Cancer detection via the lasso and customized training

Robert Tibshirani, Stanford University

Google $->$ Tibshirani
(WARNING: top link might be that imposter)

Machine learning department, CMU; 2014

Statistics- Machine learning glossary 2014

Statistics	*Machine learning*
model	hypothesis space
hypothesis testing	never heard of it!
estimation	inference
inference	never heard of it!
parameter estimation	learning
lots of parameter estimation	deep learning

An effective seminar technique	
Present slides of unnecessary details burying the assumptions so that no-one notices how unrealistic they are	Present the work at 100mph, not giving enough details for anyone to critique it

Careful study of a proposed method	
Endless pages of simulation results in a font too small to read	It works on one example that I concocted

Statistics- Machine learning glossary 2014

Statistics	*Machine learning*
model	hypothesis space
hypothesis testing	never heard of it!
estimation	inference
inference	never heard of it!
parameter estimation	learning
lots of parameter estimation	deep learning

An effective seminar technique	
Present slides of unnecessary details burying the assumptions so that no-one notices how unrealistic they are	Present the work at 100mph, not giving enough details for anyone to critique it

Careful study of a proposed method	
Endless pages of simulation results in a font too small to read	It works on one example that I concocted

Statistics- Machine learning glossary 2014

Statistics	*Machine learning*
model	hypothesis space
hypothesis testing	never heard of it!
estimation	inference
inference	never heard of it!
parameter estimation	learning
lots of parameter estimation	deep learning

An effective seminar technique	
Present slides of unnecessary details burying the assumptions so that no-one notices how unrealistic they are	Present the work at 100mph, not giving enough details for anyone to critique it

Careful study of a proposed method	
Endless pages of simulation results in a font too small to read	It works on one example that I concocted

Statistics- Machine learning glossary 2014

Statistics	*Machine learning*
model	hypothesis space
hypothesis testing	never heard of it!
estimation	inference
inference	never heard of it!
parameter estimation	learning
lots of parameter estimation	deep learning

An effective seminar technique

Present slides of unnecessary details burying the assumptions so that no-one notices how unrealistic they are	Present the work at 100mph, not giving enough details for anyone to critique it

Careful study of a proposed method

Endless pages of simulation results in a font too small to read	It works on one example that I concocted

Statistics- Machine learning glossary 2014

Statistics	*Machine learning*
model	hypothesis space
hypothesis testing	never heard of it!
estimation	inference
inference	never heard of it!
parameter estimation	learning
lots of parameter estimation	deep learning

An effective seminar technique

Present slides of unnecessary details burying the assumptions so that no-one notices how unrealistic they are	Present the work at 100mph, not giving enough details for anyone to critique it

Careful study of a proposed method

Endless pages of simulation results in a font too small to read	It works on one example that I concocted

Statistics- Machine learning glossary 2014

Statistics	*Machine learning*
model	hypothesis space
hypothesis testing	never heard of it!
estimation	inference
inference	never heard of it!
parameter estimation	learning
lots of parameter estimation	deep learning

An effective seminar technique	
Present slides of unnecessary details burying the assumptions so that no-one notices how unrealistic they are	Present the work at 100mph, not giving enough details for anyone to critique it

Careful study of a proposed method	
Endless pages of simulation results in a font too small to read	It works on one example that I concocted

Statistics- Machine learning glossary 2014

Statistics	*Machine learning*
model	hypothesis space
hypothesis testing	never heard of it!
estimation	inference
inference	never heard of it!
parameter estimation	learning
lots of parameter estimation	deep learning

An effective seminar technique

Present slides of unnecessary details burying the assumptions so that no-one notices how unrealistic they are	Present the work at 100mph, not giving enough details for anyone to critique it

Careful study of a proposed method

Endless pages of simulation results in a font too small to read	It works on one example that I concocted

Statistics- Machine learning glossary 2014

Statistics	*Machine learning*
model	hypothesis space
hypothesis testing	never heard of it!
estimation	inference
inference	never heard of it!
parameter estimation	learning
lots of parameter estimation	deep learning

An effective seminar technique	
Present slides of unnecessary details burying the assumptions so that no-one notices how unrealistic they are	Present the work at 100mph, not giving enough details for anyone to critique it

Careful study of a proposed method	
Endless pages of simulation results in a font too small to read	It works on one example that I concocted

Statistics- Machine learning glossary 2014

Statistics	*Machine learning*
model	hypothesis space
hypothesis testing	never heard of it!
estimation	inference
inference	never heard of it!
parameter estimation	learning
lots of parameter estimation	deep learning

An effective seminar technique	
Present slides of unnecessary details burying the assumptions so that no-one notices how unrealistic they are	Present the work at 100mph, not giving enough details for anyone to critique it

Careful study of a proposed method	
Endless pages of simulation results in a font too small to read	It works on one example that I concocted

Statistics- Machine learning glossary 2014

Statistics	*Machine learning*
model	hypothesis space
hypothesis testing	never heard of it!
estimation	inference
inference	never heard of it!
parameter estimation	learning
lots of parameter estimation	deep learning

An effective seminar technique	
Present slides of unnecessary details burying the assumptions so that no-one notices how unrealistic they are	Present the work at 100mph, not giving enough details for anyone to critique it

Careful study of a proposed method	
Endless pages of simulation results in a font too small to read	It works on one example that I concocted

Statistics- Machine learning glossary 2014

Statistics	*Machine learning*
model	hypothesis space
hypothesis testing	never heard of it!
estimation	inference
inference	never heard of it!
parameter estimation	learning
lots of parameter estimation	deep learning

An effective seminar technique	
Present slides of unnecessary details burying the assumptions so that no-one notices how unrealistic they are	Present the work at 100mph, not giving enough details for anyone to critique it

Careful study of a proposed method	
Endless pages of simulation results in a font too small to read	It works on one example that I concocted

Statistics- Machine learning glossary 2014

Statistics	*Machine learning*
model	hypothesis space
hypothesis testing	never heard of it!
estimation	inference
inference	never heard of it!
parameter estimation	learning
lots of parameter estimation	deep learning

An effective seminar technique	
Present slides of unnecessary details burying the assumptions so that no-one notices how unrealistic they are	Present the work at 100mph, not giving enough details for anyone to critique it

Careful study of a proposed method	
Endless pages of simulation results in a font too small to read	It works on one example that I concocted

Statistics- Machine learning glossary 2014

Statistics	*Machine learning*
model	hypothesis space
hypothesis testing	never heard of it!
estimation	inference
inference	never heard of it!
parameter estimation	learning
lots of parameter estimation	deep learning

An effective seminar technique	
Present slides of unnecessary details burying the assumptions so that no-one notices how unrealistic they are	Present the work at 100mph, not giving enough details for anyone to critique it

Careful study of a proposed method	
Endless pages of simulation results in a font too small to read	It works on one example that I concocted

Statistics- Machine learning glossary 2014

Statistics	*Machine learning*
model	hypothesis space
hypothesis testing	never heard of it!
estimation	inference
inference	never heard of it!
parameter estimation	learning
lots of parameter estimation	deep learning

An effective seminar technique	
Present slides of unnecessary details burying the assumptions so that no-one notices how unrealistic they are	Present the work at 100mph, not giving enough details for anyone to critique it

Careful study of a proposed method	
Endless pages of simulation results in a font too small to read	It works on one example that I concocted

Statistics- Machine learning glossary 2014

Statistics	*Machine learning*
model	hypothesis space
hypothesis testing	never heard of it!
estimation	inference
inference	never heard of it!
parameter estimation	learning
lots of parameter estimation	deep learning

An effective seminar technique	
Present slides of unnecessary details burying the assumptions so that no-one notices how unrealistic they are	Present the work at 100mph, not giving enough details for anyone to critique it

Careful study of a proposed method	
Endless pages of simulation results in a font too small to read	It works on one example that I concocted

Statistics- Machine learning glossary 2014

Statistics	*Machine learning*
model	hypothesis space
hypothesis testing	never heard of it!
estimation	inference
inference	never heard of it!
parameter estimation	learning
lots of parameter estimation	deep learning

An effective seminar technique	
Present slides of unnecessary details burying the assumptions so that no-one notices how unrealistic they are	Present the work at 100mph, not giving enough details for anyone to critique it

Careful study of a proposed method	
Endless pages of simulation results in a font too small to read	It works on one example that I concocted

Statistics- Machine learning glossary 2014

Statistics	*Machine learning*
model	hypothesis space
hypothesis testing	never heard of it!
estimation	inference
inference	never heard of it!
parameter estimation	learning
lots of parameter estimation	deep learning

An effective seminar technique	
Present slides of unnecessary details burying the assumptions so that no-one notices how unrealistic they are	Present the work at 100mph, not giving enough details for anyone to critique it

Careful study of a proposed method	
Endless pages of simulation results in a font too small to read	It works on one example that I concocted

Statistics-Machine learning glossary 2014

Statistics	*Machine learning*
Large scale computation	
Won't run on my old laptop	Needs Distributed computing
Typical class	
15 bored students who can't wait for class to end	100,000 students taking it voluntarily!

Statistics-Machine learning glossary 2014

Statistics	*Machine learning*
Large scale computation	
Won't run on my old laptop	Needs Distributed computing
Typical class	
15 bored students who can't wait for class to end	100,000 students taking it voluntarily!

Statistics-Machine learning glossary 2014

Statistics	*Machine learning*
	Large scale computation
Won't run on my old laptop	Needs Distributed computing
	Typical class
15 bored students who can't wait for class to end	100,000 students taking it voluntarily!

Statistics-Machine learning glossary 2014

Statistics	*Machine learning*
Large scale computation	
Won't run on my old laptop	Needs Distributed computing
Typical class	
15 bored students who can't wait for class to end	100,000 students taking it voluntarily!

Statistics-Machine learning glossary 2014

Statistics	*Machine learning*
	Large scale computation
Won't run on my old laptop	Needs Distributed computing
	Typical class
15 bored students who can't wait for class to end	100,000 students taking it voluntarily!

What to do for this talk?

1. A 100mph presentation on everything my lab has done for the past 2 years, with motion graphics zooming in and out
2. A theoretical presentation on minimax learning rates, "relaxing" an $n^{3/2}$ to $n^{4/3}$ error factor, compared to that other guy (who is not as smart as me) and who just submitted his paper to arXiv 15 minutes after me
3. A statistics talk masquerading as an Big Data ML talk, changing terms such as"regression" to "supervised learning", with Big data = 1000 observations on 250 variables

What to do for this talk?

1. A 100mph presentation on everything my lab has done for the past 2 years, with motion graphics zooming in and out
2. A theoretical presentation on minimax learning rates, "relaxing" an $n^{3/2}$ to $n^{4/3}$ error factor, compared to that other guy (who is not as smart as me) and who just submitted his paper to arXiv 15 minutes after me
3. A statistics talk masquerading as an Big Data ML talk, changing terms such as"regression" to "supervised learning", with Big data = 1000 observations on 250 variables

What to do for this talk?

1. A 100mph presentation on everything my lab has done for the past 2 years, with motion graphics zooming in and out
2. A theoretical presentation on minimax learning rates, "relaxing" an $n^{3/2}$ to $n^{4/3}$ error factor, compared to that other guy (who is not as smart as me) and who just submitted his paper to arXiv 15 minutes after me
3. A statistics talk masquerading as an Big Data ML talk, changing terms such as"regression" to "supervised learning", with Big data = 1000 observations on 250 variables

What to do for this talk?

1. A 100mph presentation on everything my lab has done for the past 2 years, with motion graphics zooming in and out
2. A theoretical presentation on minimax learning rates, "relaxing" an $n^{3/2}$ to $n^{4/3}$ error factor, compared to that other guy (who is not as smart as me) and who just submitted his paper to arXiv 15 minutes after me
3. A statistics talk masquerading as an Big Data ML talk, changing terms such as"regression" to "supervised learning", with Big data = 1000 observations on 250 variables

Outline of this talk

- An ongoing collaboration - cancer detection during surgery
- Sparsity and the lasso
- Customized training

NO THEOREMS IN THIS TALK!

My "gem" of wisdom for today

It's important to spend at least part of your time working on *real* problems, with *real* scientists. [This doesn't mean analyzing data from the UCI database].

- It keeps you grounded, reminds you what's important,
- It suggests new challenges
- It forces you to communicate the essence of your approach in a non-technical way
- Data doesn't talk! But scientists get confused, ask questions you can't answer, etc.
- In my statistical consulting/collaborations, most of the time the help that I provide is *simple* (design advice, plots, t-tests..). That's OK! But sometimes it requires state-of-the-art tools.

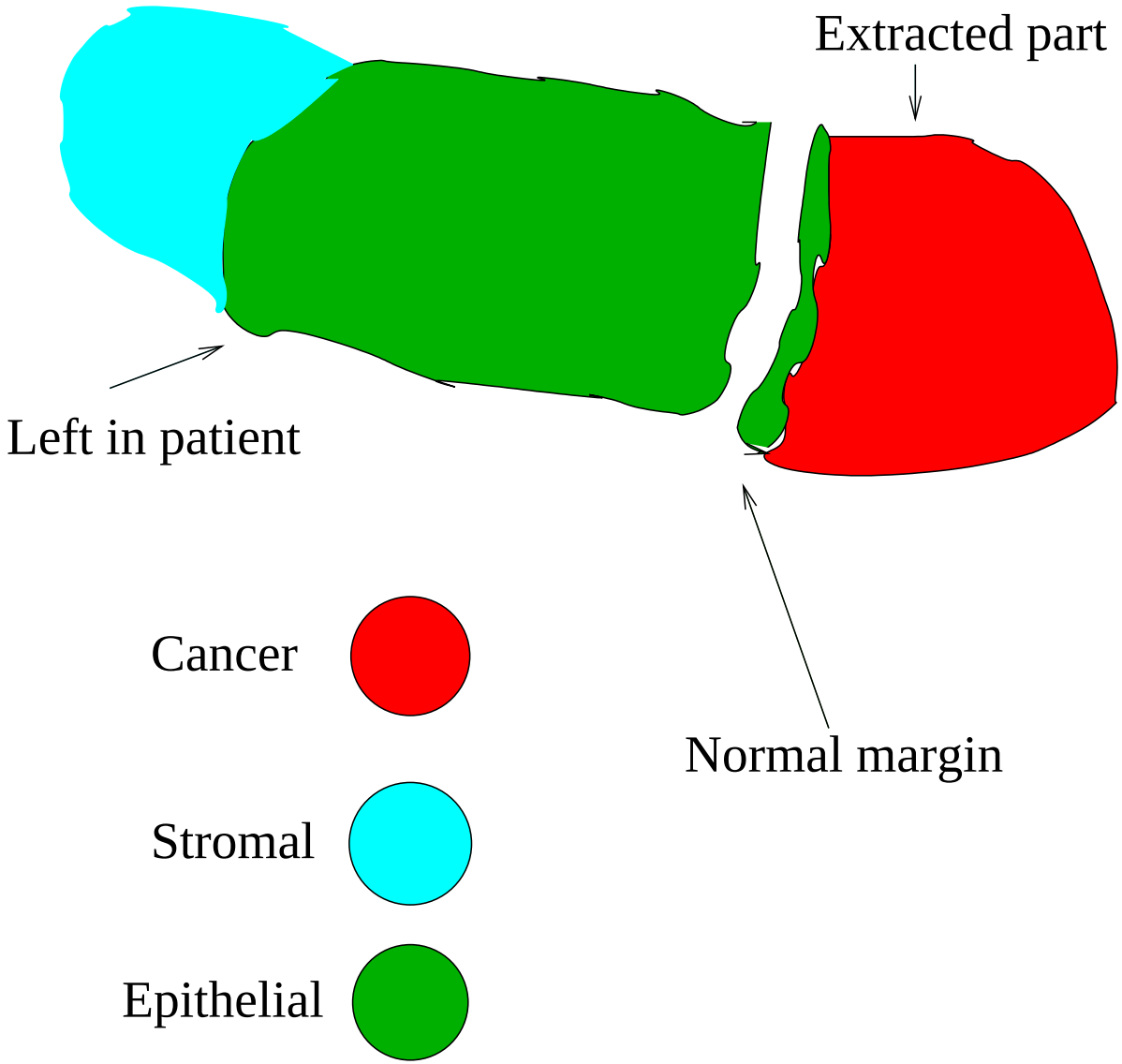
A Cancer detection problem

- I am currently working in a cancer diagnosis project with co-workers at Stanford; Livia Eberlin (PI) —PostDoc (Chemistry); Richard Zare (Chemistry) and George Poulsides (Surgery)
- Eberlin et al (2014). "Molecular assessment of surgical-resection margins of gastric cancer by mass-spectrometric imaging". Proc. Nat. Acad. Sci.
- They have collected samples of tissue from a number of patients undergoing surgery for stomach cancer.

Livia
Eberlin

Richard
Zare

George
Poulsides


Extracted part
Left in patient
Normal margin
Cancer
Stromal
Epithelial

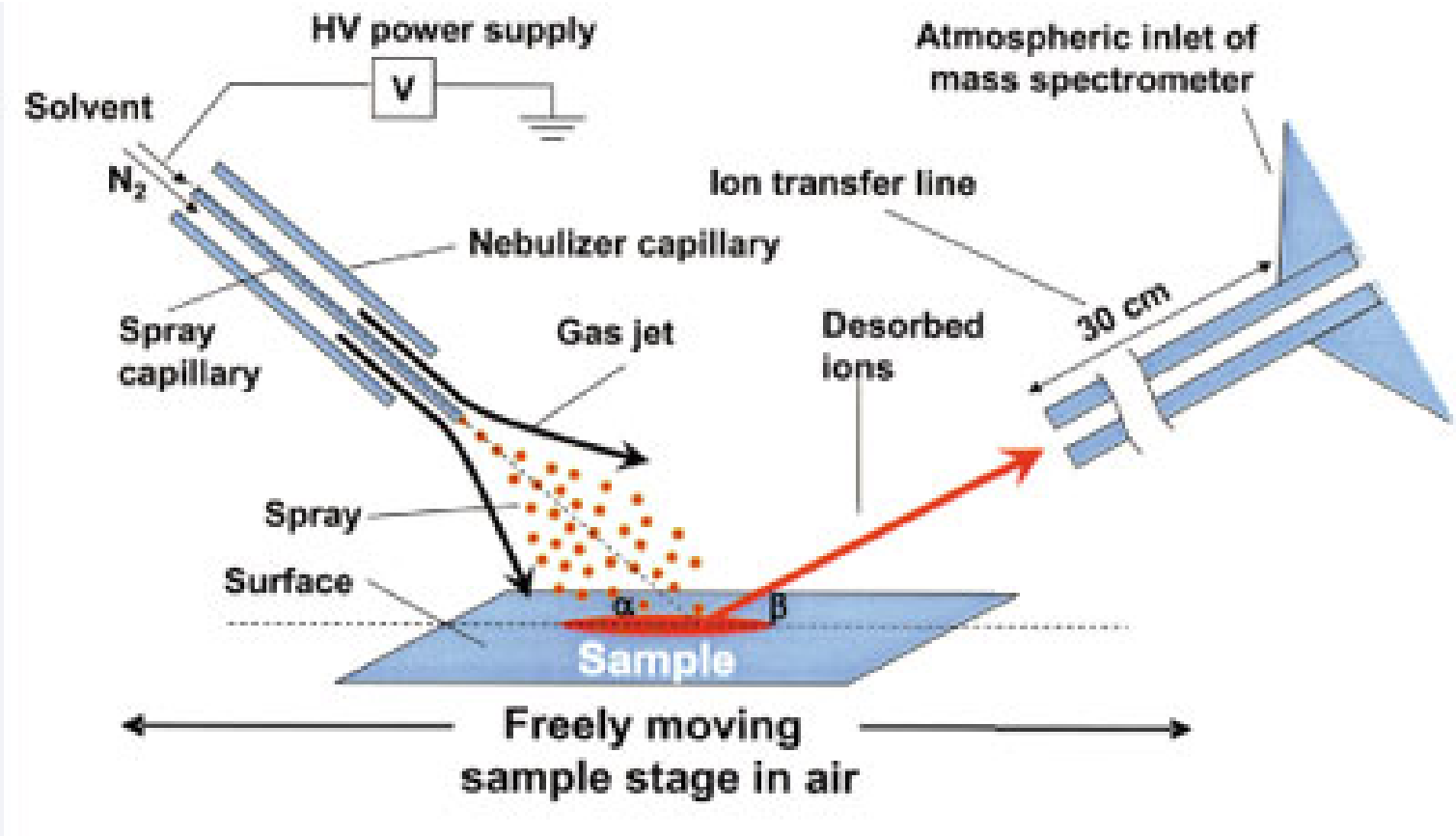
The challenge

- Build a classifier than can distinguish three kinds of tissue: normal epithelial, normal stromal and cancer.
- Such a classifier could be used to assist surgeons in determining, in real time, whether they had successfully removed all of the tumor. Current pathologist error rate for the real-time call can be as high as 20%.

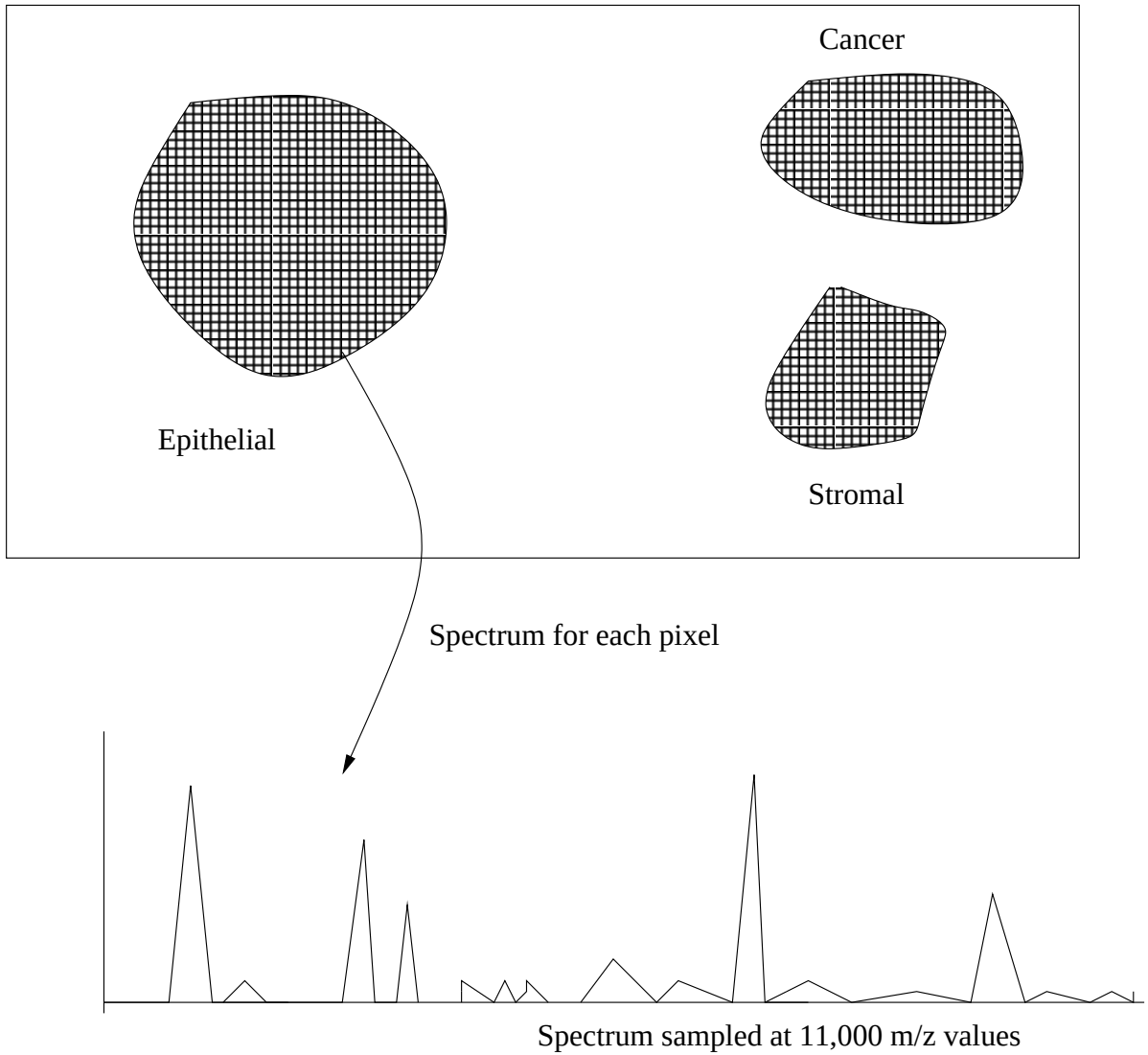
Technology to the rescue!

DESI (Desorption electrospray ionization)

An electrically charged "mist" is directed at the sample; surface ions are freed and enter the mass spec.



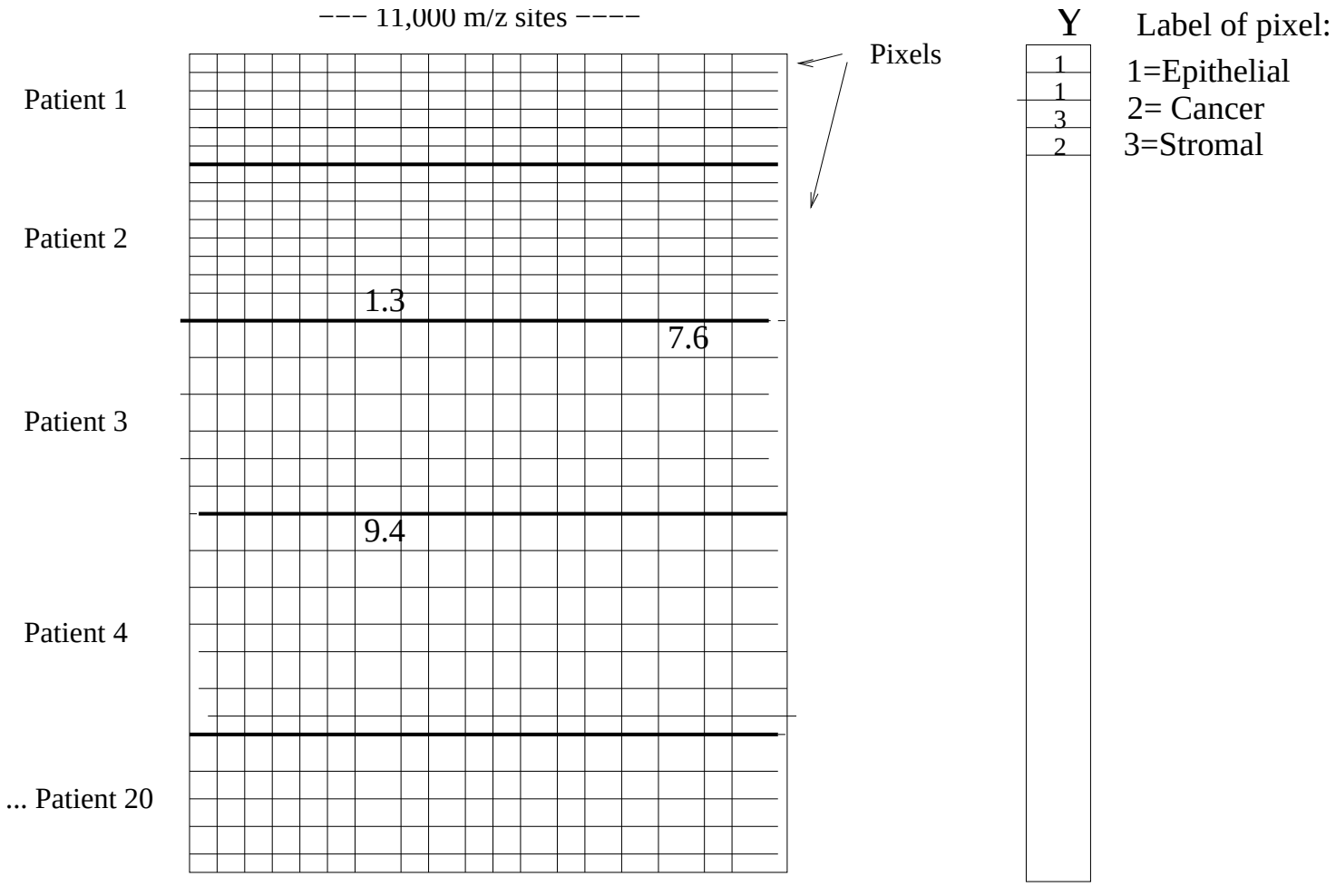
The data for one patient



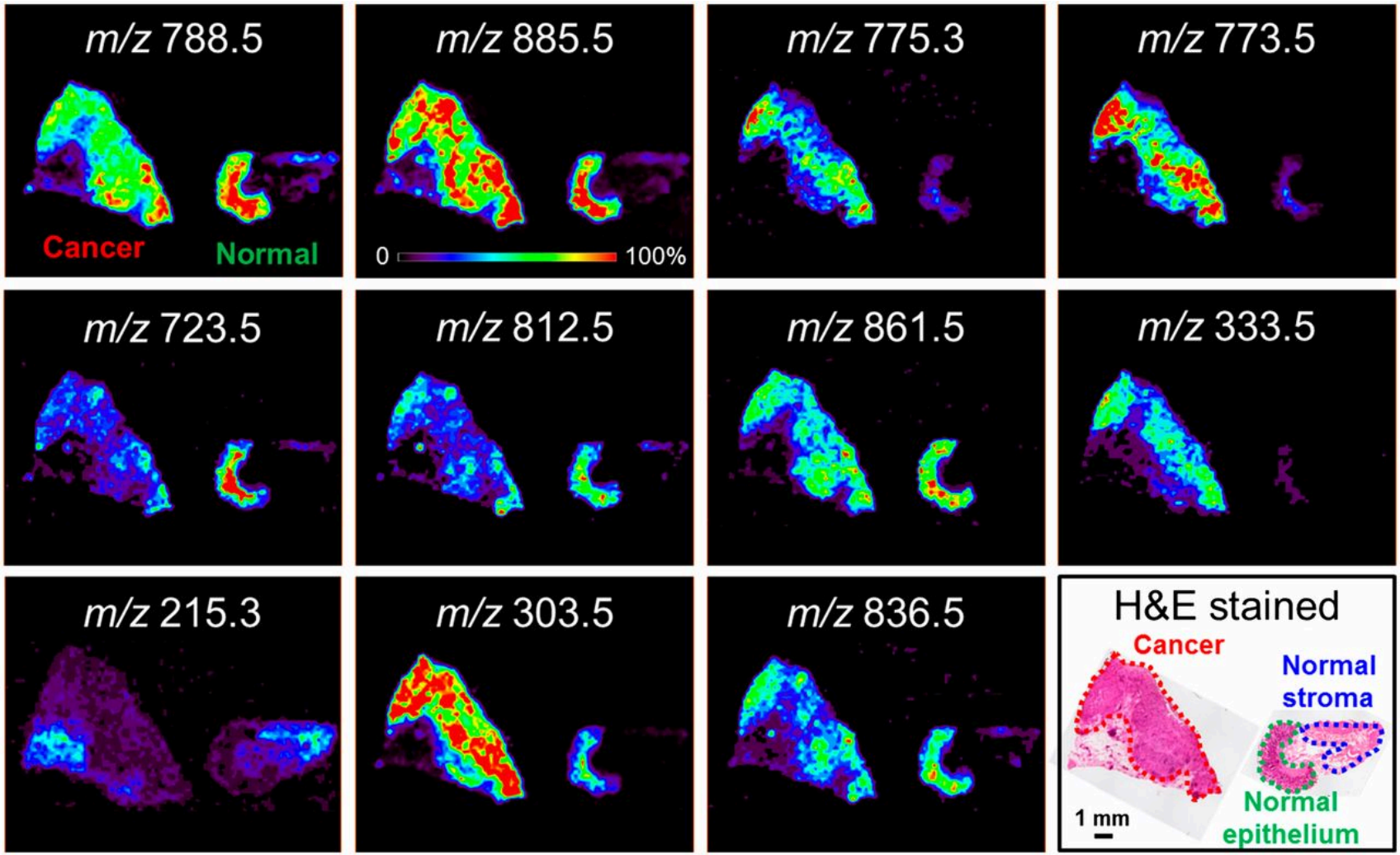
Details

- 20 patients, each contributing a sample of epithelial, stromal and cancer tissue.
- Labels determined after 2 weeks of testing in pathology lab.
- At each pixel in the image, the intensity of metabolites is measured by DESI. Peaks in the spectrum representing different metabolites.
- The spectrum has been finely sampled, with the intensity measured at about $11,000$ m/z sites across the spectrum, for each of about 8000 pixels.

The overall data



Selected negative ion mode DESI-MS ion images of sample GC727.



What we need to tackle this problem

- A statistical classifier (algorithm) that sorts through the large number of features, and finds the most informative ones: a sparse set of features.
- We are doing pixel-wise classification

The Lasso

- Regression problem: We observe n feature-response pairs $(x_i,\ y_i)$, where x_i is a p-vector and y_i is real-valued.
- Let $x_i = (x_{i1}, x_{i2}, \ldots x_{ip})$
- Consider a *linear regression model*:

$$y_i = \beta_0 + \sum_j x_{ij}\beta_j + \epsilon_i$$

where ϵ_i is an error term with mean zero. β_j is the weight given feature j
Later: y_i will take one of 3 values (epithelial, stromal, cancer) and x_{ij} will be the height of the spectrum for patient i, at m/z site j.

- Least squares fitting is defined by

$$\underset{\beta_0,\beta}{\text{minimize}}\ \frac{1}{2}\sum_i (y_i - \beta_0 - \sum_j x_{ij}\beta_j)^2$$

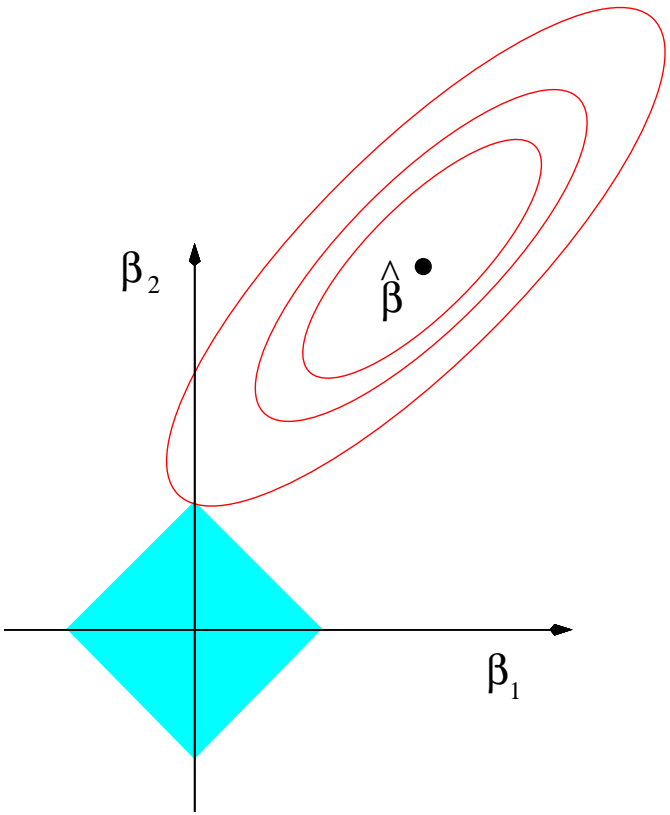
The Lasso— continued

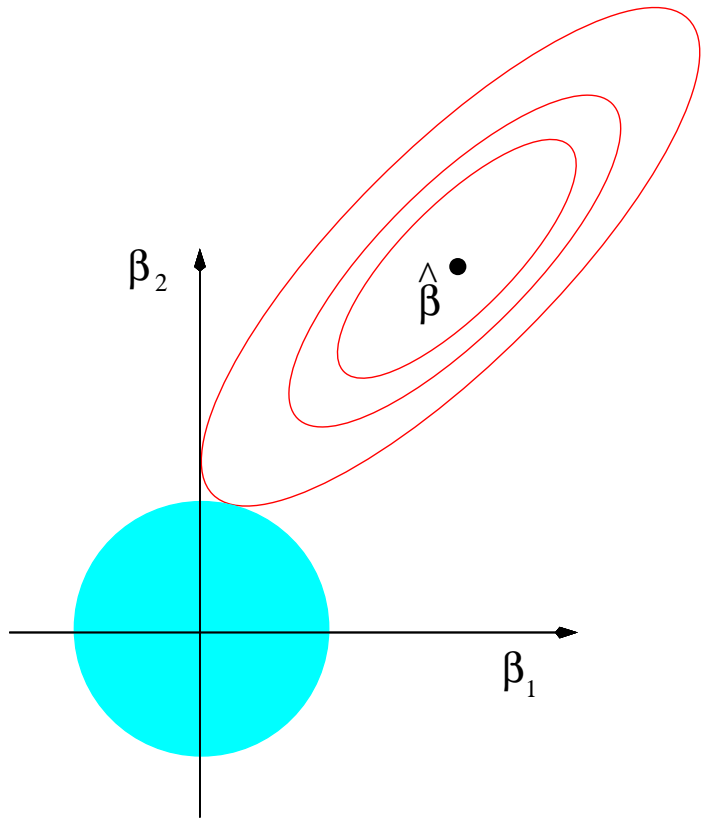
The **Lasso** is an estimator defined by the following optimization problem:

$$\underset{\beta_0,\beta}{\text{minimize}} \frac{1}{2}\sum_i (y_i - \beta_0 - \sum_j x_{ij}\beta_j)^2 \qquad \text{subject to} \quad \sum |\beta_j| \leq s$$

- Penalty $\implies$ sparsity (feature selection)
- Convex problem (good for computation and theory)
- *Ridge regression* uses penalty $\sum_j \beta_j^2 \leq s$ and does not yield sparsity

Why does the lasso give a sparse solution?



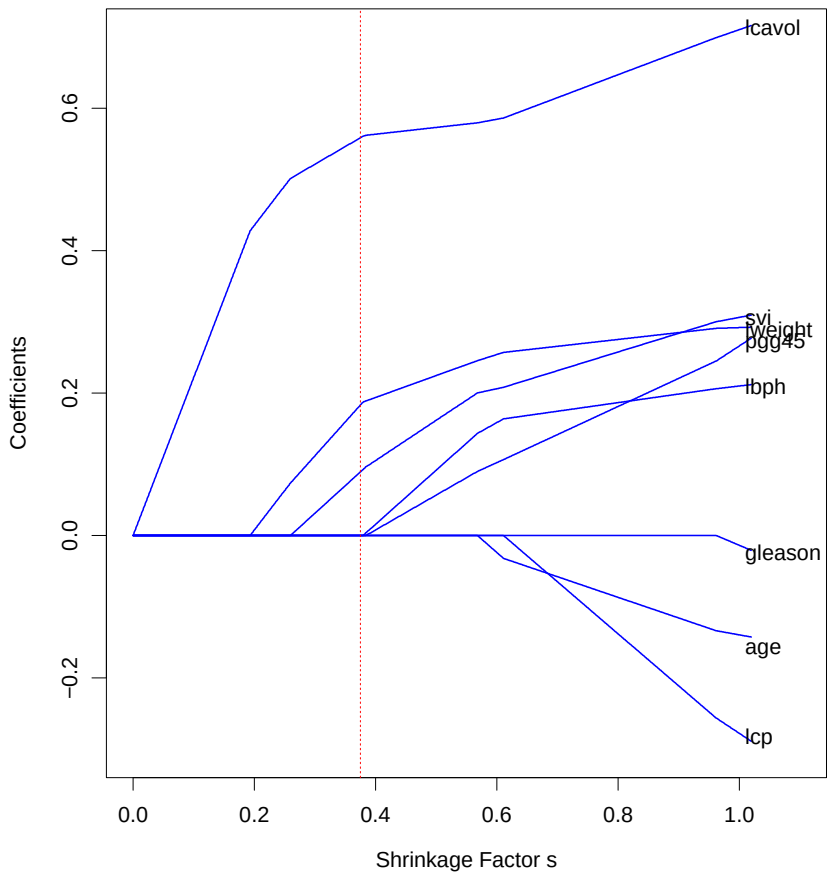


Lasso $\sum_j |\beta_j| \leq s$

Ridge $\sum_j \beta_j^2 \leq s$

Prostate cancer example

$N = 88, p = 8$. Predicting log-PSA, in men after prostate cancer surgery



Back to our problem

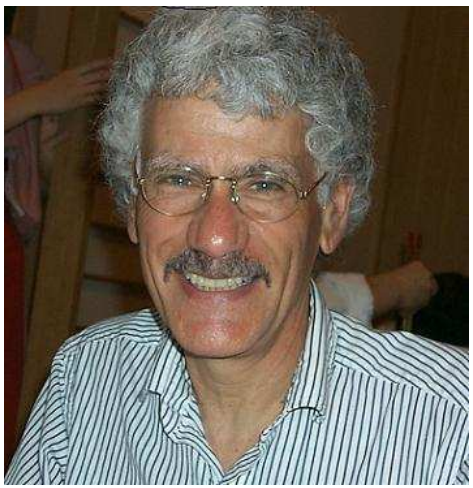
- $K = 3$ classes (epithelial, stromal, cancer): multinomial model

$$\log \frac{Pr(Y_i = k|x)}{\sum_k Pr(Y_i = k|x)} = \beta_{0k} + \sum_j x_{ij}\beta_{jk}, \ k = 1, 2, \ldots K$$

  Here x_{ij} is height of spectrum for sample i at jth m/z position
- We replace the least squares objective function by the multinomial log-likelihood
- Add lasso penalty $\sum |\beta_j| \leq s$; optimize, using cross-validation to estimate best value for budget s.
- yields a pixel classifier, and also reveals which m/z sites are informative.

Fast computation is essential

- Our lab has written a open-source R language package called `glmnet` for fitting lasso models. Written in FORTRAN!!!!!
- It is very fast- can solve the current problem in a few minutes on a PC. Some builtin parallelization too.
- Not "off-the shelf": Many clever computational tricks were used to achieve the impressive speed.
- Lots of features- Gaussian, Logistic, Poisson, Survival models; elastic net; grouping; parameter constraints; Available in R and Matlab.



Jerry Friedman

Trevor Hastie

"FORTRAN's tragic fate has been its wide acceptance, mentally chaining thousands and thousands of programmers to our past mistakes"

Edsger W. Dijkstra, The Humble Programmer, 1972 Turing Award Lecture, Communications of the ACM 15 (10), (October 1972): pp. 859?866.

Results

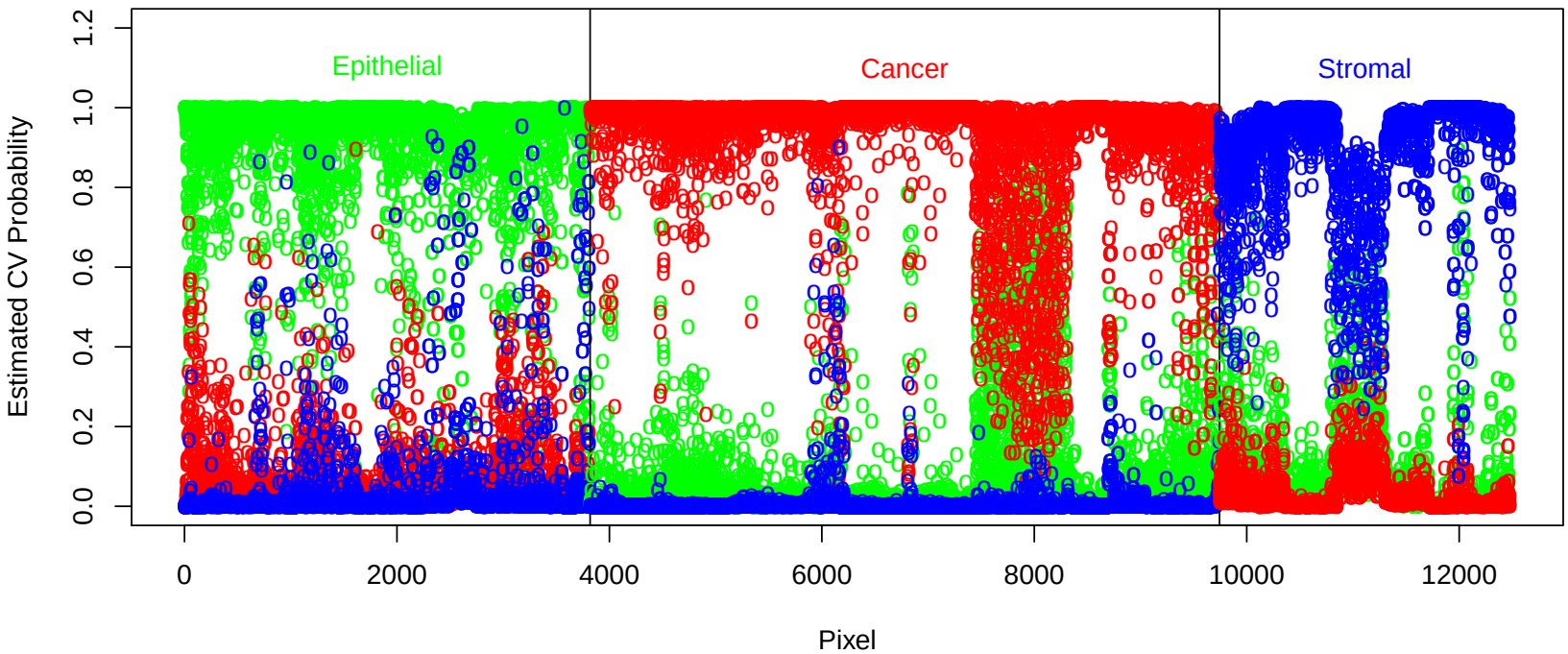
Cross-validation- min at 129 peaks; overall error rate= 4.2%

	Predicted			
true	Epi	Canc	Strom	Prop correct
Epi	3277.00	80.00	145.00	0.94
Canc	73.00	5106.00	13.00	0.98
Strom	79.00	86.00	2409.00	0.94

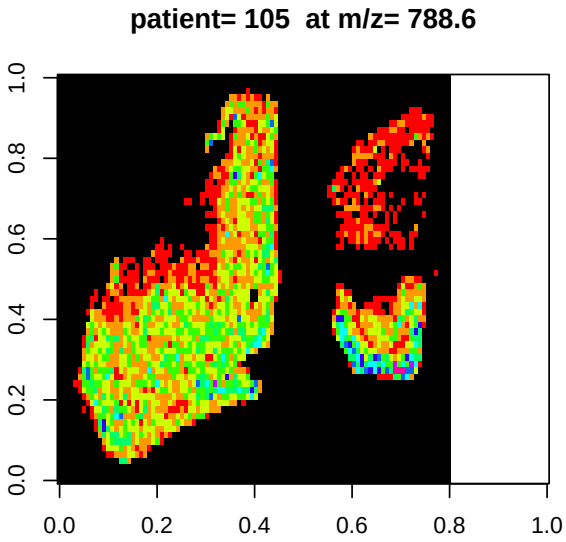
Test set: overall error rate =5.7%

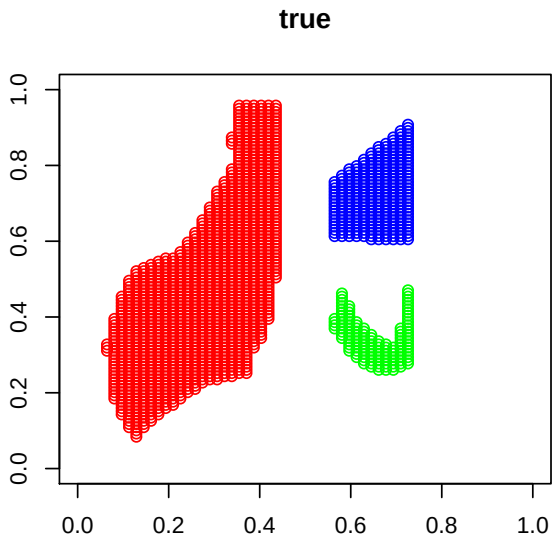
	Predicted			
true	Epi	Canc	Strom	Prop correct
Epi	1606.00	149.00	19.00	0.91
Canc	23.00	1622.00	5.00	0.98
Strom	67.00	5.00	1222.00	0.94

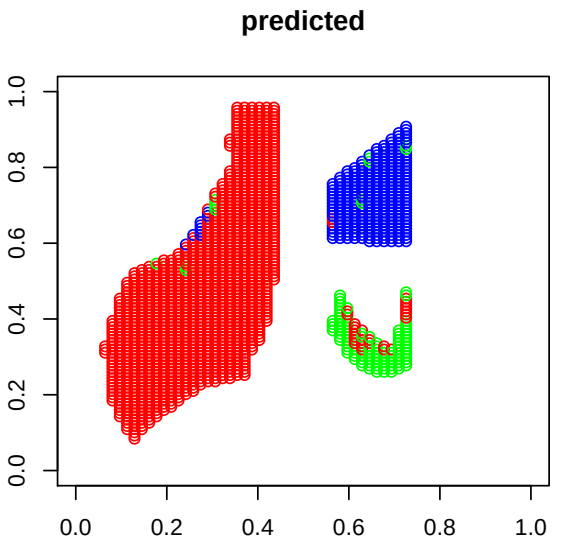
Cross-validated estimates of class probabilities

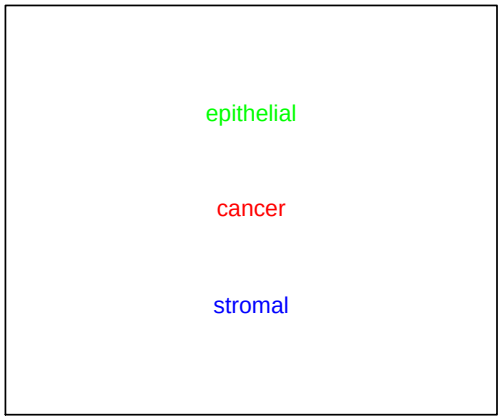


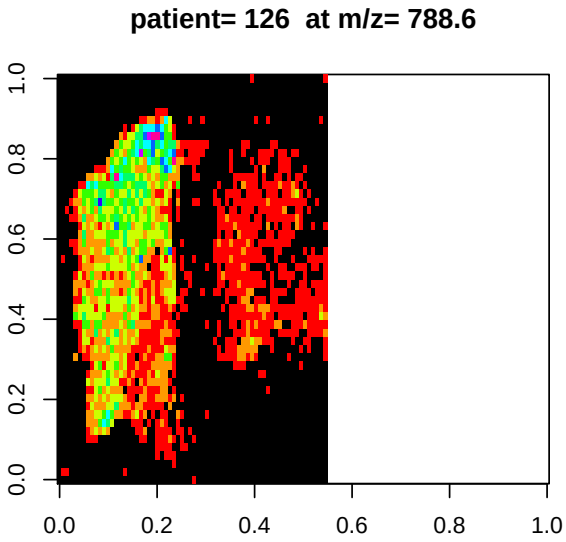
Peak #	m/z value	Epi	Canc	Strom
1	101.5			
2	107.5	0.09		
3	110.5		0.44	0.56
4	123.5			
5	132.5	0.06		
6	134.5		0.10	0.13
7	135.5		0.13	0.19
8	137.5		-0.01	0.05
9	145.5	-0.71		
10	146.5	-0.41		
11	151.5	-0.15	0.35	-0.25
12	157.5		-0.07	-0.17
13	170.5			
14	171.5			
15	174.5	0.05		
16	175.5		0.14	0.54
17	179.5			
18	188.5			
19	212.5			
20	214.5		-014	-0.13
21	215.5		-1.17	-1.07
22	222.5	0.29		
23	224.5			
24	225.5			
25	231.5			
26	244.5	-0.01		
27	247.5		-0.31	-0.41
28	258.5	0.21		
29	270.5		-0.14	-0.24
30	278.5	0.32		
31	279.5	0.39		
32	280.5			
33	285.5			
34	289.5			
35	293.5	0.25		
36	297.5			
37	299.5			
38	301.5	-0.45	0.21	0.11
39	312.5	0.10	-0.44	-0.24
40	322.5			
41	324.5	0.06		
42	325.5	0.03	-0.00	-0.00
43	333.5		0.82	0.55
44	340.5		-0.02	-0.07
45	341.5			
46	347.5	0.01		
47	349.5			
48	353.5	0.05		
49	355.5	0.02		

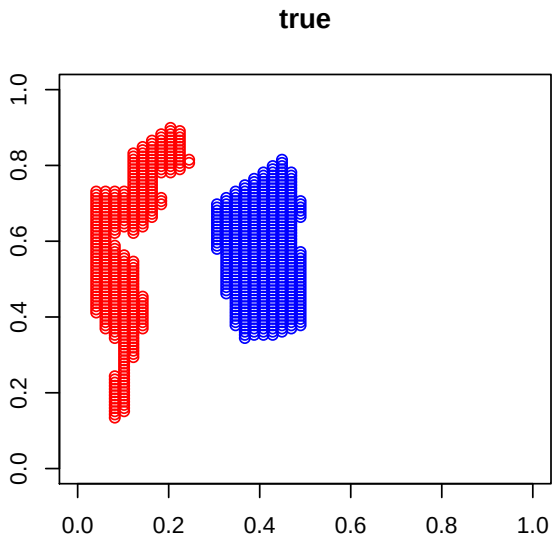

patient= 105 at m/z= 788.6
0.0
0.2
0.4
0.6
0.8
1.0

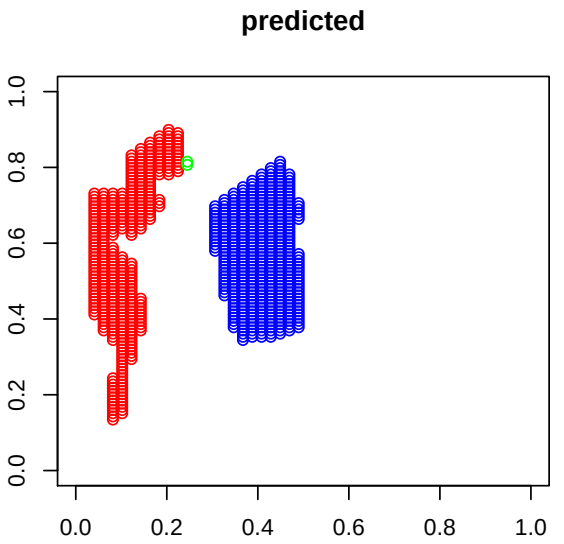

true
0.0
0.2
0.4
0.6
0.8
1.0

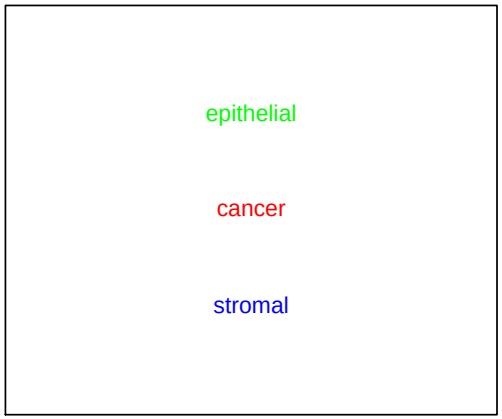

predicted
0.0
0.2
0.4
0.6
0.8
1.0

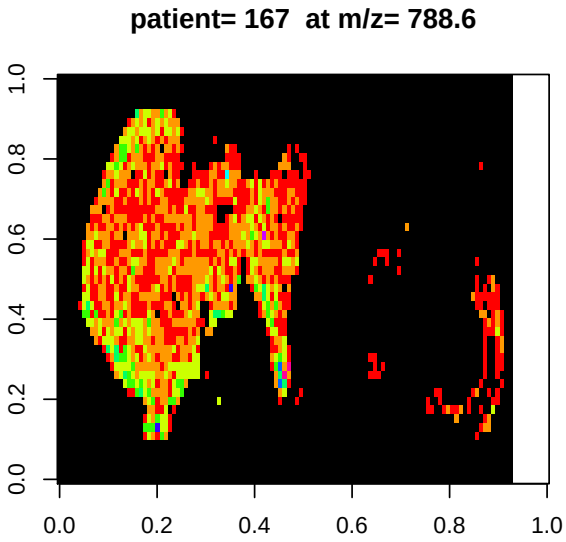

epithelial
cancer
stromal

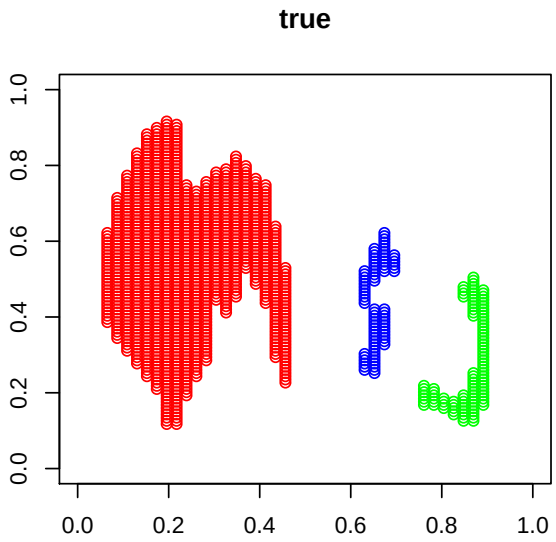

patient= 126 at m/z= 788.6
0.0
0.2
0.4
0.6
0.8
1.0

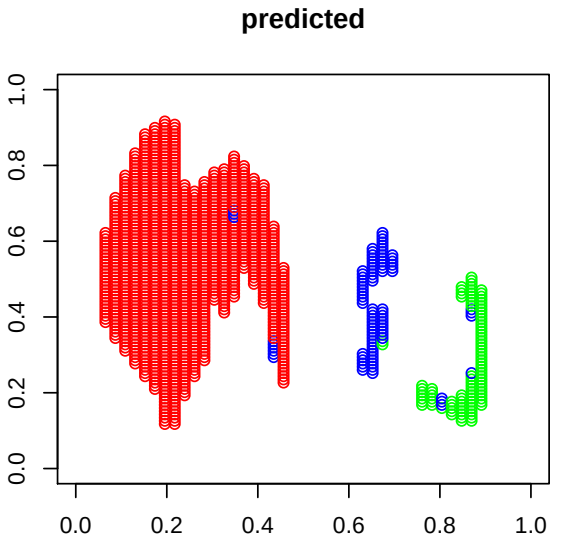

true
0.0
0.2
0.4
0.6
0.8
1.0

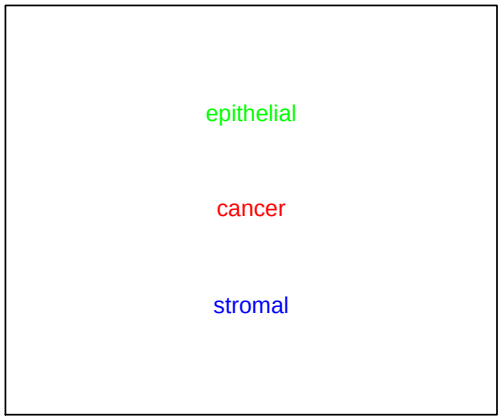

predicted
0.0
0.2
0.4
0.6
0.8
1.0


epithelial
cancer
stromal


patient= 167 at m/z= 788.6
0.0
0.2
0.4
0.6
0.8
1.0


true
0.0
0.2
0.4
0.6
0.8
1.0


predicted
0.0
0.2
0.4
0.6
0.8
1.0

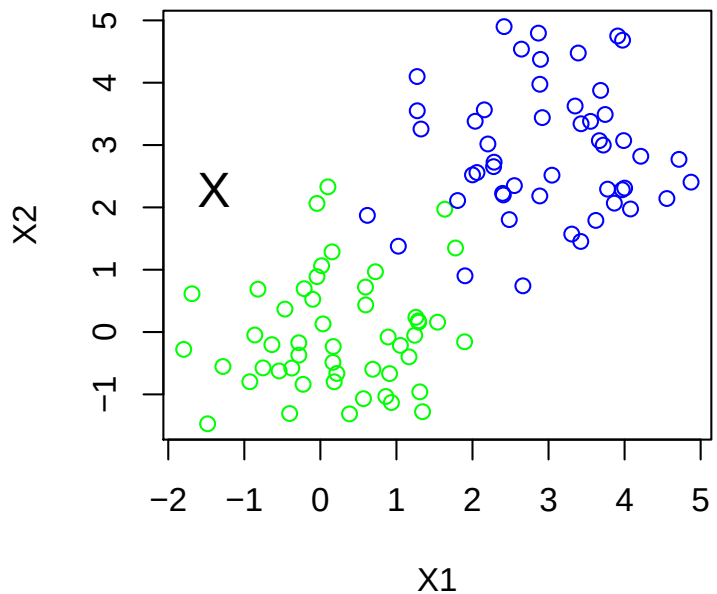

epithelial
cancer
stromal

Other approaches

- Support vector machines: classification error was a little higher than lasso; doesn't give a sparse solution easily
- Deep learning (with help from a student of Geoff Hinton): reported that it didn't work any better than lasso; thought that non-linearities were likely unimportant for this problem, and sparsity was more important

A challenge

- "Abstentions": sometimes a classifier should not make a prediction; instead it should say ""I don't know" For example, when the query feature vector is far away from the training set features.
- This problem happened in some tests of our system
- Can't rely on the fact that the largest posterior probability will be close to $1/K$ (K= number of classes):



More questions

- Should we adjust for patient effects in training? (My attempts were not successful)
- Should we use spatial proximity of pixels? If so, how?

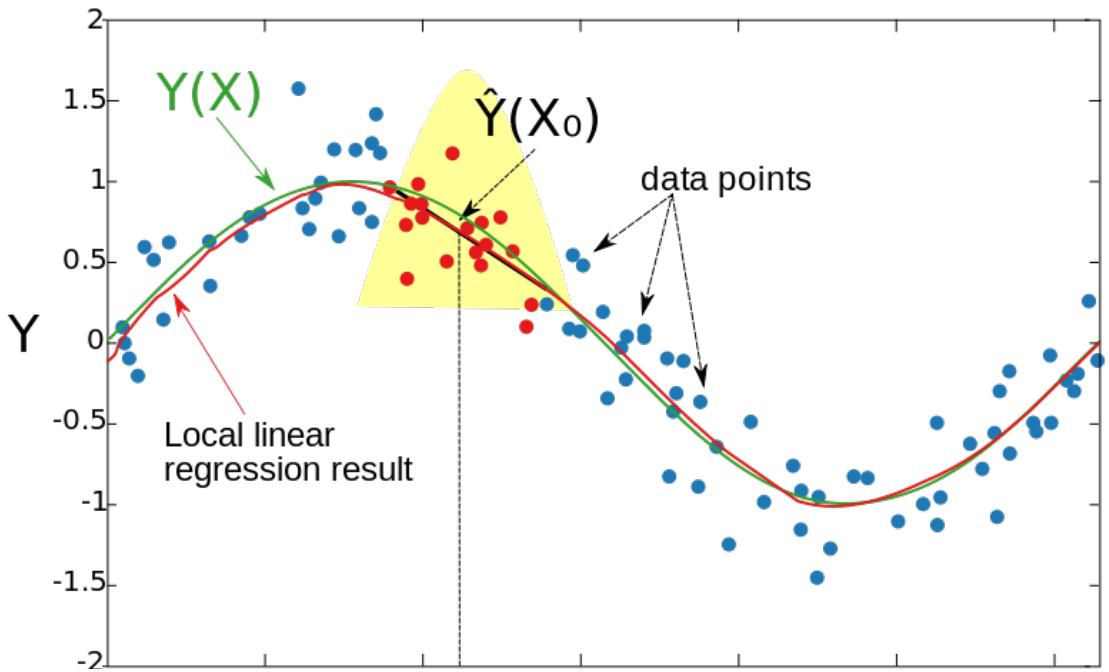
Customized Training by Data Clustering

Joint work with Scott Powers and Trevor Hastie

Motivation: Local regression

- For each test point, fits a separate regression model
- Training points weighted by distance from query point
- Example:

$$f(x_0) = \min_{\alpha(x_0),\beta(x_0)} \sum_{i=1}^{n} K_\lambda(x_0, x_i)[y_i - \alpha(x_0) - \beta(x_0)x_i]^2$$



Our idea: try to get the best of both worlds

- Idea: Cluster data (2 proposals), fit lasso model within each cluster, use this for prediction
- For high-dimensional data, combine flexibility of local learning with interpretability of sparse models.
- can apply to other learning algorithms like SVMs (not today)

Related idea: Transductive learning

- Use unlabeled data to help in training - e.g. semi-supervised learning
- Example: Transductive SVMs

Two proposed clustering methods for customized training

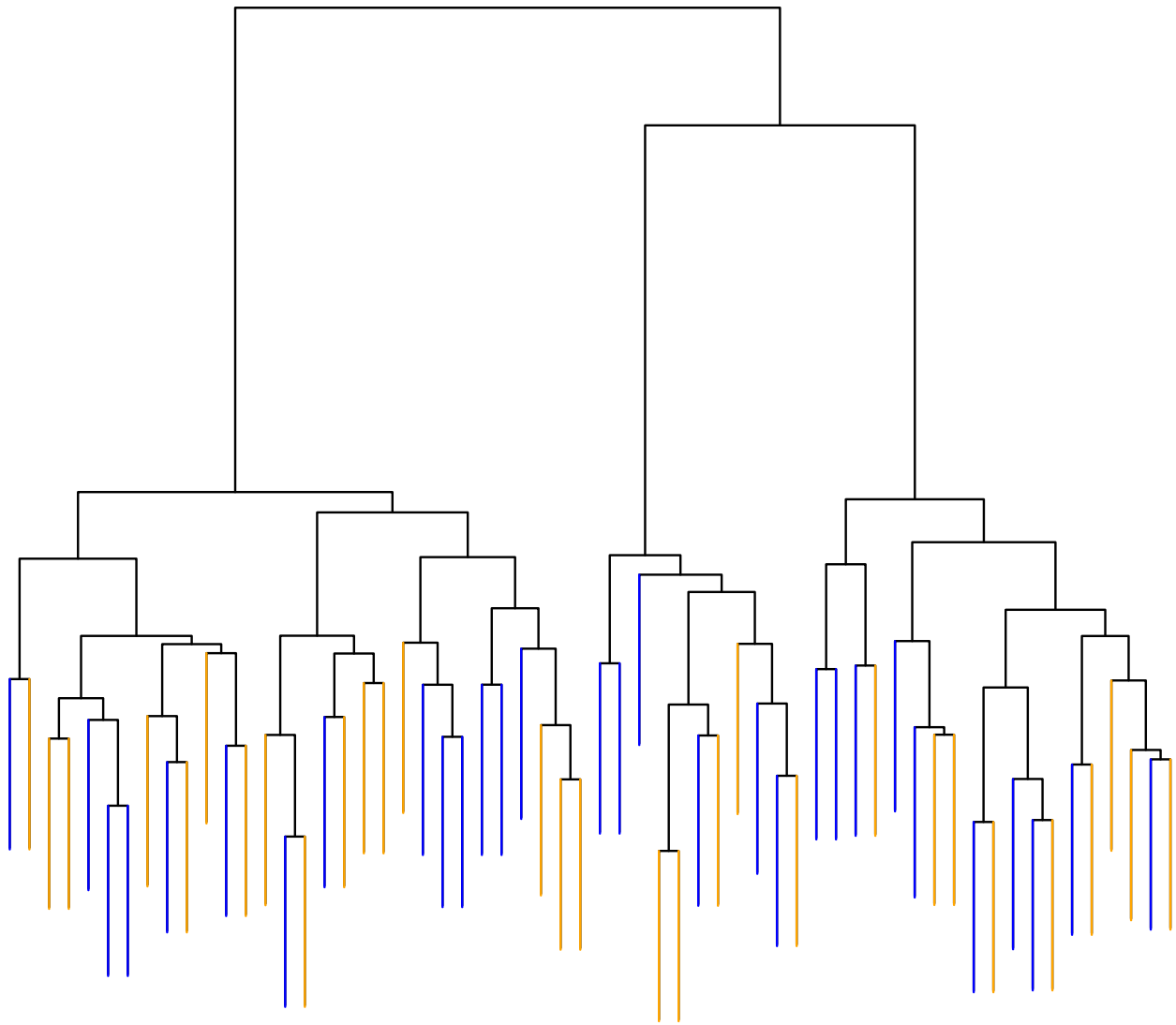
Joint clustering and Test clustering.

Joint clustering:

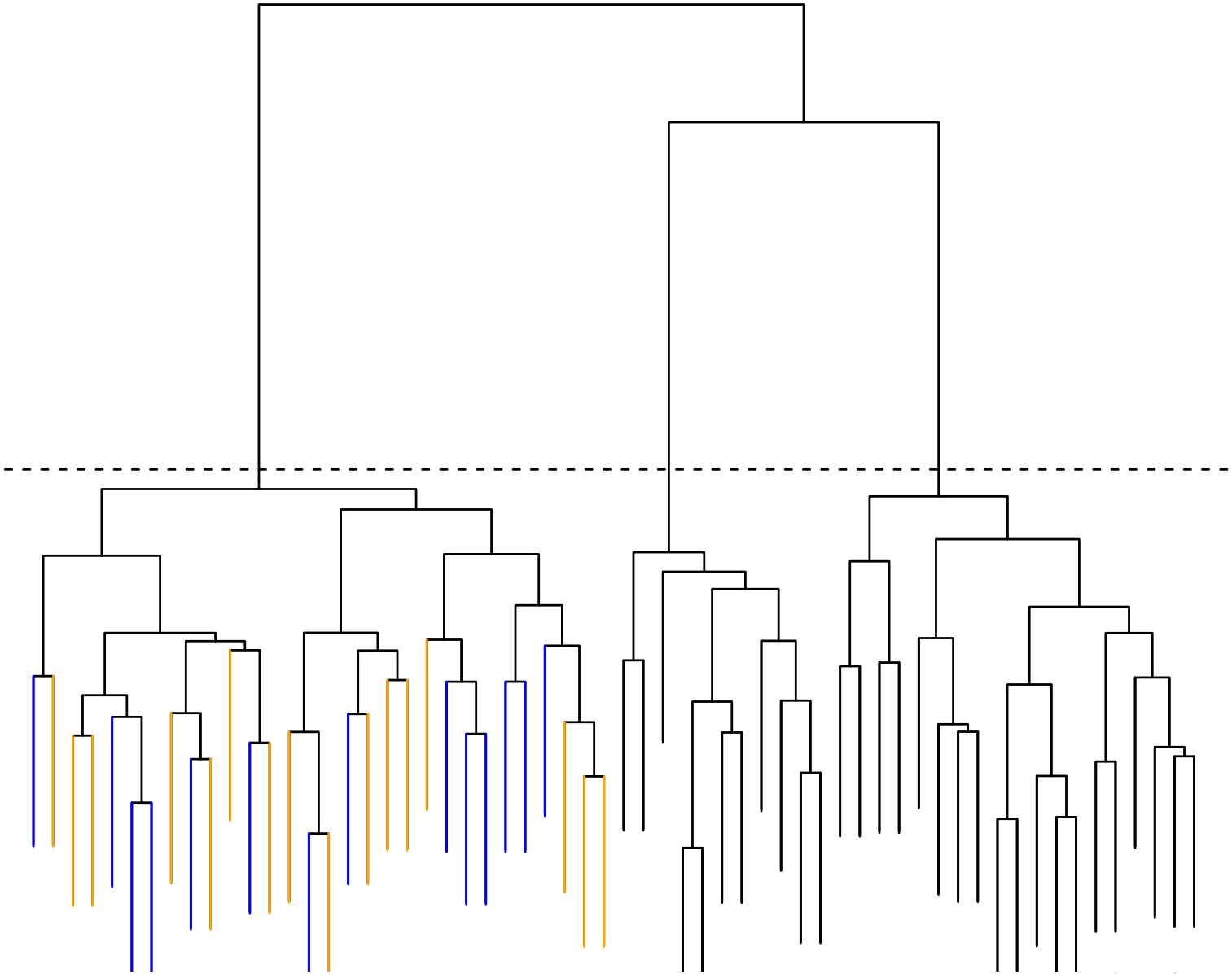
1. Hierarchically cluster training and test points together (using complete linkage).
2. Cut dendrogram off at some height determined by cross-validation
3. Fit separate models within each cluster.
4. Predict each test point using the model in the cluster where it falls.

Joint clustering

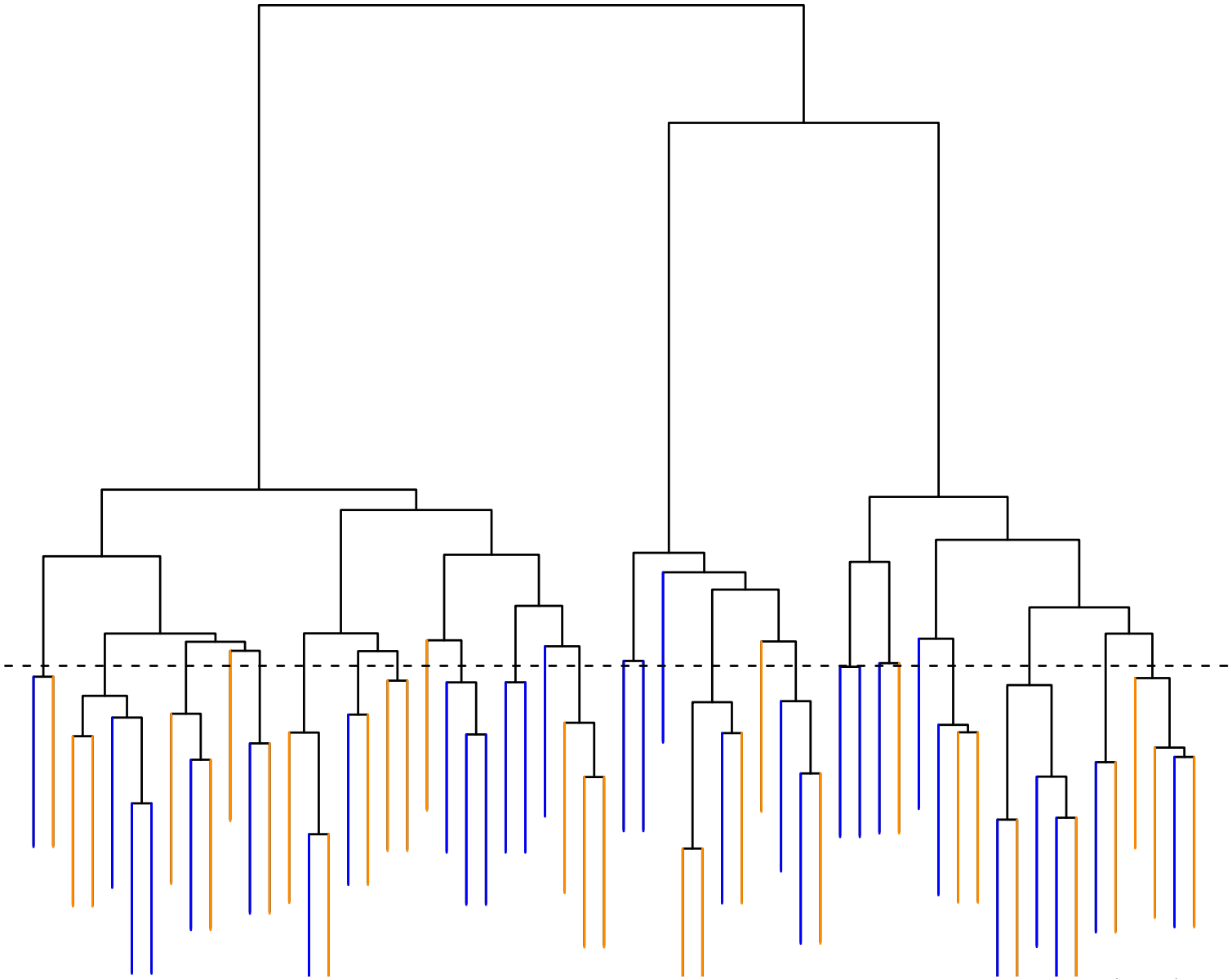
Blue: Training points; Brown: test points



Joint clustering



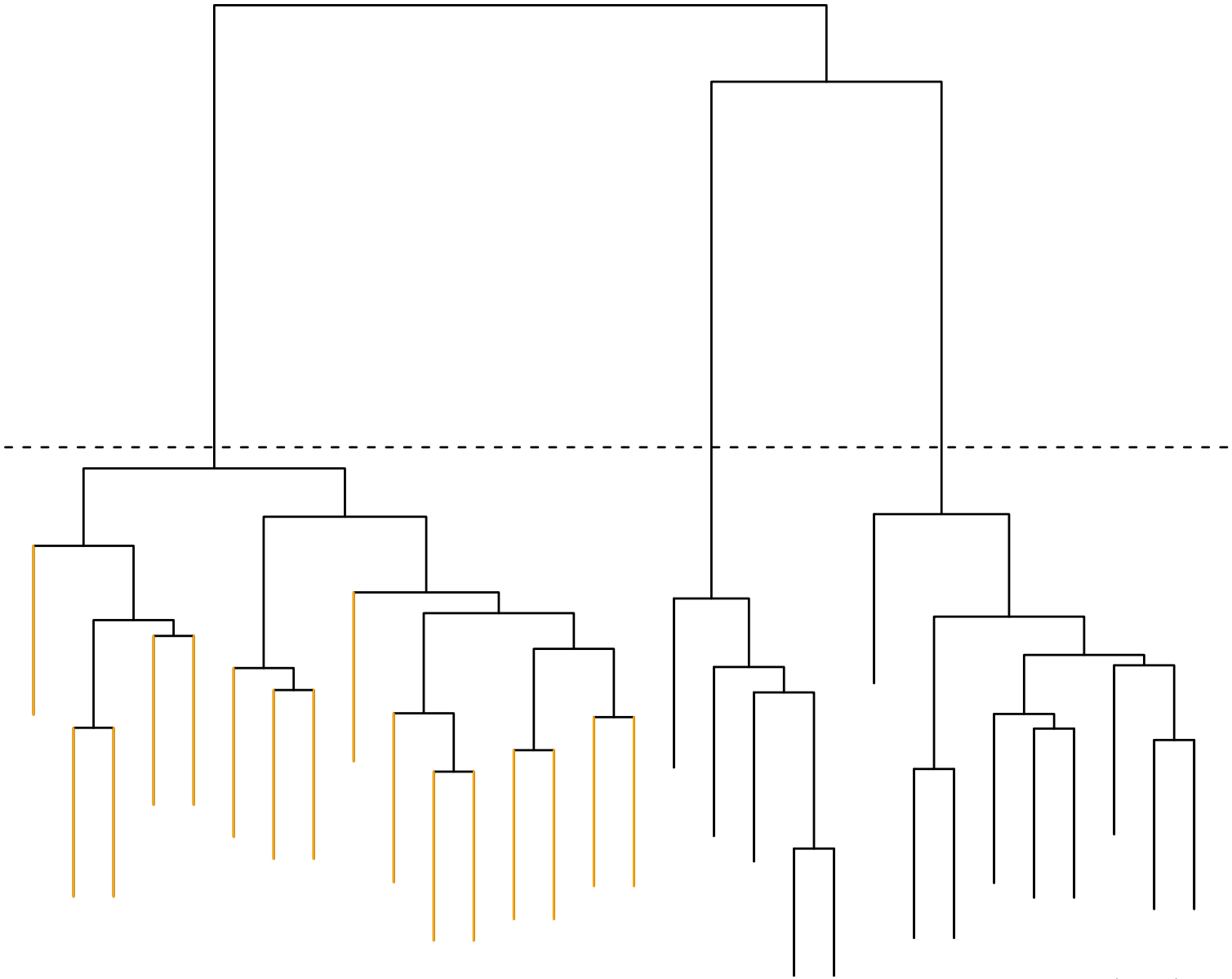
Abstentions for joint clustering



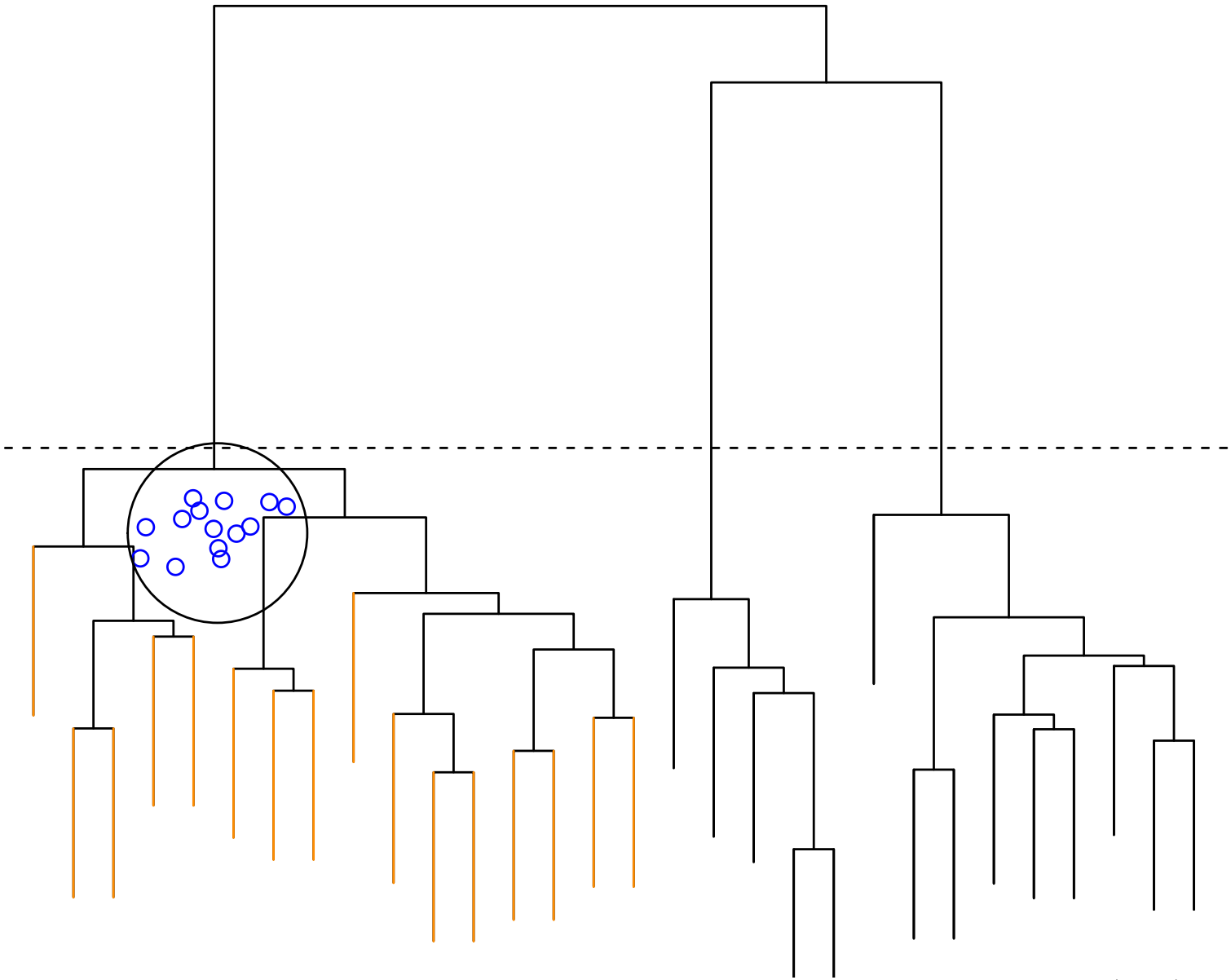
Test clustering

1. Hierarchically cluster just the test data.
2. Cut dendrogram off at some height determined by cross-validation.
3. For each cluster find, training points that are close to that cluster.
4. Fit separate models to training data within each cluster.
5. Predict each test point using the model in the cluster where it falls.

Test clustering



Test clustering



Battery of machine learning data sets

			Lasso	K-CT_J		K-CT_T		
Data set	n	p	Error	Error	K	Error	K	%Imp*
Musk (Version 1)	3299	166	.057	**.036**	3	.042	10	36.3%
Balance Scale	313	4	.102	.099	3	**.076**	10	25%
LSVT Voice Rehabilitation	63	310	.142	.142	1	**.111**	2	22.2%
seeds	105	7	.057	**.047**	2	**.047**	2	16.6%
QSAR Biodegradation	528	41	.149	**.138**	5	.153	1	7.5%
vowel	528	10	.525	**.487**	2	.523	10	7.4%
Parkinsons	98	22	.154	.154	3	**.144**	3	6.6%
Contraceptive Method Choice	737	18	.467	.455	5	**.448**	10	4.0%
Steel Plates Faults	971	27	.303	**.293**	5	.294	10	3.3%
Teaching Assistant Evaluation	76	53	.480	**.470**	10	.693	5	1.9%
Breast Cancer Wisconsin (Diagnostic)	285	30	**.035**	**.035**	2	**.035**	2	—
Fertility	50	9	**.160**	**.160**	1	**.160**	1	—
Mushroom	4062	96	**.000**	**.000**	1	**.000**	2	—
User Knowledge Modeling	258	5	**.020**	**.020**	1	.034	1	—
Optical Recognition of Handwritten Digits	3823	62	**.043**	.044	5	.044	1	-2.5%
Chess (King-Rook vs King-Pawn)	1598	38	**.024**	.026	5	.026	10	-10.2%

* %Imp: Percent improvement of the better of K-CT_J and K-CT_T, relative to lasso error rate

Back to the cancer detection problem

- clustering not necessary
- for each patient:
  - find 10 nearest neighbors of each data point
  - take union of these neighbor sets as customized training set

Test data:

Patient	Data
1 1 ... 1	
2 2 ... 2	
3 3 ... 3	
4 ...	

Mass spectrometry cancer results

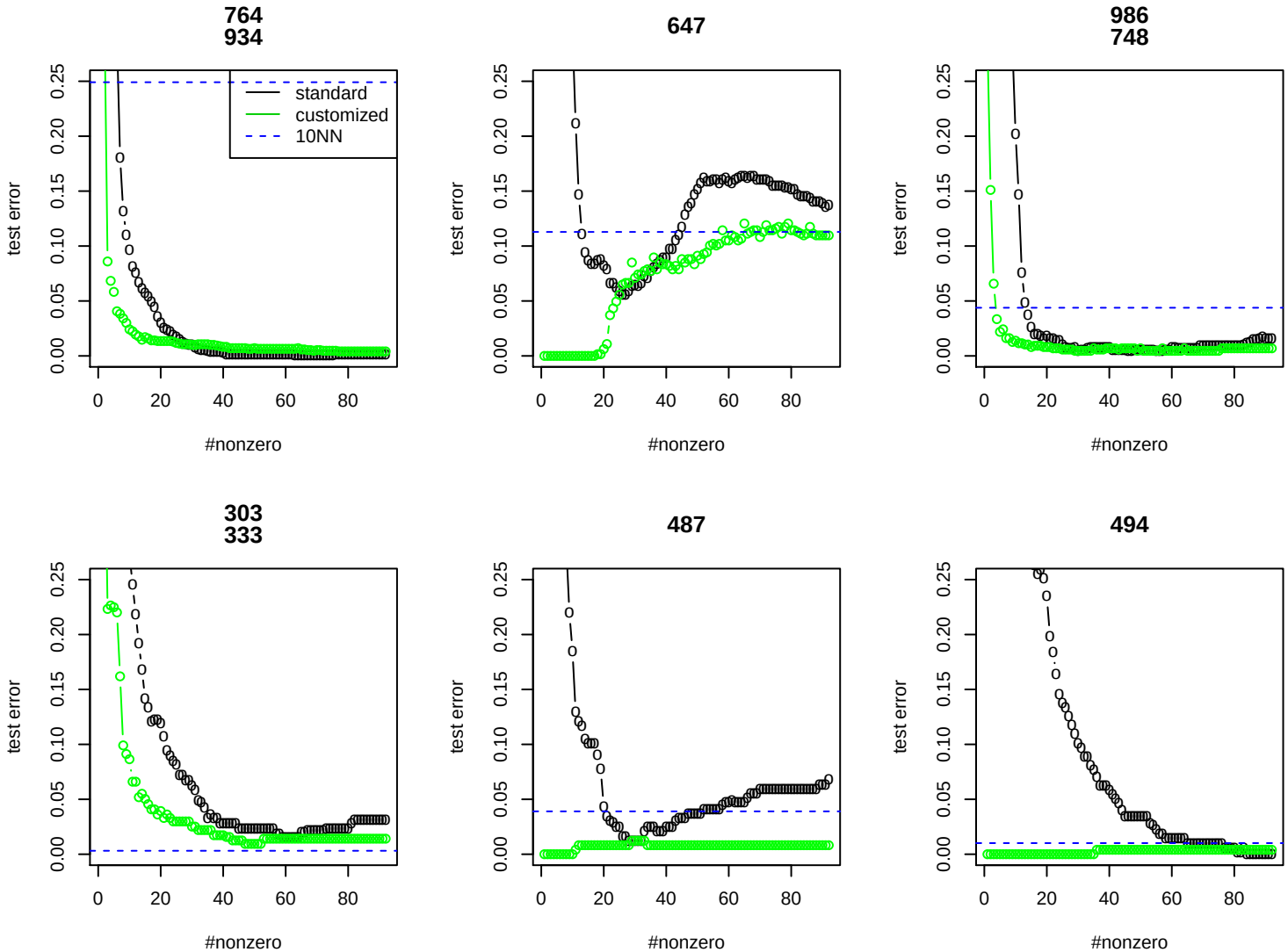
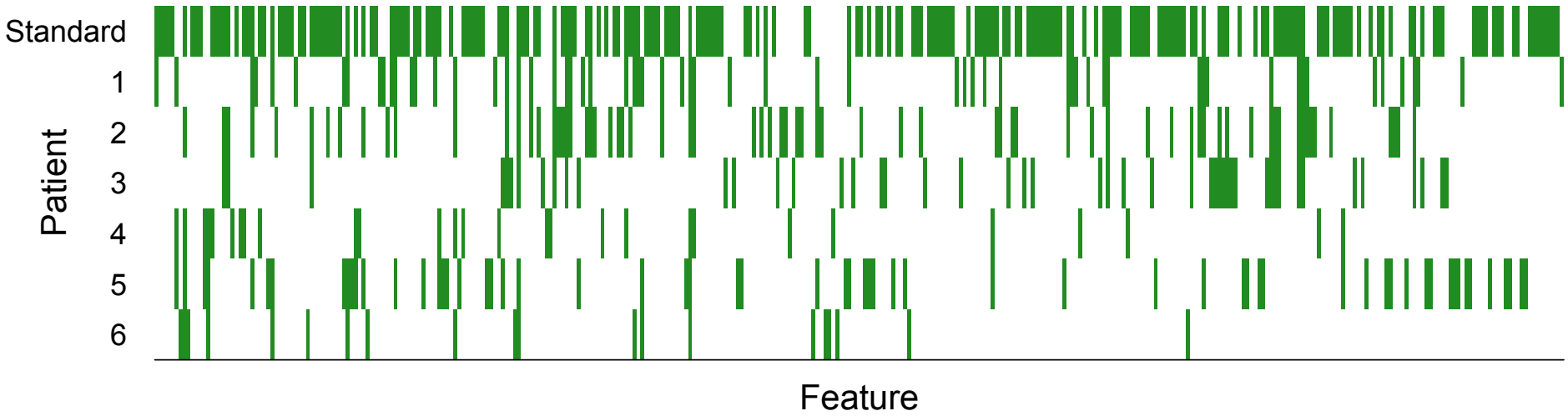


Table: Test error rates by patient for standard and customized training

Patient	1	2	3	4	5	6	Overall
Standard training	0.29%	4.56%	6.78%	0.00%	13.76%	2.77%	3.58%
Customized training	0.71%	1.89%	0.82%	0.40%	9.43%	0.92%	1.89%



Features selected by customized training for each patient (variables not selected by any model are omitted from the x-axis).

Using hierarchical clustering with Jaccard distance between the sets of selected features to split the patients into two clusters, patients 1, 2 and 3 were in one cluster, with patients 4, 5 and 6 in the other.

Another twist for customized training in this example

Train on normal data from a new patient, on the fly.

Use this information to improve predictions

Conclusions

- Working with real scientists and real problems is challenging and fulfilling
- Some scientists really seem to like the sparsity that results from ℓ_1 -regularization
- Customized training is a simple idea, that can discover hidden structure in data and potentially improve predictions.