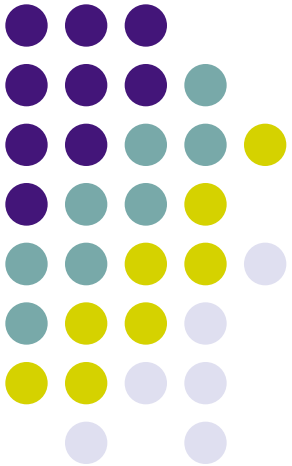
Big Data: The Science of Patterns

Dr. Lutz Hamel

Dept. of Computer Science and Statistics

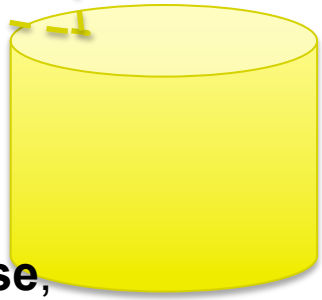
hamel@cs.uri.edu



The Blessing and the Curse: Lots of Data

Outlook	Temp	Humidity	Wind	Play
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mold	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mold	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mold	High	Strong	No

1 petabyte is
2^{50} bytes,
1024 terabytes,
or a million gigabytes



Take Yahoo Inc.'s **2-petabyte**, specially built **data warehouse**, which it uses to analyze the behavior of its half-billion Web visitors per month (2008).

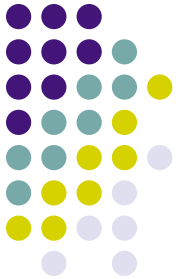
The Problem: Data vs. Information

Outlook	Temp	Humidity	Wind	Play
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mold	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mold	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mold	High	Strong	No

Data

Information

Information as Patterns



- From an AI perspective information is represented as patterns
  - patterns summarize large collections of data
  - patterns can be converted into actionable information
    - in Yahoo's case, web behavior patterns can be connected to the kinds of online ads Yahoo might show to its customers.
  - patterns come in all kinds of shapes and forms
    - graphical, rule-based, visual, numeric, *etc.*

Can You find some Patterns here?

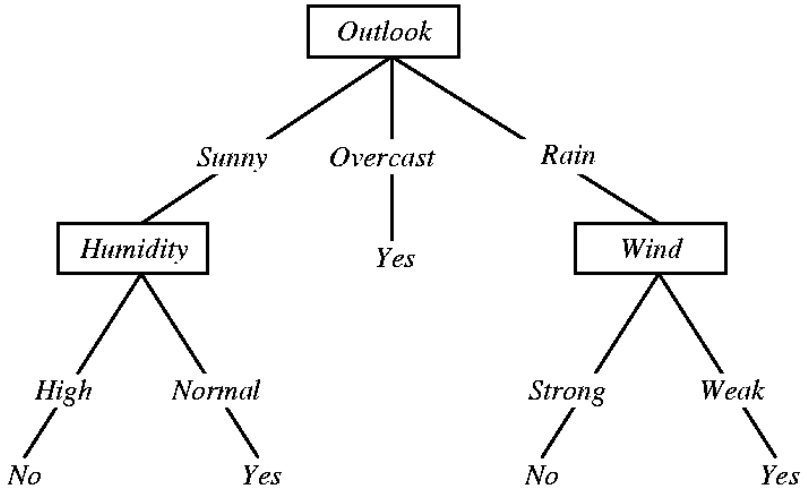
Outlook	Temp	Humidity	Wind	Play
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mold	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mold	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mold	High	Strong	No

Data

Information

A Tree based Pattern

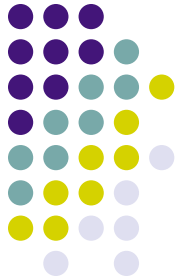
Outlook	Temp	Humidity	Wind	Play
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mold	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mold	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mold	High	Strong	No



Data

Information

Rule based Patterns



Best rules found:

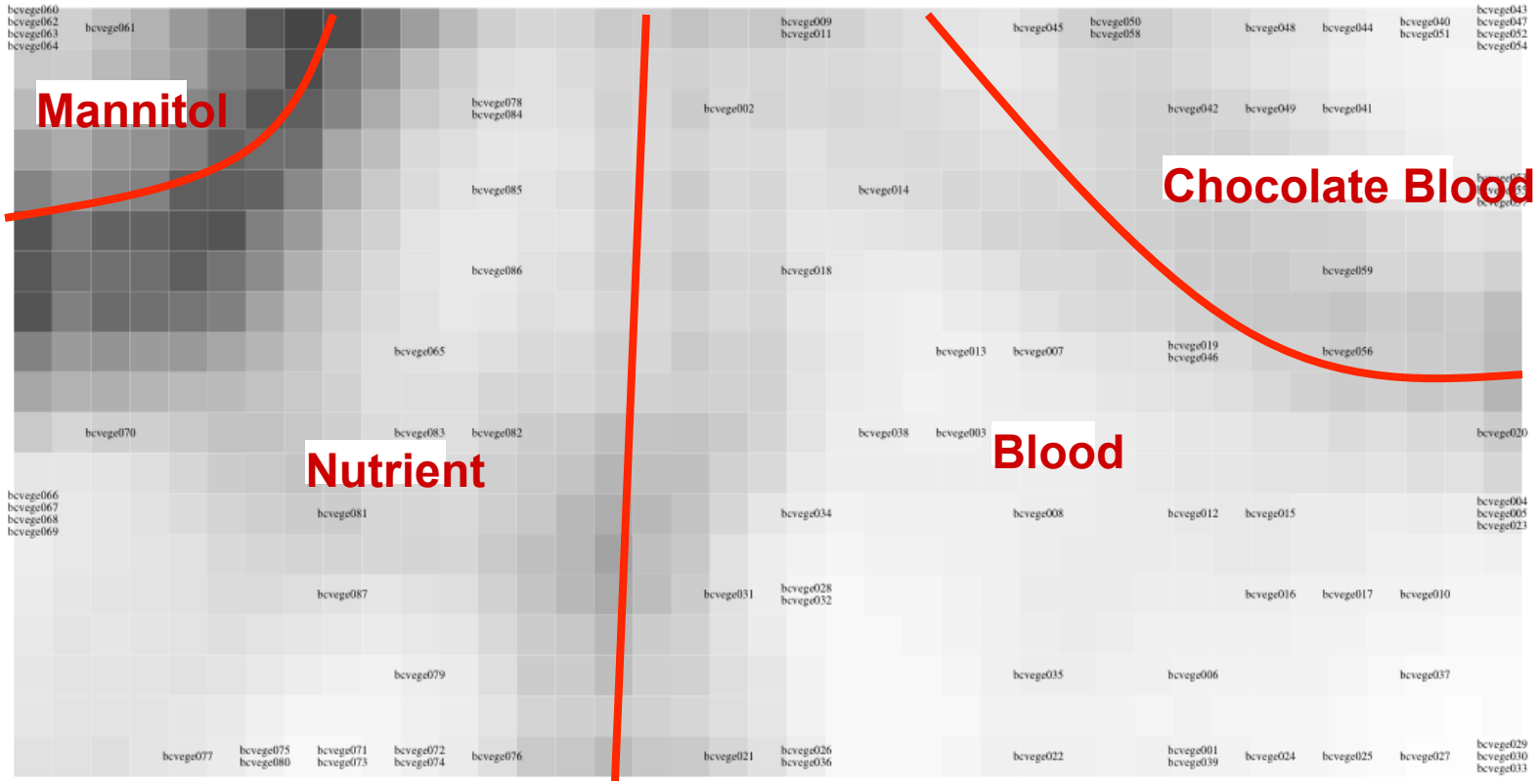
1. outlook=overcast 4 ==> play=yes 4 conf:(1)
2. temperature=cool 4 ==> humidity=normal 4 conf:(1)
3. humidity=normal windy=FALSE 4 ==> play=yes 4 conf:(1)
4. outlook=sunny play=no 3 ==> humidity=high 3 conf:(1)
5. outlook=sunny humidity=high 3 ==> play=no 3 conf:(1)
6. outlook=rainy play=yes 3 ==> windy=FALSE 3 conf:(1)
7. outlook=rainy windy=FALSE 3 ==> play=yes 3 conf:(1)
8. temperature=cool play=yes 3 ==> humidity=normal 3 conf:(1)
9. outlook=sunny temperature=hot 2 ==> humidity=high 2 conf:(1)
10. temperature=hot play=no 2 ==> outlook=sunny 2 conf:(1)

These are called “association rules.”

Visual Pattern

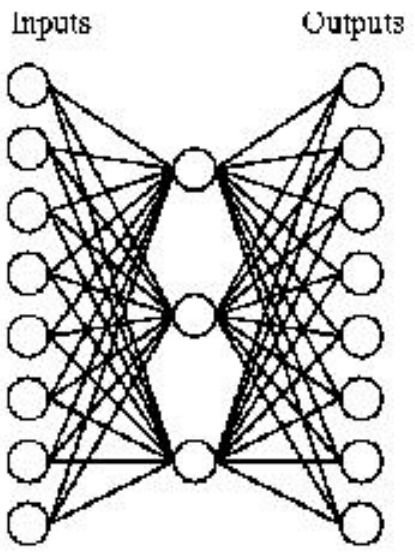
Bacterium *b-cereus* on different agars

Self-Organizing Map



"You are what you eat!"

Numeric Patterns

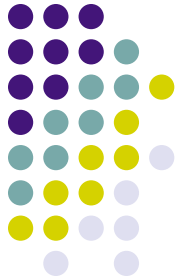


Input		Hidden Values				Output
10000000	→	.89	.04	.08	→	10000000
01000000	→	.01	.11	.88	→	01000000
00100000	→	.01	.97	.27	→	00100000
00010000	→	.99	.97	.71	→	00010000
00001000	→	.03	.05	.02	→	00001000
00000100	→	.22	.99	.99	→	00000100
00000010	→	.80	.01	.98	→	00000010
00000001	→	.60	.94	.01	→	00000001

1 0 0
0 0 1
0 1 0
1 1 1
0 0 0
0 1 1
1 0 1
1 1 0

ANNs learn numeric patterns on the weighted connections of their neurons.

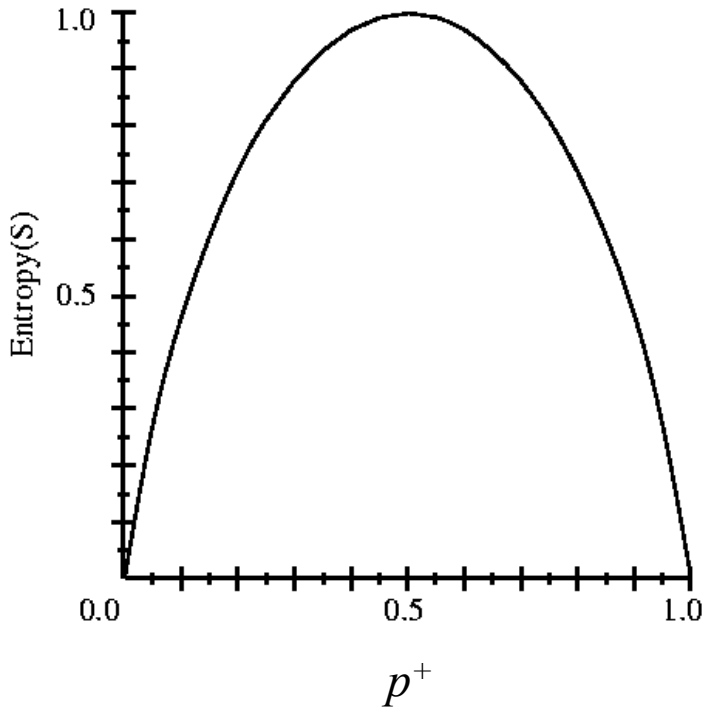
Decision Tree Learning



- “Supervised Learning” – we have a key concept (target attribute) that we want to learn, e.g. when to play tennis.
- The key idea is that the attributes and their values should be used to sort the data instances in such a way that target attribute is a non-random as possible – its *entropy* as close to 0 as possible.

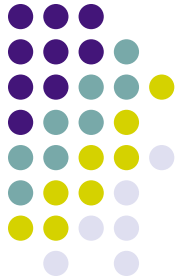
Entropy

- S is a sample of training examples
- p^+ is the proportion of positive examples in S
- p^- is the proportion of negative examples in S
- Entropy measures the impurity (randomness) of S



$$Entropy(S) \equiv - p^+ \log_2 p^+ - p^- \log_2 p^-$$

Decision Tree Learning



Recursive Algorithm

Main loop:

① Let attribute *A* be the attribute that minimizes the average entropy at the current node

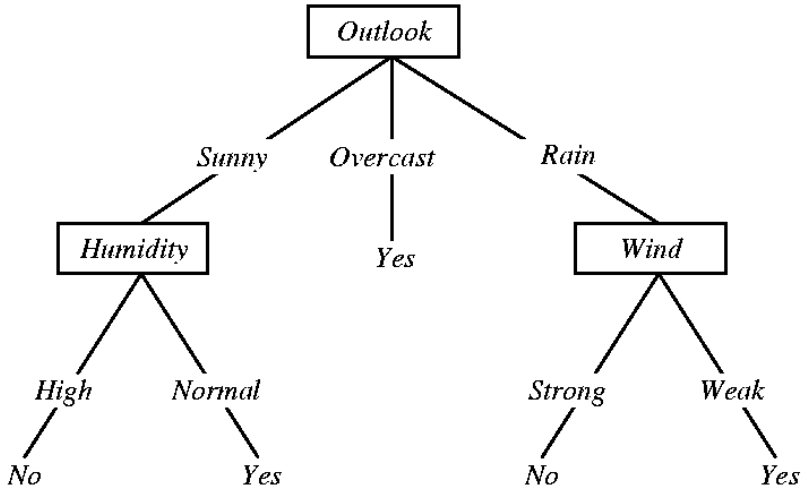
② For each attribute value of *A*, create new decendents of current node

③ Sort training examples to decendents

④ If training examples are perfectly sorted (entropy=0), then STOP, else iterate over decendents.

How does that exactly work?

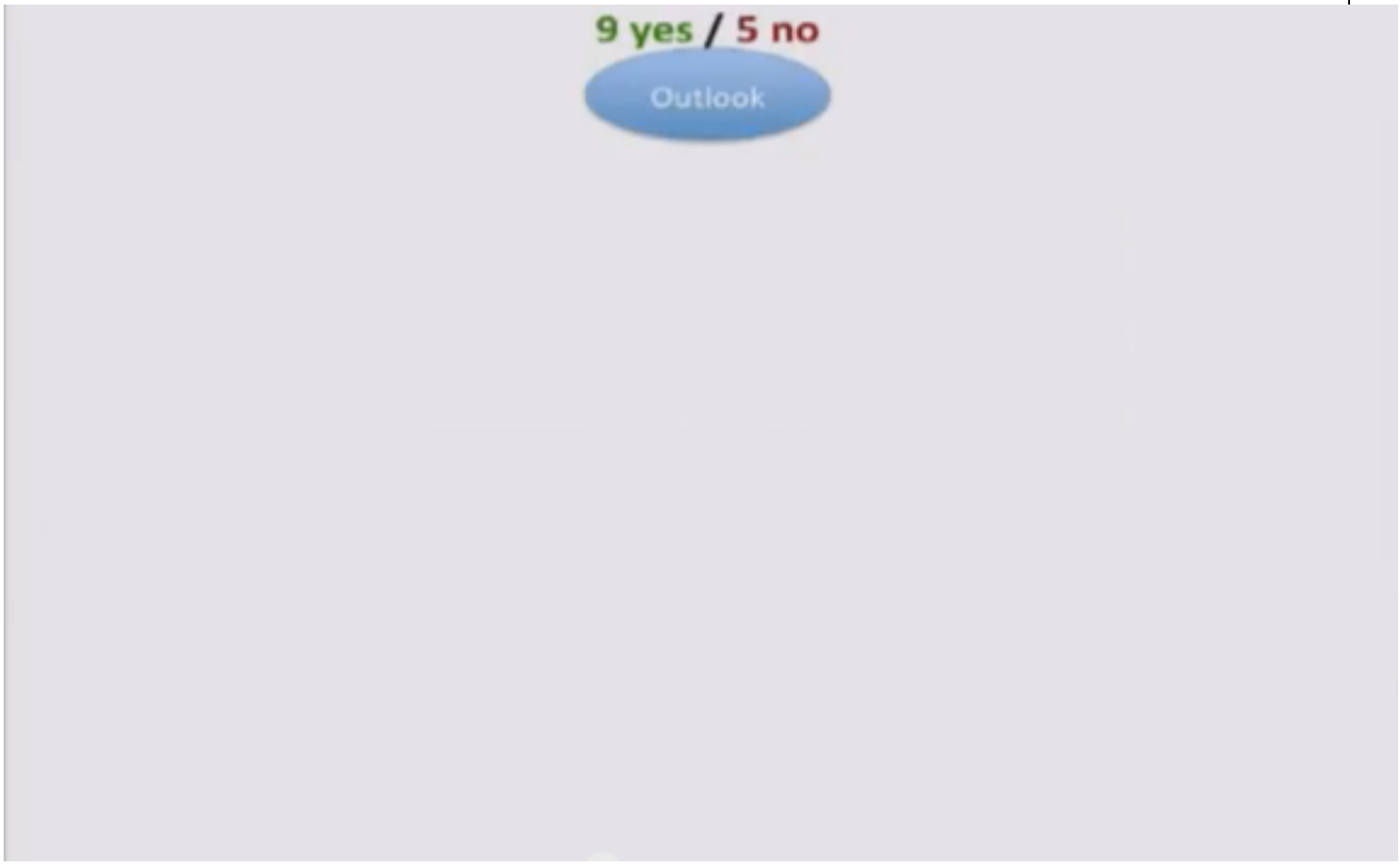
Outlook	Temp	Humidity	Wind	Play
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mold	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mold	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mold	High	Strong	No



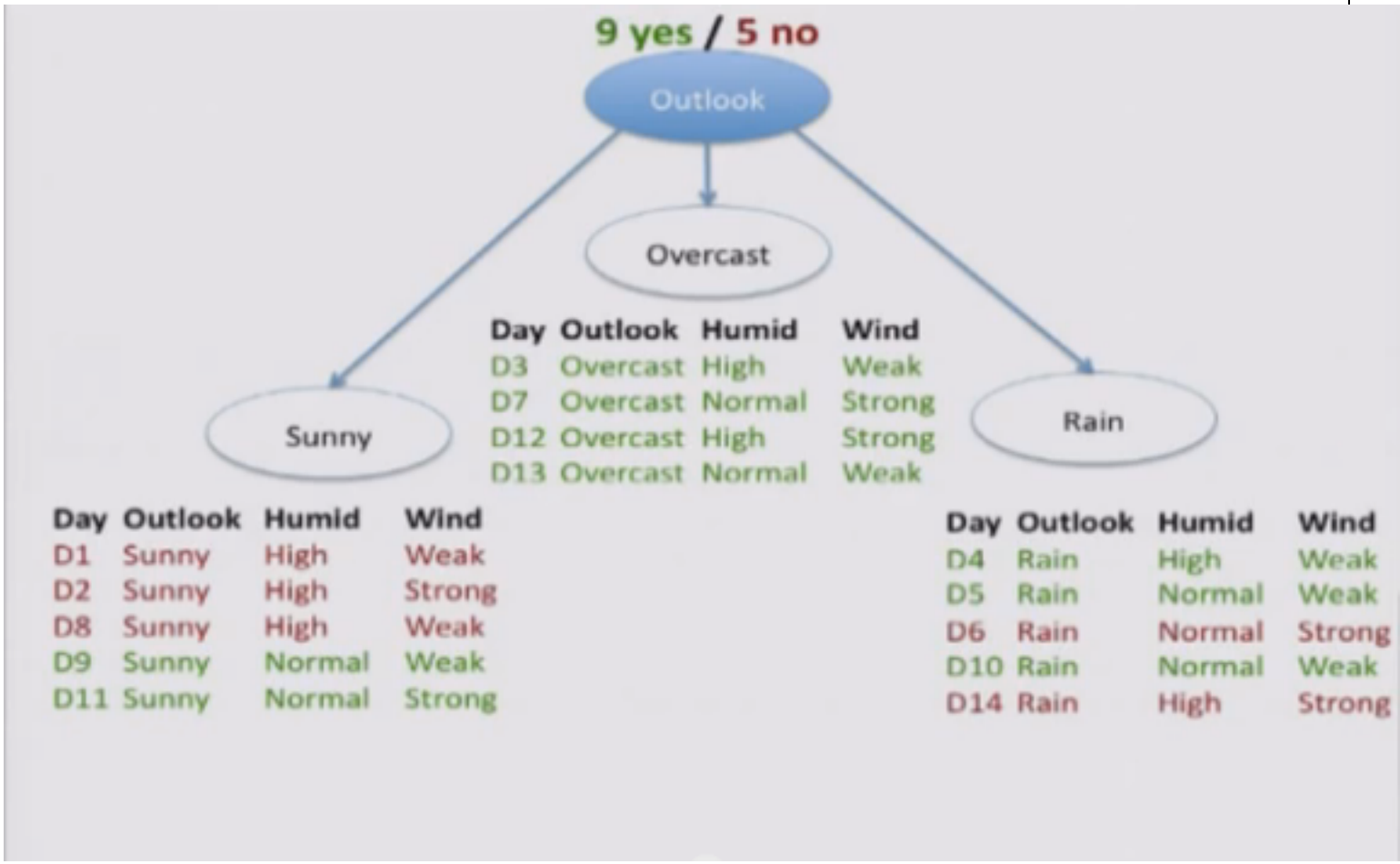
Data

Information

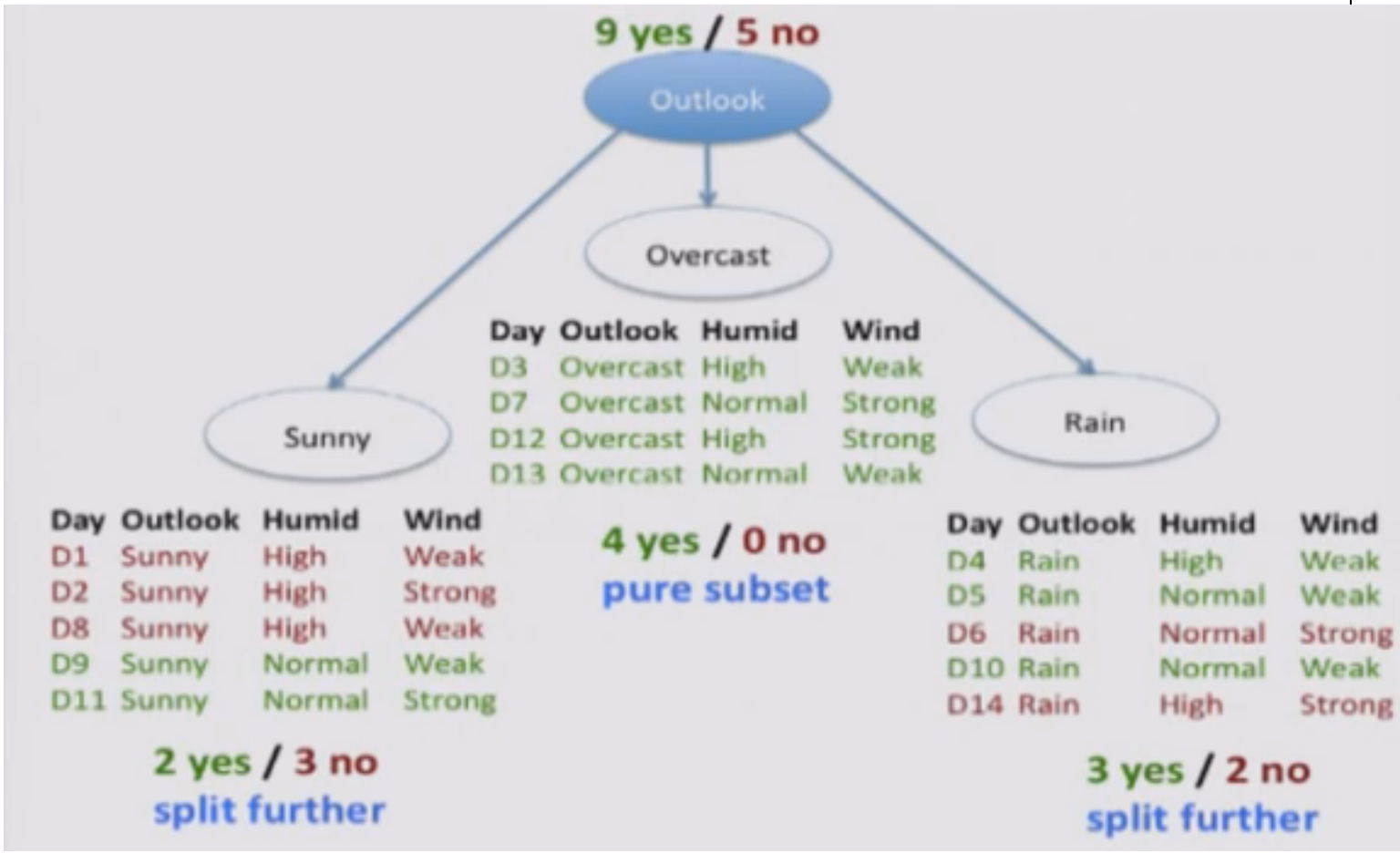
Decision Tree Learning



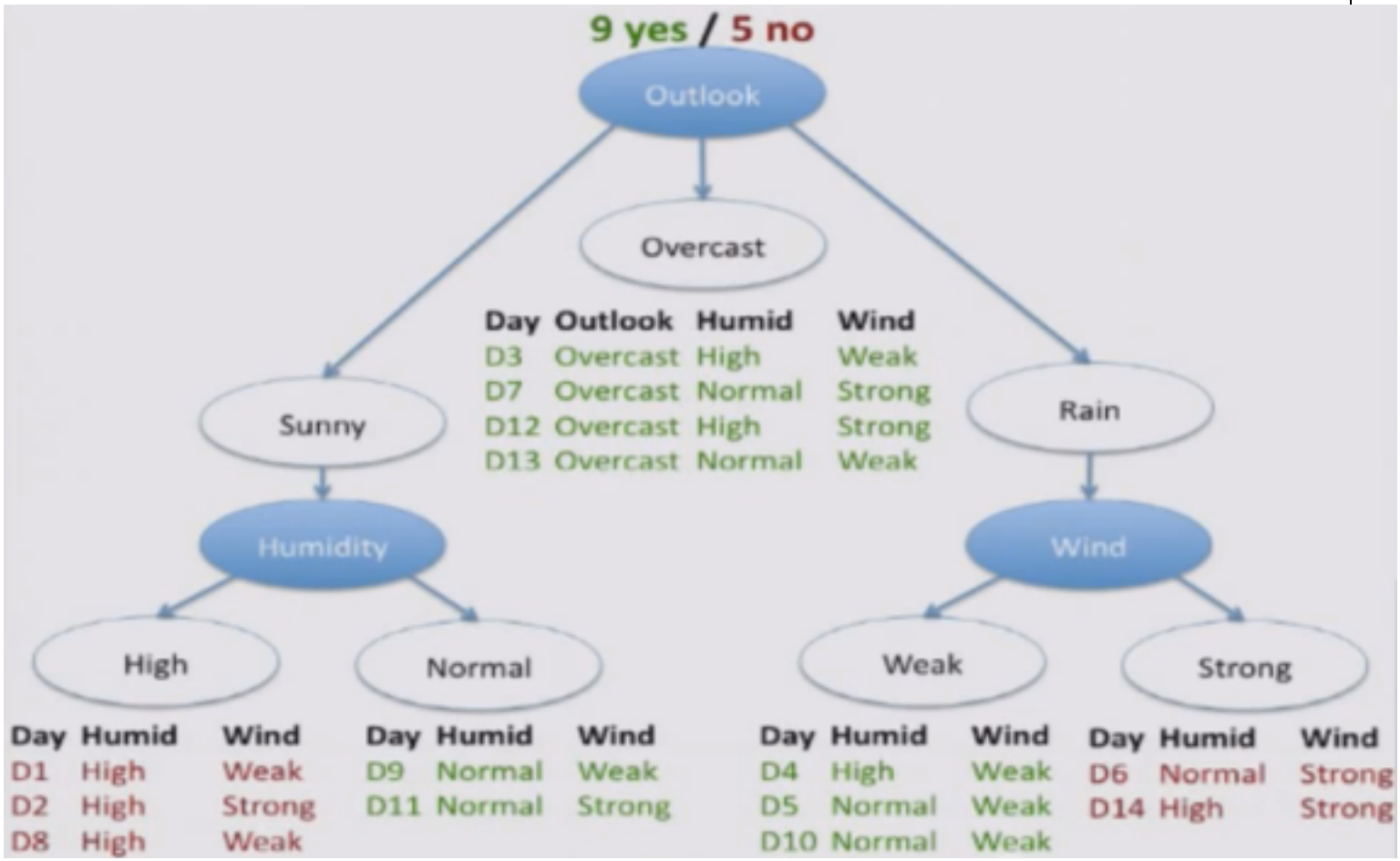
Decision Tree Learning



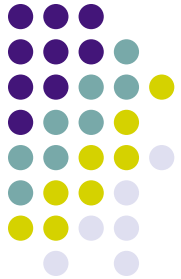
Decision Tree Learning



Decision Tree Learning

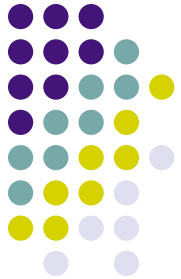


The Limits of Big Data



- There are some domains that are infinite, that is, there are an infinite number of distinct observations one could collect.
- That implies that even the biggest data warehouse will not be able to store all the observations.
- That means we are building patterns from only partial information – a problem for *inductive reasoning*.

The Limits of Big Data



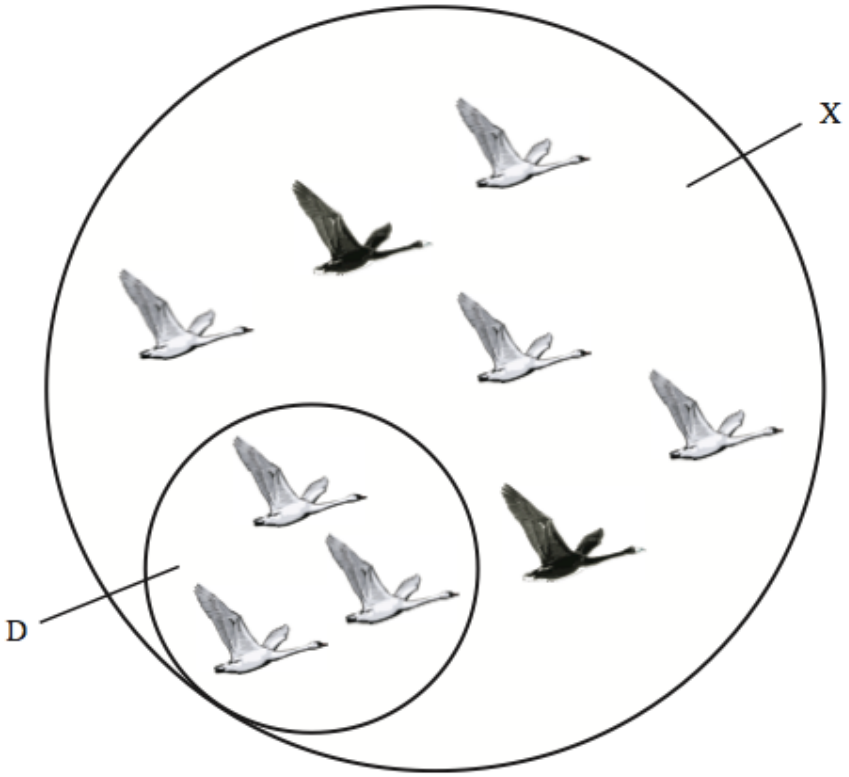
- Karl Popper (1902-1994) was a philosopher of science and first put forward the problem of inductive reasoning as the “black swan” problem.
  - “With a limited sample on swans you will most likely conclude that all swans are white. But it turns out there are black swans in Australia.”
  - Paraphrasing this: our patterns are only as good as the data which generated them.

The Limits of Big Data

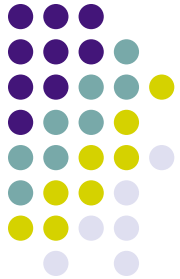
If your data warehouse only captures D of the overall data universe X, then your pattern:

"all swans are white"

will not be correct.



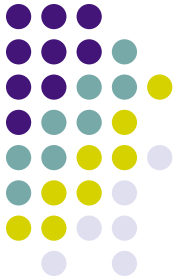
Weka Demo



- Weka is a data mining tool freely available on the web.
- It is written in Java and should probably not be used for big data projects but it is a great way to experiment with tools and techniques important to big data projects.
- http://www.cs.waikato.ac.nz/ml/weka/

- repository.seasr.org/Datasets/UCI/arff/mushroom.arff

Resources and References



- Tom Mitchell -- Machine Learning, McGraw Hill, 1997
- Karl Popper -- The Logic of Scientific Discovery, 1934 (as Logik der Forschung, English translation 1959), ISBN 0-415-27844-9
- Weka Data Mining -- http://www.cs.waikato.ac.nz/ml/weka/
- Yahoo! data warehouse -- http://www.computerworld.com/article/2535825/business-intelligence/size-matters--yahoo-claims-2-petabyte-database-is-world-s-biggest--busiest.html
- Decision tree learning -- http://youtu.be/eKD5gxPPeY0?list=PLBv09BD7ez_4temBw7vLA19p3tdQH6FYO
- Classifying Bacteria using SOM -- *Bayesian Probability Approach to Feature Significance for Infrared Spectra of Bacteria*, Lutz Hamel, Chris W. Brown, Applied Spectroscopy, Volume 66, Number 1, 2012.