Pattern Recognition

Using expert systems to classify fluorescence EEMS

Dr Lee Chapman
University of Birmingham

Outline

- Who am I?
- The problem to solve…
- Background to Pattern Recognition
- Machine Learning Algorithms & Software
  - PCA / PARAFAC
  - Artificial Neural Networks
  - Decision Trees
- Conclusions

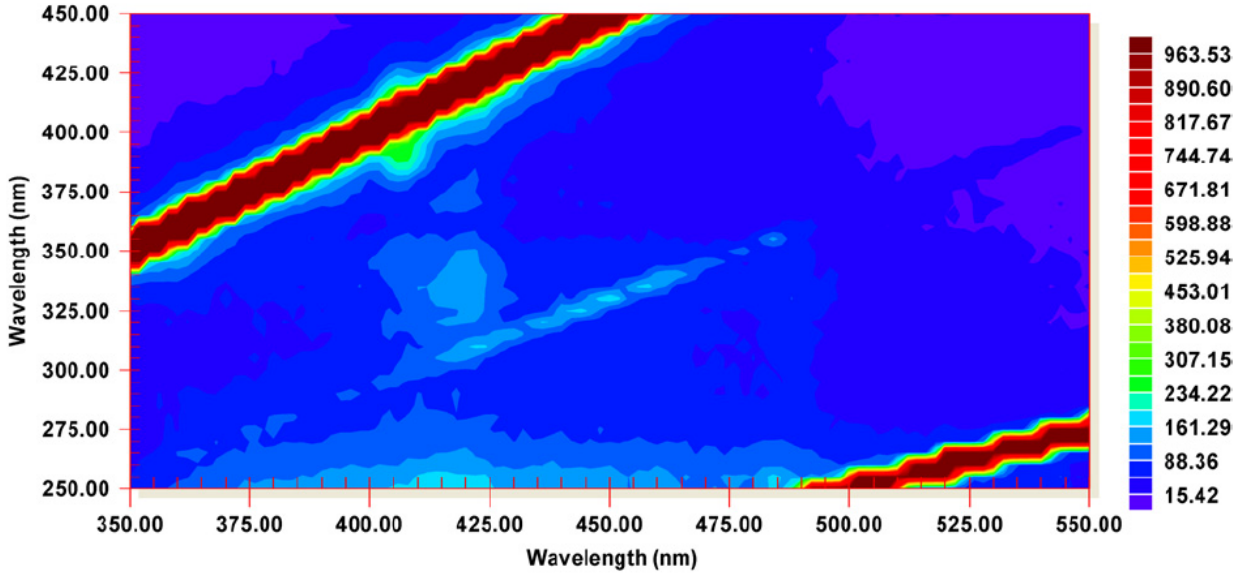
Who am I?

- Not a philosopher!
- Geographer by trade:
  - 50% Climatologist
  - 40% Geomaticist
  - 10% Computer Scientist (generous)
- Director of Lumin-S services
- Have used fluorescence in the winter maintenance market

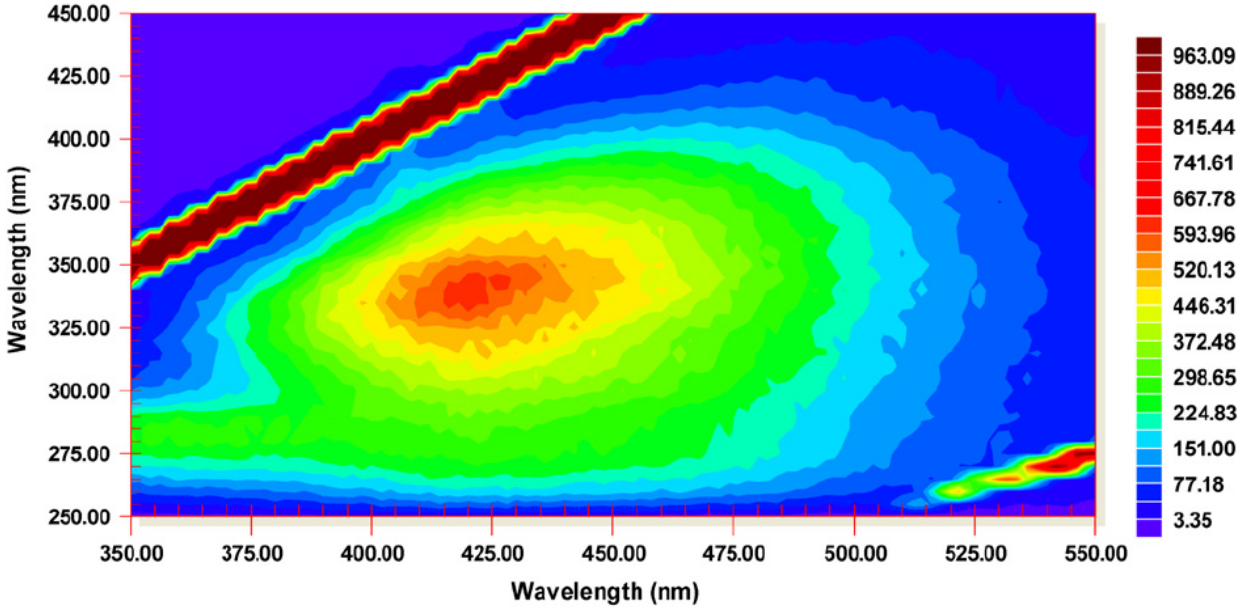
UNIVERSITY OF BIRMINGHAM

Residual Salt Monitoring

- Big problem for highway engineers
- Currently rely on a ‘point’ forecast of residual salt at an outstation
- Can currently only be measured by contact techniques
- Can it be done via remote sensing?
- Due to road salt additives (molasses), this may be possible thanks to fluorescence



EEM of Road Surface Water

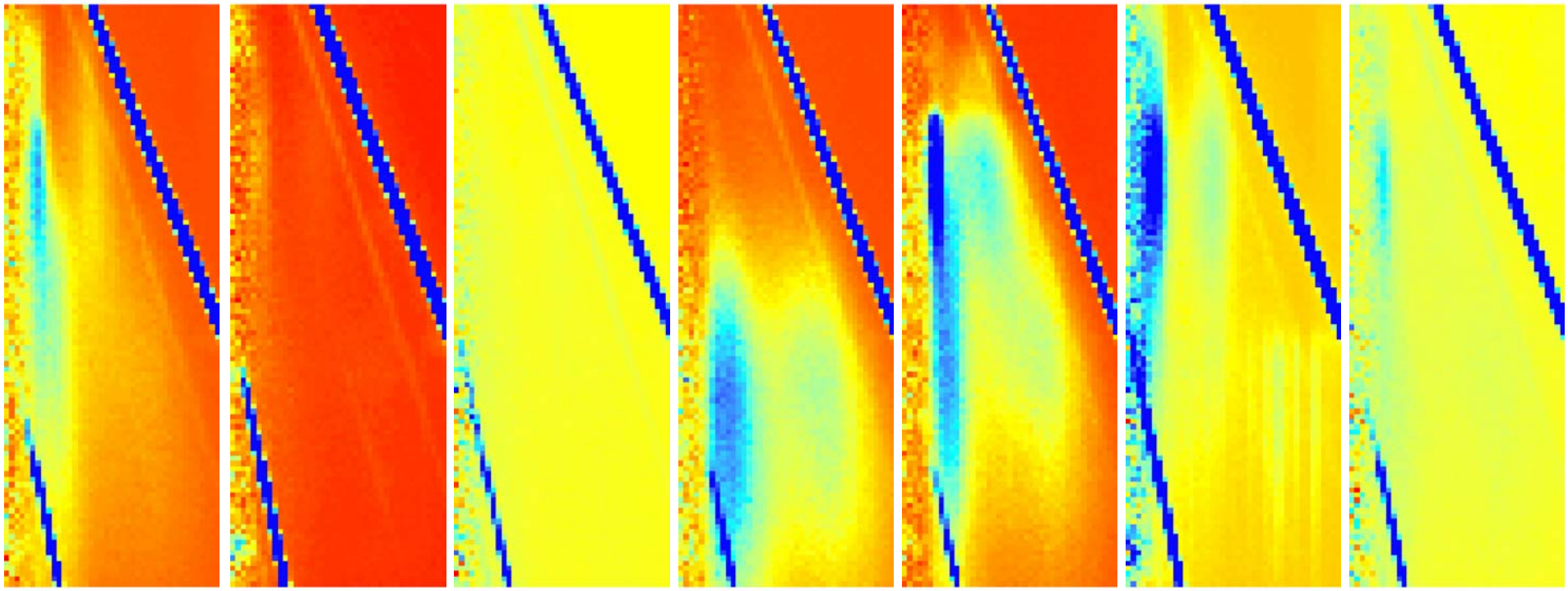


EEM of Road Surface Water doped with Rock Salt and Safecote

UNIVERSITY OF BIRMINGHAM

The problem to be solved…

- Can TLS EEMs be automatically classified using an expert system based on modern computer science techniques?
- Based on a water quality problem, subdivided into 7 classes:
  - Canal
  - Marine
  - Groundwater
  - River Water
  - Leachate
  - Raw Sewage
  - Treated Sewage



Canal | Marine | Groundwater | River Water | Leachate | Raw Sewage | Treated Sewage

Training dataset of two of each class
+ Separate testing dataset of one of each class

UNIVERSITY OF BIRMINGHAM

Pattern Recognition

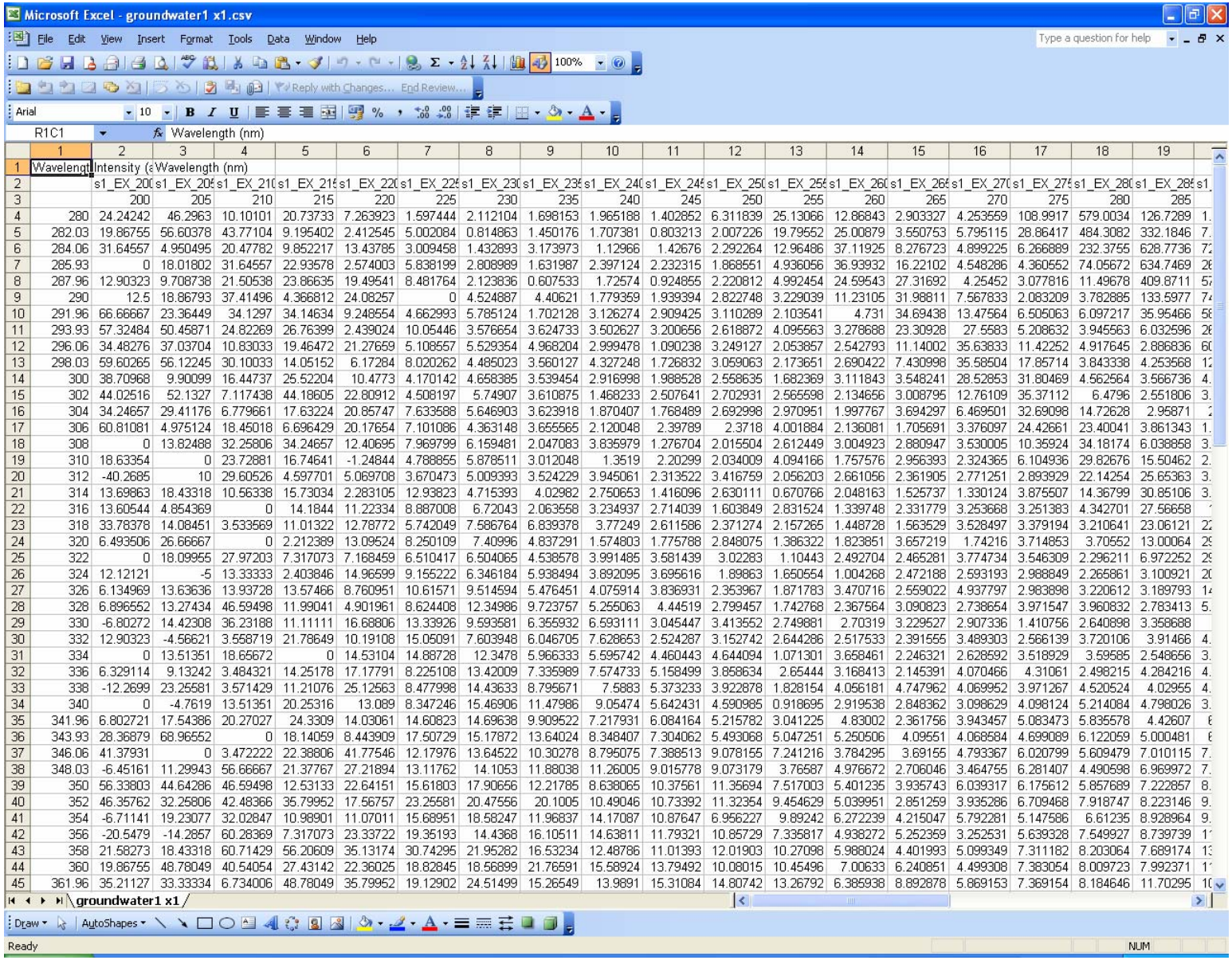
- ‘The act of taking in raw data and taking an action based on the category of that data.’
- Aims to classify data based on ‘a priori’ knowledge or statistical information extracted from the patterns.
- Supervised classification is based on a set of pre-classified patterns (training set)
- Algorithms are then used to detect statistical regularities in the data to classify new data
- There are many ‘machine learning’ algorithms out there!

Pattern Recognition

- Applications can take on many forms:
  - Speech recognition
  - Email text classification (spam / not spam)
  - University plagiarism software (more on this later!)
  - Number plate recognition
  - Human face recognition
- Last two examples are examples of image analysis.

Plagiarism Software

- Simple pattern recognition – looks for a common series of letters, words or numbers
- For example, *'Simphile' uses the common compression algorithm gzip as its pattern detection engine. Let us say that we are comparing file A and file B. We compress file A to determine how small it can get. We then compress file B to see the amount it will shrink. Finally, we compress file A+B. If gzip(A+B) is significantly less than gzip(A) + gzip(B), then that means files A and B share patterns!*
- Could it be used on the .csv files commonly outputted from fluorescence spectrophotometers?
- Lets have a go…

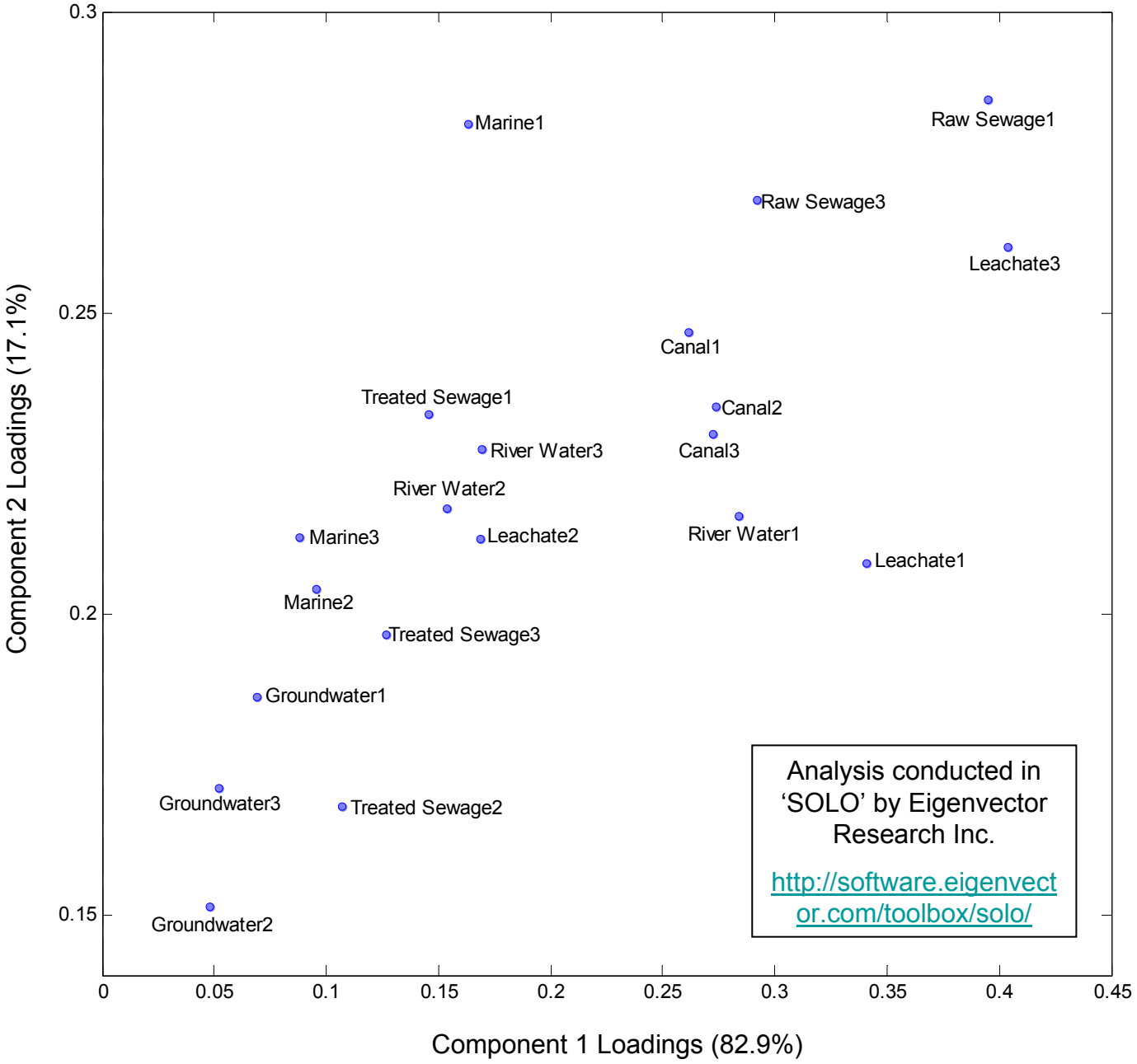


	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	
1	Wavelengt	Intensity (a	Wavelength (nm)																	
2		s1_EX_200	s1_EX_205	s1_EX_210	s1_EX_215	s1_EX_220	s1_EX_225	s1_EX_230	s1_EX_235	s1_EX_240	s1_EX_245	s1_EX_250	s1_EX_255	s1_EX_260	s1_EX_265	s1_EX_270	s1_EX_275	s1_EX_280	s1_EX_285	s1_
3		200	205	210	215	220	225	230	235	240	245	250	255	260	265	270	275	280	285	
4	280	24.24242	46.2963	10.10101	20.73733	7.263923	1.597444	2.112104	1.698153	1.965188	1.402852	6.311839	25.13066	12.86843	2.903327	4.253559	108.9917	579.0034	126.7289	1.
5	282.03	19.86755	56.60378	43.77104	9.195402	2.412545	5.002084	0.814863	1.450176	1.707381	0.803213	2.007226	19.79552	25.00879	3.550753	5.795115	28.86417	484.3082	332.1846	7.
6	284.06	31.64557	4.950495	20.47782	9.852217	13.43785	3.009458	1.432893	3.173973	1.12966	1.42676	2.292264	12.96486	37.11925	8.276723	4.899225	6.266889	232.3755	628.7736	7
7	285.93	0	18.01802	31.64557	22.93578	2.574003	5.838199	2.808989	1.631987	2.397124	2.232315	1.868551	4.936056	36.93932	16.22102	4.548286	4.360552	74.05672	634.7469	2
8	287.96	12.90323	9.708738	21.50538	23.86635	19.49541	8.481764	2.123836	0.607533	1.72574	0.924855	2.220812	4.992454	24.59543	27.31692	4.25452	3.077816	11.49678	409.8711	5
9	290	12.5	18.86793	37.41496	4.366812	24.08257	0	4.524887	4.40621	1.779359	1.939394	2.822748	3.229039	11.23105	31.98811	7.567833	2.083209	3.782885	133.5977	7
10	291.96	66.66667	23.36449	34.1297	34.14634	9.248554	4.662993	5.785124	1.702128	3.126274	2.909425	3.110289	2.103541	4.731	34.69438	13.47564	6.505063	6.097217	35.95466	5
11	293.93	57.32484	50.45871	24.82269	26.76399	2.439024	10.05446	3.576654	3.624733	3.502627	3.200656	2.618872	4.095563	3.278688	23.30928	27.5583	5.208632	3.945563	6.032596	2
12	296.06	34.48276	37.03704	10.83033	19.46472	21.27659	5.108557	5.529354	4.968204	2.999478	1.090238	3.249127	2.053857	2.542793	11.14002	35.63833	11.42252	4.917645	2.886836	6
13	298.03	59.60265	56.12245	30.10033	14.05152	6.17284	8.020262	4.485023	3.560127	4.327248	1.726832	3.059063	2.173651	2.690422	7.430998	35.58504	17.85714	3.843338	4.253568	1
14	300	38.70968	9.90099	16.44737	25.52204	10.4773	4.170142	4.658385	3.539454	2.916998	1.988528	2.558635	1.682369	3.111843	3.548241	28.52853	31.80469	4.562564	3.566736	4.
15	302	44.02516	52.1327	7.117438	44.18605	22.80912	4.508197	5.74907	3.610875	1.468233	2.507641	2.702931	2.565598	2.134656	3.008795	12.76109	35.37112	6.4796	2.551806	3.
16	304	34.24657	29.41176	6.779661	17.63224	20.85747	7.633588	5.646903	3.623918	1.870407	1.768489	2.692998	2.970951	1.997767	3.694297	6.469501	32.69098	14.72628	2.95871	2
17	306	60.81081	4.975124	18.45018	6.696429	20.17654	7.101086	4.363148	3.655565	2.120048	2.39789	2.3718	4.001884	2.136081	1.705691	3.376097	24.42661	23.40041	3.861343	1.
18	308	0	13.82488	32.25806	34.24657	12.40695	7.969799	6.159481	2.047083	3.835979	1.276704	2.015504	2.612449	3.004923	2.880947	3.530005	10.35924	34.18174	6.038858	3.
19	310	18.63354	0	23.72881	16.74641	-1.24844	4.788855	5.878511	3.012048	1.3519	2.20299	2.034009	4.094166	1.757576	2.956393	2.324365	6.104936	29.82676	15.50462	2.
20	312	-40.2685	10	29.60526	4.597701	5.069708	3.670473	5.009393	3.524229	3.945061	2.313522	3.416759	2.056203	2.661056	2.361905	2.771251	2.893929	22.14254	25.65363	3.
21	314	13.69863	18.43318	10.56338	15.73034	2.283105	12.93823	4.715393	4.02982	2.750653	1.416096	2.630111	0.670766	2.048163	1.525737	1.330124	3.875507	14.36799	30.85106	3.
22	316	13.60544	4.854369	0	14.1844	11.22334	8.887008	6.72043	2.063558	3.234937	2.714039	1.603849	2.831524	1.339748	2.331779	3.253668	3.251383	4.342701	27.56658	1
23	318	33.78378	14.08451	3.533569	11.01322	12.78772	5.742049	7.586764	6.839378	3.77249	2.611586	2.371274	2.157265	1.448728	1.563529	3.528497	3.379194	3.210641	23.06121	2
24	320	6.493506	26.66667	0	2.212389	13.09524	8.250109	7.40996	4.837291	1.574803	1.775788	2.848075	1.386322	1.823851	3.657219	1.74216	3.714853	3.70552	13.00064	2
25	322	0	18.09955	27.97203	7.317073	7.168459	6.510417	6.504065	4.538578	3.991485	3.581439	3.02283	1.10443	2.492704	2.465281	3.774734	3.546309	2.296211	6.972252	2
26	324	12.12121	-5	13.33333	2.403846	14.96599	9.155222	6.346184	5.938494	3.892095	3.695616	1.89863	1.650554	1.004268	2.472188	2.593193	2.988849	2.265861	3.100921	2
27	326	6.134969	13.63636	13.93728	13.57466	8.760951	10.61571	9.514594	5.476451	4.075914	3.836931	2.353967	1.871783	3.470716	2.559022	4.937797	2.983898	3.220612	3.189793	1
28	328	6.896552	13.27434	46.59498	11.99041	4.901961	8.624408	12.34986	9.723757	5.255063	4.44519	2.799457	1.742768	2.367564	3.090823	2.738654	3.971547	3.960832	2.783413	5.
29	330	-6.80272	14.42308	36.23188	11.11111	16.68806	13.33926	9.593581	6.355932	6.593111	3.045447	3.413552	2.749881	2.70319	3.229527	2.907336	1.410756	2.640898	3.358688	
30	332	12.90323	-4.56621	3.558719	21.78649	10.19108	15.05091	7.603948	6.046705	7.628653	2.524287	3.152742	2.644286	2.517533	3.489303	2.566139	3.720106	3.91466	4.
31	334	0	13.51351	18.65672	0	14.53104	14.88728	12.3478	5.966333	5.595742	4.460443	4.644094	1.071301	3.658461	2.246321	2.628592	3.518929	3.59585	2.548656	3.
32	336	6.329114	9.13242	3.484321	14.25178	17.17791	8.225108	13.42009	7.335989	7.574733	5.158499	3.858634	2.65444	3.168413	2.145391	4.070466	4.31061	2.498215	4.284216	4.
33	338	-12.2699	23.25581	3.571429	11.21076	25.12563	8.477998	14.43633	8.795671	7.5883	5.373233	3.922878	1.828154	4.056181	4.747962	4.069952	3.971267	4.520524	4.02955	4.
34	340	0	-4.7619	13.51351	20.25316	13.089	8.347246	15.46906	11.47986	9.05474	5.642431	4.590985	0.918695	2.919538	2.848362	3.098629	4.098124	5.214084	4.798026	3.
35	341.96	6.802721	17.54386	20.27027	24.3309	14.03061	14.60823	14.69638	9.909522	7.217931	6.084164	5.215782	3.041225	4.83002	2.361756	3.943457	5.083473	5.835578	4.42607	6
36	343.93	28.36879	68.96552	0	18.14059	8.443909	17.50729	15.17872	13.64024	8.348407	7.304062	5.493068	5.047251	5.250506	4.09551	4.068584	4.699089	6.122059	5.000481	8
37	346.06	41.37931	0	3.472222	22.38806	41.77546	12.17976	13.64522	10.30278	8.795075	7.388513	9.078155	7.241216	3.784295	3.69155	4.793367	6.020799	5.609479	7.010115	7.
38	348.03	-6.45161	11.29943	56.66667	21.37767	27.21894	13.11762	14.1053	11.88038	11.26005	9.015778	9.073179	3.76587	4.976672	2.706046	3.464755	6.281407	4.490598	6.969972	7.
39	350	56.33803	44.64286	46.59498	12.53133	22.64151	15.61803	17.90656	12.21785	8.638065	10.37561	11.35694	7.517003	5.401235	3.935743	6.039317	6.175612	5.857689	7.222857	8.
40	352	46.35762	32.25806	42.48366	35.79952	17.56757	23.25581	20.47556	20.1005	10.49046	10.73392	11.32354	9.454629	5.039951	2.851259	3.935286	6.709468	7.918747	8.223146	9.
41	354	-6.71141	19.23077	32.02847	10.98901	11.07011	15.68951	18.58247	11.96837	14.17087	10.87647	6.956227	9.89242	6.272239	4.215047	5.792281	5.147586	6.61235	8.928964	9.
42	356	-20.5479	-14.2857	60.28369	7.317073	23.33722	19.35193	14.4368	16.10511	14.63811	11.79321	10.85729	7.335817	4.938272	5.252359	3.252531	5.639328	7.549927	8.739739	1
43	358	21.58273	18.43318	60.71429	56.20609	35.13174	30.74295	21.95282	16.53234	12.48786	11.01393	12.01903	10.27098	5.988024	4.401993	5.099349	7.311182	8.203064	7.689174	1
44	360	19.86755	48.78049	40.54054	27.43142	22.36025	18.82845	18.56899	21.76591	15.58924	13.79492	10.08015	10.45496	7.00633	6.240851	4.499308	7.383054	8.009723	7.992371	1
45	361.96	35.21127	33.33334	6.734006	48.78049	35.79952	19.12902	24.51499	15.26549	13.9891	15.31084	14.80742	13.26792	6.385938	8.892878	5.869153	7.369154	8.184646	11.70295	1

UNIVERSITY OF BIRMINGHAM

Machine Learning: PCA / PARAFAC

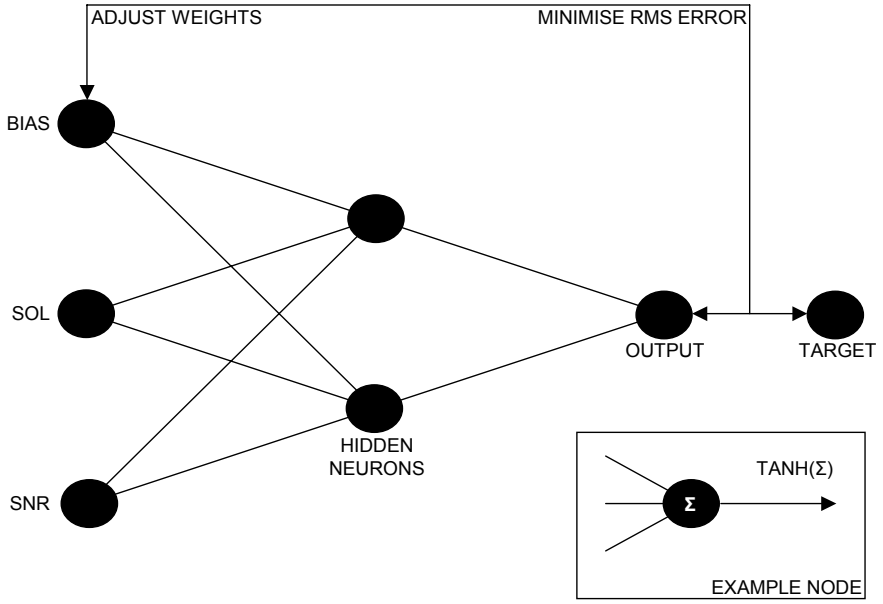
- A commonly used multivariate technique which acts *unsupervised*.
- Reduces the size of the dataset to aid management and understanding of the main patterns
- EEMs are high dimensional datasets and therefore there is a high redundancy in the data
- PARAFAC / PCA is used for feature extraction by extracting the axes in which the data shows the highest variability
- When plotted, PARAFAC loadings can reveal natural patterns or clusters in the data
- Can be used as a classifier by themselves or to produce a reduced number of inputs to other machine learning algorithms (e.g. Scott *et al*, 2003)



UNIVERSITY OF BIRMINGHAM

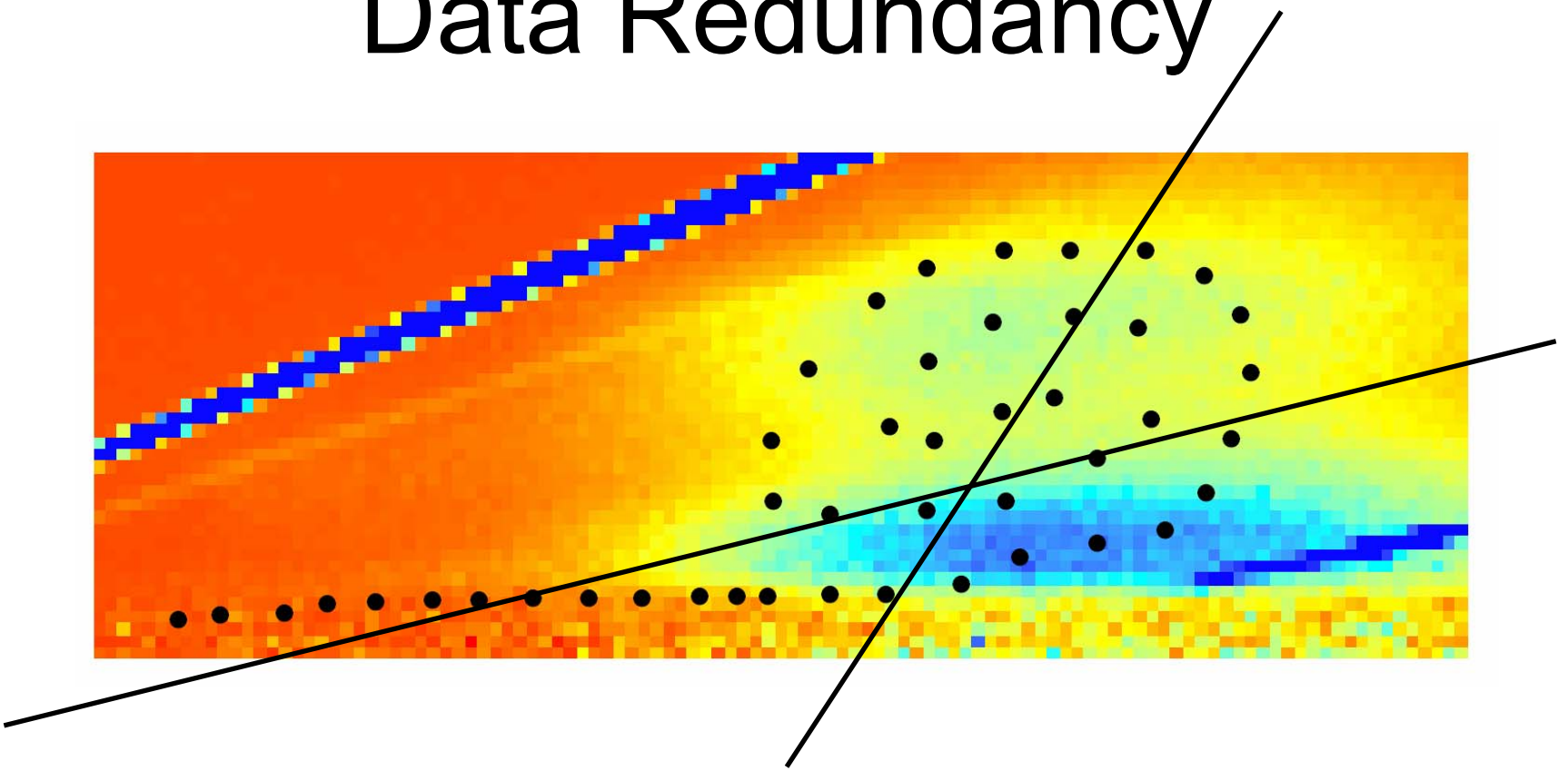
Machine Learning: ANNs

- A supervised approach so needs a training dataset
- Numerous algorithms exist, but the feed forward back propagation algorithm is the most commonly used for classification
- EEMs contain a lot of redundant information, so there is a challenge in removing this (e.g. PCA), but…
- …is this really needed?



UNIVERSITY OF BIRMINGHAM

Data Redundancy



- Can a simple function extract the useful data for a ANN?
- How do we determine what the function is?

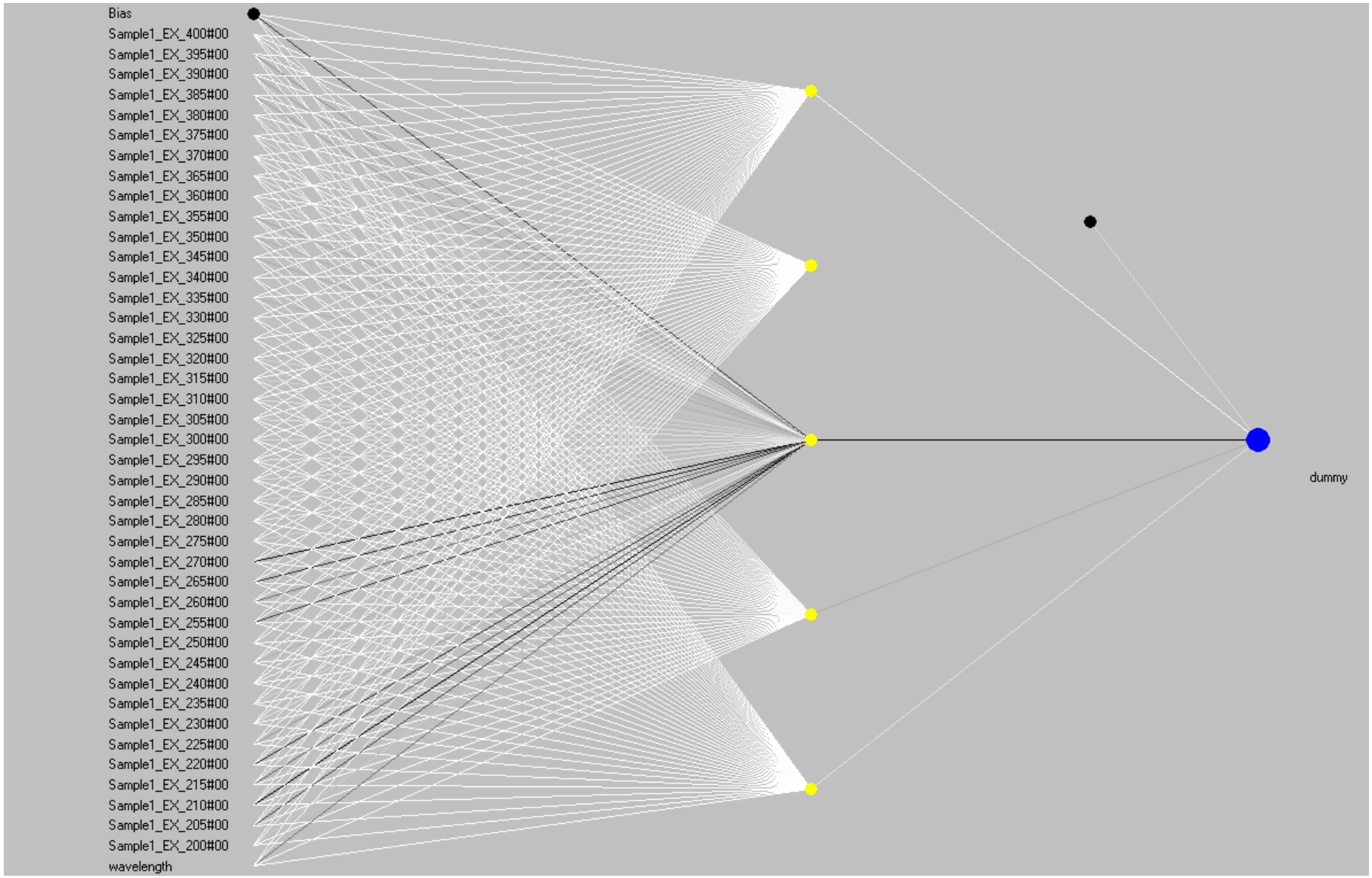
ANN - Tiberius

- Software used is shareware and available from here: http://www.philbrierley.com/
- Very simple and easy to use data mining suite
- Attempts to build a successful ANN using swirl and linear functions were not good!
- Although this removes data redundancy, is it needed? EEMs are not that big.

Entire EEM ANN

- Why not include all the data in the .csv file?
- ANN should be intelligent enough to cope
- A little complicated, each class requires training via it's own individual classification net
- The emissions at each excitation wavelength are then read in and the ANN then decides if the response of that wavelength is representative of the class in the training data.
- Hence a score out of 111 Is obtained for each classification net (7 in this example)
- How did it get on?

Sample ANN: Treated Sewage

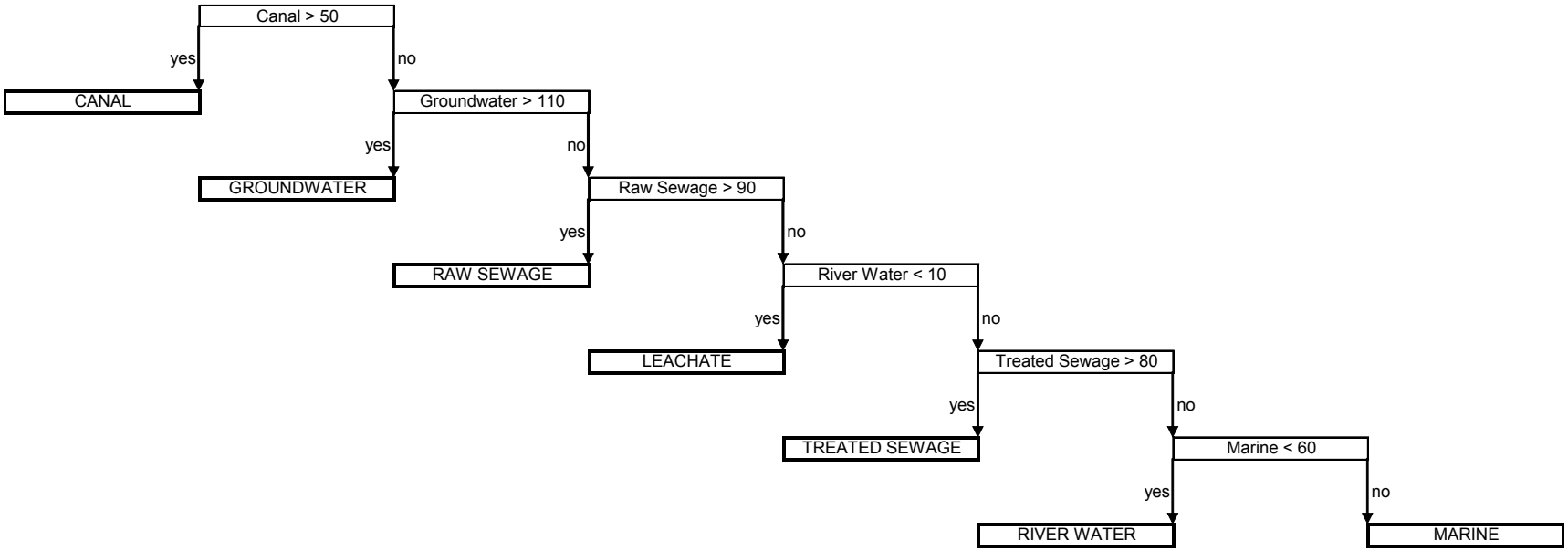


Results

	Testing Dataset						
	Canal	Groundwater	Leachate	Marine	Raw Sewage	River Water	Treated
Canal ANN	94	1	0	6	11	0	0
Groundwater ANN	0	111	0	107	0	0	109
Leachate ANN	110	0	34	8	4	26	8
Marine ANN	0	55	0	111	0	7	111
Raw Sewage ANN	45	3	0	7	110	0	0
River ANN	6	25	1	34	0	35	60
Treated Sewage ANN	0	27	0	37	0	0	104

- Not too bad, but…
- With a bit of tinkering with decision trees we can achieve 100% classification.
- Decision trees are predictive models which map observation data to target values
- Can be thought of as classification by if / then

Decision tree for ANN



- Trees can be developed using software called ‘R’
- Achieved 100% classification on additional testing set of 5 EEMs.
- Looks good, but far too complicated
- Can trees alone be used to simplify matters?

PiXiT

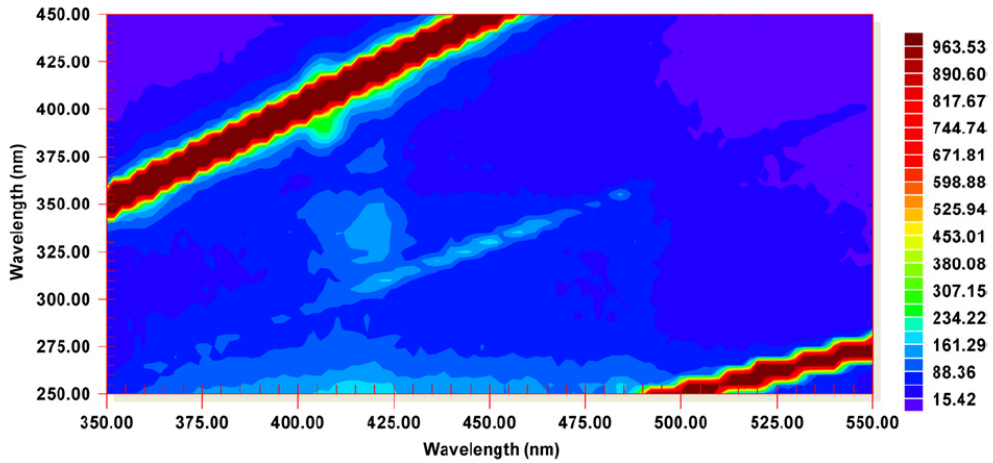
- Automatic image categorisation software
- Based on random subwindow extraction and extra trees
- Freely downloadable from http://www.montefiore.ulg.ac.be/~maree/pixit.html
- Very user friendly
- You need to build a learning database of images (not csv files) before training an algorithm for testing.

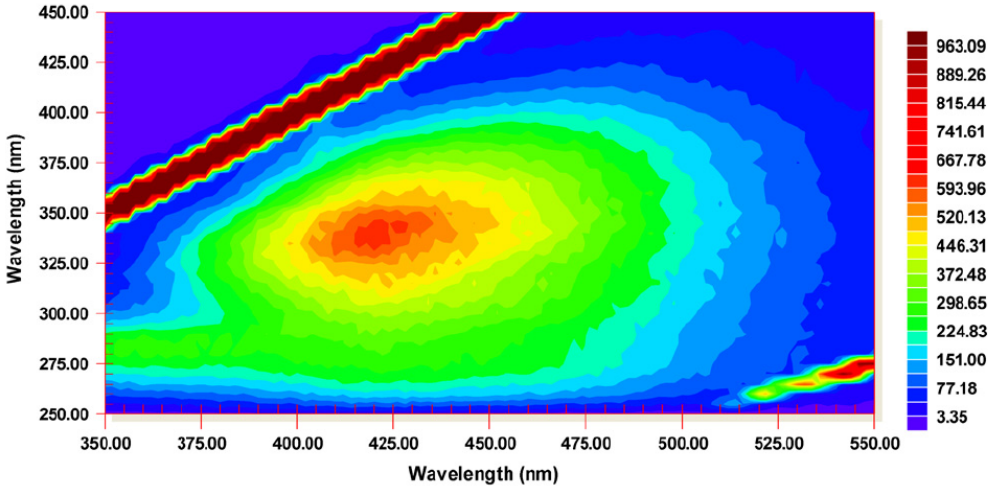
PiXiT Results

Actual	Class
Canal	Canal
Groundwater	Marine
Leachate	Canal
Marine	Leachate
Raw Sewage	Raw Sewage
River Water	River Water
Treated Sewage	Marine

- Not as good as the ANN, but should improve with a bigger training dataset
- Much slower than the ANN as far as computer processing is concerned, but…
- …quicker as images can ‘theoretically’ be processed in real-time direct from the spectrophotometer

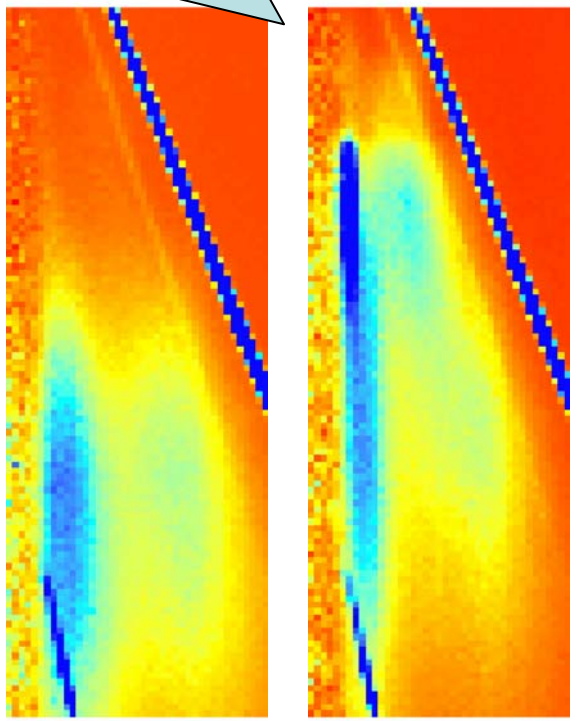
Why theoretically?





Subtle differences exist between raw outputs

This is why processed data has been used throughout



UNIVERSITY OF BIRMINGHAM

Conclusions

- This is just a small pilot study and the number of samples needs increasing before the results have any true meaning.
- A combination of approaches will probably yield the best results
- Many other algorithms out there still to be tested – Genetic Algorithms could be the answer to slow computer processing.
- For the meantime, PiXiT seems the most user-friendly solution.