# Intrusion Detection:
# Base Rate Fallacy

Prof. Ravi Sandhu
Executive Director and Endowed Chair

Lecture 8-1

ravi.utsa@gmail.com
www.profsandhu.com

# Testing Outcomes

S: Patient is **S**ick
(has the disease)

R: Test **R**esult
is positive

| | S | ¬S |
|---|---|---|
| R | R ∧ S<br>True positive | R ∧ ¬S<br>False positive |
| ¬R | ¬R ∧ S<br>False negative | ¬R ∧ ¬S<br>True negative |

# Testing Outcomes

S: Patient is **<u>S</u>**ick
(has the disease)
System is under attack

R: Test **<u>R</u>**esult
is positive
Alarm is raised

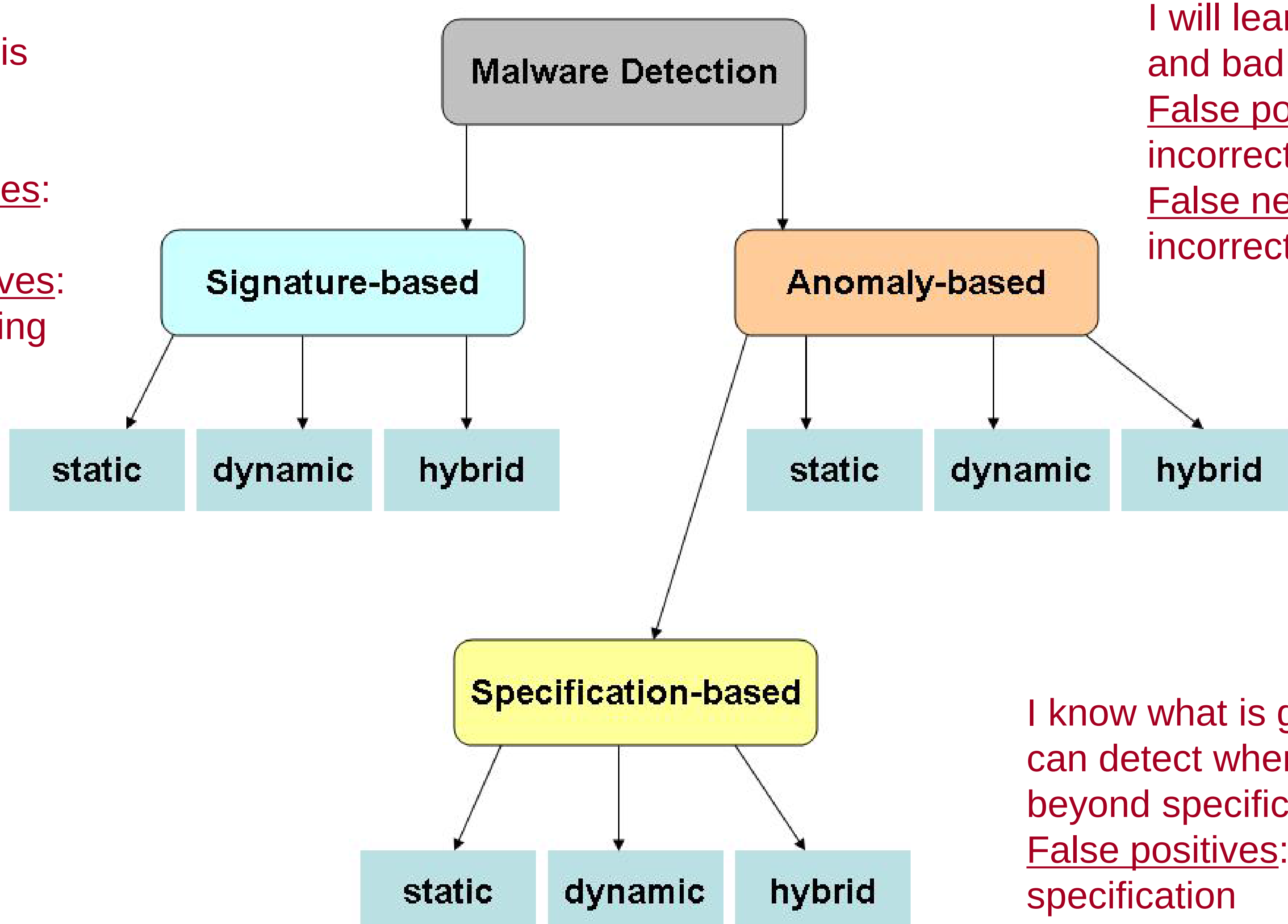
| | S | ¬S |
|---|---|---|
| R | R ∧ S<br>True positive | R ∧ ¬S<br>False positive |
| ¬R | ¬R ∧ S<br>False negative | ¬R ∧ ¬S<br>True negative |

# Malware Detection Techniques



I know what is bad and can detect it
False positives: none
False negatives: ever increasing

I will learn what is good and bad
False positives: incorrect learning
False negatives: incorrect learning




I know what is good and can detect when you go beyond specification
False positives: incomplete specification
False negatives: incorrect specification

Nwokedi Idika and Aditya Mathur, A Survey of Malware Detection Techniques, Purdue University, Feb 2007.

# Testing Outcomes

S: Patient is **<u>S</u>**ick
(has the disease)

R: Test **<u>R</u>**esult
is positive

| | S | ¬S |
|---|---|---|
| R | $R \wedge S$<br>True positive | $R \wedge \neg S$<br>False positive |
| ¬R | $\neg R \wedge S$<br>False negative | $\neg R \wedge \neg S$<br>True negative |

I·C·S
The Institute for Cyber Security

# Testing Outcomes

UTSA

S: Patient is **S**ick
(has the disease)

| | S | ¬S |
|---|---|---|
| R | $R \wedge S$<br>True positive<br>$P(R \mid S) = 0.99$ | $R \wedge \neg S$<br>False positive<br>$P(R \mid \neg S) = 0.01$ |
| ¬R | $\neg R \wedge S$<br>False negative<br>$P(\neg R \mid S) = 0.01$ | $\neg R \wedge \neg S$<br>True negative<br>$P(\neg R \mid \neg S) = 0.99$ |

R: Test **R**esult
is positive

Rows must total between 0 and 2

These probabilities can be empirically estimated

Columns must total 1

# Estimating P(R|S) etc

2000 sick

1000 not sick

| Test R is positive | Test R is negative | Test R is positive | Test R is negative |
| --- | --- | --- | --- |
| 1980 | 20 | 10 | 990 |

estimate $P(R|S) = 0.99$ $P(\neg R|S) = 0.01$ $P(R|\neg S) = 0.01$ $P(\neg R|\neg S) = 0.99$

Coincidentally equal

I·C·S
The Institute for Cyber Security

# Estimating P(R|S) etc

UTSA

2000 sick

1000 not sick

Test R is positive

1980

Test R is negative

20

Test R is positive

30

Test R is negative

970

estimate P(R|S) = 0.99 P(¬R|S) = 0.01 P(R|¬S) = 0.03 P(¬R|¬S) = 0.97

In general will not be equal

# Base-Rate Fallacy

S: Patient is **S**ick
(has the disease)

R: Test **R**esult
is positive

| | S | ¬S |
|---|---|---|
| R | R ∧ S<br>True positive<br>P(R\|S) = 0.99 | R ∧ ¬S<br>False positive<br>P(R\|¬S) = 0.03 |
| ¬R | ¬R ∧ S<br>False negative<br>P(¬R\|S) = 0.01 | ¬R ∧ ¬S<br>True negative<br>P(¬R\|¬S) = 0.97 |

Rows must total between 0 and 2

These probabilities can be empirically estimated

Columns must total 1

# Base-Rate Fallacy



We will continue with these numbers

S: Patient is **S**ick (has the disease)

R: Test **R**esult is positive

| | S | ¬S |
|---|---|---|
| R | R ∧ S<br>True positive<br>P(R\|S) = 0.99 | R ∧ ¬S<br>False positive<br>P(R\|¬S) = 0.01 |
| ¬R | ¬R ∧ S<br>False negative<br>P(¬R\|S) = 0.01 | ¬R ∧ ¬S<br>True negative<br>P(¬R\|¬S) = 0.99 |

These probabilities can be empirically estimated

# Real Interest

S: Patient is **S**ick
(has the disease)

R: Test **R**esult
is positive

| | S | ¬S |
|---|---|---|
| R | R ∧ S<br>True positive<br>P(S\|R) = ?? | R ∧ ¬S<br>False positive<br>P(¬S\|R) = ?? |
| ¬R | ¬R ∧ S<br>False negative<br>P(S\|¬R) = ?? | ¬R ∧ ¬S<br>True negative<br>P(¬S\|¬R) = ?? |

Rows must total 1

Columns must total between 0 and 2

These probabilities can be computed by Bayes' theorem if we know P(S)

# Bayes' Theorem

- $P(S|R) = (P(S) \times P(R|S))/(P(S) \times P(R|S)+P(\neg S)\ ) \times P(R|\neg S))$

- $P(\neg S|R) = 1 - P(S|R)$

- $P(S|\neg R) = (P(S) \times P(\neg R|S))/(P(S) \times P(\neg R|S)+P(\neg S)\ ) \times P(\neg R|\neg S))$

- $P(\neg S|\neg R) = 1 - P(S|\neg R)$

# Test Outcomes



We will continue with these numbers

S: Patient is **S**ick (has the disease)

R: Test **R**esult is positive

| | S | ¬S |
|---|---|---|
| R | R ∧ S<br>True positive<br>P(R\|S) = 0.99 | R ∧ ¬S<br>False positive<br>P(R\|¬S) = 0.01 |
| ¬R | ¬R ∧ S<br>False negative<br>P(¬R\|S) = 0.01 | ¬R ∧ ¬S<br>True negative<br>P(¬R\|¬S) = 0.99 |

These probabilities can be empirically estimated

# Base Rate

Assume
P(S)=0.0001
1 in 10,000 has disease

S: Patient is **S**ick
(has the disease)

R: Test **R**esult is positive

| | S | ¬S |
|---|---|---|
| R | R ∧ S<br>True positive<br>P(S\|R) = 0.009804 | R ∧ ¬S<br>False positive<br>P(¬S\|R) = 0.990196 |
| ¬R | ¬R ∧ S<br>False negative<br>P(S\|¬R) = 0.000001 | ¬R ∧ ¬S<br>True negative<br>P(¬S\|¬R) = 0.999999 |

Rows must total 1

These probabilities can be computed by Bayes' theorem if we know P(S)

Columns must total between 0 and 2

# False Alarms Predominate!

Assume
P(S)=0.0001
1 in 10,000 has disease

| P(S\|R) | requires | P(R\|¬S) |
|---|---|---|
| 0.01 | | 0.01 |
| 0.09 | | 0.001 |
| 0.5 | | 0.0001 |
| 0.9 | | 0.00001 |
| 0.99 | | 0.000001 |

# Base-Rate Fallacy

S: Patient is **S**ick
(has the disease)

Total population = 1,000,000
1 in 10,000 has disease

R: Test **R**esult
is positive

R is 99% accurate
for sick and non-sick
populations

| | S<br>100 | ¬S<br>999,900 |
|---|---|---|
| R | R ∧ S<br>True positive | R ∧ ¬S<br>False positive |
| ¬R | ¬R ∧ S<br>False negative | ¬R ∧ ¬S<br>True negative |

# Base-Rate Fallacy

S: Patient is **S**ick
(has the disease)

Total population = 1,000,000
1 in 10,000 has disease

| | S<br>100 | ¬S<br>999,900 |
|---|---|---|
| R | R ∧ S<br>True positive<br>99 | R ∧ ¬S<br>False positive<br>9,999 |
| ¬R | ¬R ∧ S<br>False negative<br>1 | ¬R ∧ ¬S<br>True negative<br>989,901 |

R: Test **R**esult is positive

R is 99% accurate for sick and non-sick populations