

# INTRO TO DATA SCIENCE

## LECTURE 9: CLUSTERING & DECISION TREE CLASSIFIERS

YUCHEN ZHAO / DAT-14

# HOMEWORK 2 REVIEW

**I. CLUSTER ANALYSIS**

**II. K-MEANS CLUSTERING**

**III. INTERPRETING RESULTS**

**IV. LAB: K-MEANS CLUSTERING (PART 1)**

## AGENDA

---

**I. CLUSTERING SUMMARY**

**II. DECISION TREES**

**III. BUILDING DECISION TREES**

**LAB:**

**IV. HOMEWORK 2 REVIEW**

**V. K-MEANS CLUSTERING (PART 2)**

---

## K-Means Clustering

---

K-Means is the **most** popular clustering algorithm.

---

## K-Means Clustering

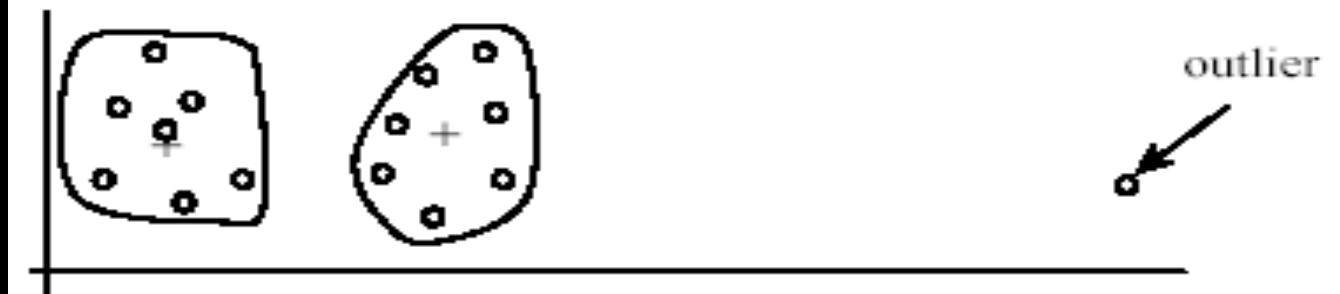
---

but sensitive to outliers...

## K-Means Clustering



(A): Undesirable clusters



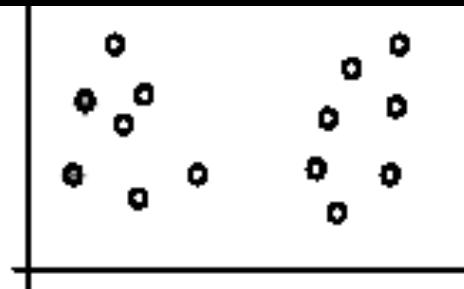
(B): Ideal clusters

## K-Means Clustering

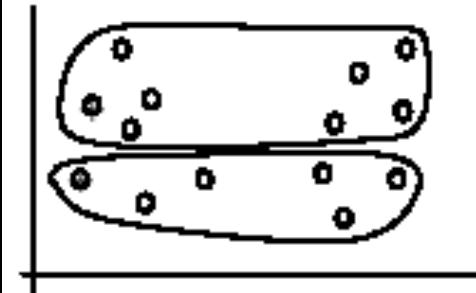
---

also sensitive to **initial seeds**

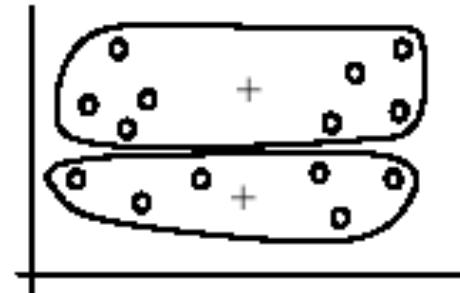
## K-Means Clustering



(A). Random selection of seeds (centroids)



(B). Iteration 1



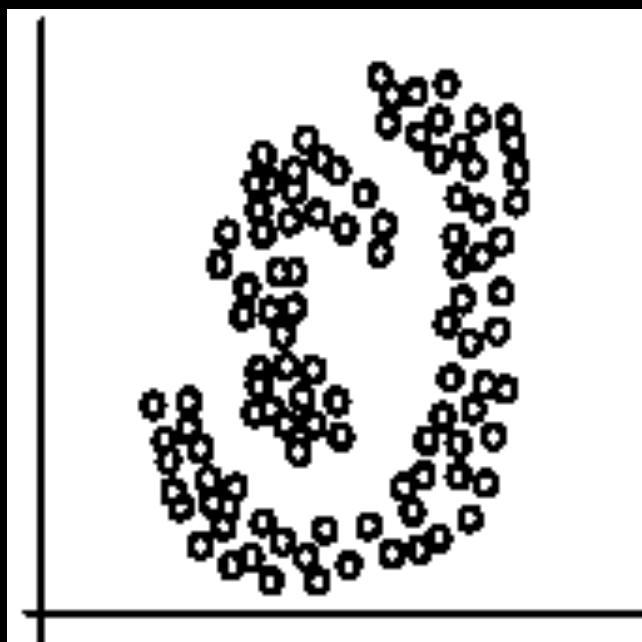
(C). Iteration 2

## K-Means Clustering

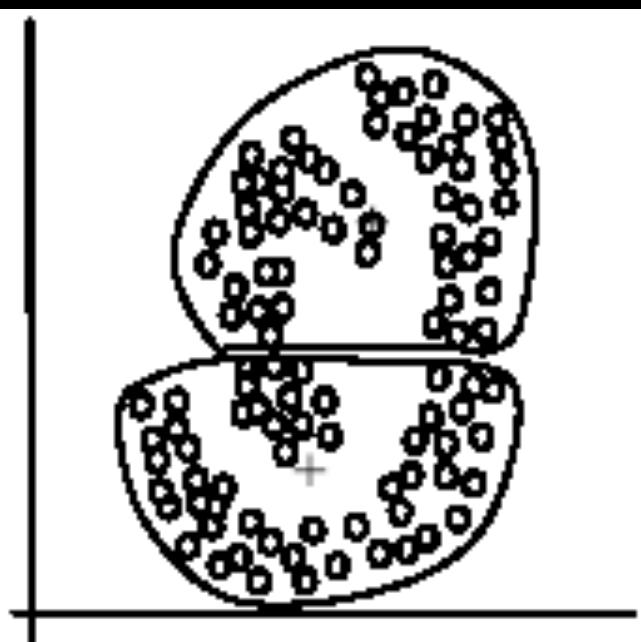
---

and not suitable for discovering  
**non-hyper-sphere clusters**

## K-Means Clustering



(A): Two natural clusters



(B):  $k$ -means clusters

## K-Means Clustering

---

No clear evidence that  
any other clustering algorithm  
performs better in general

# How to choose a clustering algorithm?

## K-Means Clustering

---

choosing the best algorithm  
is a challenge

## K-Means Clustering

---

Every algorithm has limitations and works well with certain data distributions.

## K-Means Clustering

---

In practice, It is very hard, if not impossible, to know what distribution the application data follow.

---

## K-Means Clustering

---

The common practice  
is to...

## K-Means Clustering

---

Run several algorithms using  
different distance functions and  
parameter settings

## K-Means Clustering

---

then carefully  
analyze and compare the results

## K-Means Clustering

---

Clustering is highly  
**application dependent** and  
to certain extent **subjective**  
(personal preferences).

---

## K-Means Clustering

---

Clustering in Practice:  
learning machine data

---

## K-Means Clustering

---

what is machine data?

## Clustering Application



what does the  
machine data look like?

## 1. Shell

```
NAPSHOT requires scala version: 2.9.1
[WARNING] com.sumologic.collector-interchange:collector-interchange:18.0-SNAPSHOT requires scala version: 2.9.1
[WARNING] com.sumologic.interchange:interchange:18.0-SNAPSHOT requires scala version: 2.9.1
[WARNING] com.sumologic.meta-client:meta-client:18.0-SNAPSHOT requires scala version: 2.9.1
[WARNING] org.neo4j:neo4j-cypher:1.4.1 requires scala version: 2.9.0-1
[WARNING] Multiple versions of scala libraries detected!
[INFO] includes = [**/*.scala,**/*.java,]
[INFO] excludes = []
[INFO] Nothing to compile - all classes are up to date
[INFO] [compiler:compile {execution: default}]
[INFO] Nothing to compile - all classes are up to date
[INFO] Preparing exec:java
[INFO] No goals needed for project - skipping
[INFO] [exec:java {execution: default-cli}]
[INFO]
```

---

```
[INFO] BUILD SUCCESSFUL
[INFO]
```

---

```
[INFO] Total time: 4 seconds
[INFO] Finished at: Tue May 22 08:03:06 PDT 2012
[INFO] Final Memory: 46M/95M
[INFO]
```

```
API_HOME=/Users/christian/Development/sumo/system/..../ops/assemblies/latest/api-18.0-SNAPSHOT
BILL_HOME=/Users/christian/Development/system/system/.../ops/assemblies/latest/bill-18.0-SNAPSHOT
COLLECTOR_HOME=/Users/christian/Development/sumo/system/..../ops/assemblies/latest/collector-18.0-SNAPSHOT
CONFIG_HOME=/Users/christian/Development/sumo/system/..../ops/assemblies/latest/config-18.0-SNAPSHOT
KATTA_SUMO_HOME=/Users/christian/Devlopment/sumo/system/.../ops/assemblies/latest/katta-sumo-18.0-SNAPSHOT
META_HOME=/Users/christian/Development/sumo/system/.../ops/assemblies/latest/meta-18.0-SNAPSHOT
NOVA_HOME=/Users/christian/Development/system/system/.../ops/assemblies/latest/nova-18.0-SNAPSHOT
OPS_HOME=/Users/christian/Development/sumo/system/..../ops/assemblies/latest/ops-18.0-SNAPSHOT
RAW_HOME=/Users/christian/Development/system/system/..../ops/assemblies/latest/raw-18.0-SNAPSHOT
RECEIVER_HOME=/Users/christian/Development/system/system/.../ops/assemblies/latest/receiver-18.0-SNAPSHOT
SEARCH_HOME=/Users/christian/Development/sumo/system/.../ops/assemblies/latest/search-18.0-SNAPSHOT
SERVICE_HOME=/Users/christian/Development/sumo/system/.../ops/assemblies/latest/service-18.0-SNAPSHOT
STREAM_HOME=/Users/christian/Development/sumo/system/.../ops/assemblies/latest/stream-18.0-SNAPSHOT
idoru: christian $
```

```
[GC 71949K->49884K(98816K), 0.0048185 secs]
[GC 71484K->48844K(96192K), 0.0036761 secs]
[GC 71244K->48548K(98890K), 0.0041013 secs]
[GC 70436K->48713K(98944K), 0.0042911 secs]
[GC 78601K->48678K(99944K), 0.0054358 secs]
[GC 78559K->48681K(95488K), 0.00550188 secs]
[GC 70569K->48689K(99844K), 0.0041172 secs]
[GC 70641K->48697K(98880K), 0.00539677 secs]
[GC 70649K->48738K(99008K), 0.0148833 secs]
[GC 70682K->48541K(99944K), 0.0106239 secs]
[GC 70685K->48652K(99072K), 0.0076121 secs]
[GC 71116K->48581K(98944K), 0.0073889 secs]
```

```
2012-05-22 08:44:37,967 -0700 INFO [module=RECEIVER] [logger=util.scala.http.GlobalHttpTrackerList$] [thread=MTP-MessagePipeline-8] [auth=Collector:localId=0000000000000322:0000000000000005C:false] [remote_ip=127.0.0.1] [web_session=uTM1xkz...]
Recovery: unhealthy: Receiver-blockProduction Check: In since ping > healthy
2012-05-22 08:44:37,970 -0700 INFO [module=RECEIVER] [logger=util.scala.http.GlobalHttpTrackerList$] [thread=MTP-MessagePipeline-6] [auth=Collector:localId=00000000000322:00000000000005C:false] [remote_ip=127.0.0.1] [web_session=uTM1xkz...]
Pile for customer: '00000000000005C', ID: '8000000327', block: '8000000000000A', msg count: '53', size: '6807', collector: '0000000000000322'
-
```

```
ds-1421773886]] [auth=User:daddy@demo.com:0000000000000000000005C:false] [remote_ip=0:0:0:0:0:10] [web_session=3bdquom...] [session_id=396F9C8405CB8099] [customer=00000000000000000005C] [call=InboundRawProtocol.getMessages] [session_path=067722E6EF2B66C] getMessages(sessionId=396F9C8405CB8099, requestId=0FF940C103786C74, blockId=00000000000000000002
2012-05-22 08:44:38,001 -0700 INFO [module=RAW] [logger=scala.raw.MessagesProtocolHandler] [thread=Thread-31] [auth=HornetQ-client-global-threads-1421773886] [auth=Customer:00000000000000000005C:00000000000000000005C:00000000000000000005C:false] [customer=00000000000000000005C] [call=messageProtocol.publishMessageBlock] Block for customer: '00000000000000000009', msg count: '10020', size: '1253828'
-
```

```
er#MetaDataLookupCallback] [thread=MTP-SearchQueryHandler-5] [auth=User:daddy@demo.com:0000000000000000000005C:false] [remote_ip=0:0:0:0:0:10] [web_session=3bdquom...] [session_id=5E61589566768730] [customer=00000000000000000005C] [call=InboundSearchProtocol.startSearch] [session_path=067722E6EF2B66C] Getting 310 hits from 2 indices [92-13377013065258-675443514294204376, 92-13377013849636-3688725643144198513] for session 5E61589566768730
2012-05-22 08:44:38,316 -0700 INFO [module=SEARCH] [logger=scala.meta_client.protocol.message.IndexLookupProtocol$Stream] [thread=Thread-26] [group:HornetQ-client-global-threads-15719147293] [auth=User:daddy@demo.com:00000000000000000005C:00000000000000000005C:false] [customer=00000000000000000005C] [call=OutboundMetaQueryProtocol.IndicesPage] [session_id=067722E6EF2B66C/5E61589566768730] Received a batch of indices,_
-
```

```
456528779558801], "node": "", "nodes": [], "blockIds": [], "startReceiptTime": -1, "endReceiptTime": -1}
2012-05-22 08:44:38,365 -0700 INFO [module=META] [logger=scala.meta.CassandraIndexPersistor] [thread=Thread-24] [group:HornetQ-client-global-threads-1253374690] Added index metadata for customer: '00000000000000000005C', in doc id: E218223966E000, block {"id": 3221460265933791453, "customerId": 92, "path": "search.index.s3:lucene-index/92-1337701472826-627678167736475201", "partitionId": "deprecated", "startTimestamp": 1336899476000, "endTimestamp": 1336899419800, "count": 28, "name": "92-1337701472826-627678167736745201"}, "node": "", "nodes": [], "blockIds": [], "startReceiptTime": -1, "endReceiptTime": -1}
-
```

```
2012-05-22 08:44:35,713 [thread=4] [group:HornetQ-client-global-threads-69248303] INFO com.sumologic.scala.collector.CommonsHTTPSender Publishing message piles: '18', messages: '1827', bytes: '335958', encoded: '335944', threshold: 'false', compressed: '24505'
2012-05-22 08:44:36,816 [thread=2] [group:HornetQ-client-global-threads-69248303] INFO com.sumologic.scala.collector.CommonsHTTPSender Publishing message piles: '22', messages: '2223', bytes: '428435', encoded: '428421', threshold: 'false', compressed: '38468'
2012-05-22 08:44:37,776 [thread=4] [group:HornetQ-client-global-threads-69248303] INFO com.sumologic.scala.collector.CommonsHTTPSender Publishing message piles: '20', messages: '2025', bytes: '355193', encoded: '355179', threshold: 'false', compressed: '31755'
-
```

```
saging.DefaultHornetQConsumerTracker] [thread=Thread-19] [group:HornetQ-client-global-threads-1707803799] Dropping message 103421
2012-05-22 08:44:35,380 -0700 WARN [module=CONFIG] [logger=avrox.scala.messaging.DefaultHornetQConsumerTracker] [thread=Thread-19] [group:HornetQ-client-global-threads-1707803799] After depletion 1 messages left in queue notification-input-queue, customerId=0000000000000005C, sessionId=68784677183F51C5.
2012-05-22 08:44:35,385 -0700 INFO [module=CONFIG] [logger=scala.interceptor.session.server.ServerQueueSession] [thread=Thread-19] [group:HornetQ-client-global-threads-1707803799] Stopped queue session with session ID: '68784677183F51C5', organization: '0000000000000005C', ancestors: '' in ms: '108'
-
```

```
2012-05-22 08:37:28,483 -0700 INFO [module=OPS] [logger=ops.scala.util.ThirdPartyRegistrar$] [thread=main] New services:
    search_jmx (192.168.242.139)
    stream_jmx (192.168.242.139)
2012-05-22 08:37:43,499 -0700 INFO [module=OPS] [logger=ops.scala.util.ThirdPartyRegistrar$] [thread=main] New services:
    service_http (192.168.242.139)
    service_jmx (192.168.242.139)
2012-05-22 08:38:43,554 -0700 INFO [module=OPS] [logger=ops.scala.util.ThirdPartyRegistrar$] [thread=main] New services:
    collector_jmx (192.168.242.139)
-
```

```
ch.scala.katta.KattaIndexStore] [thread=MTP-IndexDeployer-1] Deploying index 92-1337701472826-627678167736745201
2012-05-22 08:44:38,033 -0700 INFO [module=SEARCH] [logger=search.ch.katta.DefaultIndexDeployer] [thread=MTP-IndexDeployer-1] Finished deploying index, name=92-1337701472826-627678167736745201
2012-05-22 08:44:38,333 -0700 INFO [module=SEARCH] [logger=search.ch.katta.KattaIndexStore] [thread=MTP-IndexDeployer-1] Deploying index 92-1337701472830-8554356854656354614
2012-05-22 08:44:38,338 -0700 INFO [module=SEARCH] [logger=search.ch.katta.DefaultIndexDeployer] [thread=MTP-IndexDeployer-1] Finished deploying index, name=92-1337701472830-8554356854656354614
-
```

```
[remote_ip=0:0:0:0:0:10] [web_session=3bdquom...] [session_id=5E61589566768730] [customer=00000000000000000005C] [call=OutboundSearchProtocol.messages]
[session_path=067722E6EF2B66C] Callback for session 067722E6EF2B66C received 106 tuples
2012-05-22 08:44:38,312 -0700 INFO [module=STREAM] [logger=stream.scala.delegates.SearchDelegate] [thread=Thread-18] [group:HornetQ-client-global-threads-1088549637] [auth>User:daddy@demo.com:00000000000000000005C:false] [remote_ip=0:0:0:0:0:10] [web_session=3bdquom...] [session_id=5E61589566768730] [customer=00000000000000000005C] [call=OutboundSearchProtocol.messages]
[session_path=067722E6EF2B66C] Acknowledging 106 tuples received for session 5E61589566768730
-
```

# Log Example

```
2012-05-22 18:47:26,807 -0700 INFO [tId=long-frontend-1] [module=RECEIVER]  
[logger=scala.receiver.MessageBlocker] [thread=MTP-MessagePilePipeline-3]  
[auth=Collector:prod-cass-raw-8:000000000000483D:0000000000000005:false]  
[remote_ip=184.73.74.54] [web_session=MepMG8CS...] Pile for customer:  
'0000000000000005', ID: '800000006407637B', block: '80000000004C9A11', msg  
count: '1', size: '264', collector: '000000000000483D'
```

- Timestamp with time zone!

# Log Example

- Timestamp with time zone!
  - Log level

# Log Example

```
2012-05-22 18:47:26,807 -0700 INFO  [hostId=long-frontend-1] [module=RECEIVER]
[logger=scala.receiver.MessageBlocker] [thread=MTP-MessagePilePipeline-3]
[auth=Collector:prod-cass-raw-8:000000000000483D:0000000000000005:false]
[remote_ip=184.73.74.54] [web_session=MepMG8CS...] Pile f customer:
'0000000000000005', ID: '800000006407637B', block: '80000000004C9A11', msg
count: '1', size: '264', collector: '000000000000483D'
```



- Timestamp with time zone!
- Log level
- Host ID & module name (process/service)

# Log Example

```
2012-05-22 18:47:26,807 -0700 INFO  [hostId=long-frontend-1] [module=RECEIVER]
[logger=scala.receiver.MessageBlocker] [tpName=TP-MessagePilePipeline-3]
[auth=Collector:prod-cass-raw-8:000000000000483D:0000000000000005:false]
[remote_ip=184.73.74.54] [web_session=MepMG8CS...] Pile for customer:
'0000000000000005', ID: '800000006407637B', block: '80000000004C9A11', msg
count: '1', size: '264', collector: '000000000000483D'
```

- Timestamp with time zone!
- Log level
- Host ID & module name (process/service)
- Code location or class

# Log Example

```
2012-05-22 18:47:26,807 -0700 INFO  [hostId=long-frontend-1] [module=RECEIVER]
[logger=scala.receiver.MessageBlocker] [thread=MTP-MessagePilePipeline-3]
[auth=Collector:prod-cass-raw-8:000000000000483D:0000000000000005:false]
[remote_ip=184.73.74.54] [web_session=MepMG8CS...] Pile for customer:
'0000000000000005', ID: '800000006407...', block: '80000000004C9A11', msg
count: '1', size: '264', collector: '0000000000483D'
```



- Timestamp with time zone!
- Log level
- Host ID & module name (process/service)
- Code location or class
- Authentication context

# Log Example

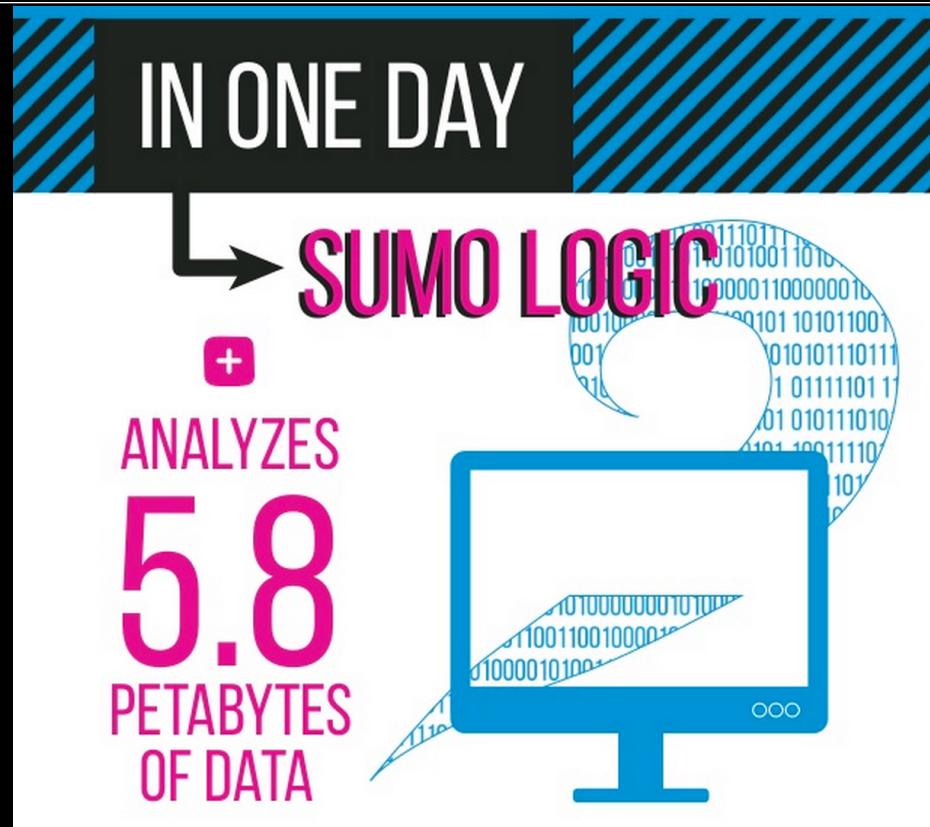
```
2012-05-22 18:47:26,807 -0700 INFO  [hostId=long-frontend-1] [module=RECEIVER]
[logger=scala.receiver.MessageBlocker] [thread=MTP-MessagePilePipeline-3]
[auth=Collector:prod-cass-raw-8:000000000000483D:0000000000000005:false]
[remote_ip=184.73.74.54] [web_session=MepMG8CS...] Pile for customer:
'0000000000000005', ID: '800000006407637B', block: '80000000004C9A11', msg
count: '1', size: '264', collector: '000000000000483D'
```



- Timestamp with time zone!
- Log level
- Host ID & module name (process/service)
- Code location or class
- Authentication context
- Key-value pairs

what is the size of  
machine data?

## Clustering Application



## Using Clustering - Log Reduce

The screenshot shows the Sumo Logic web interface. At the top, there's a navigation bar with links for Welcome, Search, Status, Collectors, Users, Account, Dashboards, Help, and Sign out. A message indicates an account expiration in 160 days. Below the navigation is a search bar containing the query `_sourceCategory=stock_trader_ben | summarize`. The search results are displayed in two main sections: a histogram and a list of log entries.

**Histogram:** The histogram shows event counts over time from 03/15/2013 01:49:02 PM to 01:59:02 PM. The Y-axis ranges from 0 to 300. The X-axis shows time intervals every minute. The data shows several spikes, notably around 01:53 PM, 01:54 PM, and 01:55 PM, reaching values between 150 and 300.

**Log Entries:** The log entries are listed below the histogram. The first entry is a complex object named `*instance of Win32_NTLogEvent` with various properties like Category, ComputerName, EventCode, etc. The second entry is a log message from a PIX firewall about a user executing a command. The third entry is a Microsoft IIS configuration message.

```
6 10 9.53 *instance of Win32_NTLogEvent
{
    Category = 0;
    ComputerName = "AMAZONA-5CC5A52";
    EventCode = 0;
    EventIdentifier = 0;
    EventType = 3;
    InsertionStrings = {"***** de *****"};
    Logfile = "Application";
    Message = "***** de *****";
    RecordNumber = ****;
    SourceName = ".NET StockTrader Web Application";
    TimeGenerated = "2013 *****.000000-000";
    TimeWritten = "2013 *****.000000-000";
    Type = "Information";
};

7 1 9.53 *** <13>1 $DATE***** pix1.be.sumologic.net *** - - [meta sequenceId="*****"] *****4:02: %PIX-5-111008: User pixadmin executed the command access-list acl_inside deny tcp 10.0.0.0 255.0.0.0 any range 1000 2000
8 5 9.53 #Software: Microsoft Internet Information Services 7.5
```

At the bottom, there are video controls for playback and a timestamp of 2:43 / 3:28. The URL <https://www.youtube.com/watch?v=ZF0zbCu9aps> is visible at the very bottom.

Discussion:  
what are the challenges?

## II. DECISION TREES

*Q: What is a decision tree?*

*Q: What is a decision tree?*

*A: A non-parametric hierarchical classification technique.*

*Q: What is a decision tree?*

*A: A non-parametric hierarchical classification technique.*

**non-parametric:** *no parameters, no distribution assumptions*

*Q: What is a decision tree?*

*A: A non-parametric hierarchical classification technique.*

**non-parametric:** *no parameters, no distribution assumptions*

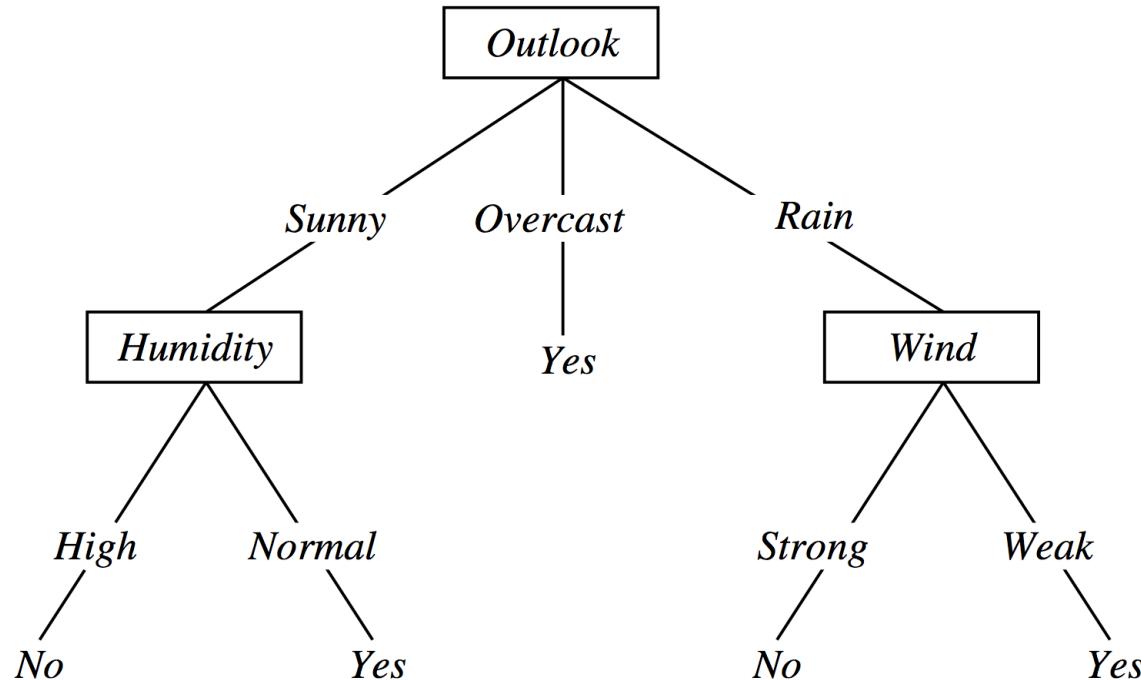
**hierarchical:** *consists of a sequence of questions which yield a class label when applied to any record*

*Q: How is a decision tree represented?*

*Q: How is a decision tree represented?*

*A: Using a configuration of nodes and edges.*

## DECISION TREE CLASSIFIERS



Classify an instance: <**outlook**=Sunny, **temp** = Hot, **humidity**=High, **wind** = Strong>

## DECISION TREE CLASSIFIERS

---

*Q: How is a decision tree represented?*

*A: Using a configuration of nodes and edges.*

*More concretely, as a multiway tree, which is a type of (directed acyclic) graph.*

## DECISION TREE CLASSIFIERS

---

*Q: How is a decision tree represented?*

*A: Using a configuration of nodes and edges.*

*More concretely, as a multiway tree, which is a type of (directed acyclic) graph.*

*In a decision tree, the nodes represent questions (test conditions) and the edges are the answers to these questions.*

---

## TYPES OF NODES

---

*The top node of the tree is called the root node. This node has 0 incoming edges, and 2+ outgoing edges.*

## TYPES OF NODES

---

*The top node of the tree is called the **root node**. This node has 0 incoming edges, and 2+ outgoing edges.*

*An **internal node** has 1 incoming edge, and 2+ outgoing edges. Internal nodes represent test conditions.*

*The top node of the tree is called the **root node**. This node has 0 incoming edges, and 2+ outgoing edges.*

*An **internal node** has 1 incoming edge, and 2+ outgoing edges. Internal nodes represent test conditions.*

*A **leaf node** has 1 incoming edge and, 0 outgoing edges. Leaf nodes correspond to class labels.*

*The top node of the tree is called the **root node**. This node has 0 incoming edges, and 2+ outgoing edges.*

*An **internal node** has 1 incoming edge, and 2+ outgoing edges. Internal nodes represent test conditions.*

*A **leaf node** has 1 incoming edge and, 0 outgoing edges. Leaf nodes correspond to class labels.*

**NOTE**

The nodes in our tree are connected by directed edges.

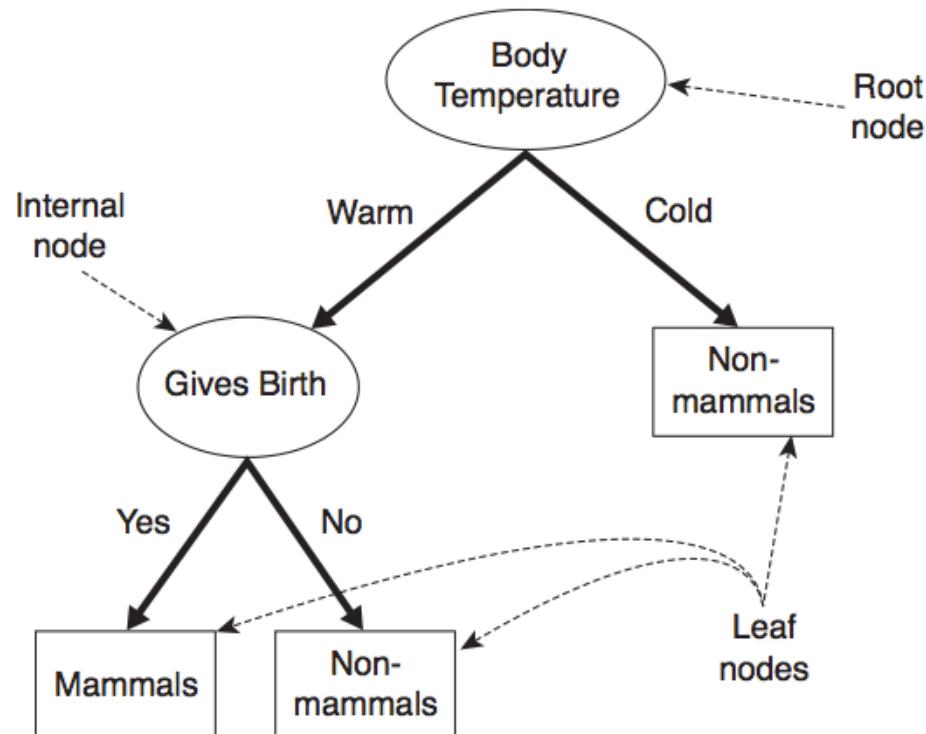
These directed edges lead from parent nodes to child nodes.

**Table 4.1.** The vertebrate data set.

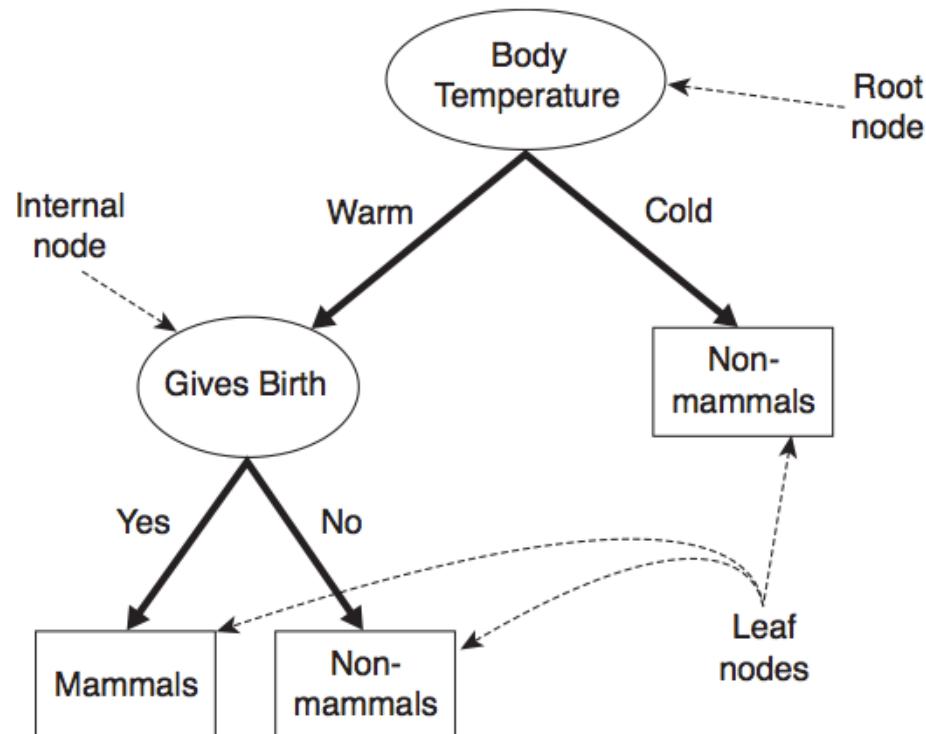
Name	Body Temperature	Skin Cover	Gives Birth	Aquatic Creature	Aerial Creature	Has Legs	Hibernates	Class Label
human	warm-blooded	hair	yes	no	no	yes	no	mammal
python	cold-blooded	scales	no	no	no	no	yes	reptile
salmon	cold-blooded	scales	no	yes	no	no	no	fish
whale	warm-blooded	hair	yes	yes	no	no	no	mammal
frog	cold-blooded	none	no	semi	no	yes	yes	amphibian
komodo dragon	cold-blooded	scales	no	no	no	yes	no	reptile
bat	warm-blooded	hair	yes	no	yes	yes	yes	mammal
pigeon	warm-blooded	feathers	no	no	yes	yes	no	bird
cat	warm-blooded	fur	yes	no	no	yes	no	mammal
leopard	cold-blooded	scales	yes	yes	no	no	no	fish
shark								
turtle	cold-blooded	scales	no	semi	no	yes	no	reptile
penguin	warm-blooded	feathers	no	semi	no	yes	no	bird
porcupine	warm-blooded	quills	yes	no	no	yes	yes	mammal
eel	cold-blooded	scales	no	yes	no	no	no	fish
salamander	cold-blooded	none	no	semi	no	yes	yes	amphibian

## EXAMPLE – DECISION TREE

51



**Figure 4.4.** A decision tree for the mammal classification problem.



**Figure 4.4.** A decision tree for the mammal classification problem.

**NOTE**  
Internal nodes represent test conditions which partition the records at that node.

# III. BUILDING DECISION TREES

*Q: How do we build a decision tree?*

*Q: How do we build a decision tree?*

*A: One possibility would be to evaluate all possible decision trees (eg, all permutations of test conditions) for a given dataset.*

*Q: How do we build a decision tree?*

*A: One possibility would be to evaluate all possible decision trees (eg, all permutations of test conditions) for a given dataset.*

*But this is generally too complex to be practical →  $O(2^n)$ .*

*Q: How do we build a decision tree?*

*A: One possibility would be to evaluate all possible decision trees (eg, all permutations of test conditions) for a given dataset.*

*But this is generally too complex to be practical  $\rightarrow O(2^n)$ .*

*Q: How do we find a practical solution that works?*

## BUILDING A DECISION TREE

---

*Q: How do we build a decision tree?*

*A: One possibility would be to evaluate all possible decision trees (eg, all permutations of test conditions) for a given dataset.*

*But this is generally too complex to be practical  $\rightarrow O(2^n)$ .*

*Q: How do we find a practical solution that works?*

*A: Use a heuristic algorithm.*

*The basic method used to build (or “grow”) a decision tree is Hunt’s algorithm.*

*The basic method used to build (or “grow”) a decision tree is Hunt’s algorithm.*

*This is a greedy recursive algorithm that leads to a local optimum.*

*The basic method used to build (or “grow”) a decision tree is Hunt’s algorithm.*

*This is a greedy recursive algorithm that leads to a local optimum.*

**greedy** – *algorithm makes locally optimal decision at each step*

**recursive** – *splits task into subtasks, solves each the same way*

**local optimum** – *solution for a given neighborhood of points*

*Hunt's algorithm builds a decision tree by recursively partitioning records into smaller & smaller subsets.*

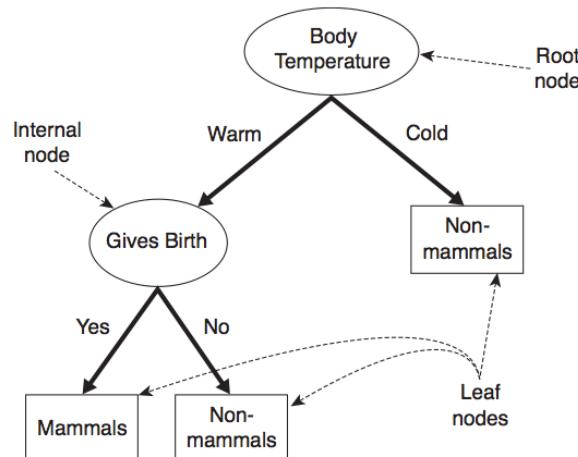


Figure 4.4. A decision tree for the mammal classification problem.

## BUILDING A DECISION TREE

---

*Hunt's algorithm builds a decision tree by recursively partitioning records into smaller & smaller subsets.*

*The partitioning decision is made at each node according to a metric called purity.*

## BUILDING A DECISION TREE

---

*Hunt's algorithm builds a decision tree by recursively partitioning records into smaller & smaller subsets.*

*The partitioning decision is made at each node according to a metric called purity.*

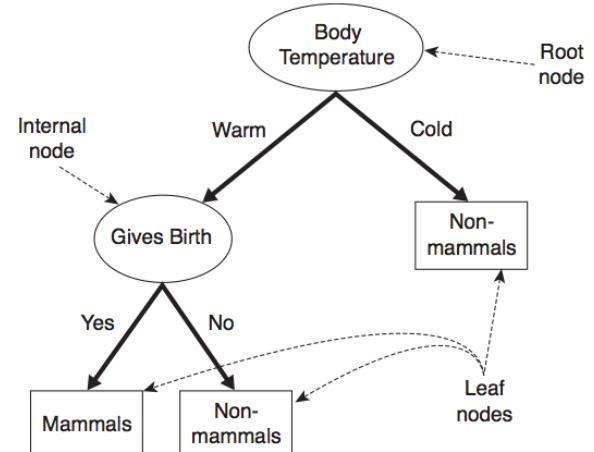
*A partition is 100% pure when all of its records belong to a single class.*

# PURITY

---

**Table 4.1.** The vertebrate data set.

Name	Body Temperature	Skin Cover	Gives Birth	Aquatic Creature	Aerial Creature	Has Legs	Hibernates	Class Label
human	warm-blooded	hair	yes	no	no	yes	no	mammal
python	cold-blooded	scales	no	no	no	no	yes	reptile
salmon	cold-blooded	scales	no	yes	no	no	no	fish
whale	warm-blooded	hair	yes	yes	no	no	no	mammal
frog	cold-blooded	none	no	semi	no	yes	yes	amphibian
komodo dragon	cold-blooded	scales	no	no	no	yes	no	reptile
bat	warm-blooded	hair	yes	no	yes	yes	yes	mammal
pigeon	warm-blooded	feathers	no	no	yes	yes	no	bird
cat	warm-blooded	fur	yes	no	no	yes	no	mammal
leopard	cold-blooded	scales	yes	yes	no	no	no	fish
shark	cold-blooded	scales	no	semi	no	yes	no	reptile
turtle	cold-blooded	scales	no	semi	no	yes	no	bird
penguin	warm-blooded	feathers	no	semi	no	yes	no	mammal
porcupine	warm-blooded	quills	yes	no	no	yes	yes	fish
eel	cold-blooded	scales	no	yes	no	no	no	amphibian
salamander	cold-blooded	none	no	semi	no	yes	yes	



**Figure 4.4.** A decision tree for the mammal classification problem.

We'll discuss the details of  
tree building in the next lecture

**LAB:**

**HOMEWORK 2 REVIEW**

**K-MEANS CLUSTERING (PART 2)**

---

**INTRO TO DATA SCIENCE**

---

**DISCUSSION**