THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Data Mining with Weka

Class 5 – Lesson 1

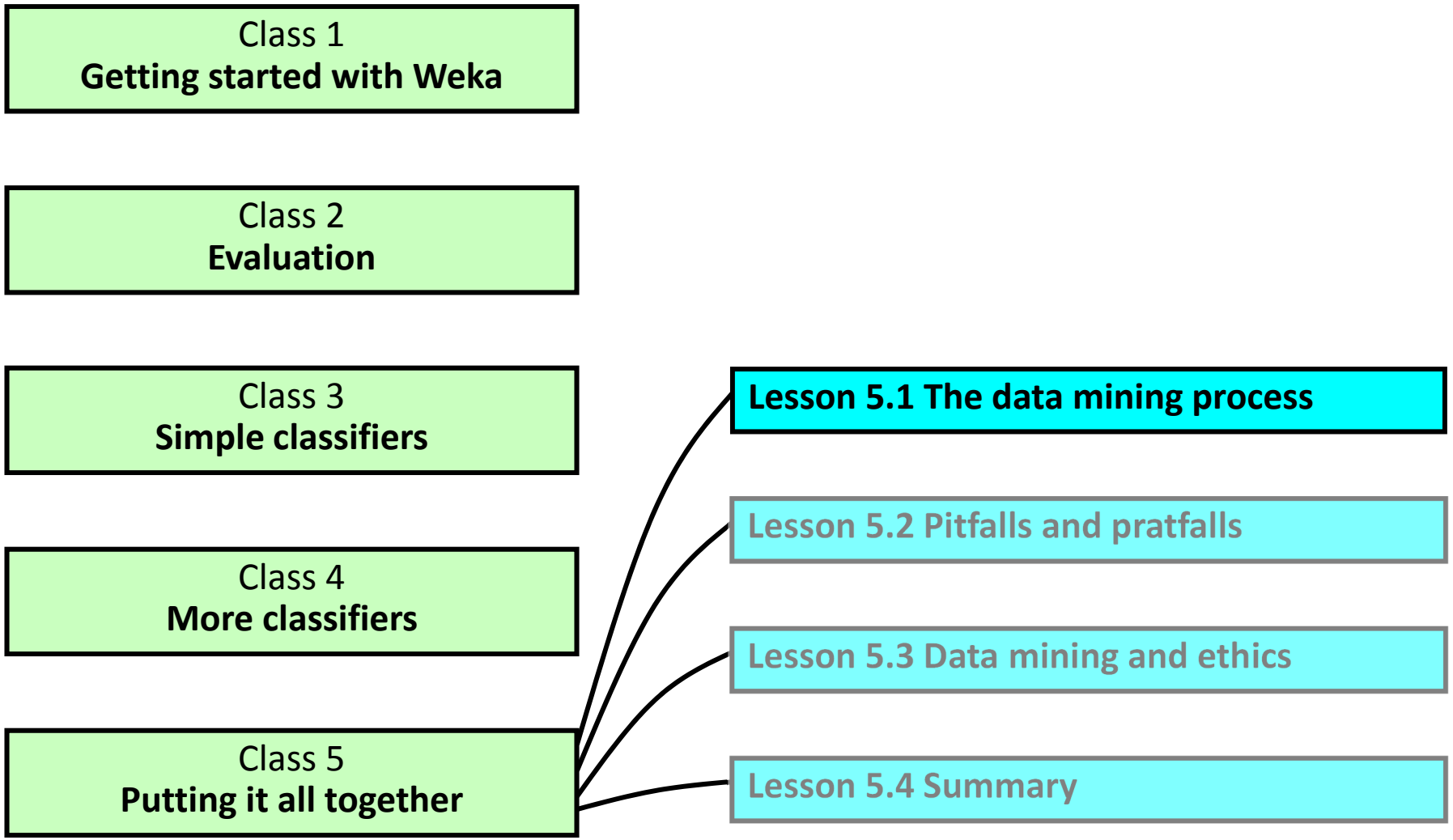
The data mining process

Ian H. Witten

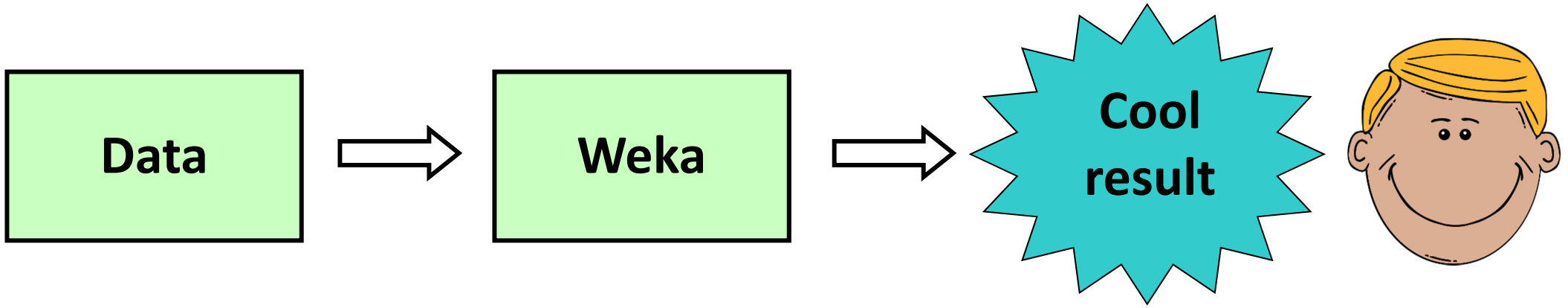
Department of Computer Science
University of Waikato
New Zealand

weka.waikato.ac.nz

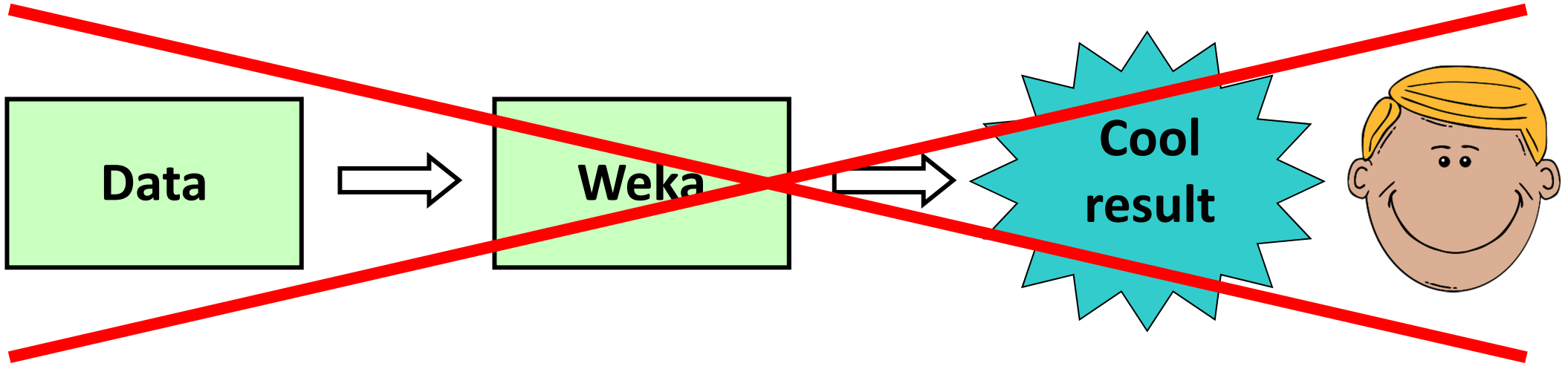
Lesson 5.1 The data mining process



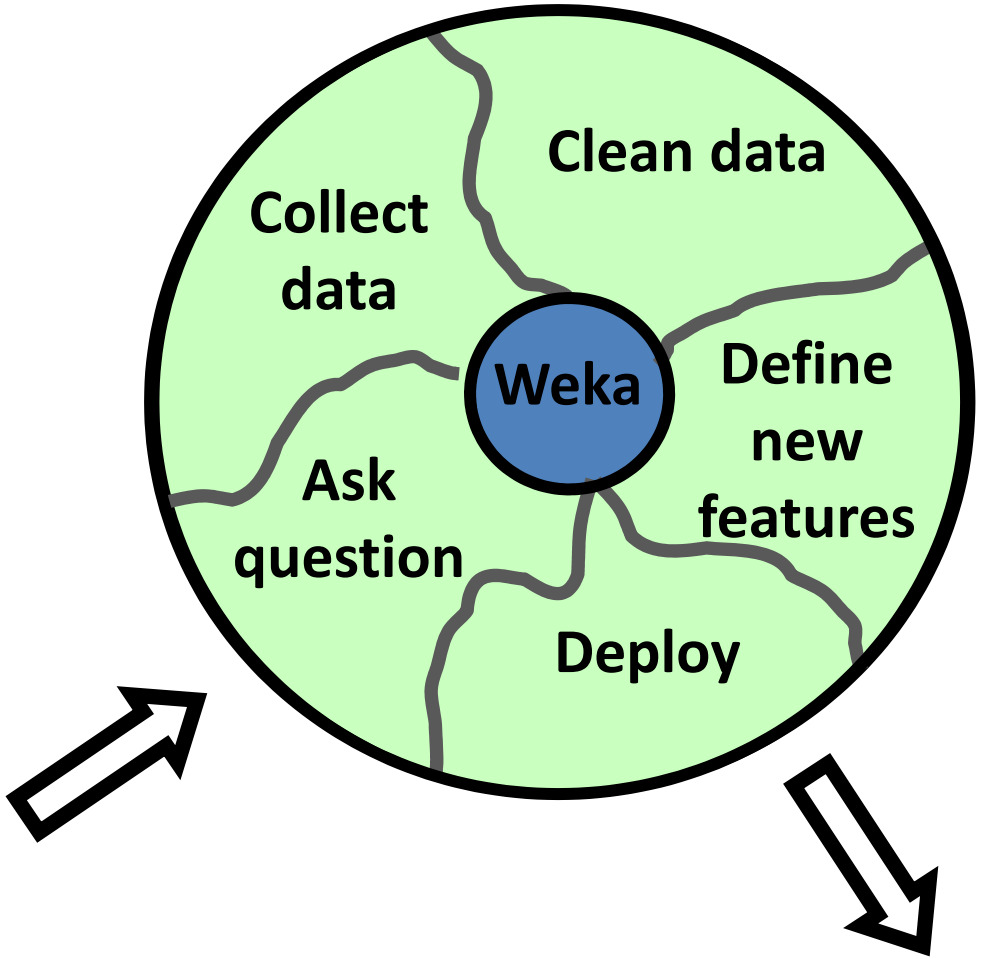
Lesson 5.1 The data mining process



Lesson 5.1 The data mining process



Lesson 5.1 The data mining process



Lesson 5.1 The data mining process

- Ask a question
  - *what do you want to know?*
  - *“tell me something cool about the data” is not enough!*
- Gather data
  - *there’s soooo much around …*
  - *… **but** … we need (expert?) classifications*
  - *more data beats a clever algorithm*
- Clean the data
  - *real data is very mucky*
- Define new features
  - *feature engineering—the key to data mining*
- Deploy the result
  - *technical implementation*
  - *convince your boss!*

Lesson 5.1 The data mining process

(Selected) filters for feature engineering

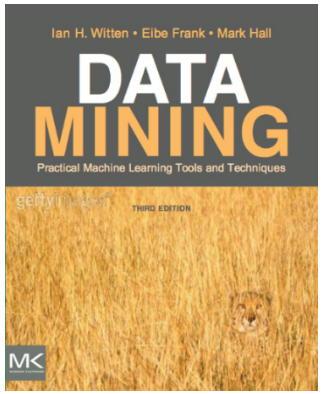
- AddExpression (MathExpression)
  - *Apply a math expression to existing attributes to create new one (or modify existing one)*
- Center (Normalize) (Standardize)
  - *Transform numeric attributes to have zero mean (or into a given numeric range) (or to have zero mean and unit variance)*
- Discretize (also supervised discretization)
  - *Discretize numeric attributes to have nominal values*
- PrincipalComponents
  - *Perform a principal components analysis/transformation of the data*
- RemoveUseless
  - *Remove attributes that do not vary at all, or vary too much*
- TimeSeriesDelta, TimeSeriesTranslate
  - *Replace attribute values with successive differences between this instance and the next*

Lesson 5.1 The data mining process

- Weka is only a small part (unfortunately) …
- … and it's the easy part

  "may all your problems be technical ones"

  – old programmer's blessing



Course text

- Section 1.3 *Fielded applications*

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Data Mining with Weka

Class 5 – Lesson 2

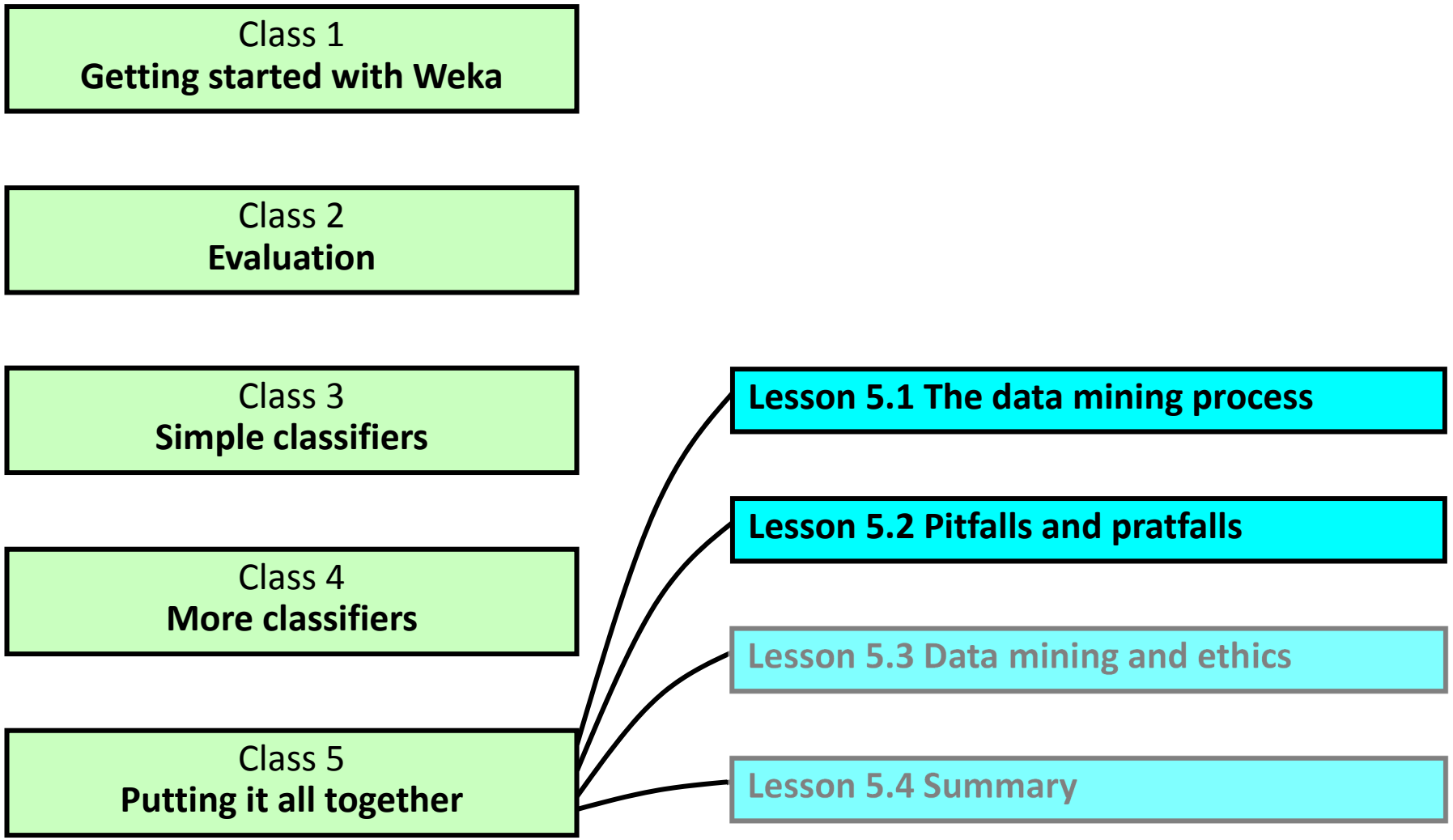
Pitfalls and pratfalls

Ian H. Witten

Department of Computer Science
University of Waikato
New Zealand

weka.waikato.ac.nz

Lesson 5.2 Pitfalls and pratfalls



Lesson 5.2 Pitfalls and pratfalls

Pitfall: A hidden or unsuspected danger or difficulty

Pratfall: A stupid and humiliating action

Lesson 5.2 Pitfalls and pratfalls

Be skeptical

- In data mining, it's very easy to cheat
  - *whether consciously or unconsciously*
- For reliable tests, use a completely fresh sample of data that has never been seen before

Overfitting has many faces

- Don't test on the training set (of course!)
- Data that has been used for development (in any way) is tainted
- Leave some evaluation data aside for the *very* end

Lesson 5.2 Pitfalls and pratfalls

Missing values

"Missing" means what …

- ❖ Unknown?
- ❖ Unrecorded?
- ❖ Irrelevant?

Should you: 1. Omit instances where the attribute value is missing?

or 2. Treat "missing" as a separate possible value?

Is there significance in the fact that a value is missing?

Most learning algorithms deal with missing values

– but they may make different assumptions about them

Lesson 5.2 Pitfalls and pratfalls

OneR and J48 deal with missing values in different ways

- Load weather-nominal.arff
- OneR gets 43%, J48 gets 50% (using 10-fold cross-validation)
- Change the outlook value to unknown on the first four no instances
- OneR gets 93%, J48 still gets 50%
- Look at OneR's rules: it uses "?" as a fourth value for outlook

Lesson 5.2 Pitfalls and pratfalls

No free lunch

- 2-class problem with 100 binary attributes
- Say you know a million instances, and their classes (training set)
- You don't know the classes of $2^{100} - 10^6$ examples!

  (that's 99.9999…% of the data set)
- How could you possibly figure them out?

In order to generalize, every learner must embody some knowledge or assumptions beyond the data it's given

A learning algorithm implicitly provides a set of assumptions

There can be no "universal" best algorithm (no free lunch)

Data mining is an experimental science

Lesson 5.2 Pitfalls and pratfalls

- Be skeptical
- Overfitting has many faces
- Missing values – different assumptions
- No “universal” best learning algorithm
- Data mining is an experimental science
- It’s very easy to be misled

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Data Mining with Weka

Class 5 – Lesson 3

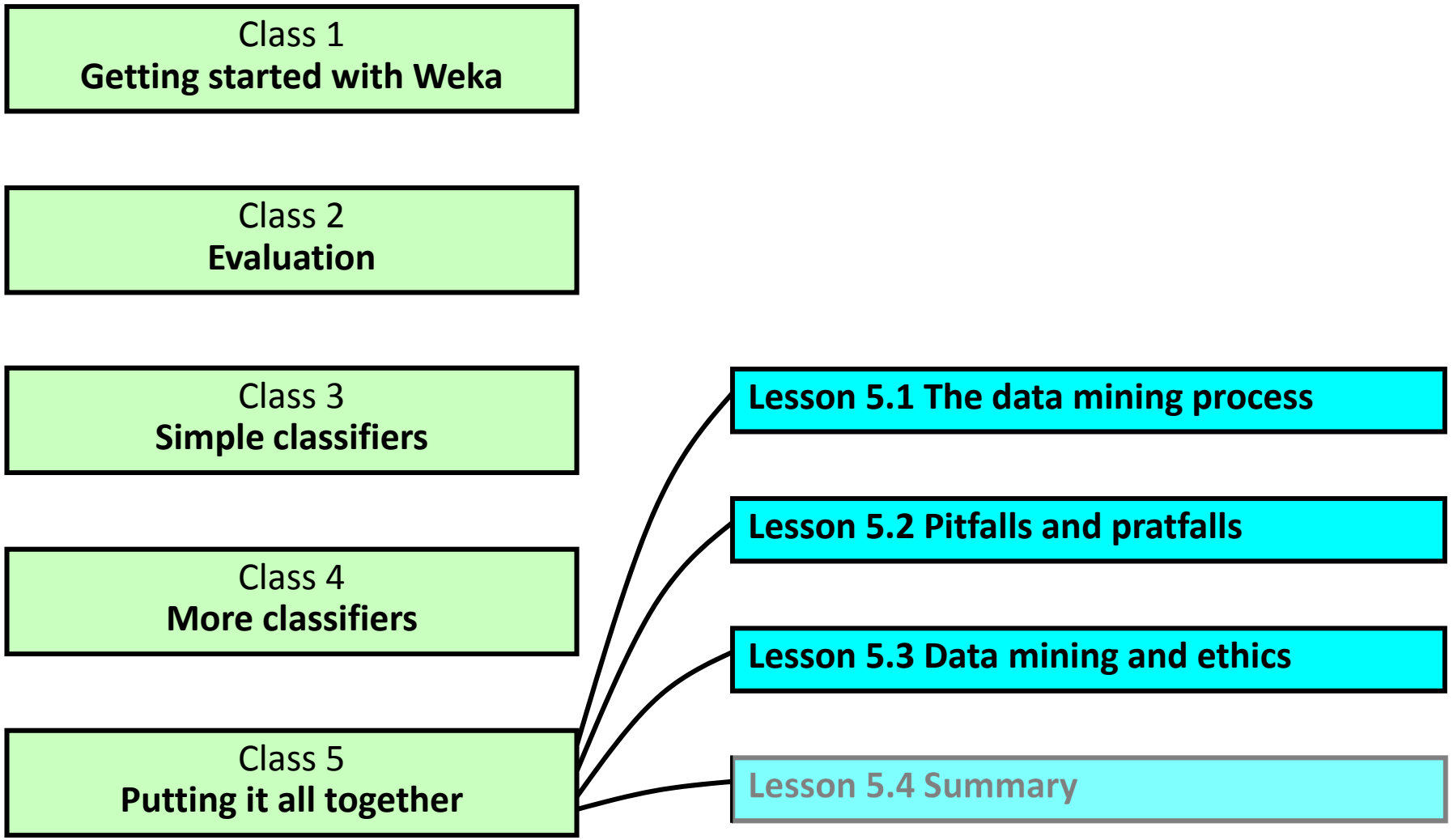
Data mining and ethics

Ian H. Witten

Department of Computer Science
University of Waikato
New Zealand

weka.waikato.ac.nz

Lesson 5.3 Data mining and ethics



Lesson 5.3 Data mining and ethics

Information privacy laws (in Europe, but not US)

- A purpose must be stated for any personal information collected
- Such information must not be disclosed to others without consent
- Records kept on individuals must be accurate and up to date
- To ensure accuracy, individuals should be able to review data about themselves
- Data must be deleted when it is no longer needed for the stated purpose
- Personal information must not be transmitted to locations where equivalent data protection cannot be assured
- Some data is too sensitive to be collected, except in extreme circumstances (e.g., sexual orientation, religion)

Lesson 5.3 Data mining and ethics

Anonymization is harder than you think

When Massachusetts released medical records summarizing every state employee's hospital record in the mid-1990s, the governor gave a public assurance that it had been anonymized by removing all identifying information such as name, address, and social security number. He was surprised to receive his own health records (which included diagnoses and prescriptions) in the mail.

Reidentification techniques. Using publicly available records:

- 50% of Americans can be identified from city, birth date, and sex
- 85% can be identified if you include the 5-digit zip code as well

Netflix movie database: 100 million records of movie ratings (1–5)

- Can identify 99% of people in the database if you know their ratings for 6 movies and approximately when they saw the movies (± one week)
- Can identify 70% if you know their ratings for 2 movies and roughly when they saw them

Lesson 5.3 Data mining and ethics

The purpose of data mining is to discriminate …

- who gets the loan
- who gets the special offer

Certain kinds of discrimination are unethical, and illegal

- racial, sexual, religious, …

Lesson 5.3 Data mining and ethics

The purpose of data mining is to discriminate …

- who gets the loan
- who gets the special offer

Certain kinds of discrimination are unethical, and illegal

- racial, sexual, religious, …

But it depends on the context

- sexual discrimination is usually illegal
- … except for doctors, who are *expected* to take gender into account

… and information that appears innocuous may not be

- ZIP code correlates with race
- membership of certain organizations correlates with gender

Lesson 5.3 Data mining and ethics

Correlation does not imply causation

As icecream sales increase, so does the rate of drownings.

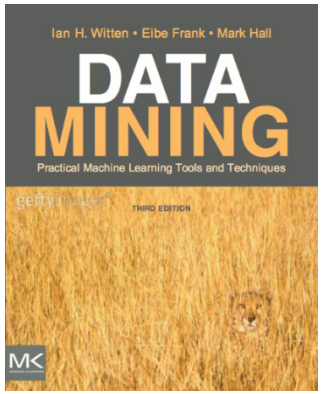
Therefore icecream consumption causes drowning???

Data mining reveals correlation, not causation

but really, we want to predict the effects of our actions

Lesson 5.3 Data mining and ethics

- Privacy of personal information
- Anonymization is harder than you think
- Reidentification from supposedly anonymized data
- Data mining and discrimination
- Correlation does not imply causation



Course text

- Section 1.6 *Data mining and ethics*

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Data Mining with Weka

Class 5 – Lesson 4

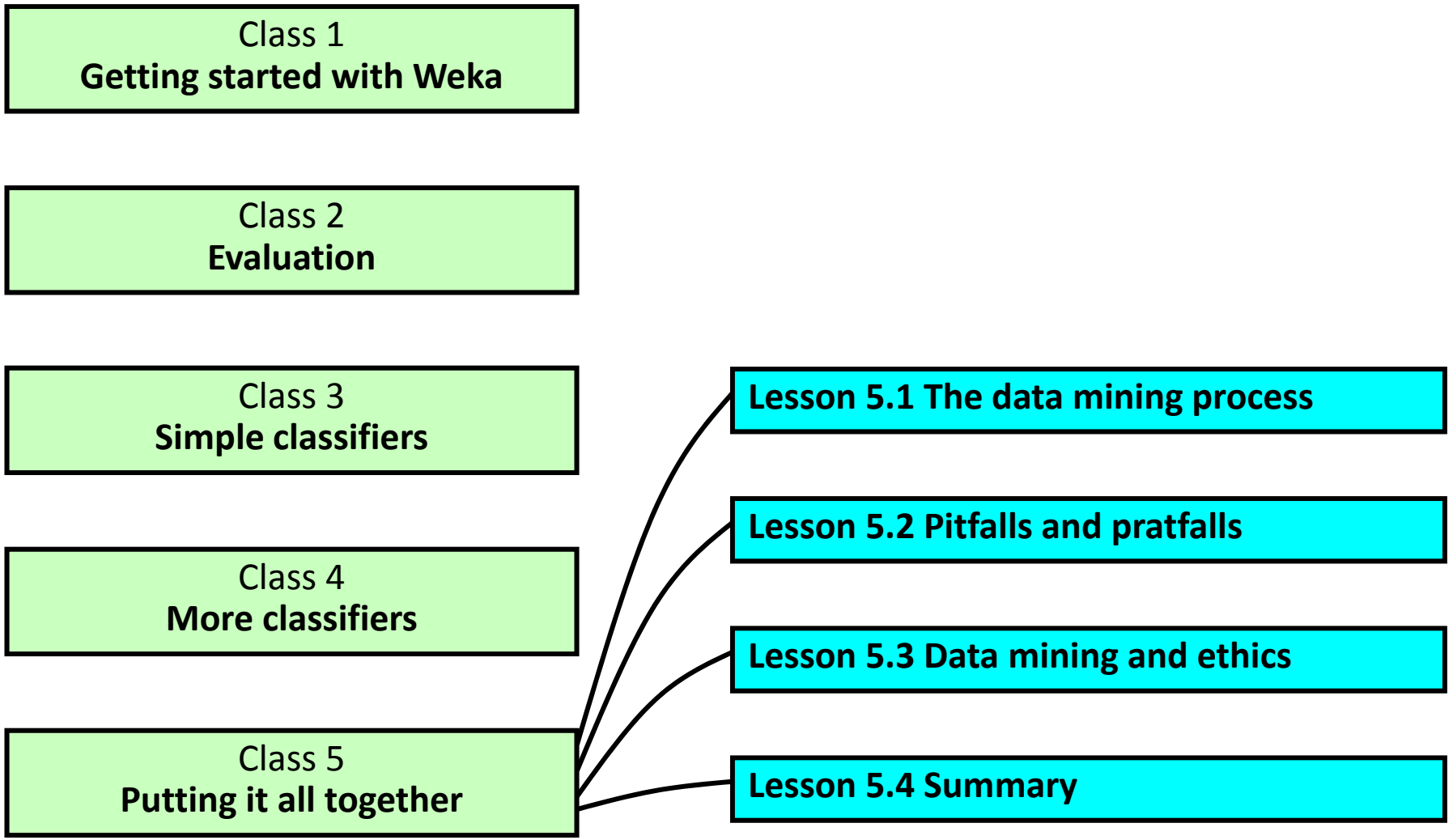
Summary

Ian H. Witten

Department of Computer Science
University of Waikato
New Zealand

weka.waikato.ac.nz

Lesson 5.4 Summary



Lesson 5.4 Summary

- **There's no magic in data mining**
  - *Instead, a huge array of alternative techniques*
- **There's no single universal "best method"**
  - *It's an experimental science!*
  - *What works best on your problem?*
- **Weka makes it easy**
  - *… maybe too easy?*
- **There are many pitfalls**
  - *You need to understand what you're doing!*
- **Focus on evaluation … and significance**
  - *Different algorithms differ in performance – but is it significant?*

Lesson 5.4 Summary

What have we missed?

- **Filtered classifiers**

  Filter training data but not test data – during cross-validation

- **Cost-sensitive evaluation and classification**

  Evaluate and minimize cost, not error rate

- **Attribute selection**

  Select a subset of attributes to use when learning

- **Clustering**

  Learn something even when there's no class value

- **Association rules**

  Find associations between attributes, when no "class" is specified

- **Text classification**

  Handling textual data as words, characters, n-grams

- **Weka Experimenter**

  Calculating means and standard deviations automatically … + more

Lesson 5.4 Summary

What have we missed?

- **Filtered classifiers**

  Filter training data but not test data – during cross-validation

- **Cost-sensitive evaluation and classification**

  Evaluate and minimize cost, not error rate

- **Attribute selection**

  Select a subset of attributes to use when learning

- **Clustering**

  Learning even when there's no class value

- **Association rules**

  Find associations between attributes, when no "class" is specified

- **Text classification**

  Handling textual data as words, characters, n-grams

- **Weka Experimenter**

  Calculating means and standard deviations automatically … + more

Advanced Data Mining with Weka??

Lesson 5.4 Summary

- **Data**
  - *recorded facts*
- **Information**
  - *patterns, or expectations, that underlie them*
- **Knowledge**
  - *the accumulation of your set of expectations*
- **Wisdom**
  - *the value attached to knowledge*

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

Data Mining with Weka

Department of Computer Science
University of Waikato
New Zealand

weka.waikato.ac.nz