Malware mining

Dr. Igor Muttik – McAfee Labs™

Barcelona
2011

Agenda

- What makes a great AV product
- What is data-mining?
- Data-mining methods
  - Decision trees
  - Support Vector Machines
  - ROCs
- Data-mining as a heuristic/generic method
- Extracting features
- Practical uses of malware data-mining
- Conclusions and questions

What would make a great AV product?

McAfee

- Proactively detects as many threats as possible
- Creates a very low number of false alarms
- Requires as little maintenance as possible
- Runs quickly and introduces little overhead

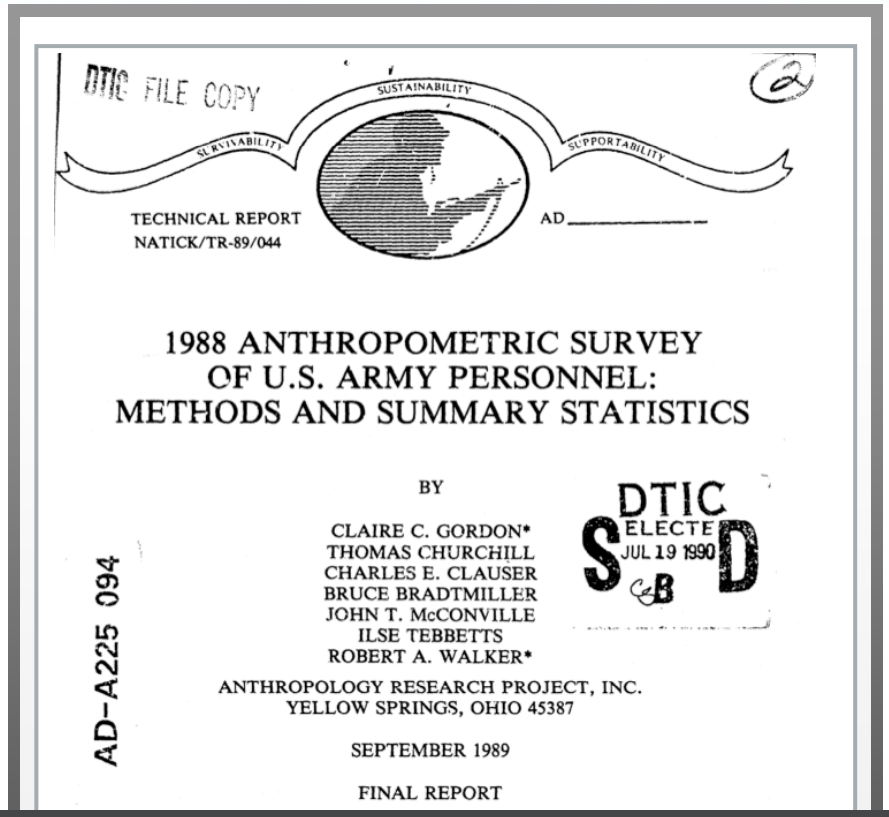
But how can this be done?

- Traditional (specific) detection methods (strings, hashes)
- Generic detection methods
- Heuristic detection methods
- Cloud-based methods
- Reputation methods
- Behavioural methods

Detection method	Proactive capability	Cost to update	Available for
Specific (manual)	None	Medium	Many years
Specific (automated)	None	Low	Many years
Generic (manual)	High	High	Many years
Generic (automated)	Medium	Medium	Many years
Heuristic (manual)	Medium	High	Many years
Heuristic (automated)	Medium	Low	???

What is data-mining?

Data Mining—An Example



DTIC FILE COPY

SUSTAINABILITY SURVIVABILITY SUPPORTABILITY

TECHNICAL REPORT
NATICK/TR-89/044

AD _____

1988 ANTHROPOMETRIC SURVEY
OF U.S. ARMY PERSONNEL:
METHODS AND SUMMARY STATISTICS

BY

CLAIRE C. GORDON*
THOMAS CHURCHILL
CHARLES E. CLAUSER
BRUCE BRADTMILLER
JOHN T. McCONVILLE
ILSE TEBBETTS
ROBERT A. WALKER*

ANTHROPOLOGY RESEARCH PROJECT, INC.
YELLOW SPRINGS, OHIO 45387

DTIC ELECTE JUL 19 1990 S D

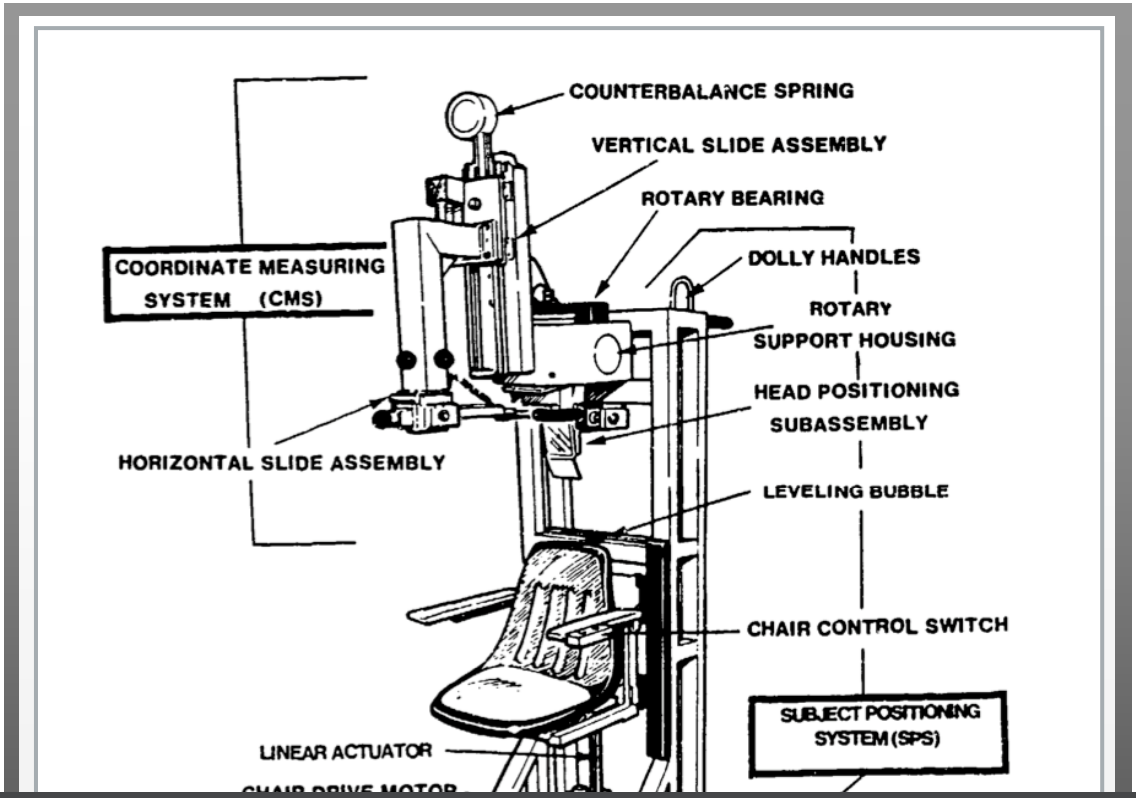
AD-A225 094

SEPTEMBER 1989

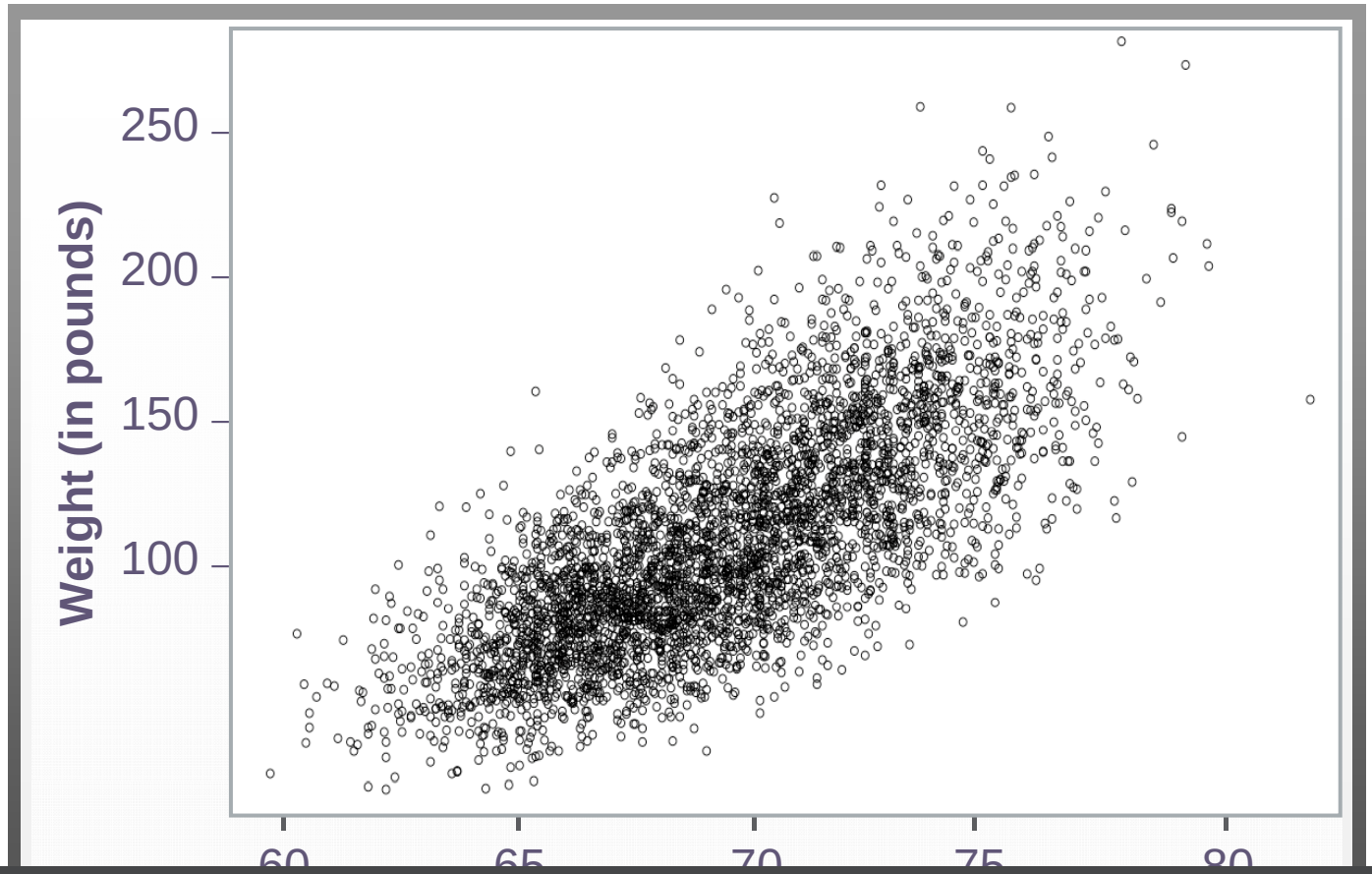
FINAL REPORT

Source: http://mreed.umtri.umich.edu/mreed/downloads.html#anthro, http://www.dtic.mil/dticasd/docs-a/anthro_military.html

Measurements

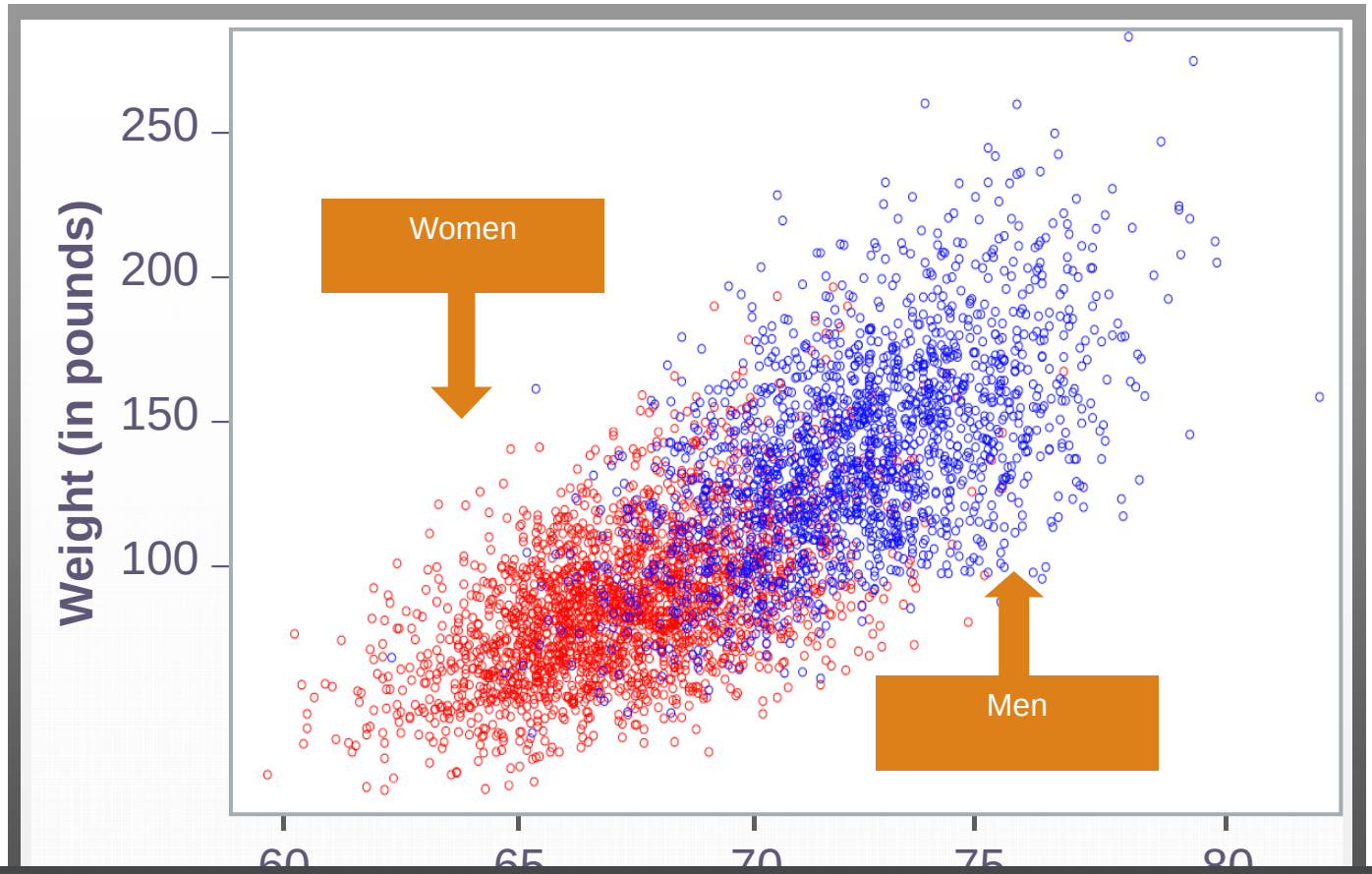


Height Versus Weight

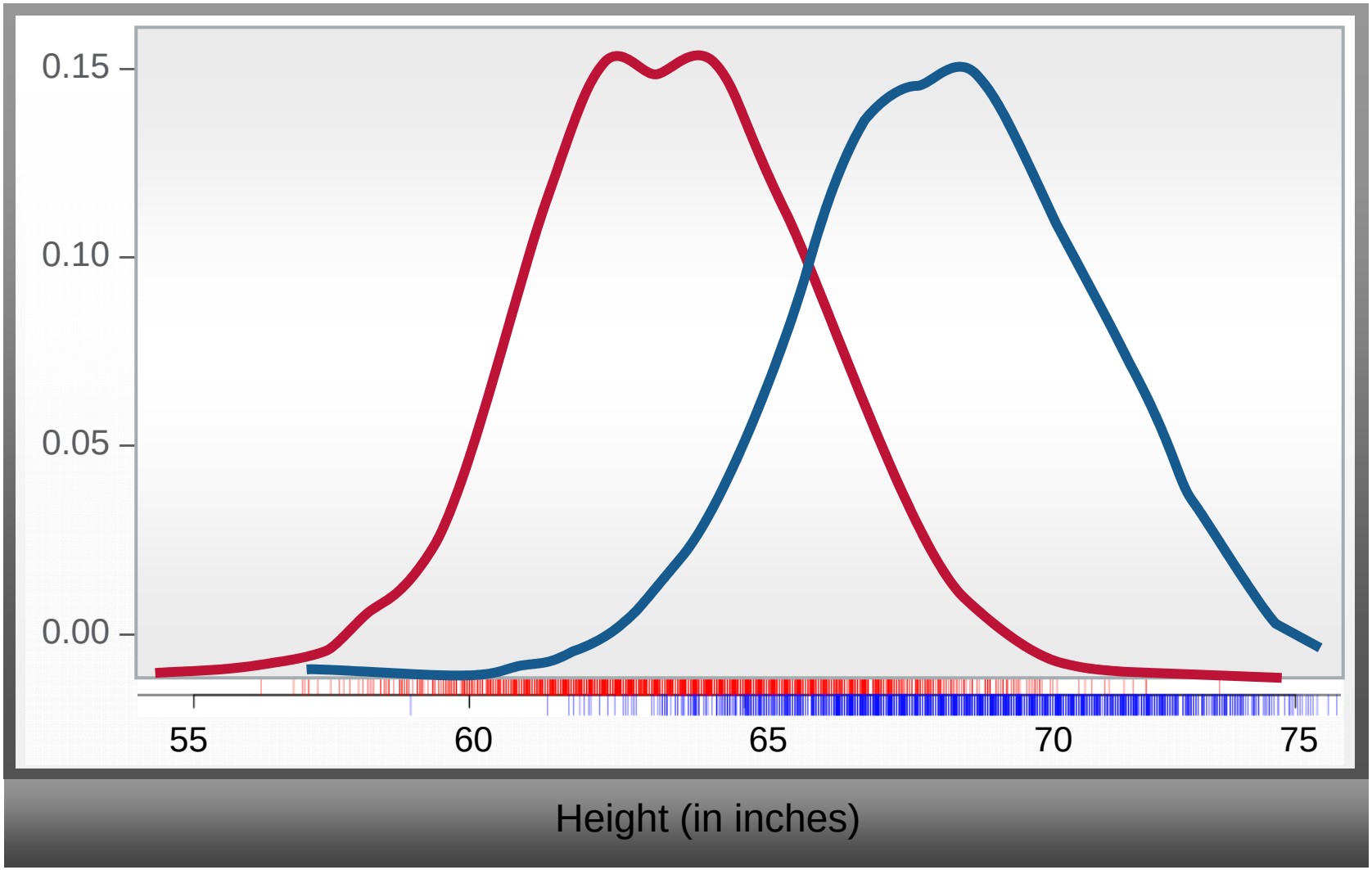


Height Versus Weight + 1 Boolean feature

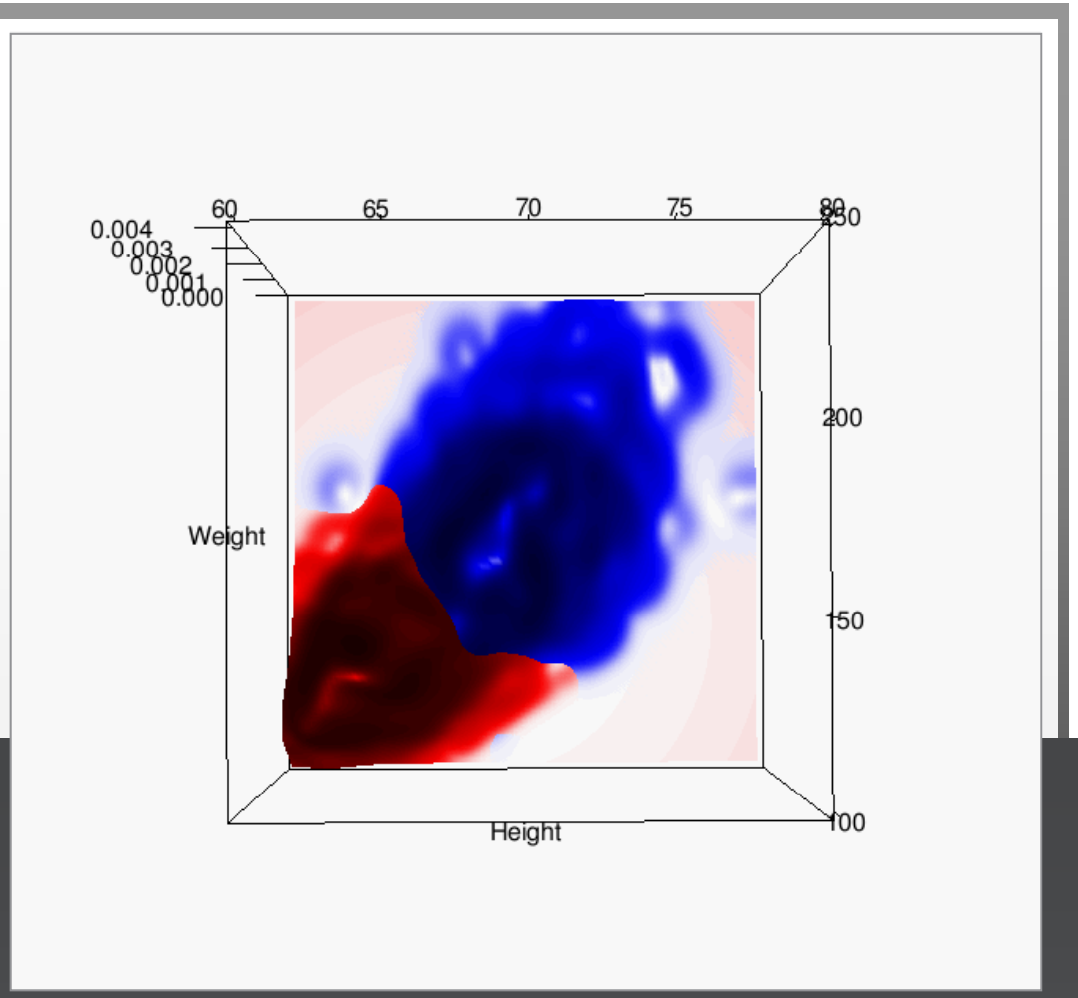
McAfee



One Dimension Only (1 feature)

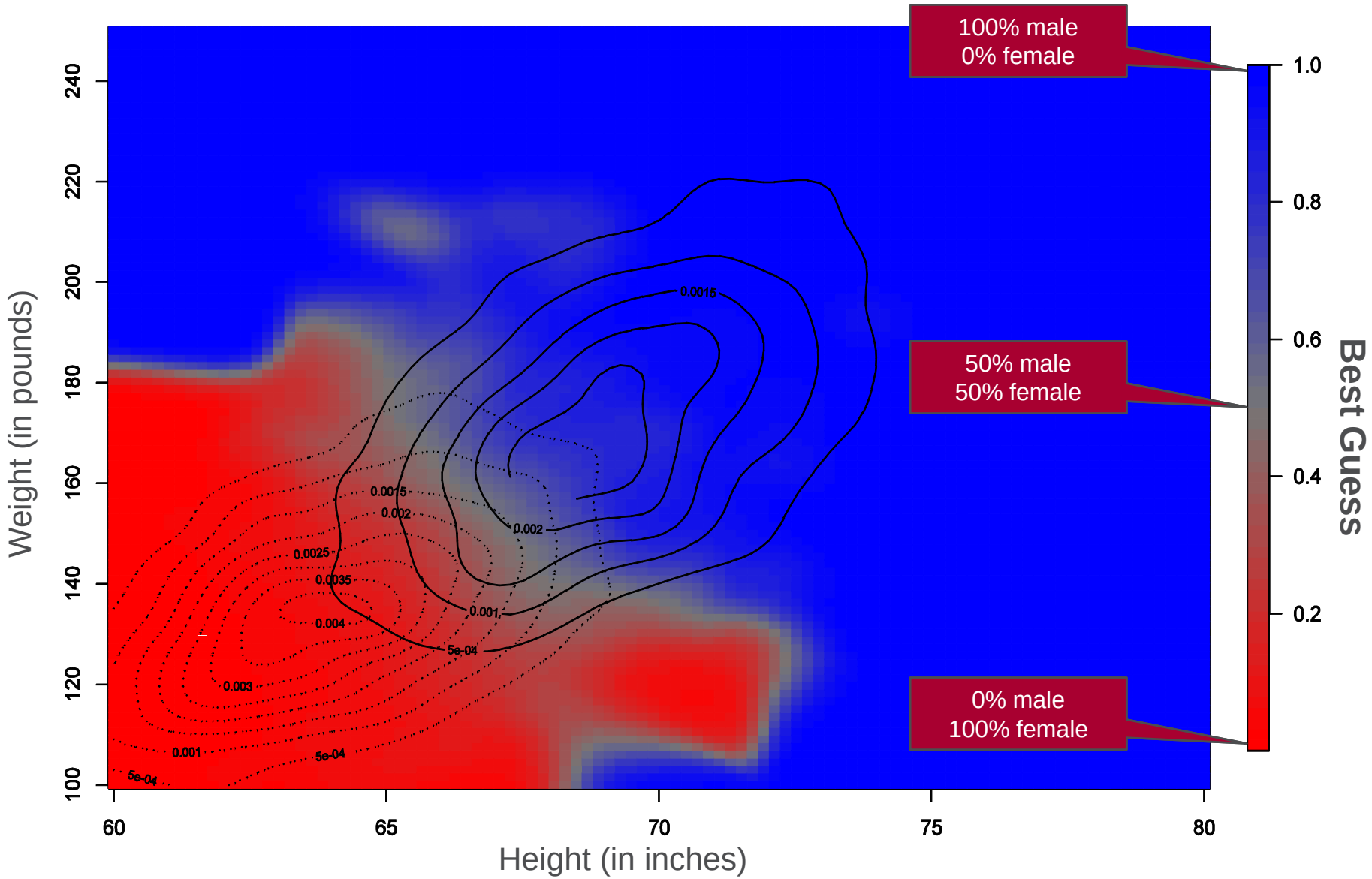


Putting Weight and Height Into Perspective

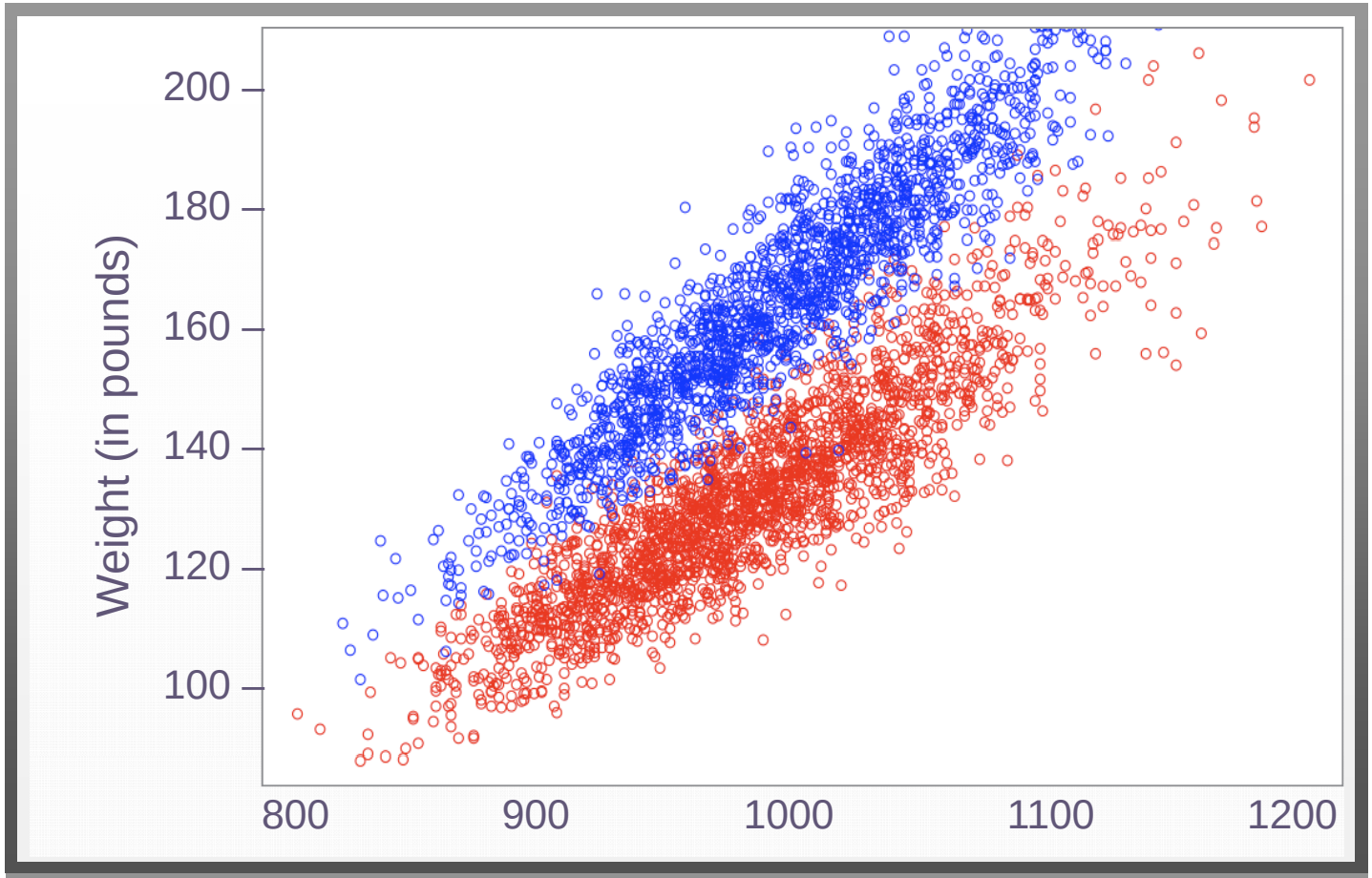


Best Guess for Gender

McAfee

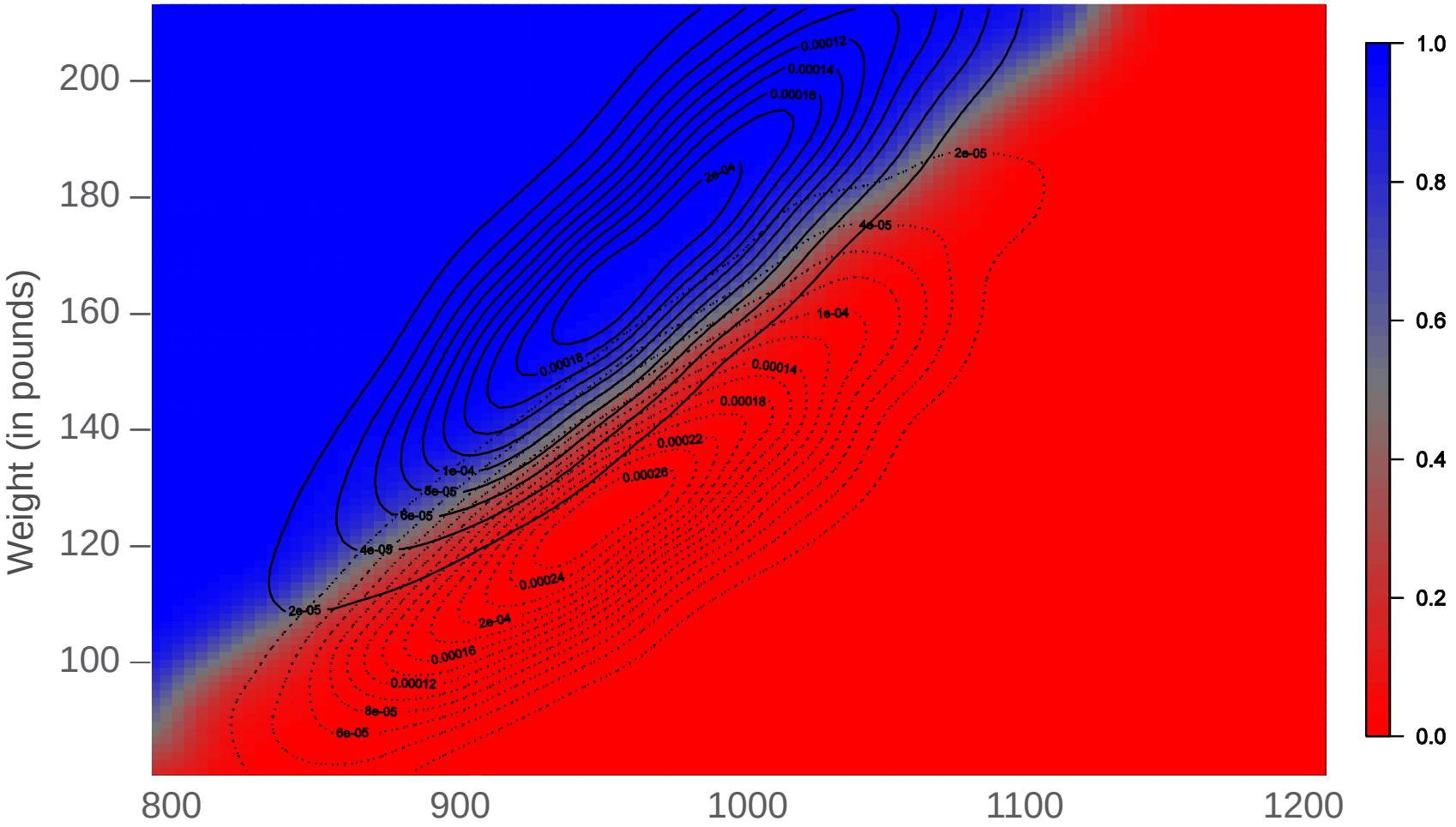


Better Features



Buttock Circumference: The circumference of the body measured at the level of the maximum posterior protuberance of the buttocks.

Best Guess for Revised Features



Buttock Circumference: The circumference of the body measured at the level of the maximum posterior protuberance of the buttocks.

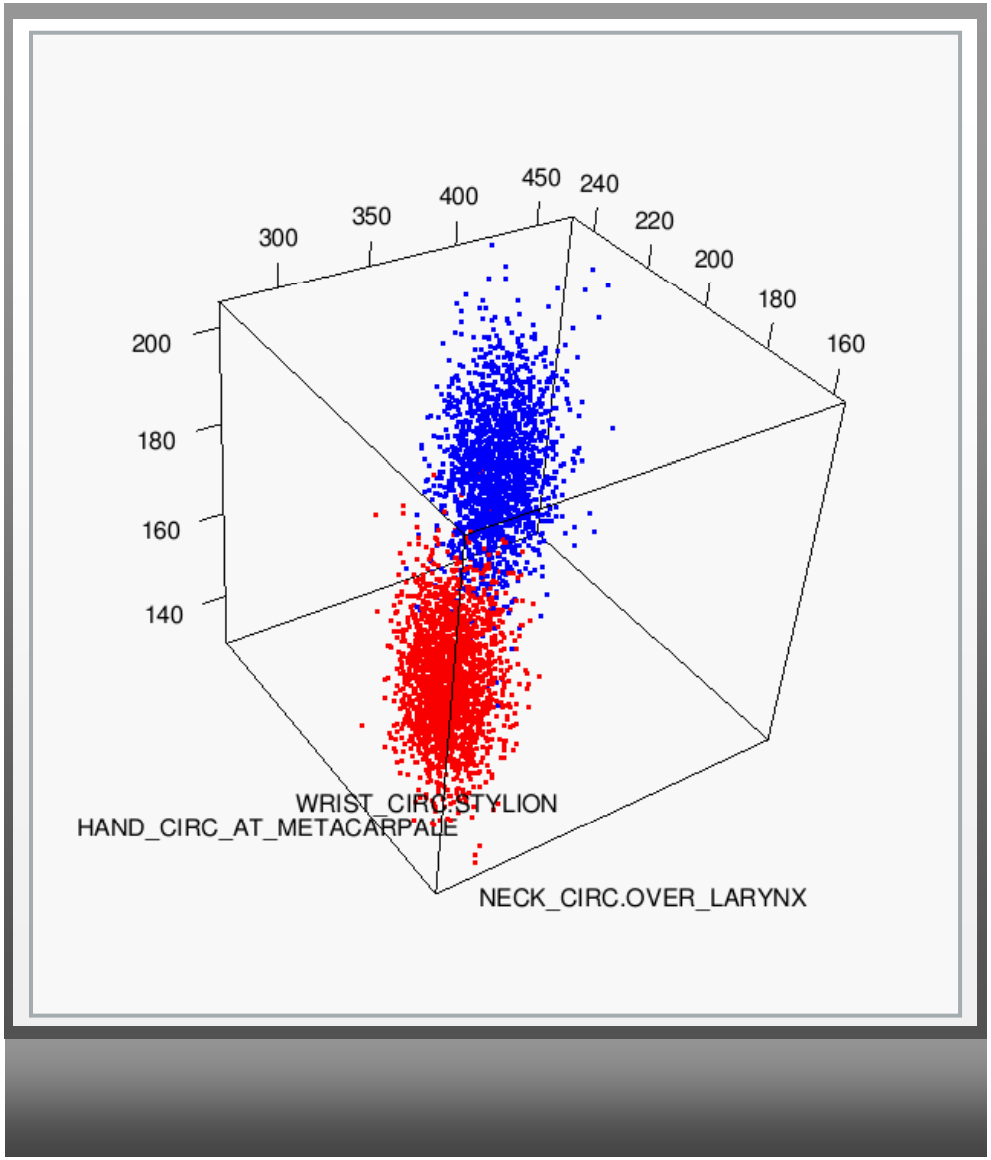
Further Improving the Separation

McAfee

Signal to Noise
Features with very different distribution per class

Correlation
Features with low correlation

Dimensionality
Consider more features at the same time

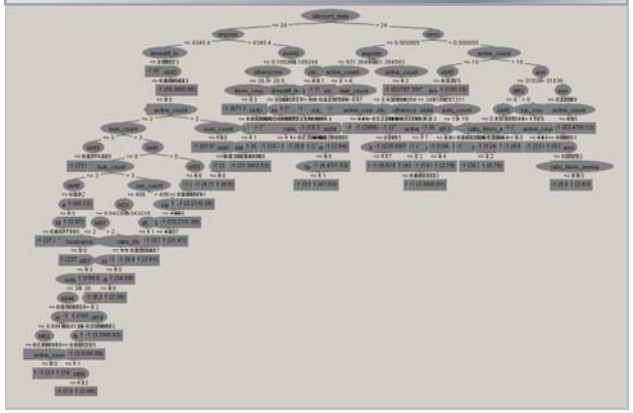


Data-mining methods

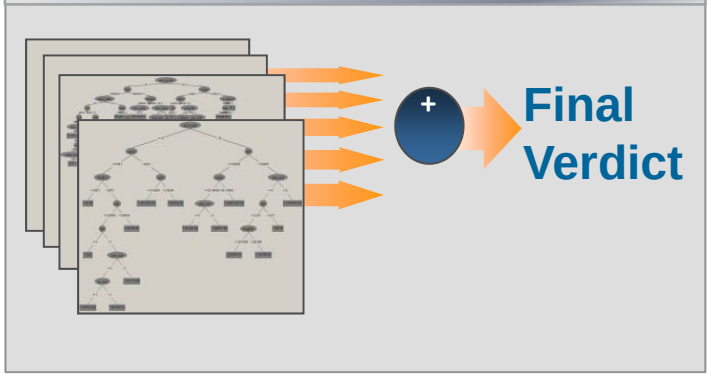
Classification Algorithms

McAfee

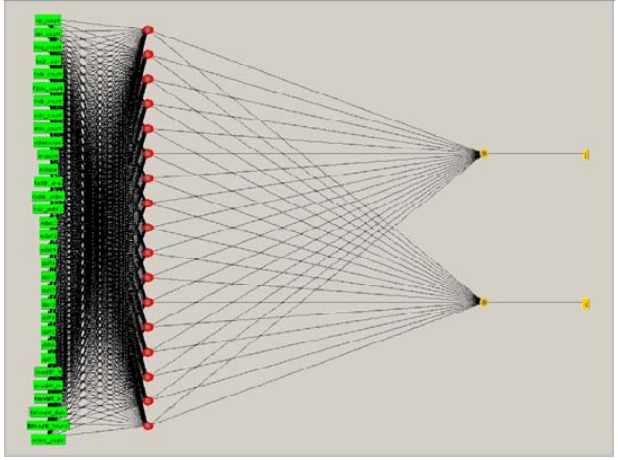
Decision Trees



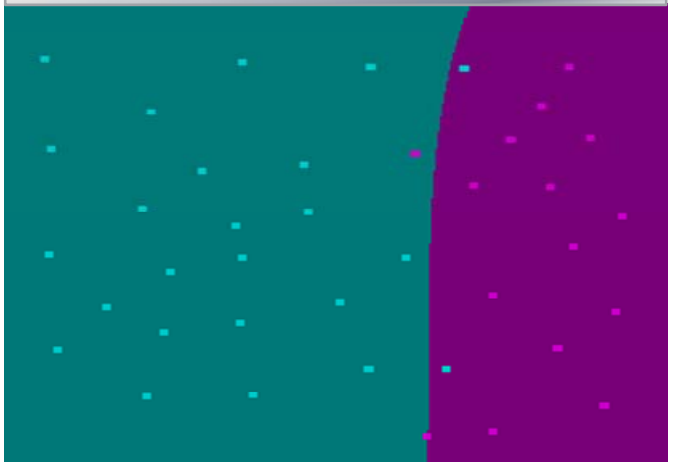
Decision Forests



Neural Networks

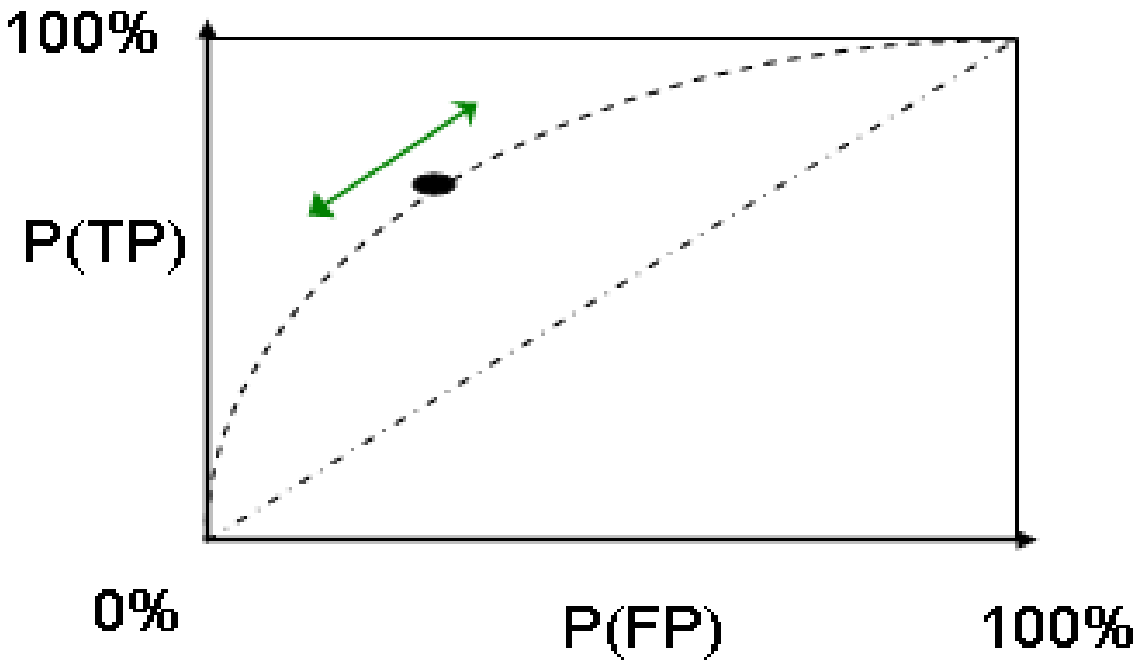


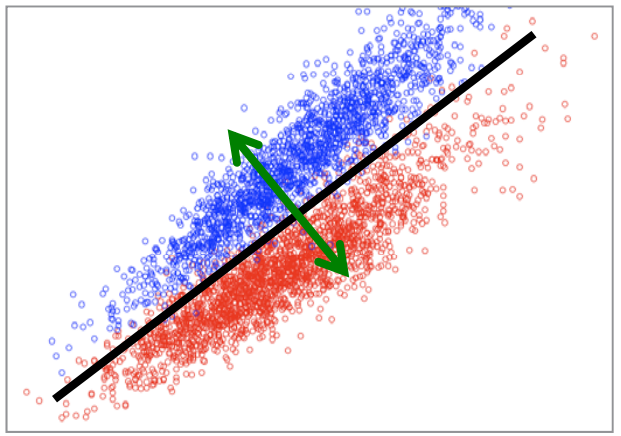
Support Vector Machines



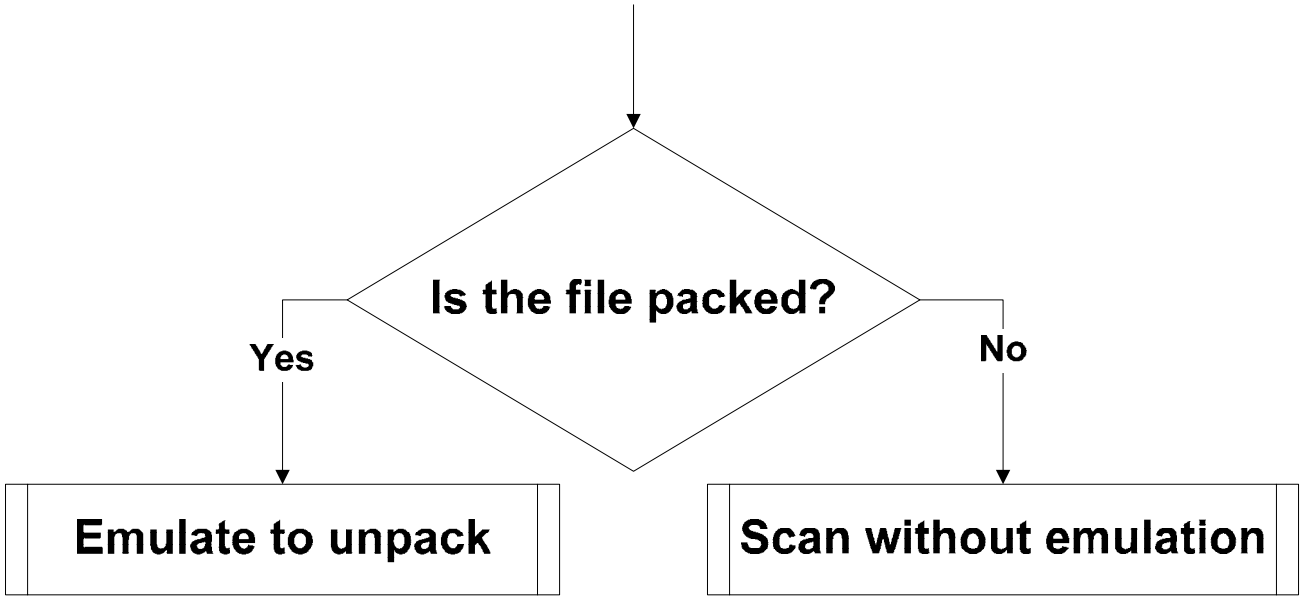
Receiver Operating Characteristic (ROC)

- ROCs graphically represent the quality of separation achieved by data-mining

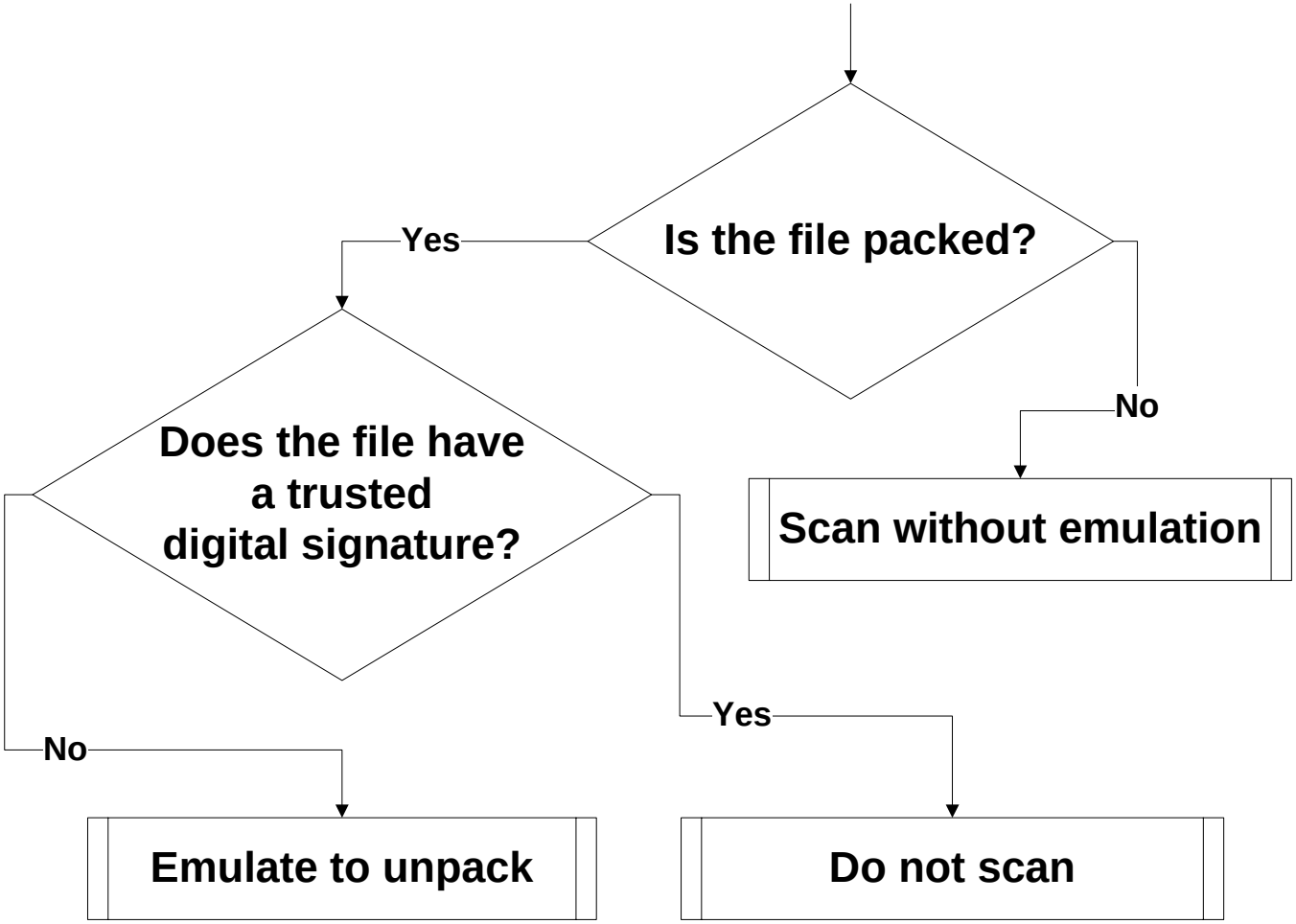




Primitive decision tree



Manually-constructed decision tree (logic)



Feature extraction for PE files

McAfee

- Is file packed = Boolean
- Is file a DLL = Boolean
- HLL language = Integer {1=VB, 2=MSVC, 3=Delphi, 4=.NET, 5=…}
- How many PE sections = Integer
- Are section names standard = Boolean
- Are characteristics of sections standard = Boolean
- Is entry point in the 1st section = Boolean
- Is file digitally signed = Boolean
- Is DOS header standard = Boolean
- Is PE timestamp old = Boolean
- How many resources = Integer
- How many languages in resources = Integer
- Do timestamps in PE header/resources/digisig match = Boolean
- Number of imports = Integer
- Number of exports = Integer
- Number of bound imports = Integer
- Number of imports from WSOCK32.DLL

Steps to build a decision tree

McAfee

Build collections

- Build two TP and FP sets (training sets)
- Setting aside collections for testing (e.g. 10-25%)

Extract features

- From TP set
- From FP set

Model

- Can be expensive for large collections and lots of features
- Pruning (if there is overfitting)

Test

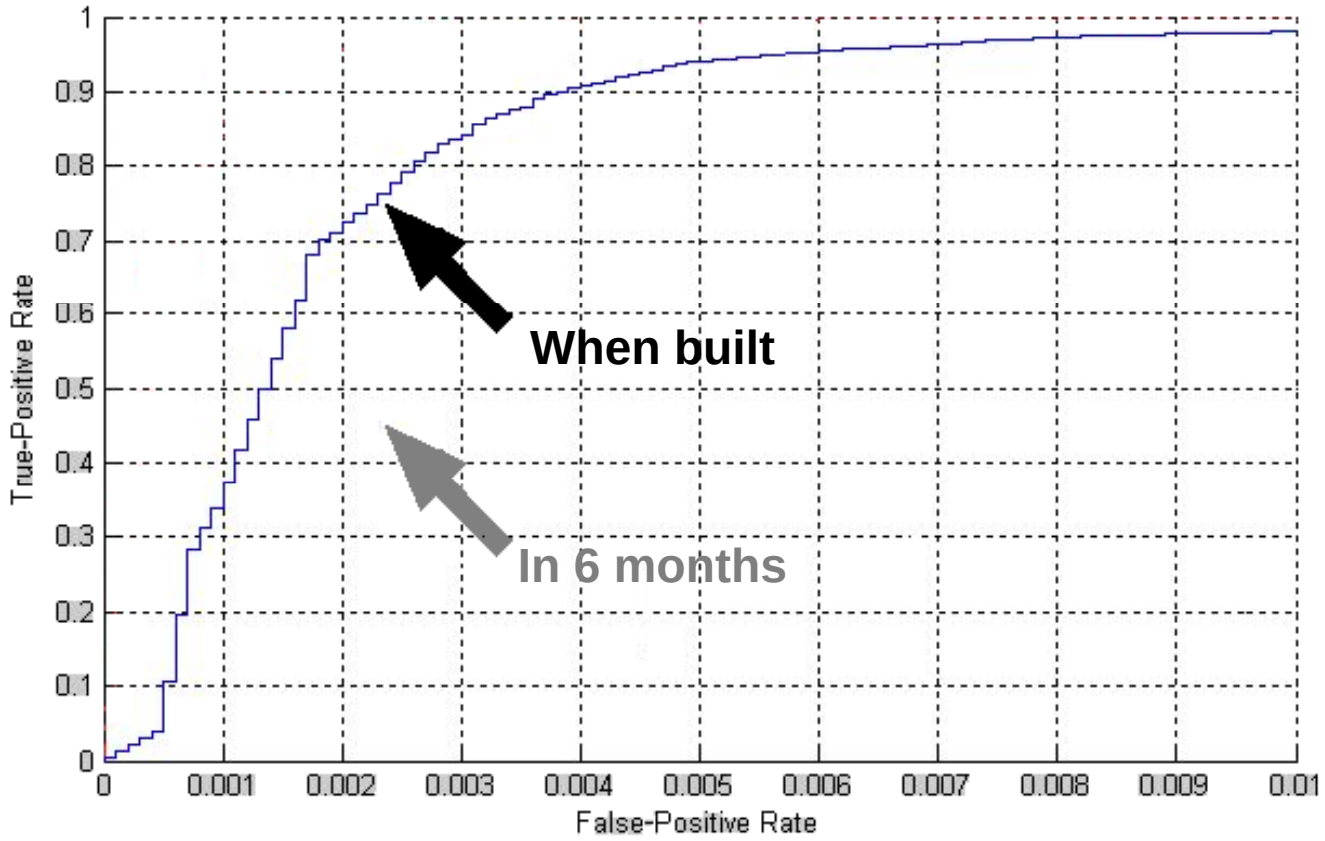
- If the result is not good – make changes and repeat

Convert into usable form

- Decision tree logic into C or C++

ROC for a decision trees we use

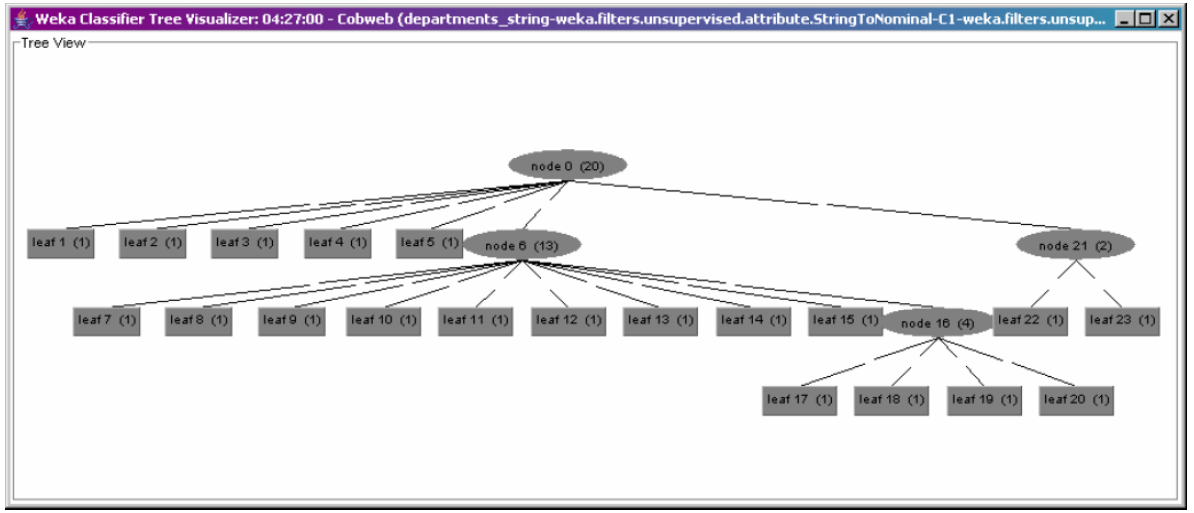
- Decision tree (DT) for PE files with 20 features



Converting logic into code

McAfee

- DT visualized in Weka



- DT as C code

```
if (LSEX <= 0) {
  if (PACK <= 0) {
    if (TIME <= 1174454487) {
      if (TIME <= 708992669) {
        if (DLLB <= 0) {
          if (SIZE <= 1043955) {
            if (CONS <= 0) {
              if (RESD <= 40) {confidence=9892; }
              else {
                if (URLD <= 0) {
                  if (WNET <= 0) {
                    if (OVLY <= 905728) {
                      if (SECS <= 8) {
                        if (WSCK <= 0) {
                          if (IMSZ <= 212) {
                            if (OVLY <= 399050) {
                              if (SIZE <= 66240) {confidence=-5500;}
                              else { confidence=9255;}
                              ...
```

Practical uses of malware data-mining

Practical uses

- To prioritize and de-prioritize samples in research queues
- To drive the depth of the sample analysis on the endpoint, gateway or a server
- To check the most suspicious samples using cloud-based security
- To exclude less suspicious samples from cloud communication
- Applies to anti-malware, anti-spam
- It is an automatic heuristic!

Conclusions

- Data-mining is very useful in malware analysis
- Easy to automate
- Can be used for both “strong” and “weak” heuristic verdicts
  - Detections
  - Limiting
  - Prioritization

Questions