THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Advanced Data Mining with Weka

Class 2 – Lesson 1

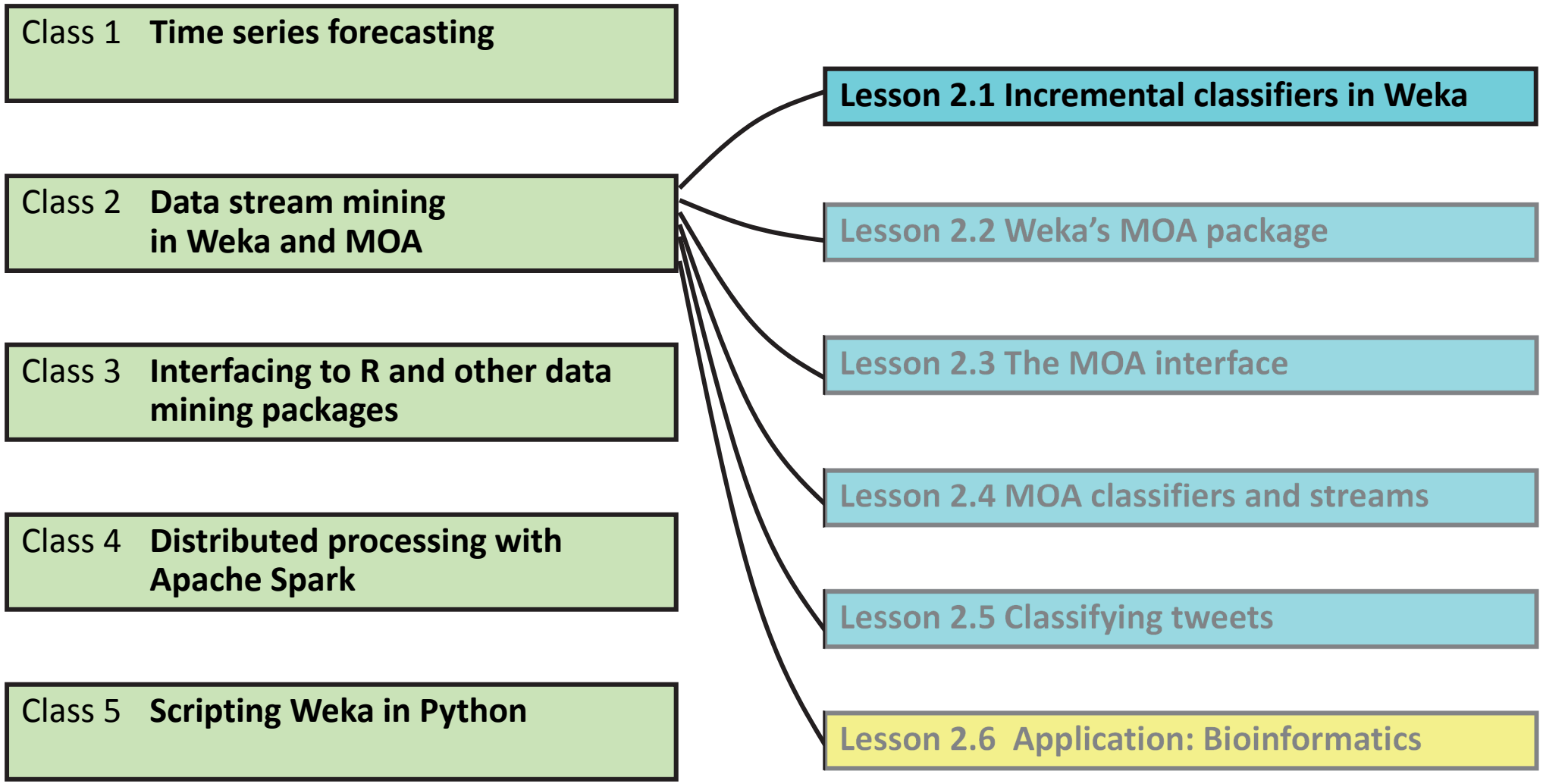
Incremental classifiers in Weka

Albert Bifet

Department of Computer Science
University of Waikato
New Zealand

weka.waikato.ac.nz

Lesson 2.1: Incremental classifiers in Weka



Incremental classifiers in Weka

Batch Setting

- Build a classifier using a **dataset** in memory

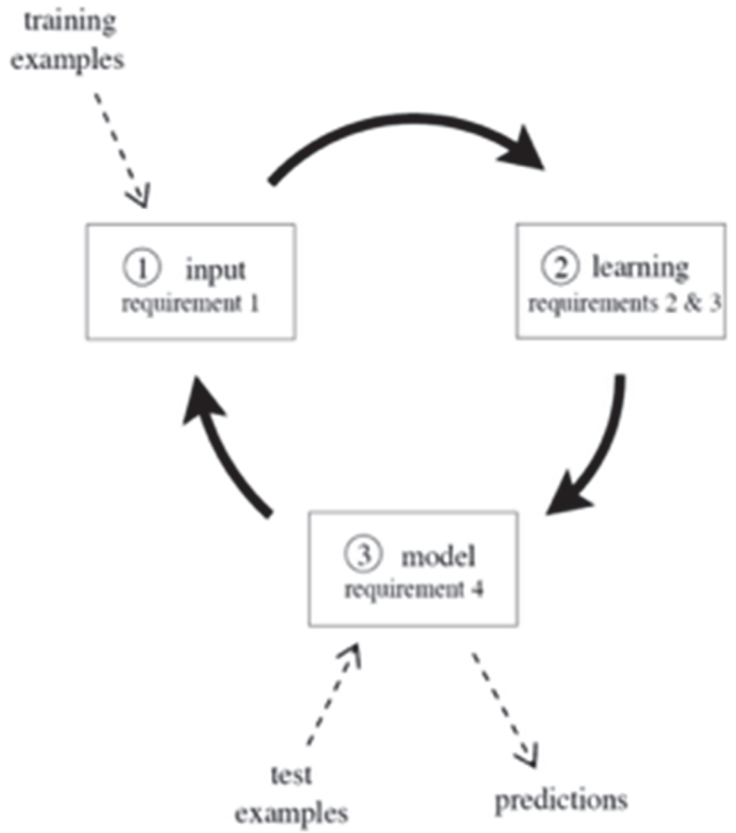
Incremental Setting

- Update a classifier using an **instance**

Incremental classifiers in Weka

Incremental Setting

- Process an example at a time,and inspect it only once (at most)
- Use a limited amount of memory
- Work in a limited amount of time
- Be ready to predict at any point



Incremental classifiers in Weka

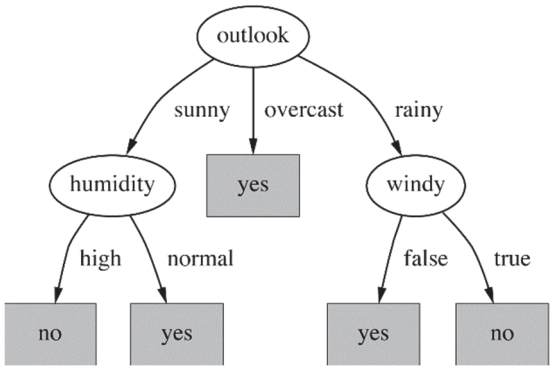
Incremental Methods (UpdateableClassifier)

- Bayes
  - NaiveBayes
  - NaiveBayesMultinomial
- Lazy
  - IBk: k-Nearest Neighbours
- Functions
  - SGD
  - SGDText
- Trees
  - Hoeffding Tree

Incremental classifiers in Weka

Hoeffding Tree

- Sample of stream enough for near optimal decision
- Estimate merit of alternatives from prefix of stream
- Choose sample size based on statistical principles
- When to expand a leaf?
  - Hoeffding bound: split if



$$G(\text{Best Attr.}) - G(\text{2nd best}) > \sqrt{\frac{R^2 \ln 1/\delta}{2n}}$$

Incremental classifiers in Weka

Batch Setting

- Build a classifier using a **dataset** in memory
  - `buildClassifier(Instances)`

Incremental Setting

- Update a classifier using an **instance**
  - `updateClassifier(Instance)`
- **Less Resources**
  - **Uses less memory:** don't need to store the dataset in memory
  - **Faster:** as data is seen only in one pass

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Advanced Data Mining with Weka

Class 2 – Lesson 2

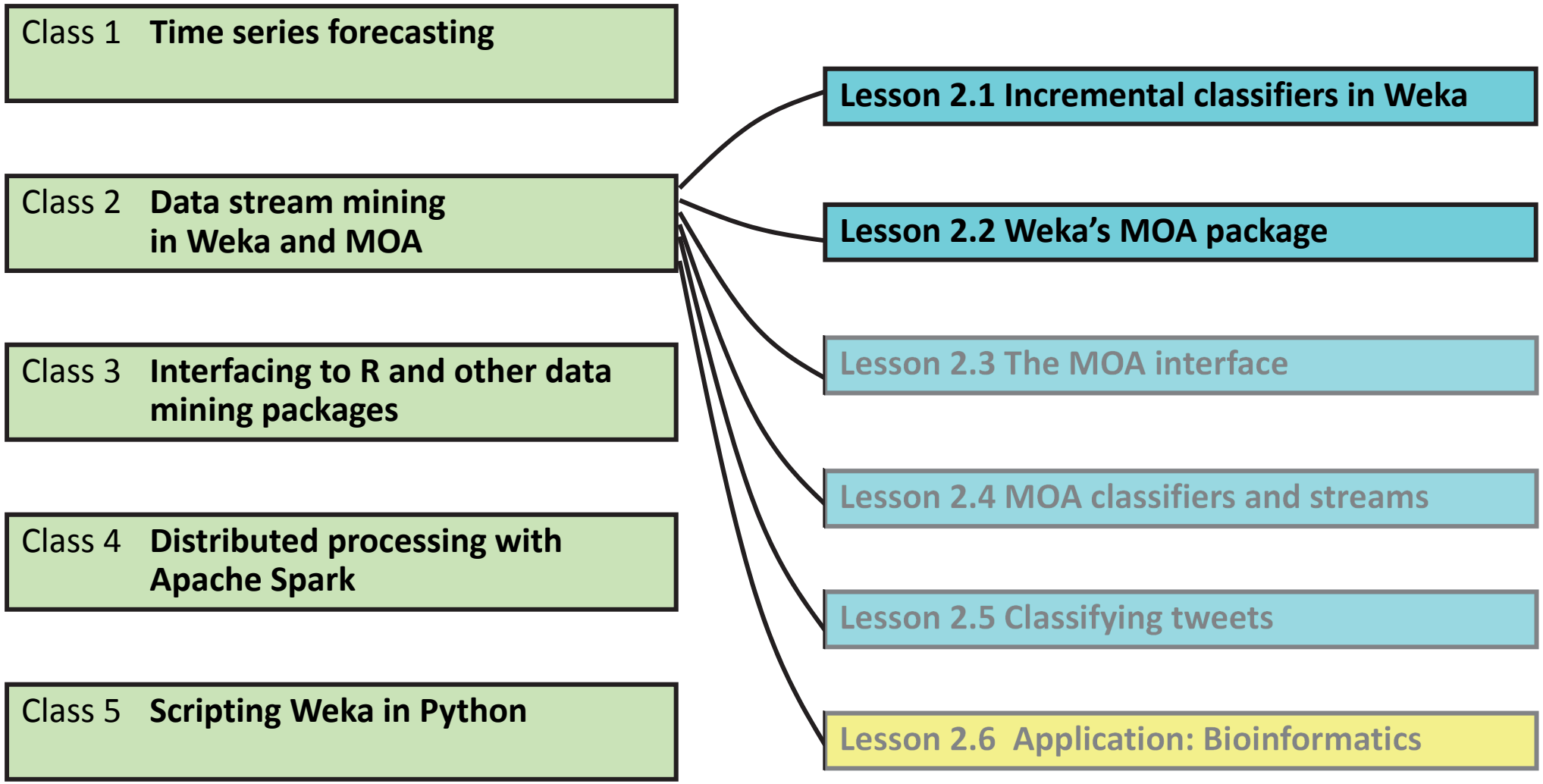
Weka's MOA package

Albert Bifet

Department of Computer Science
University of Waikato
New Zealand

weka.waikato.ac.nz

Lesson 2.2: Weka's MOA package



Weka's MOA package

MOA: Massive Online Analysis

- {M}assive {O}nline {A}nalysis is a framework for online learning from data streams.
- It handles **evolving** data streams, streams with **concept drift**.
- It includes a collection of offline and online as well as tools for evaluation:
  - classification, regression
  - clustering, frequent pattern mining
  - outlier detection, concept drift
- Easy to extend, design and run experiments

Weka's MOA package

MOA: Massive Online Analysis

- MOA can be used with
  - ADAMS: The Advanced Data mining And Machine learning System, a novel, flexible workflow engine aimed at quickly building and maintaining real-world, complex knowledge workflows.
    - https://adams.cms.waikato.ac.nz/
  - MEKA: **Multi-label** learning and evaluation open source framework
    - http://meka.sourceforge.net/

Weka's MOA package

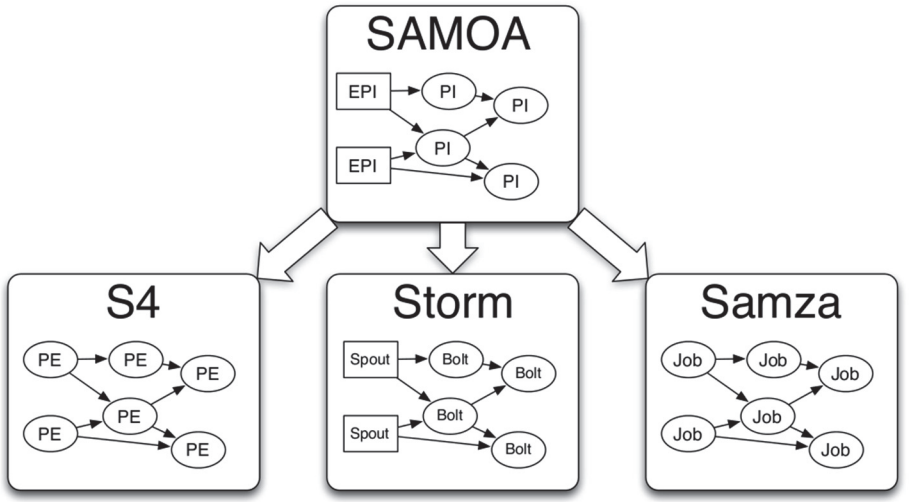
SAMOA: Scalable Advanced Massive Online Analysis

Apache SAMOA enables development of new ML algorithms over distributed stream processing engines (DSPEe, such as Apache Storm, Apache S4, and Apache Samza).

Apache SAMOA users can develop distributed streaming ML algorithms once and execute them on multiple DSPEs.

Apache SAMOA started at Yahoo Labs.

https://samoa.incubator.apache.org/

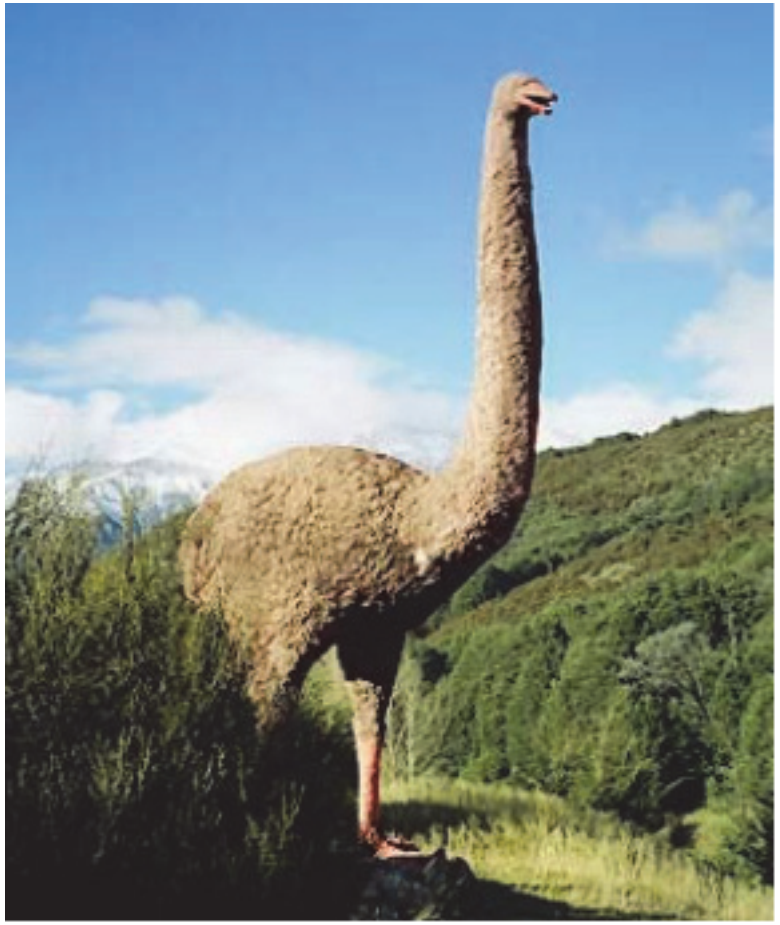


Weka's MOA package

Weka : the bird

Weka's MOA package

MOA : the bird



The MOA is another native NZ bird, flightless but extinct.

Weka's MOA package

MOA : the bird

Weka's MOA package

MOA : the bird

Weka's MOA package

Install the *massiveOnlineAnalysis* package

Weka's MOA package

MOA: Massive Online Analysis

- {M}assive {O}nline {A}nalysis is a framework for online learning from data streams.
- It handles **evolving** data streams, streams with **concept drift**.
- It includes a collection of offline and online as well as tools for evaluation:
  - classification, regression
  - clustering, frequent pattern mining
  - outlier detection, concept drift
- Easy to extend, design and run experiments

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Advanced Data Mining with Weka

Class 2 – Lesson 3

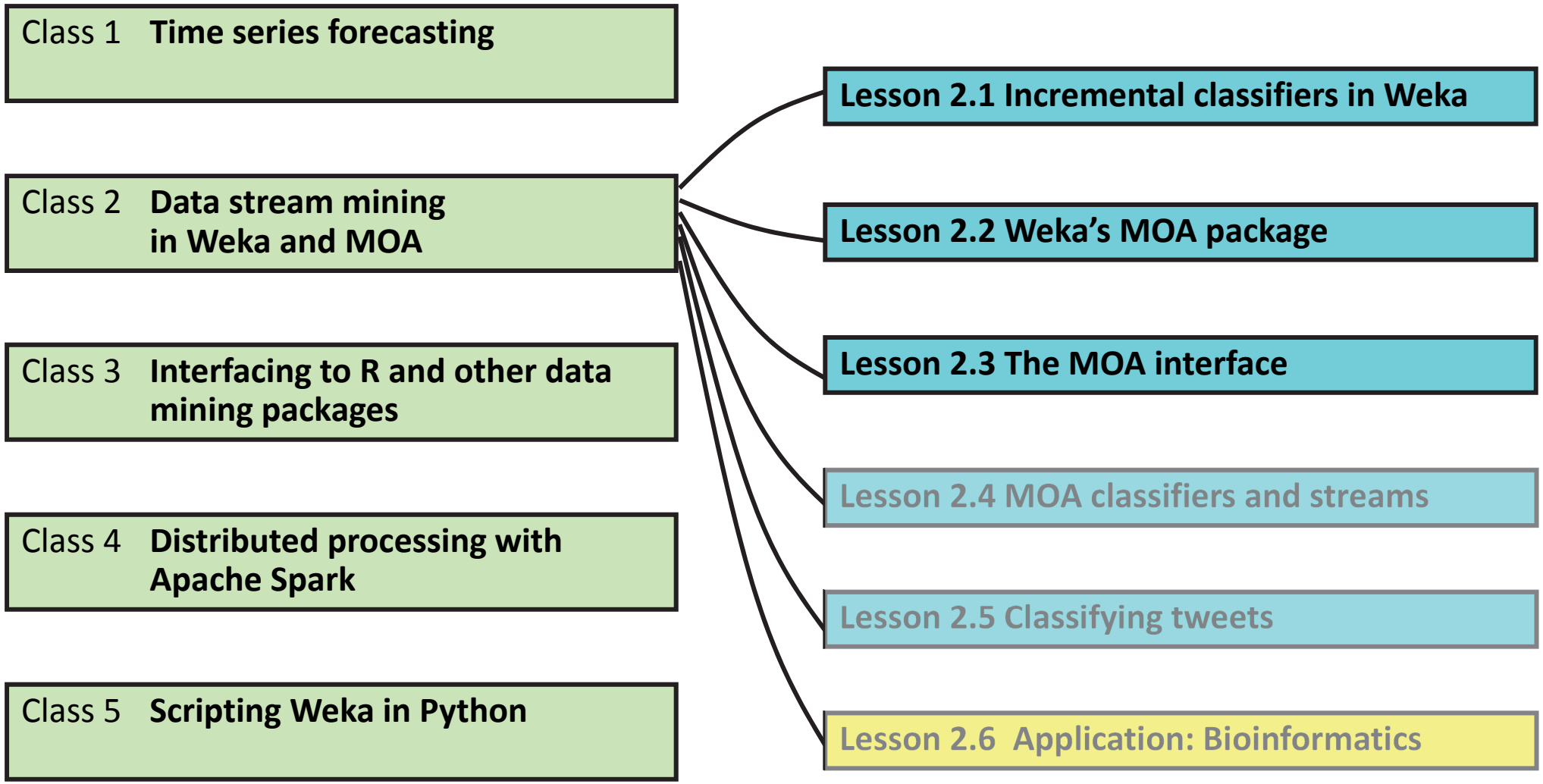
The MOA interface

Albert Bifet

Department of Computer Science
University of Waikato
New Zealand

weka.waikato.ac.nz

Lesson 2.3: The MOA interface



The MOA interface

MOA

- Graphical User Interface
- Command Line
- Java API

The MOA interface

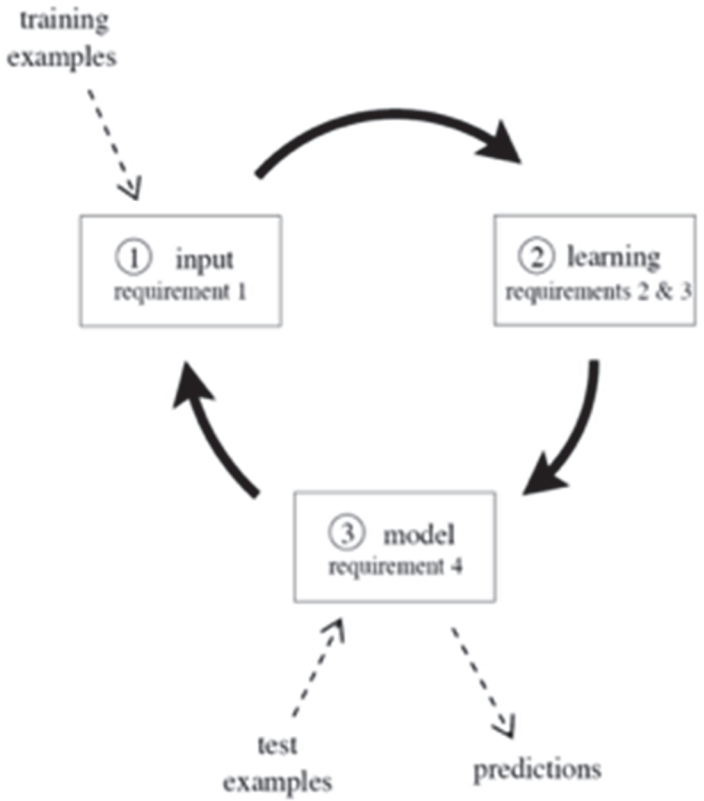
Classification Evaluation

- Holdout Evaluation
- Interleaved Test-Then-Train or Prequential

The MOA interface

Holdout an independent test set

- Apply the current decision model to the test set, at regular time intervals
- The loss estimated in the holdout is an unbiased estimator

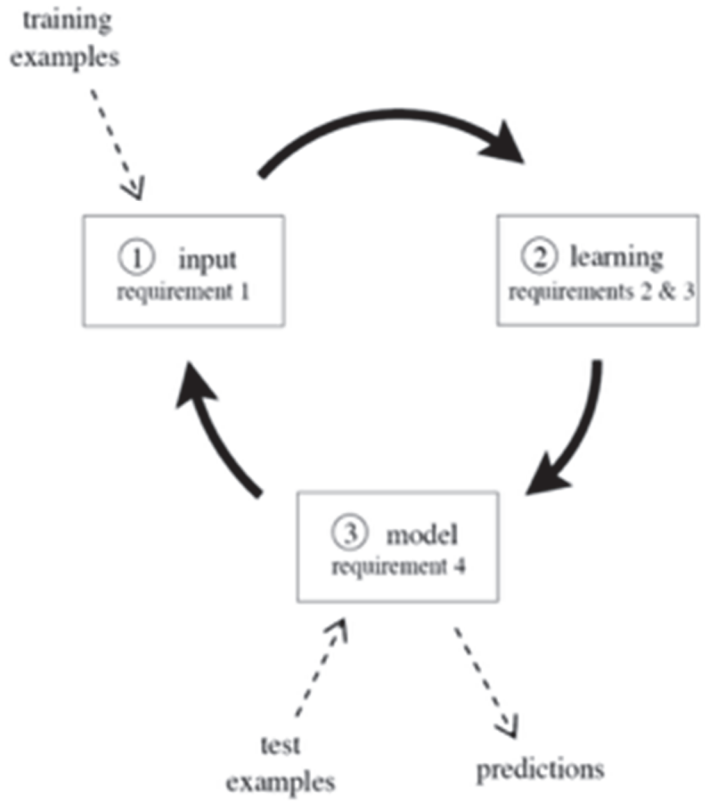


The MOA interface

Prequential Evaluation

- The error of a model is computed from the sequence of examples.
- For each example in the stream, the actual model makes a prediction based only on the example attribute-values.

$$S = \sum_{i=1}^{n} L(y_i, \hat{y}_i).$$



The MOA interface

Command Line

```
java -cp .:moa.jar:weka.jar -javaagent:sizeofag.jar
moa.DoTask "EvaluatePeriodicHeldOutTest -l
DecisionStump -s generators.WaveformGenerator -n
100000 -i 100000000 -f 1000000" > dsresult.csv
```

- This command creates a comma separated values file:
  - training the DecisionStump classifier on the WaveformGenerator data,
  - using the first 100 thousand examples for testing,
  - training on a total of 100 million examples,
  - and testing every one million examples

The MOA interface

MOA

- Graphical User Interface
- Command Line
- Java API

- Evaluation
  - Holdout
  - Prequential

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Advanced Data Mining with Weka

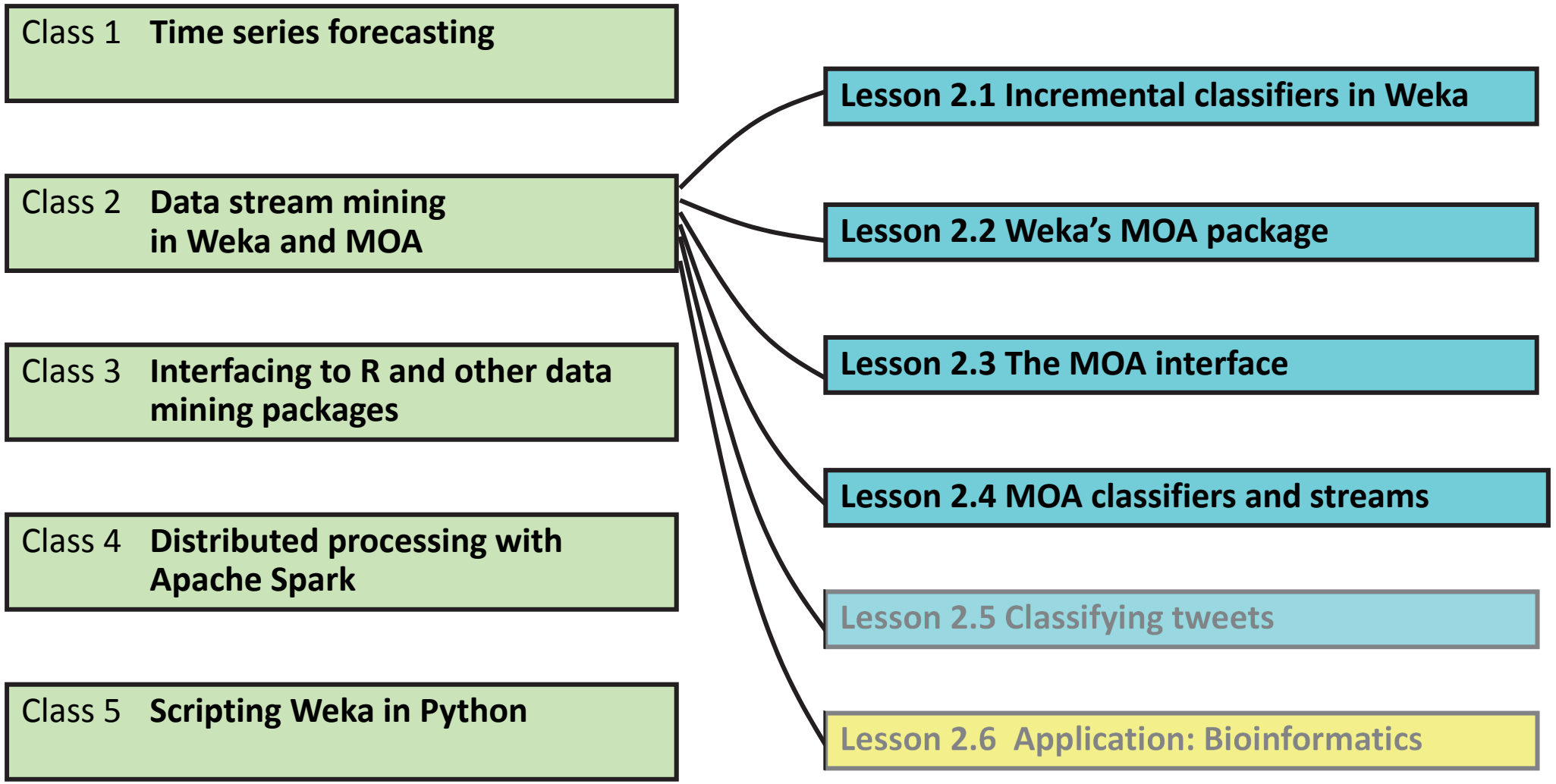
Class 2 – Lesson 4

MOA classifiers and streams

Bernhard Pfahringer

Department of Computer Science
University of Waikato
New Zealand

Lesson 2.4: MOA classifiers and streams



MOA classifiers and streams

ADWIN

- An adaptive sliding window whose size is recomputed online according to the rate of change observed.
- ADWIN has rigorous guarantees (theorems)
  - On ratio of false positives and negatives
  - On the relation of the size of the current window and change rates

MOA classifiers and streams

Hoeffding Adaptive Tree

- construct “alternative branches” as preparation for changes
- if the alternative branch becomes more accurate, switch of tree branches occurs
- checks the substitution of alternate subtrees using a change detector with theoretical guarantees (ADWIN)

MOA classifiers and streams

Bagging

- Dataset of 4 Instances : A, B, C, D
  - Classifier 1: B, A, C, B
  - Classifier 2: D, B, A, D
  - Classifier 3: B, A, C, B
  - Classifier 4: B, C, B, B
  - Classifier 5: D, C, A, C
- Bagging builds a set of M base models, with a bootstrap sample created by drawing random samples with replacement.

MOA classifiers and streams

Bagging

- Dataset of 4 Instances : A, B, C, D
  - Classifier 1: A, B, B, C
  - Classifier 2: A, B, D, D
  - Classifier 3: A, B, B, C
  - Classifier 4: B, B, B, C
  - Classifier 5: A, C, C, D
- Bagging builds a set of M base models, with a bootstrap sample created by drawing random samples with replacement.

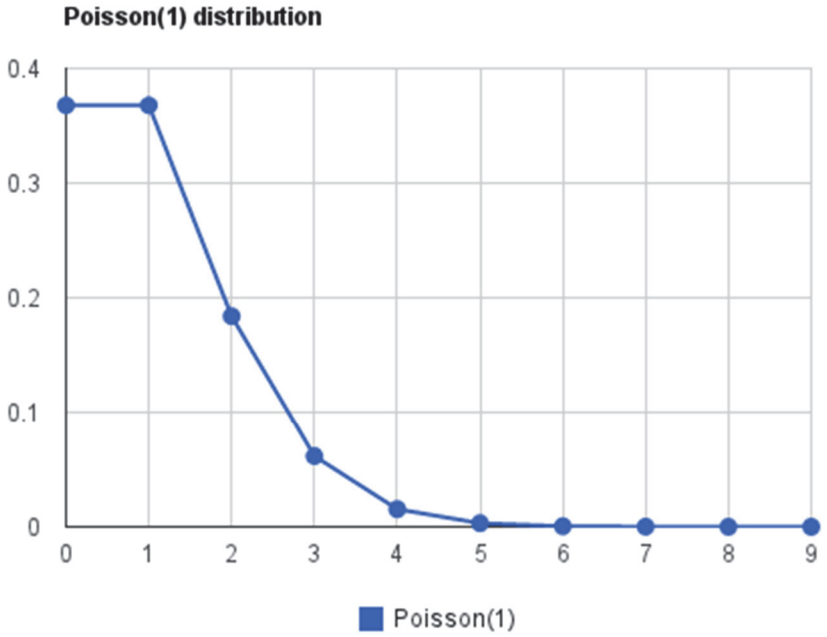
MOA classifiers and streams

Bagging

- Dataset of 4 Instances : A, B, C, D
  - Classifier 1: A, B, B, C: A(1) B(2) C(1) D(0)
  - Classifier 2: A, B, D, D: A(1) B(1) C(0) D(2)
  - Classifier 3: A, B, B, C: A(1) B(2) C(1) D(0)
  - Classifier 4: B, B, B, C: A(0) B(3) C(1) D(0)
  - Classifier 5: A, C, C, D: A(1) B(0) C(2) D(1)
- Each base model's training set contains each of the original training example K times where P(K = k) follows a binomial distribution.

MOA classifiers and streams

Bagging



- Each base model's training set contains each of the original training example K times where P(K = k) follows a binomial distribution.

MOA classifiers and streams

ADWIN Bagging

- Uses Poisson(1) to weight new instances to do online sampling
- When a change is detected, the worst classifier is removed and a new classifier is added.

Leveraging Bagging

- Uses Poisson(> 1) to weight new instances to do online sampling
- When a change is detected, the worst classifier is removed and a new classifier is added.

MOA classifiers and streams

- Evolving classifiers
  - Hoeffding Adaptive Tree
  - ADWIN Bagging
  - Leveraging Bagging

- Evolving artificial data streams
  - RandomRBF with drift
  - SEA concepts
  - LED
  - Wave
  - STAGGER concepts

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Advanced Data Mining with Weka

Class 2 – Lesson 5

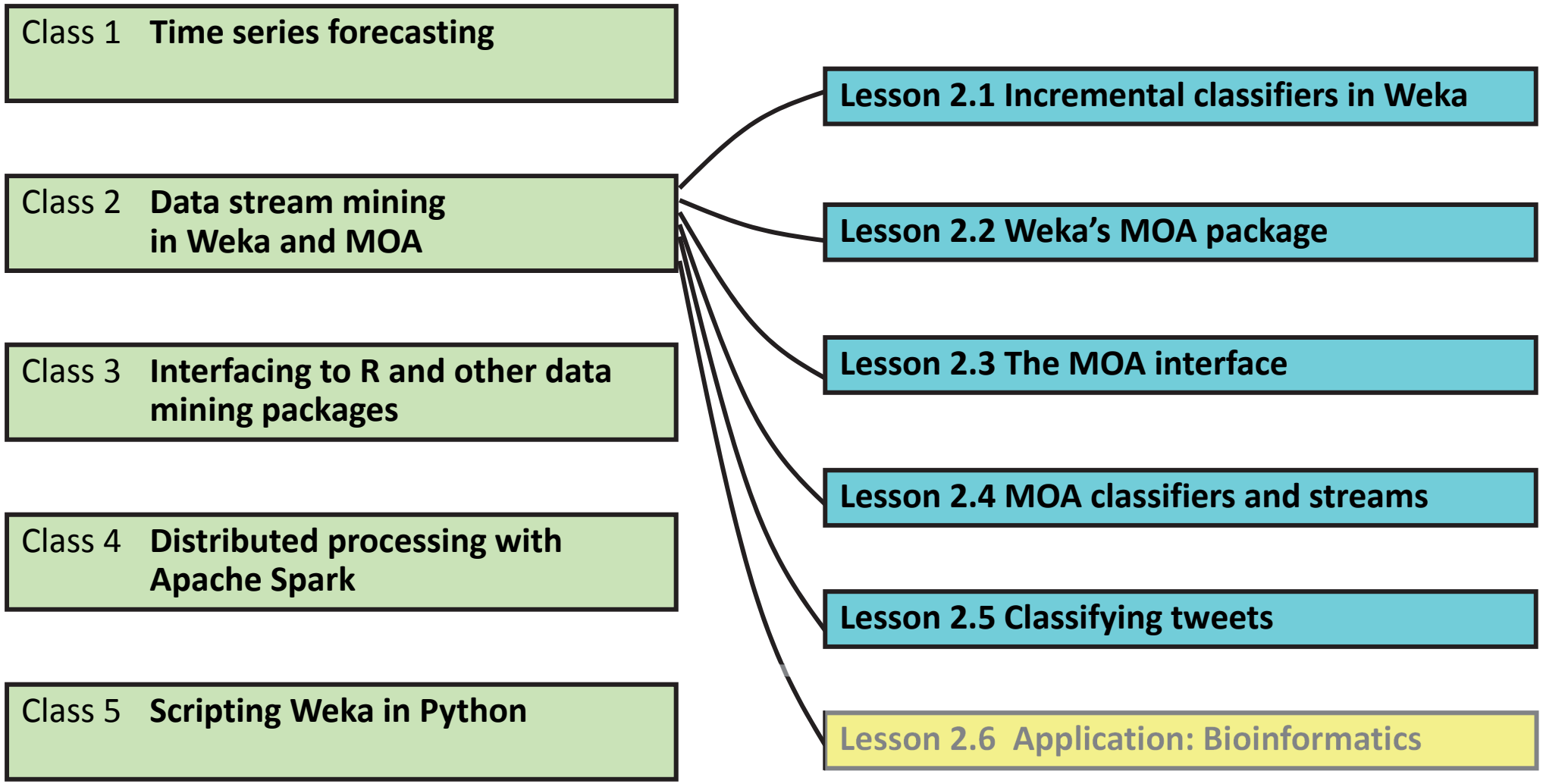
Classifying tweets

Albert Bifet

Department of Computer Science
University of Waikato
New Zealand

weka.waikato.ac.nz

Lesson 2.5: Classifying tweets



Classifying tweets

- Micro-blogging service
- Built to discover what is happening at any moment in time, anywhere in the world.
- 316 million registered users
- 2100 million search queries per day
- 3 billion requests a day via its API.

Classifying tweets

- Sentiment Analysis
  - Classifying messages into two categories depending on whether they convey positive or negative feelings
  - **Emoticons** are visual cues associated with emotional states, which can be used to define class labels for sentiment classification

Positive Emoticons	Negative Emoticons
:)	:(
:-)	:-(
:)	: (
:D	
=)	

Classifying tweets

	Predicted Class+	Predicted Class-	Total
Correct Class+	75	8	83
Correct Class-	7	10	17
Total	82	18	100

Table: Simple confusion matrix example

	Predicted Class+	Predicted Class-	Total
Correct Class+	68.06	14.94	83
Correct Class-	13.94	3.06	17
Total	82	18	100

Table: Confusion matrix for chance predictor

Classi g tweets

Kappa Statistic

- p_0: classifier's prequential accuracy
- p_c: probability that a chance classifier makes a correct prediction.
- κ statistic

$$\kappa = \frac{p_0 - p_c}{1 - p_c}$$

- $\kappa = 1$ if the classifier is always correct
- $\kappa = 0$ if the predictions coincide with the correct ones as often as those of the chance classifier

	Predicted Class+	Predicted Class-	Total
Correct Class+	75	8	83
Correct Class-	7	10	17
Total	82	18	100

Table: Simple confusion matrix example

	Predicted Class+	Predicted Class-	Total
Correct Class+	68.06	14.94	83
Correct Class-	13.94	3.06	17
Total	82	18	100

Table: Confusion matrix for chance predictor

Classifying tweets

Twitter Sentiment Corpus

- `twittersentiment.appspot.com`
- Alec Go, Richa Bhayani, Karthik Raghunathan, and Lei Huang
- Website to research the sentiment for a brand, product, or topic.
- Training dataset with messages between April 2009 and June 25, 2009
  - 800,000 tweets with positive emoticons
  - 800,000 tweets with negative emoticons
- Test dataset manually annotated
  - 177 negative tweets
  - 182 positive ones

Classifying tweets

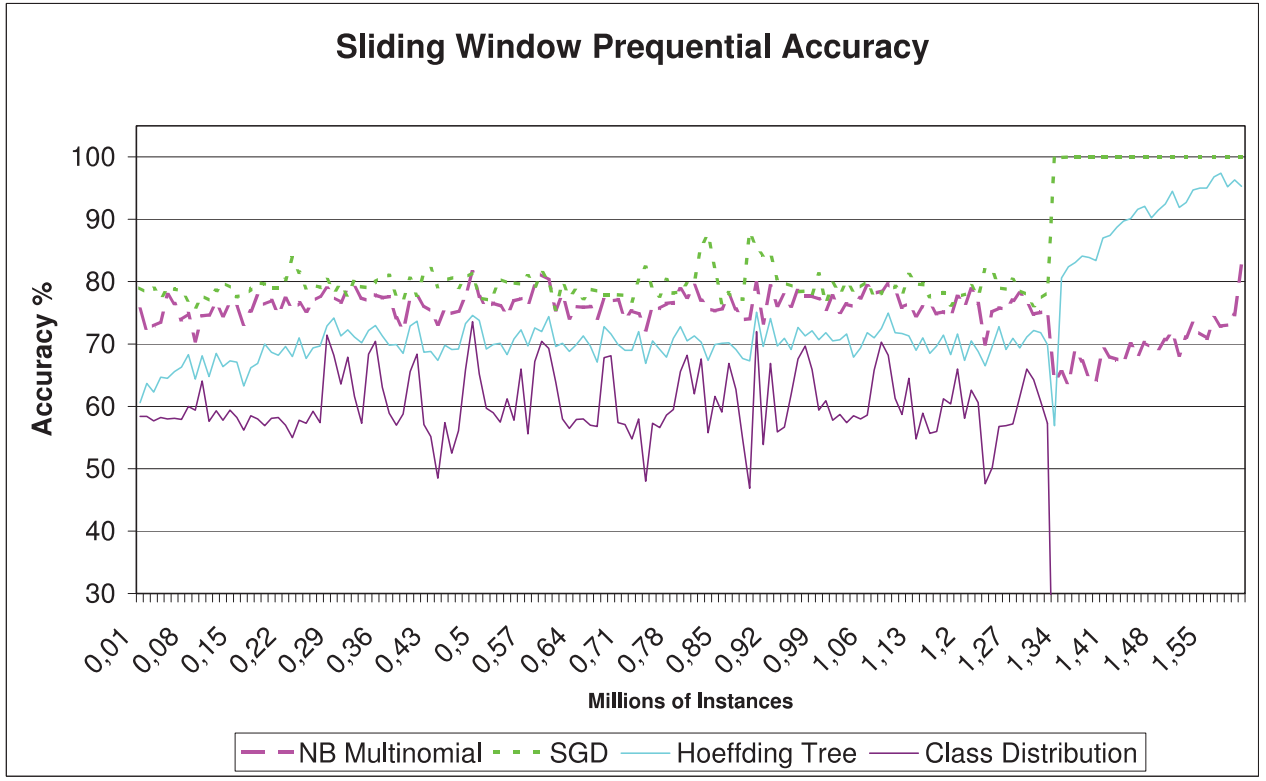


Figure: Accuracy and Kappa Statistic on `twittersentiment` corpus

Classifying tweets

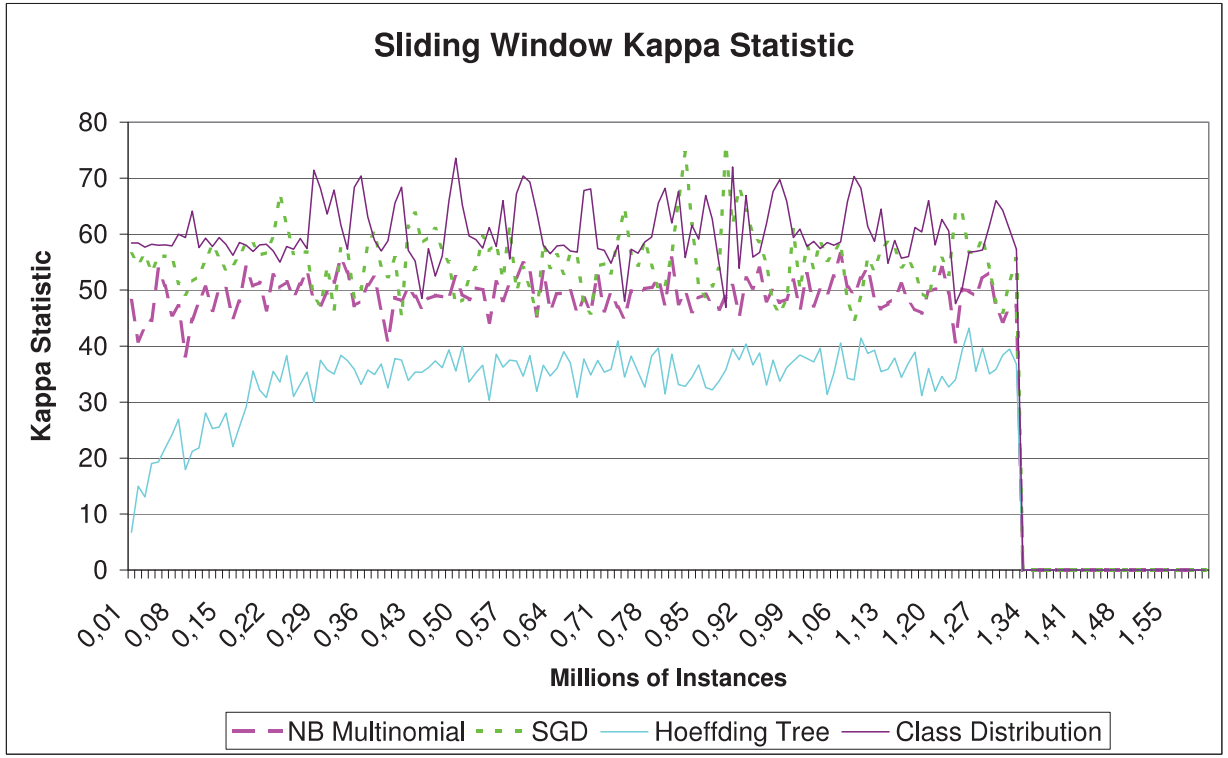


Figure: Accuracy and Kappa Statistic on `twittersentiment` corpus

Classifying tweets

Prequential Accuracy and Kappa

	Accuracy	Kappa	Time
Multinomial Naïve Bayes	75.05%	50.10%	116.62 sec.
SGD	82.80%	62.60%	219.54 sec.
Hoeffding Tree	73.11%	46.23%	5525.51 sec.

Total prequential accuracy and Kappa measured on the `twittersentiment` data stream

Classifying tweets

- Twitter: Micro-blogging streaming service
- Built to discover what is happening at any moment in time, anywhere in the world
- Data may be unbalanced
  - Accuracy is not enough
  - Use Kappa Statistic

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Advanced Data Mining with Weka

Class 2 – Lesson 6

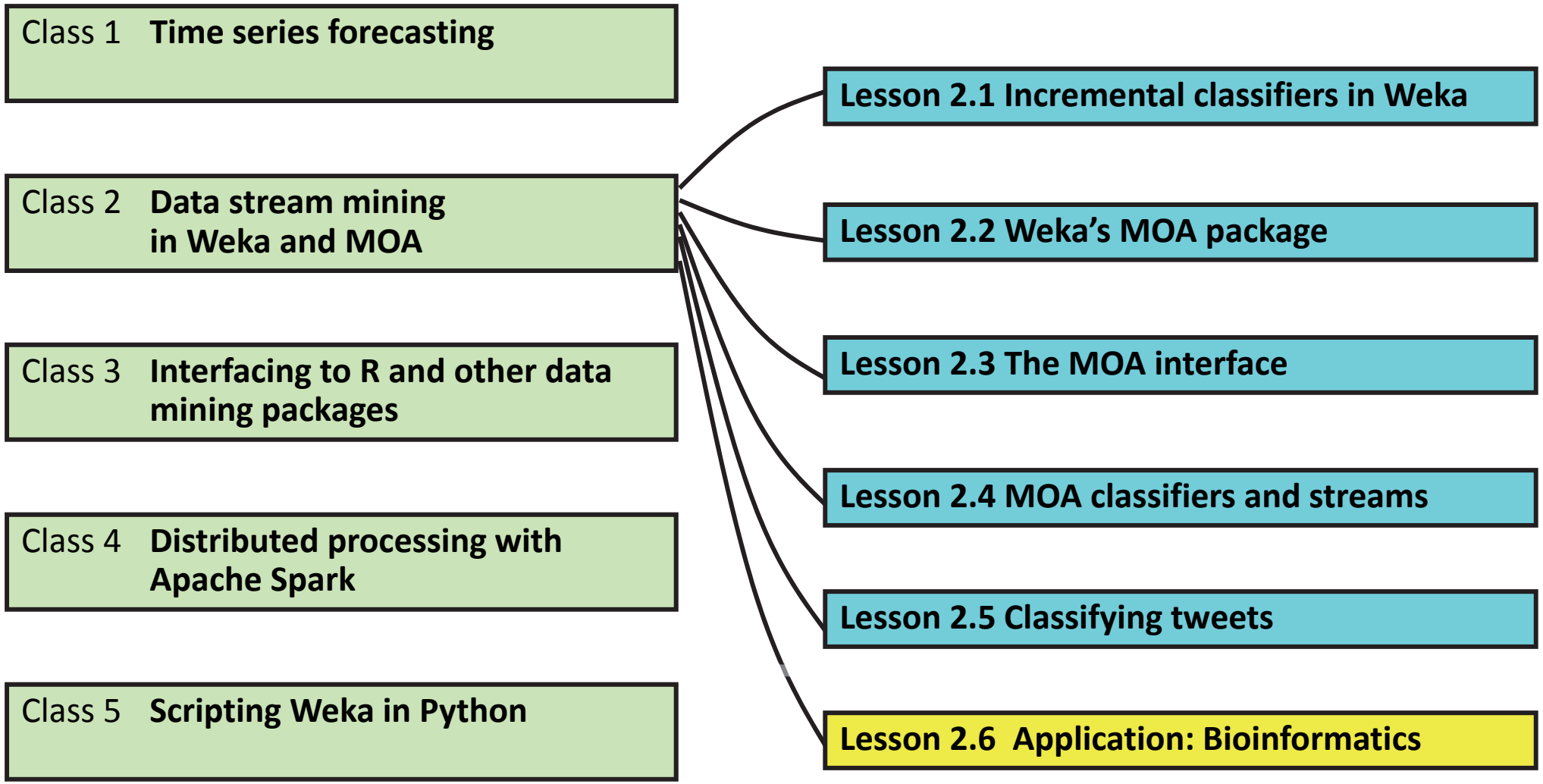
Application to Bioinformatics – Signal peptide prediction

Tony Smith

Department of Computer Science
University of Waikato
New Zealand

weka.waikato.ac.nz

Lesson 2.6: Application to Bioinformatics – Signal peptide prediction



Bioinformatics

Computation with biological data.

- Site prediction (e.g. oglycosylation points)
- Microarray analysis (e.g. gene expression)
- Genetic epidemiology (e.g. variant call correlations)
- Mass spectrum analysis (e.g. post-translational modifications)
- Sequence analysis (e.g. taxonomic classification)
- Structure prediction (e.g. fold properties)

Signal peptide – a sequence data problem

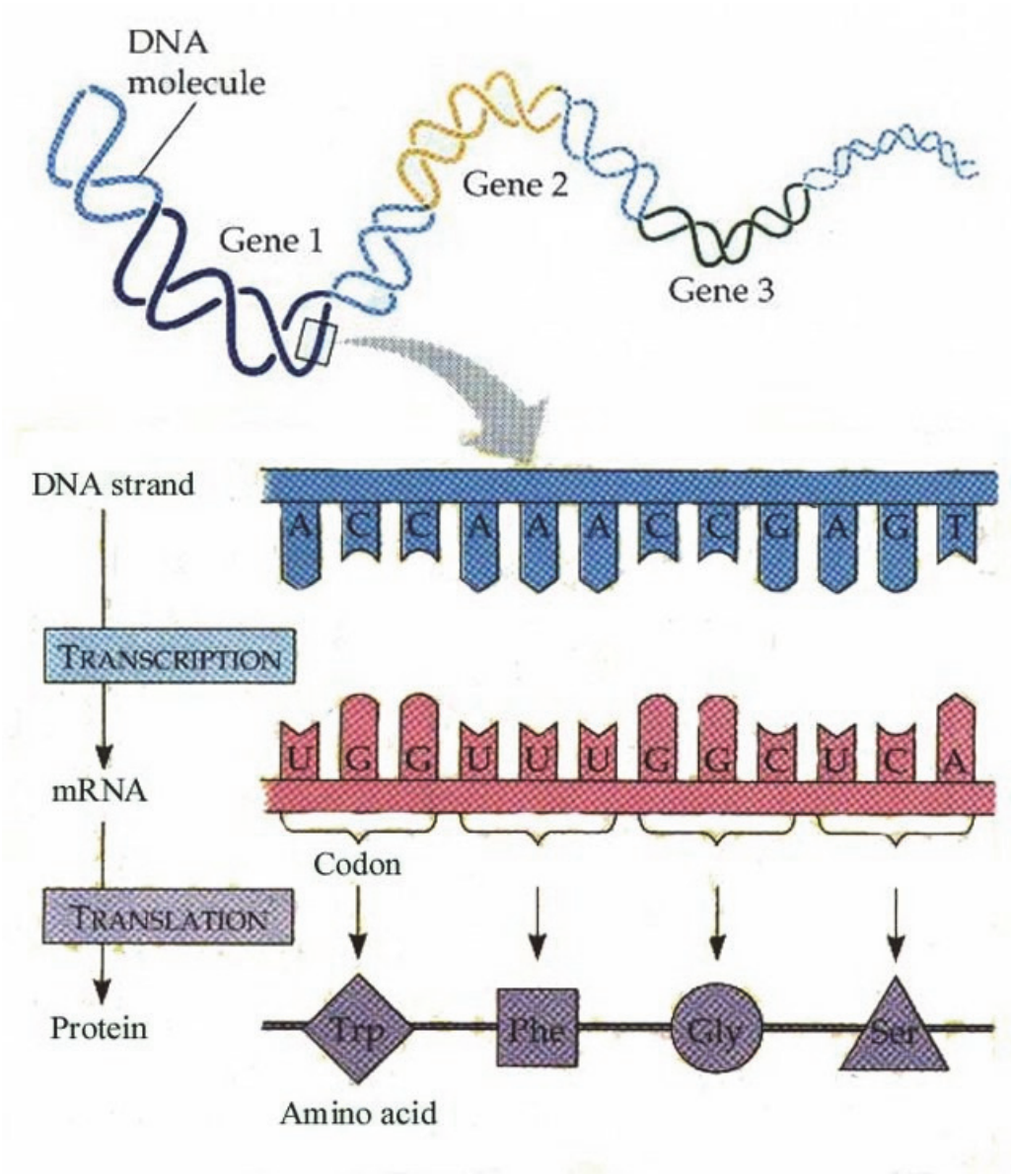
An example of an *easily stated* problem for protein sequence data

Given a freshly produced protein …

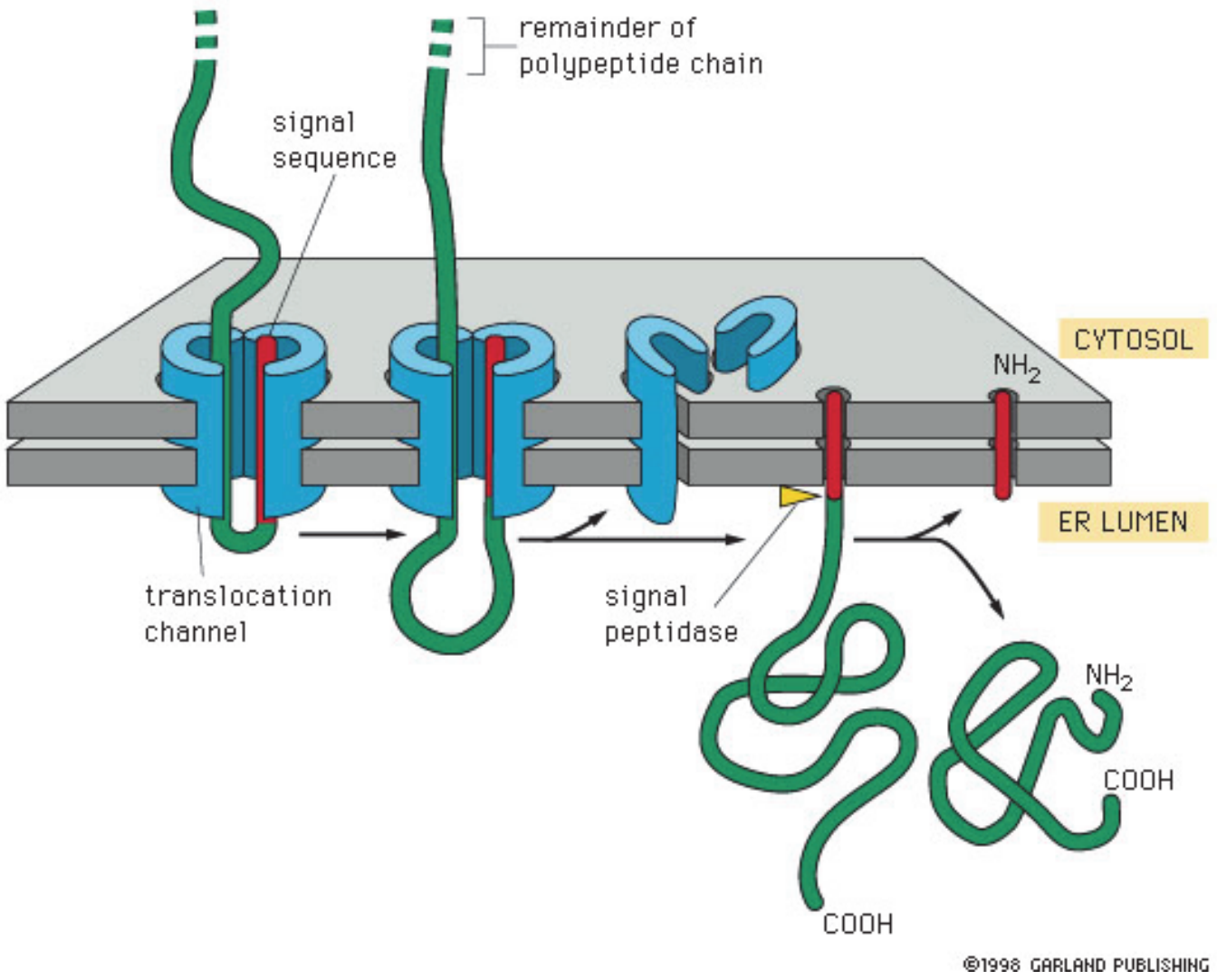
… which portion is the signal peptide?

What does this mean?

Central dogma – gene to transcript to protein



Signal peptide cleavage



Signal peptide cleavage

Where does the signal peptide end? Where is the cleavage point?

Issues:

- What is the goal - accurate prediction or an explanatory model?
- What features are relevant – how do we prepare data for success?
- What approach - predict length of peptide or the position of cleavage site?
- How will we know if we were successful?

The raw data — unstructured

MASKATLLLAFTLLFATCIARHQQRQQQQNQCQLQNIEALEPIEVIQAEA…

MARSSLFTFLCLAVFINGCLSQIEQQSPWEFQGSEVWQQHRYQSPRACRLE…

MLVMAPRTVLLLLSAALALTETWAGSHSMRYFYTSVSRPGRGEPRFISVGYVDD…

MKLSKSTLVFSALLVILAAASAAPANQFIKTSCTLTTYPAVCEQSLSAYAKT…

MANKLFLVCATLALCFLLTNASIYRTVVEFEEDDASNPVGPRQRCQKEFQQ…

MARFSIVFAAAGVLLLVAMAPVSEASTTTIITTIIEENPYGRGRTESGCYQQMEE…

MAKISVAAAALLVLMALGHATAFRATVTTTVVEEENQEECREQMQRQQMLSH…

MGNNCYNVVVIVLLLVGCEKVGAVQNSCDNCQPGTFCRKYNPVCKSCPPSTFSS…

MPRVPSASATGSSALLSLLCAFSLGRAAPFQLTILHTNDVHARVEETNQDSGKCFTQSFA…

MCPRAARAPATLLLALGAVLWPAAGAWELTILHTNDVHSRLEQTSEDSSKCVNASR…

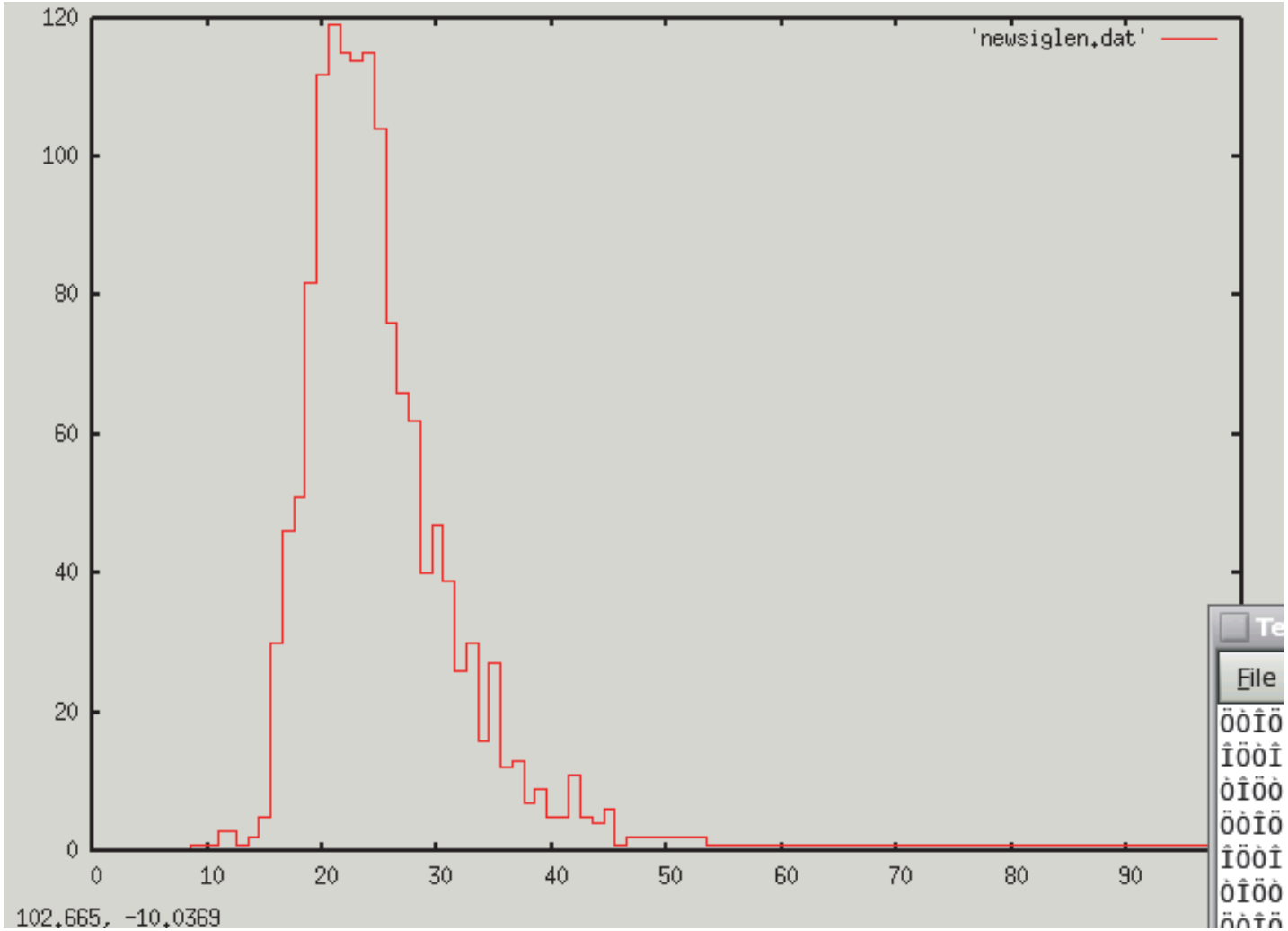
What structure? What features?

What do we think is relevant?

- Properties of the entire signal peptide?
- Properties of the cleavage site?

- ❖ Typically get some domain knowledge from the experts
- ❖ Trial and error – *ad hoc* statistical analysis

Signal peptide length



1410 samples; μ-length = 24

Patterns around cleavage site

Upstream	*Start*	*Downstream*
CIA	R	HQQ
CLS	Q	IEQ
TWA	G	SHS
ASA	A	PAN
TNA	S	IYR
SEA	S	TTT
ATA	F	RAT
GAV	Q	NSC
APF	Q	LTI
AGA	W	ELT
AFA	Y	SPR
SDS	V	TPT
VIS	S	IQD
LEA	Q	NPE
IMA	E	DAQ
AMA	A	VTN
VTS	H	LTE
FLA	E	DVQ
SLA	G	VLQ
…		

Frequency of patterns upstream of the cleavage site

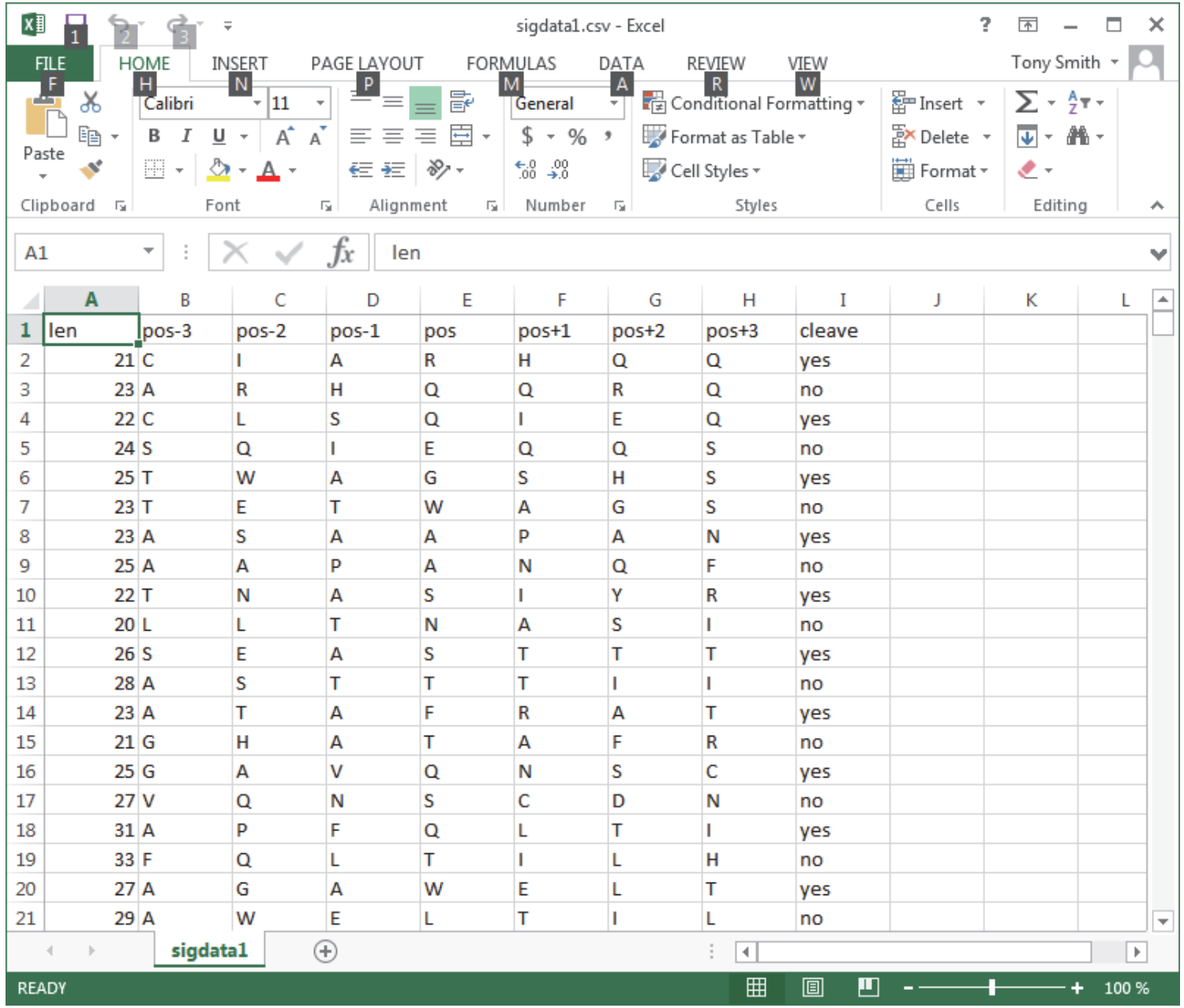
30	LAA
23	QAA
20	SAA
19	LAQ
19	HAA
17	FAA
14	NAA
13	EAA
13	AAA
11	QAE
10	TAA
10	SAS
10	LAE
9	VAA
9	LAD
8	SAL
8	RAA
8	MAA

…

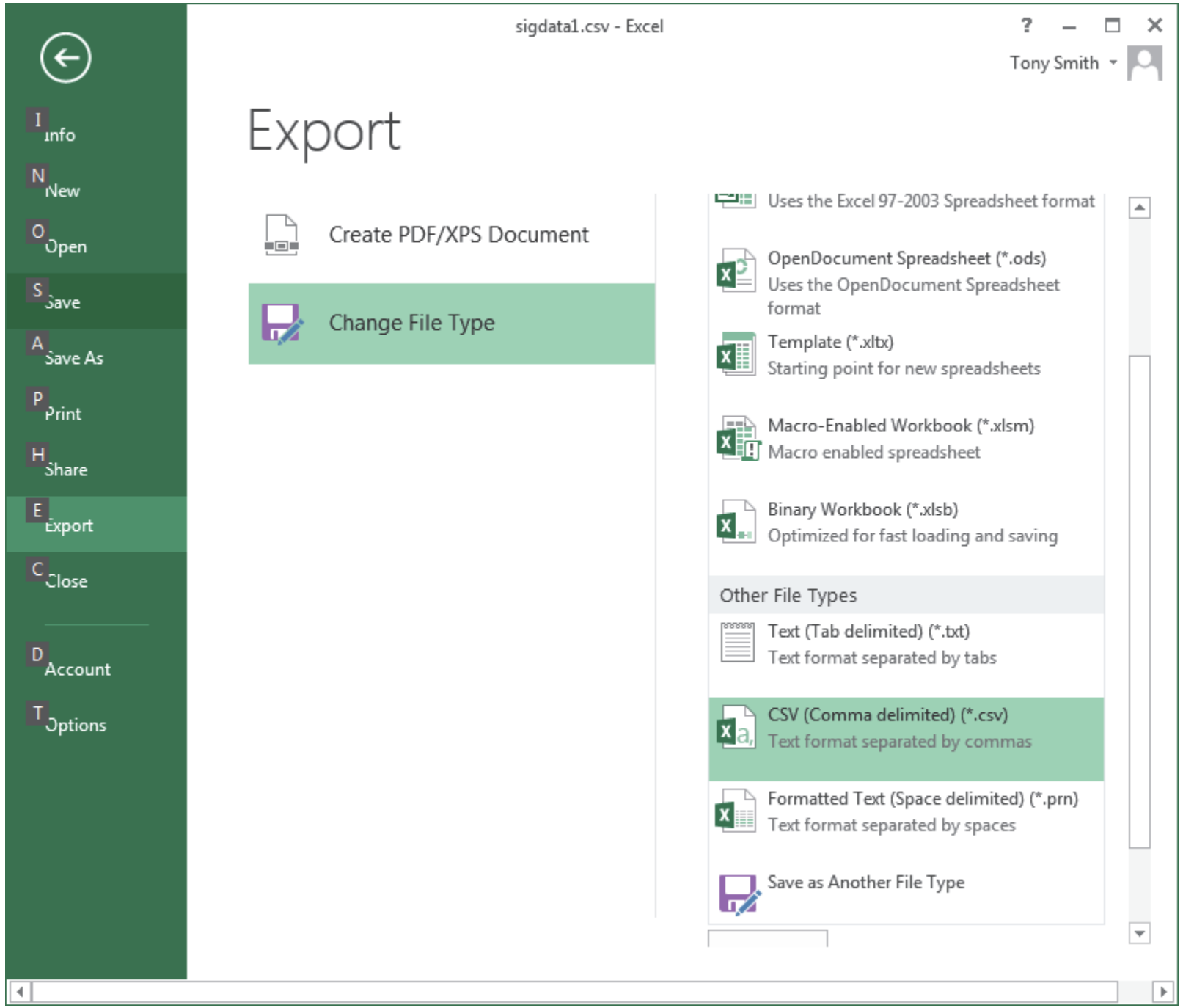
First guess at potential features

When we don't have much domain knowledge

- Position of residue being considered (i.e. length of peptide)
- Residue at each position, three either side of cleavage point
- The class (binary: cleavage or not)
- Obtain negative instances using randomly chosen residues



	A	B	C	D	E	F	G	H	I	J	K	L
1	len	pos-3	pos-2	pos-1	pos	pos+1	pos+2	pos+3	cleave			
2	21	C	I	A	R	H	Q	Q	yes			
3	23	A	R	H	Q	Q	R	Q	no			
4	22	C	L	S	Q	I	E	Q	yes			
5	24	S	Q	I	E	Q	Q	S	no			
6	25	T	W	A	G	S	H	S	yes			
7	23	T	E	T	W	A	G	S	no			
8	23	A	S	A	A	P	A	N	yes			
9	25	A	A	P	A	N	Q	F	no			
10	22	T	N	A	S	I	Y	R	yes			
11	20	L	L	T	N	A	S	I	no			
12	26	S	E	A	S	T	T	T	yes			
13	28	A	S	T	T	T	I	I	no			
14	23	A	T	A	F	R	A	T	yes			
15	21	G	H	A	T	A	F	R	no			
16	25	G	A	V	Q	N	S	C	yes			
17	27	V	Q	N	S	C	D	N	no			
18	31	A	P	F	Q	L	T	I	yes			
19	33	F	Q	L	T	I	L	H	no			
20	27	A	G	A	W	E	L	T	yes			
21	29	A	W	E	L	T	I	L	no			


sigdata1.csv - Excel
Tony Smith
Info
New
Open
Save
Save As
Print
Share
Export
Close
Account
Options
Export
Create PDF/XPS Document
Change File Type
Uses the Excel 97-2003 Spreadsheet format
OpenDocument Spreadsheet (*.ods)
Uses the OpenDocument Spreadsheet format
Template (*.xltx)
Starting point for new spreadsheets
Macro-Enabled Workbook (*.xlsm)
Macro enabled spreadsheet
Binary Workbook (*.xlsb)
Optimized for fast loading and saving
Other File Types
Text (Tab delimited) (*.txt)
Text format separated by tabs
CSV (Comma delimited) (*.csv)
Text format separated by commas
Formatted Text (Space delimited) (*.prn)
Text format separated by spaces
Save as Another File Type

Great results …. so what went wrong?

Often very easy for machine learning to find a model that works.

- Two common (related) causes for a spurious positive outcome:
  - *Sparseness of the data*: potential instance space is huge
  - *Over-fitting the data*: complex model splits data into very small subsets

Data sparseness – a form of over-fitting

- Consider two dice and one coin, and a few random outcomes …

Die 1	Die 2	Coin
3	5	H
6	4	H
2	5	T
1	1	T

6 X 6 X 2 = 72 possible random outcomes, of which we have 4

Predict the coin toss from the dice rolls: WEKA finds:

if Die1 > 2 then Coin = H else Coin = T

100% correct for our data; but additional instances should reveal no correlation.

Signal peptide: 7 residues having one of 20 values (20^7 patterns), 60 different lengths, and 2 class values = 153 billion possible instances

Overfitting

Model is so complex it practically identifies instances uniquely

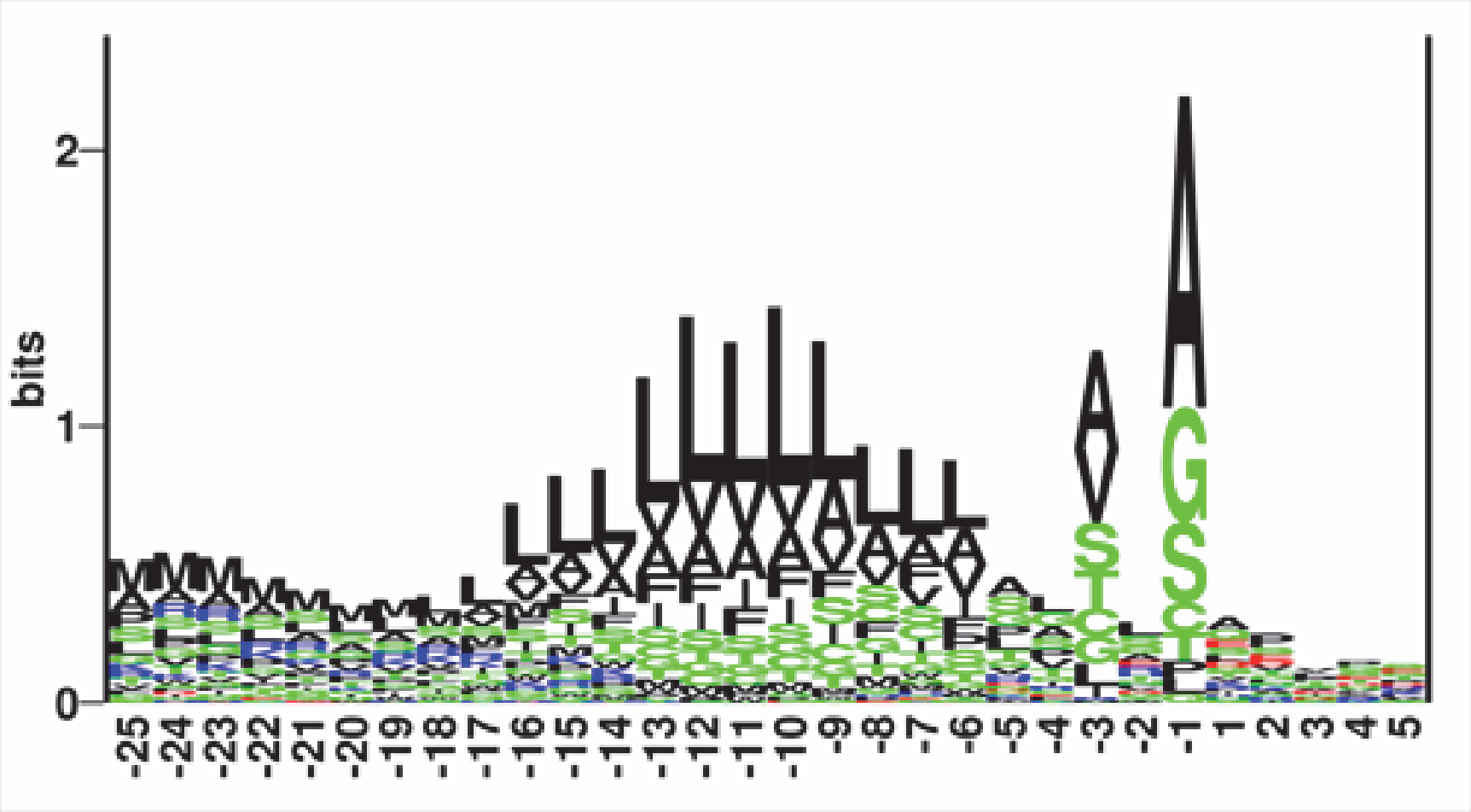
- Model splits instances into lots of very small subsets to get high predictive accuracy on the available data
- A tell-tale sign is an extremely complex model (e.g. highly branching tree)
- New data should yield poor performance
- (Actually, data sparseness is really a cause of over-fitting.)

Characteristic features – more general

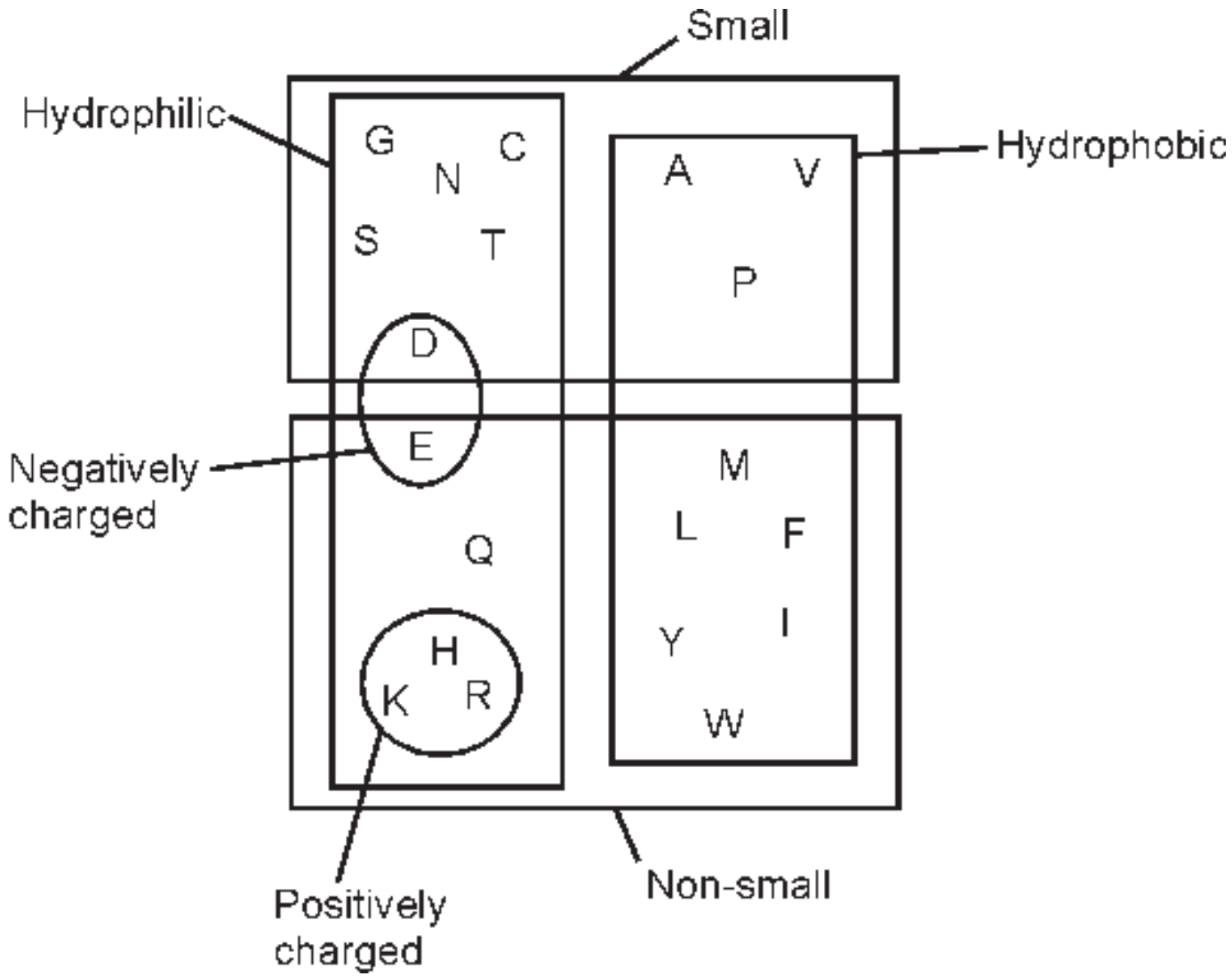
A more informed approach

- Cleavage occurs because of physical forces at molecular level
- Create features that capture physicochemical properties
- Get some domain knowledge from the experts!

Logogram


bits
2
1
0
-25 -24 -23 -22 -21 -20 -19 -18 -17 -16 -15 -14 -13 -12 -11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 1 2 3 4 5

Residue properties



Physicochemical regularities of signal peptides

MKLSKPVMTSTVASASALLVILAAAASA …

- **C-region**
  - About 4 to 6 residues, immediately upstream of cleavage site
  - Uncharged at -3 position; small side chain at -1 position
- **H-region**
  - About 8 residues, immediately upstream of C-region
  - Hydrophobic
- **N-region**
  - About 5-15 residues, immediately upstream of H-region
  - Positively charged

Characteristic features

Possible features

- Size, charge, polarity and hydrophobicity, esp. at pos-1 and pos-3
- Total hydrophobicity in approximate H-region
- Total charge, polarity and hydrophobicity in C-region
- The class (cleavage or not)

We'll just use four features:

1. position (same as length of peptide)
2. hydropathy of approximate H-region
3. side-chain size for the *-1* residue, and
4. charge of the *-3* residue.

Summary

Considerations for data mining with biological data

- Goal: predictive accuracy vs explanatory power
- Data preparation: relevant/characteristic features
- Evaluation: accuracy may be a fluke
- Collaboration: expert advice

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Advanced Data Mining with Weka

Department of Computer Science
University of Waikato
New Zealand

weka.waikato.ac.nz