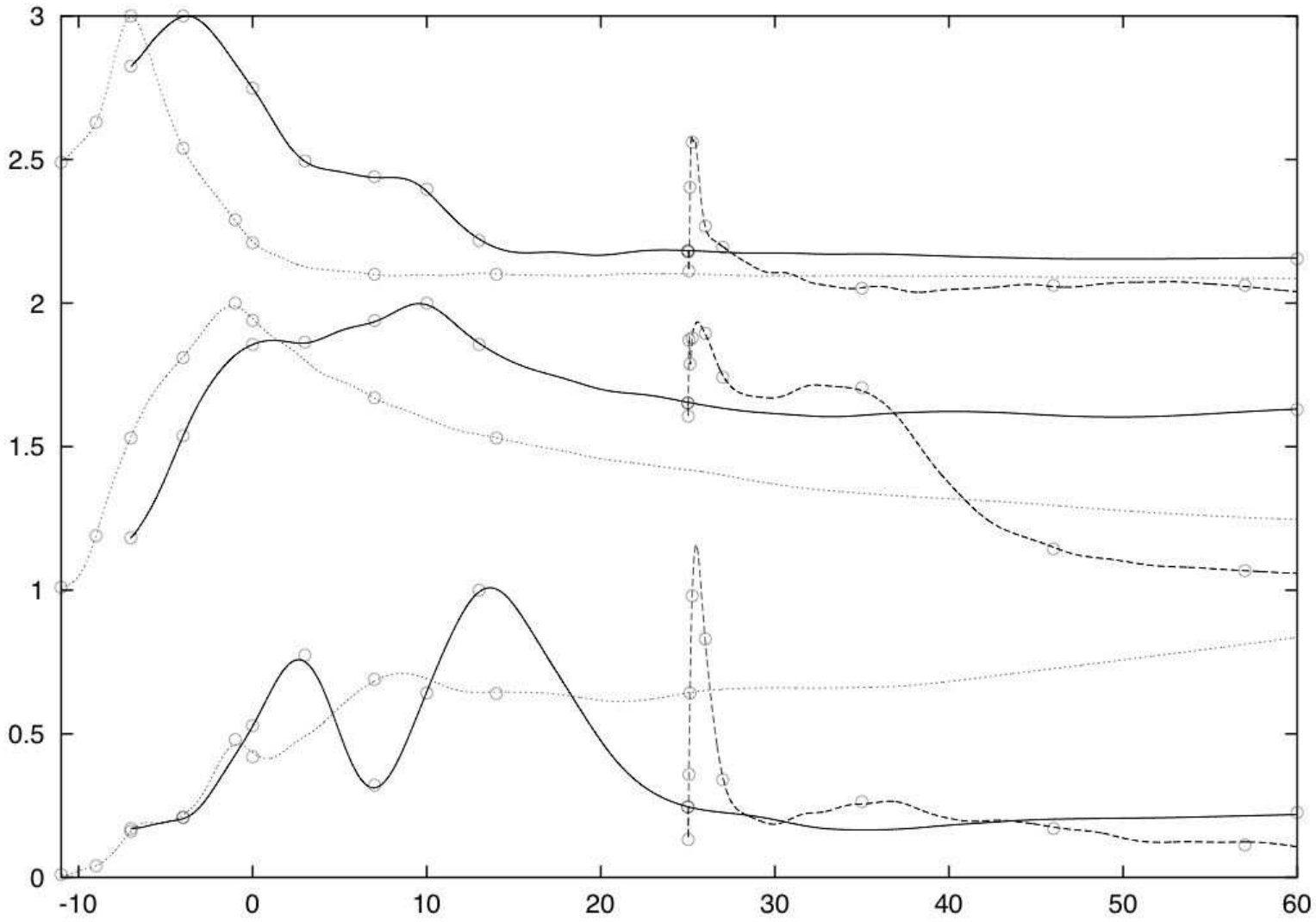
We'll study four problem types:

1. Summarization
2. Prediction
3. Probabilistic modeling
4. Modeling dynamical systems, e.g.
   - Explain changes in gene expression over time (e.g., during development, in response to environmental disturbance, in response to a drug).

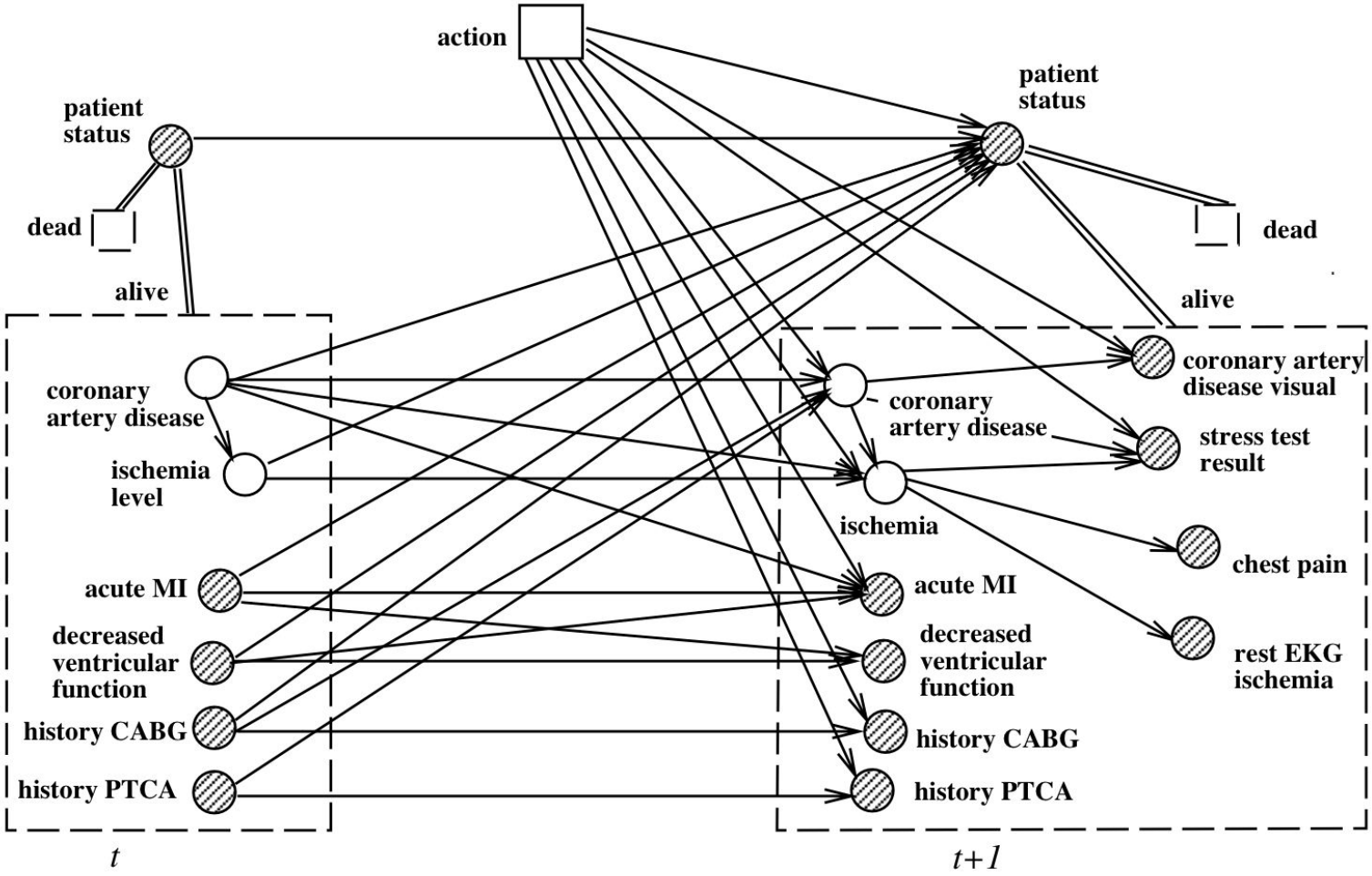
We'll study four problem types:

1. Summarization
2. Prediction
3. Probabilistic modeling
4. Modeling dynamical systems, e.g.
   - Explain changes in gene expression over time (e.g., during development, in response to environmental disturbance, in response to a drug).
   - Predict disease progression given current test results.


GAD65
GAD67
GRg2
GRa3
GRa2
GRa1
preGAD67
GRb3
GRa4
GRb1
GRb2
GRg1
G67I8086
GRg3
G67I86
nestin
GRa5



From D’Haeseleer, Wen, Fuhrman, Somogyi. PSB’99.



From Hauskrecht, Fraser. Proceedings of AIME, 1997.

Some methods we'll cover

	Summarization	Prediction	Probabilistic Modeling	Modeling Dynamics
Clustering	X			
PCA (Principal components analysis)	X			
ICA (Independent components analysis)	X			
MDS (Multi-dimensional scaling)	X			
Linear Models		X		X
Trees		X	~	
ANNs (artificial neural networks)	X	X	X	X
Bayesian networks	X	X	X	X
SVMs (support vector machines)		X		

Homework 0

If you are registered or auditing:

- Email me (at perkins@mcb.mcgill.ca) from your preferred email account, so I can construct my class mailing list.
- So I have a sense of the backgrounds of who is attending, also tell me:
  - Your home department (if any)
  - Ugrad / Grad / Postdoc / Prof / Other
  - Approximate names or subjects of most advanced courses taken in math, stats, comp sci which may pertain to this course.