GENERAL ASSEMBLY

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

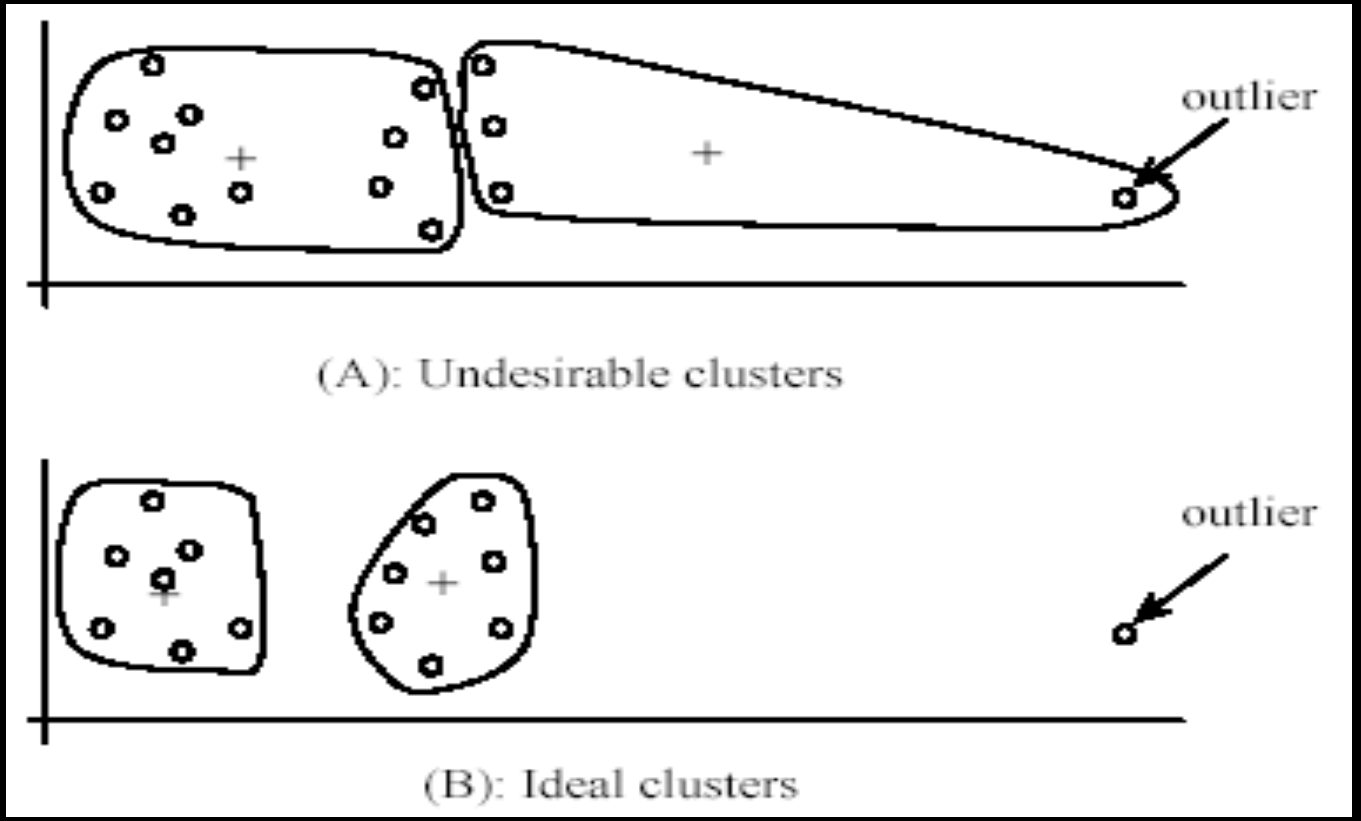
LAB:
IV. HOMEWORK 2 REVIEW
V. K-MEANS CLUSTERING (PART 2)

K-Means is the most popular clustering algorithm.

Source: http://www.cs.uic.edu/~liub/teach/cs583-fall-05/CS583-unsupervised-learning.ppt

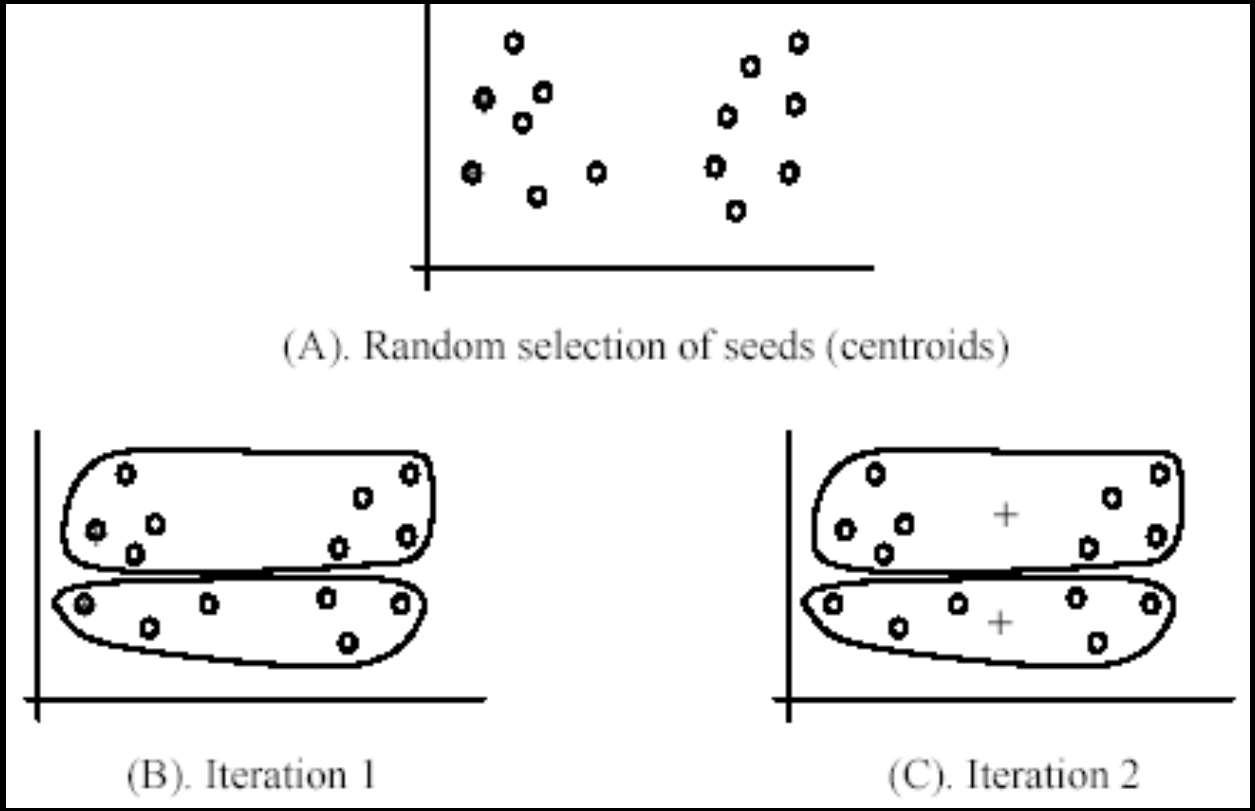
but sensitive to outliers…

K-Means Clustering



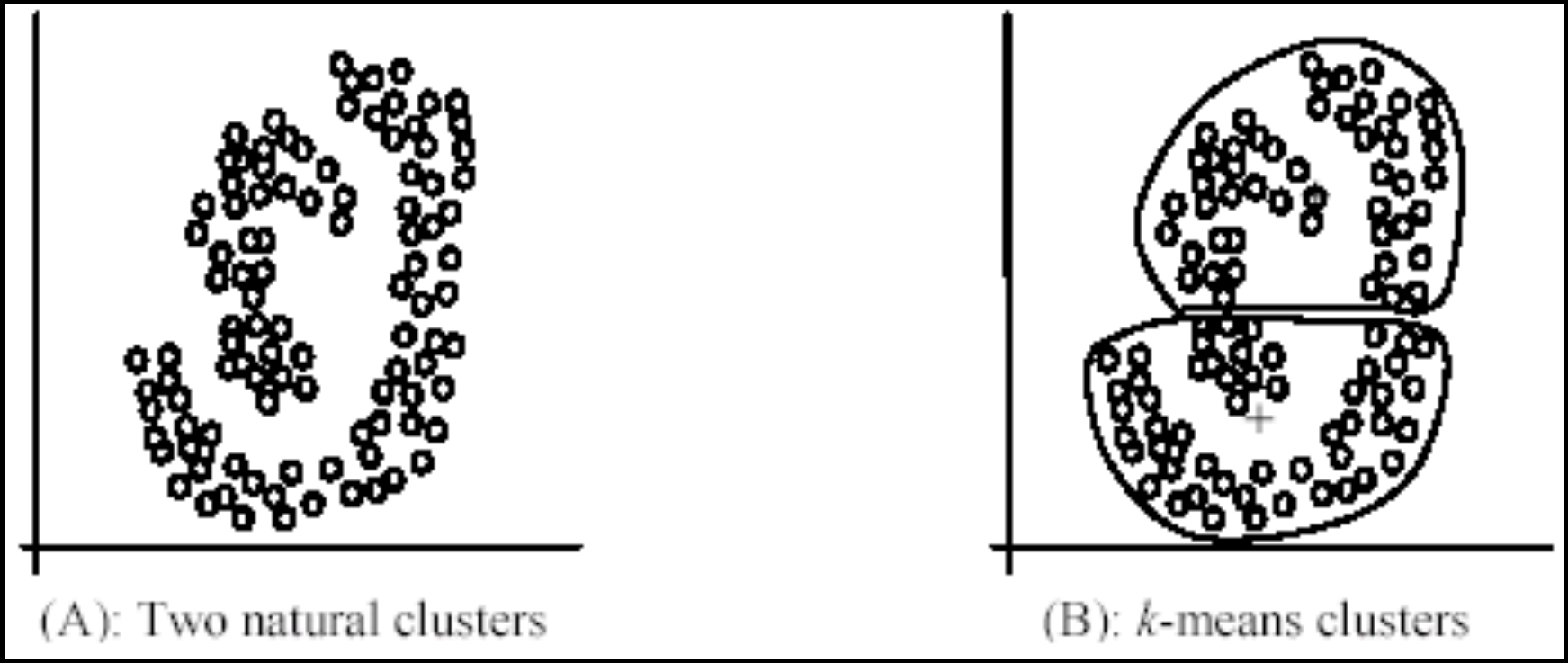
also sensitive to initial seeds

K-Means Clustering



and not suitable for discovering
non-hyper-sphere clusters

K-Means Clustering



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

How to choose a clustering algorithm?

choosing the best algorithm is a challenge

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

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

The common practice is to…

Run several algorithms using different distance functions and parameter settings

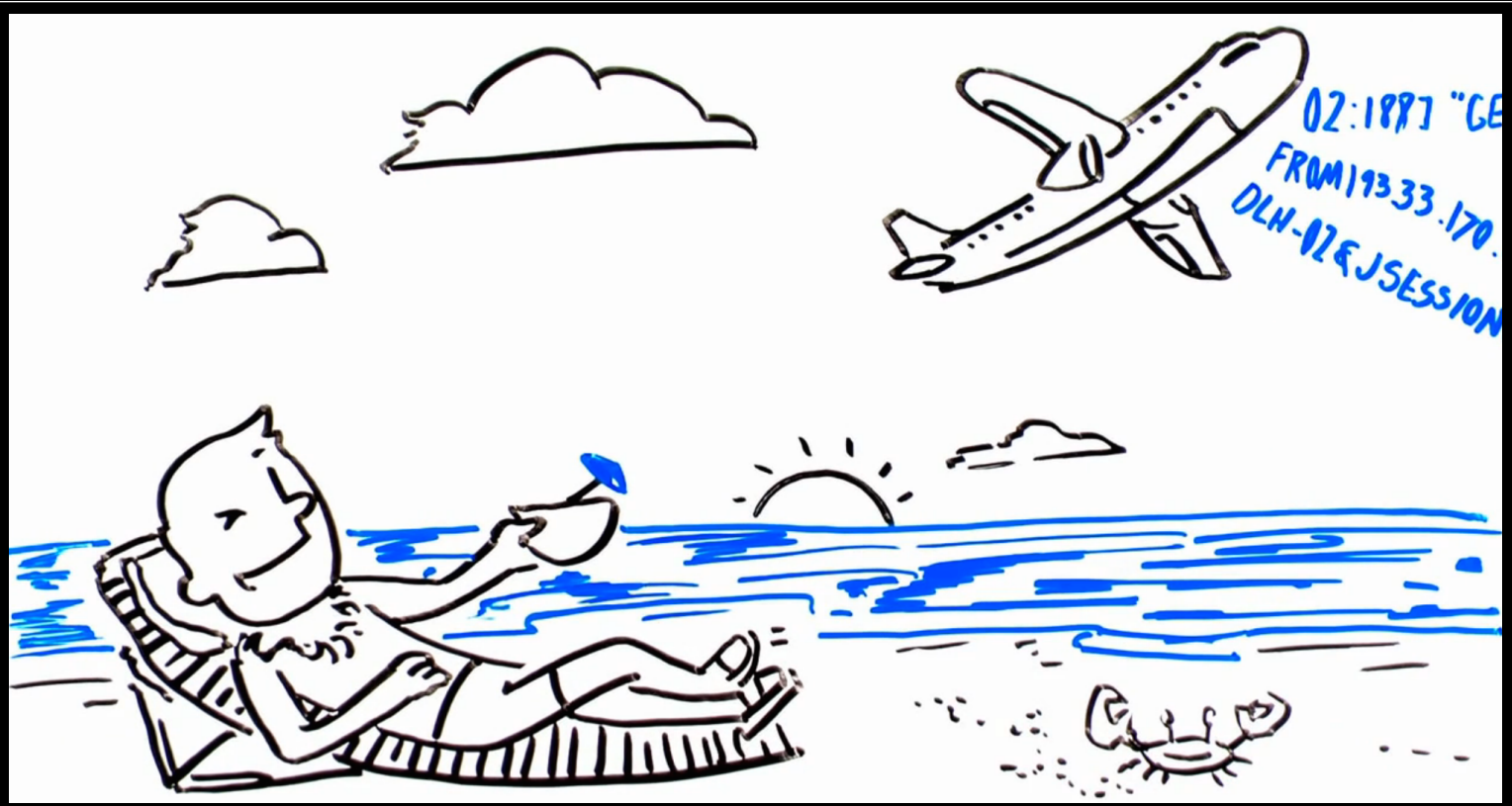
then carefully
analyze and compare the results

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

Clustering in Practice: learning machine data

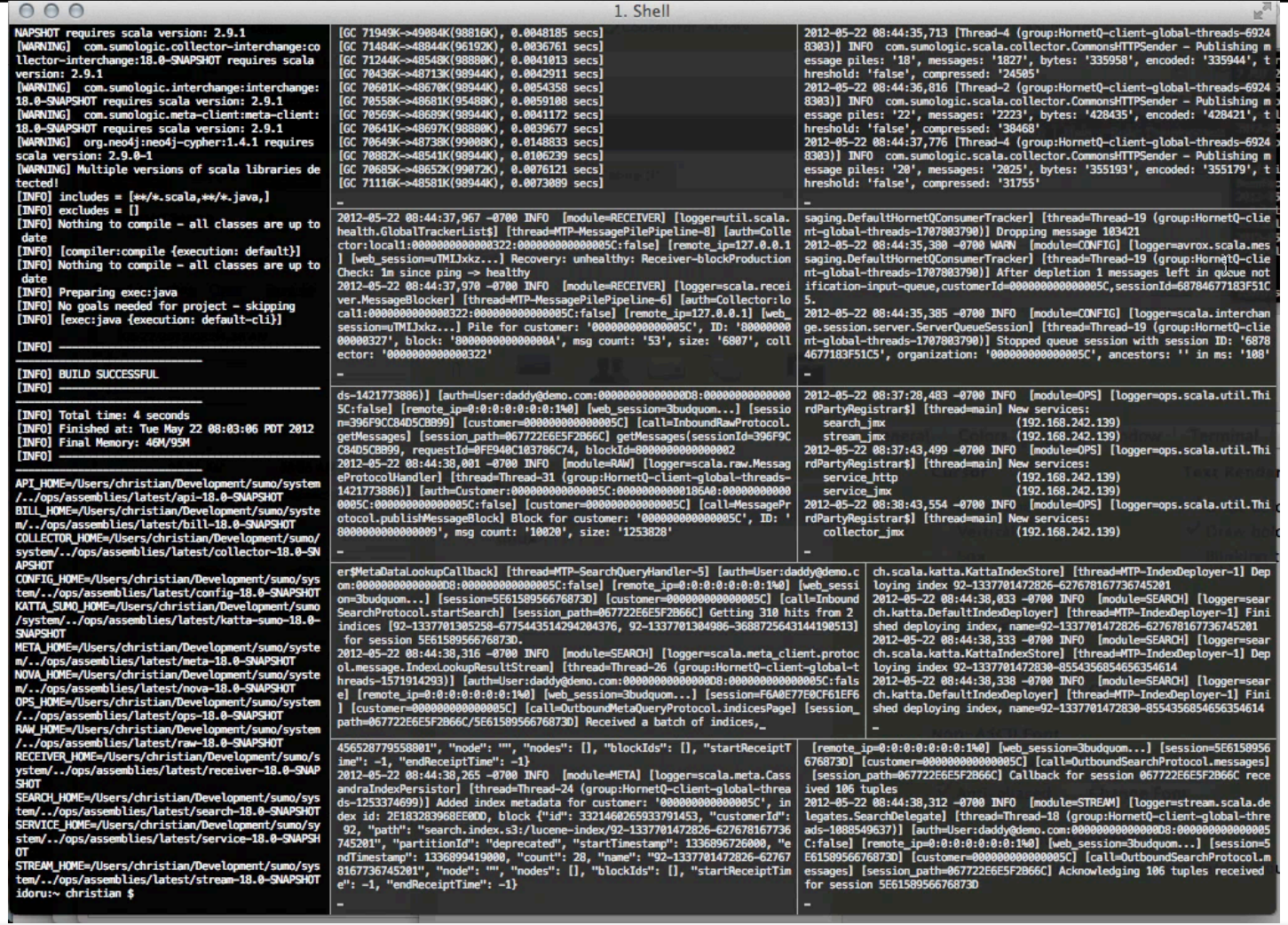
what is machine data?

Clustering Application



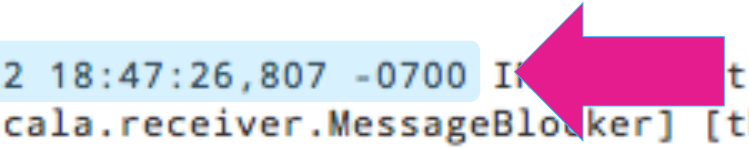
http://www.splunk.com/en_us/resources/machine-data.html

what does the machine data look like?



source: http://www.slideshare.net/Sumo_Logic/can-you-process-10-trillion-logs-per-day-software-architecture-conference-2015

Log Example



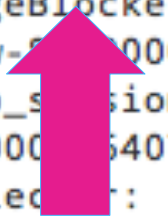
```
2012-05-22 18:47:26,807 -0700 I[illegible]tId=long-frontend-1] [module=RECEIVER]
[logger=scala.receiver.MessageBlo[illegible]ker] [thread=MTP-MessagePilePipeline-3]
[auth=Collector:prod-cass-raw-8:0000000000000483D:0000000000000005:false]
[remote_ip=184.73.74.54] [web_session=MepMG8CS...] Pile for customer:
'0000000000000005', ID: '8000000006407637B', block: '80000000004C9A11', msg
count: '1', size: '264', collector: '0000000000000483D'
```

- Timestamp with time zone!

source: http://www.slideshare.net/Sumo_Logic/can-you-process-10-trillion-logs-per-day-software-architecture-conference-2015

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-0000000000483D:0000000000000005:false]
[remote_ip=184.73.74.54] [web_session=MepMG8CS...] Pile for customer:
'0000000000000005', ID: '800000[illegible]5407637B', block: '80000000004C9A11', msg
count: '1', size: '264', collector: '0000000000483D'
```

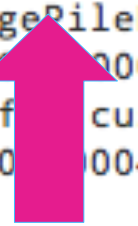


- Timestamp with time zone!
- Log level

source: http://www.slideshare.net/Sumo_Logic/can-you-process-10-trillion-logs-per-day-software-architecture-conference-2015

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-8:0000000000483D:0000000 00005:false]
[remote_ip=184.73.74.54] [web_session=MepMG8CS...] Pile f customer:
'0000000000000005', ID: '8000000006407637B', block: '80000 004C9A11', msg
count: '1', size: '264', collector: '0000000000483D'
```



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

source: http://www.slideshare.net/Sumo_Logic/can-you-process-10-trillion-logs-per-day-software-architecture-conference-2015

Log Example

```
2012-05-22 18:47:26,807 -0700 INFO  [host id=long-frontend-1] [module=RECEIVER]
[logger=scala.receiver.MessageBlocker]  TP-MessagePilePipeline-3]
[auth=Collector:prod-cass-raw-8:0000000000483D:0000000000000005:false]
[remote_ip=184.73.74.54] [web_session=MepMG8CS...] Pile for customer:
'0000000000000005', ID: '8000000006407637B', block: '80000000004C9A11', msg
count: '1', size: '264', collector: '0000000000000483D'
```

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

source: http://www.slideshare.net/Sumo_Logic/can-you-process-10-trillion-logs-per-day-software-architecture-conference-2015

Log Example

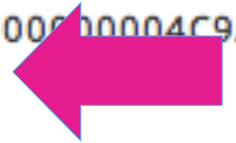
```
2012-05-22 18:47:26,807 -0700 INFO  [hostId=long-frontend-1] [module=RECEIVER]
[logger=scala.receiver.MessageBlocker] [thread=MTP-MessagePilePipeline-3]
[auth=Collector:prod-cass-raw-8:000000000000483D:0000000000000005:false]
[remote_ip=184.73.74.54] [web_session=MepMG8CS...] Pile for customer:
'0000000000000005', ID: '8000000064072', 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

source: http://www.slideshare.net/Sumo_Logic/can-you-process-10-trillion-logs-per-day-software-architecture-conference-2015

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: '8000000006407637B', block: '8000000000004C9A11', msg
count: '1', size: '264', collector: '000000000000483D'
```

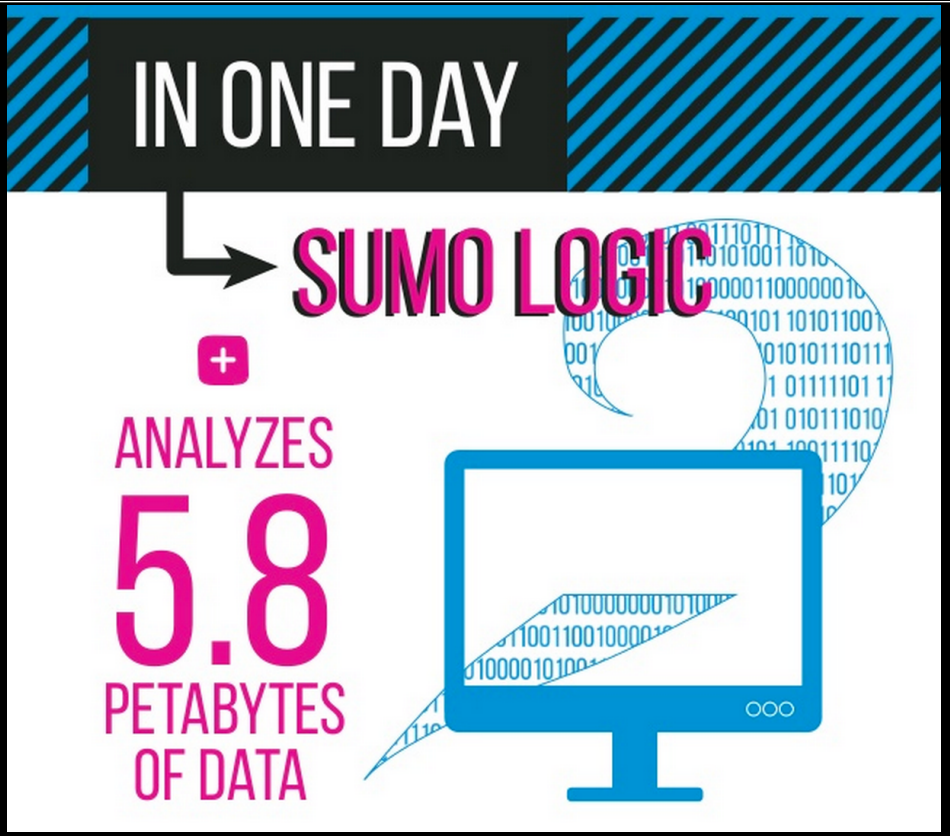


- Timestamp with time zone!
- Log level
- Host ID & module name (process/service)
- Code location or class
- Authentication context
- Key-value pairs

source: http://www.slideshare.net/Sumo_Logic/can-you-process-10-trillion-logs-per-day-software-architecture-conference-2015

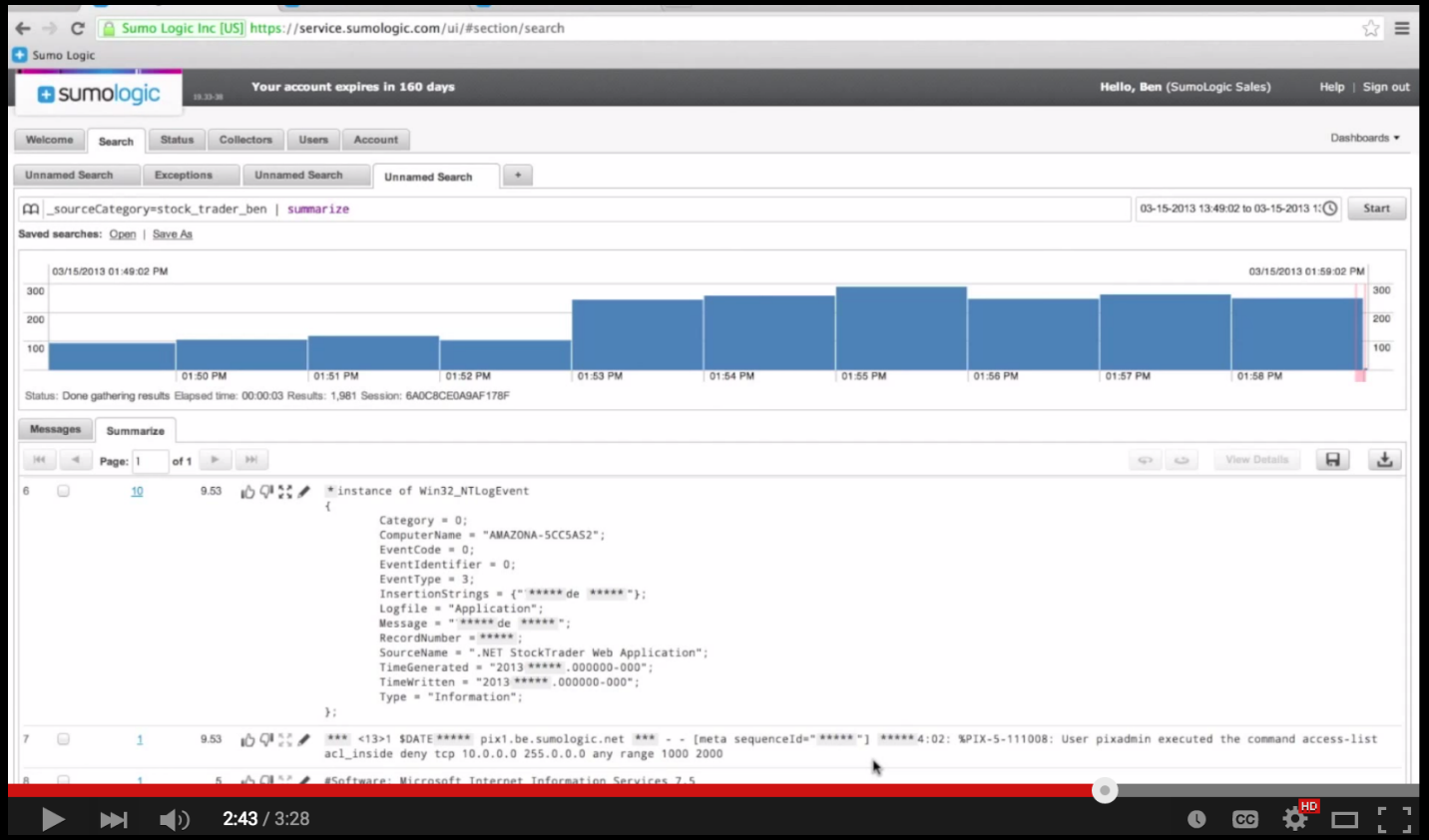
what is the size of
machine data?

Clustering Application



http://www.slideshare.net/Sumo_Logic/sumo-logic-datainoneday

Using Clustering – Log Reduce



https://www.youtube.com/watch?v=ZF0zbCu9aps

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*

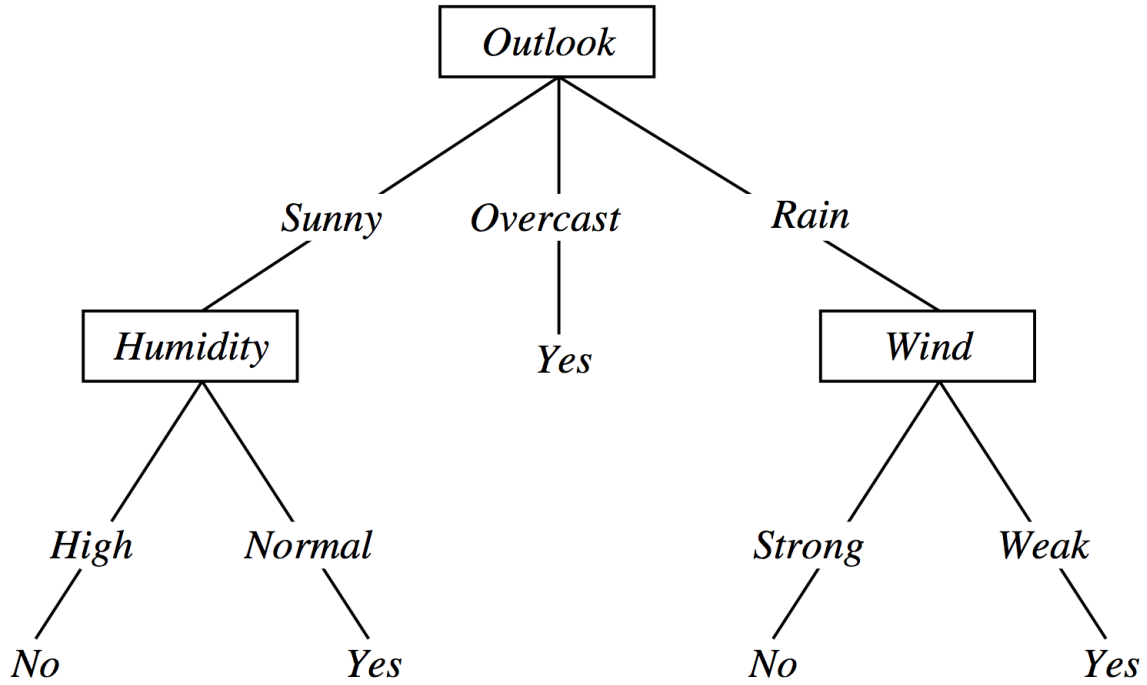
hierarchical*: consists of a sequence of questions which yield a class label when applied to any record*

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>

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**.

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.*

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. L 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	Hiber-nates	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 shark	cold-blooded	scales	yes	yes	no	no	no	fish
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

source: http://www-users.cs.umn.edu/~kumar/dmbook/ch4.pdf

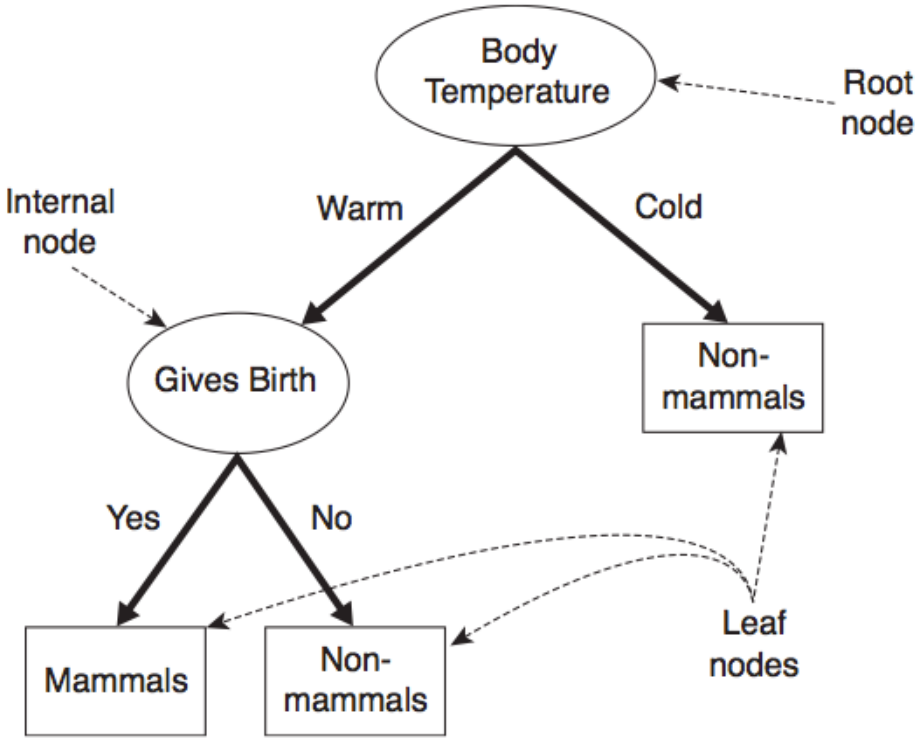


Figure 4.4. A decision tree for the mammal classification problem.

source: http://www-users.cs.umn.edu/~kumar/dmbook/ch4.pdf

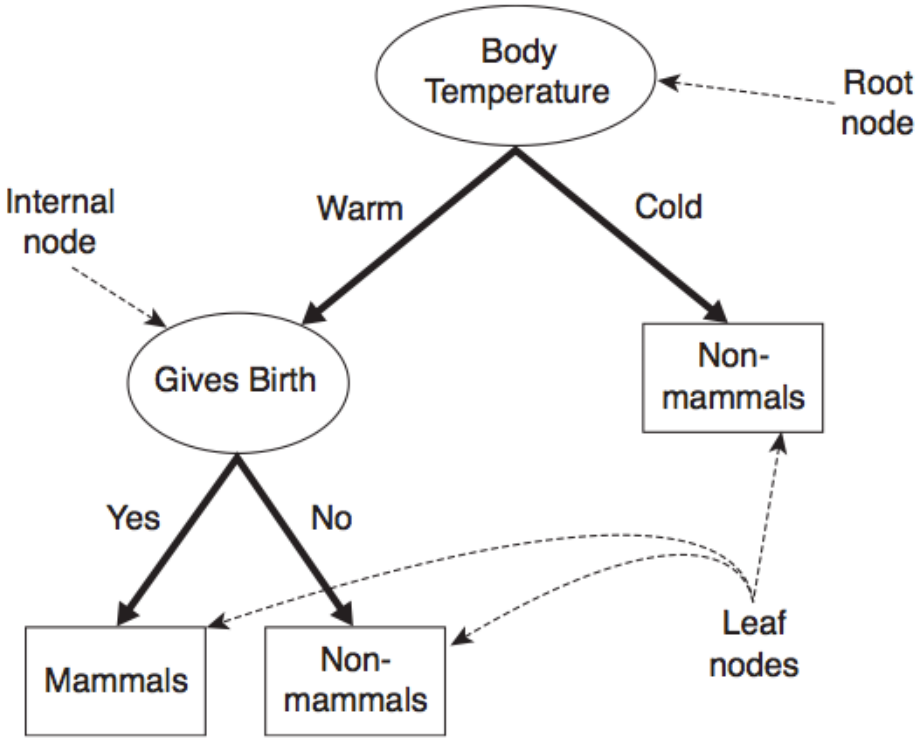


Figure 4.4. A decision tree for the mammal classification problem.

source: http://www-users.cs.umn.edu/~kumar/dmbook/ch4.pdf

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 → $O(2^n)$.

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 → $O(2^n)$.

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 → $O(2^n)$.

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.

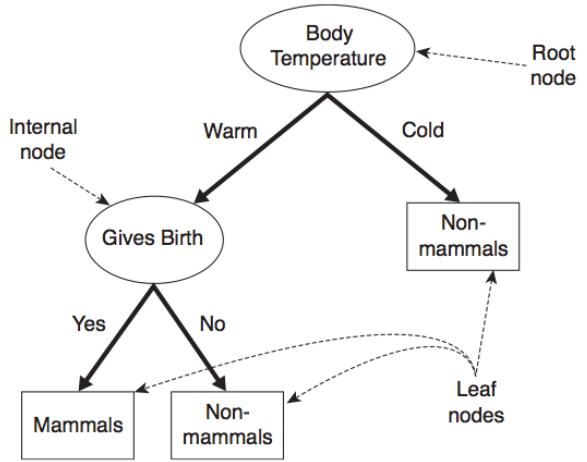


Figure 4.4. A decision tree for the mammal classification problem.

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.

Table 4.1. The vertebrate data set.

Name	Body Temperature	Skin Cover	Gives Birth	Aquatic Creature	Aerial Creature	Has Legs	Hiber-nates	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 shark	cold-blooded	scales	yes	yes	no	no	no	fish
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

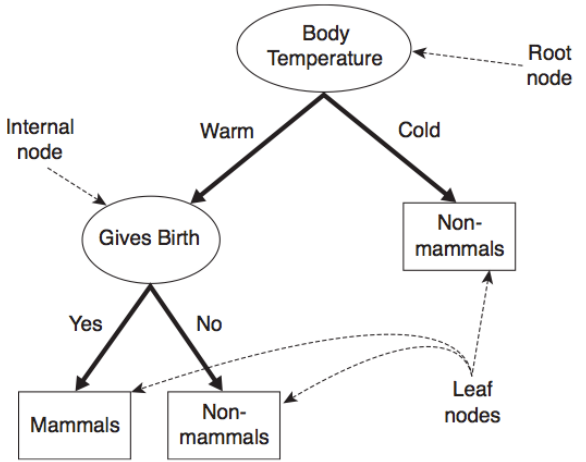


Figure 4.4. A decision tree for the mammal classification problem.

source: http://www-users.cs.umn.edu/~kumar/dmbook/ch4.pdf

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

LAB:
HOMEWORK 2 REVIEW
K-MEANS CLUSTERING (PART 2)

DISCUSSION