# Intrusion Detection: Base Rate Fallacy

Prof. Ravi Sandhu
Executive Director and Endowed Chair

Lecture 11

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

# Base-Rate Fallacy

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

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

R: Test **R**esult
is positive

# Base-Rate Fallacy

S: Patient is **S**ick
(has the disease)
System is under attack

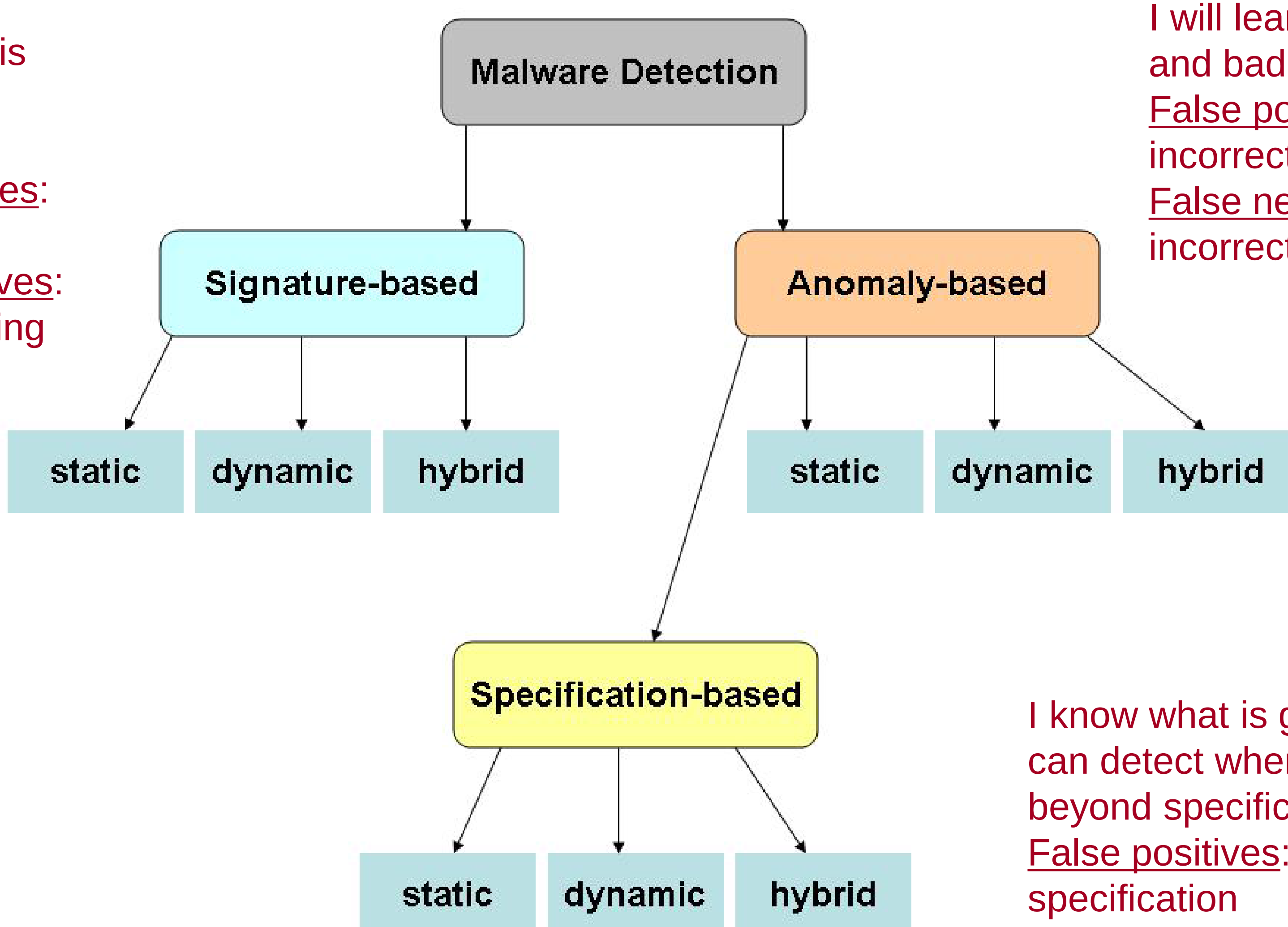
R: Test **R**esult
is positive
Alarm is raised

| | 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
<u>False positives</u>: none
<u>False negatives</u>: ever increasing

I will learn what is good and bad
<u>False positives</u>: incorrect learning
<u>False negatives</u>: incorrect learning




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

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

# 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 | 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)

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

# 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

# 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 | 30 | 970 |

estimate $P(R|S) = 0.99$ $P(\neg R|S) = 0.01$ $P(R|\neg S) = 0.03$ $P(\neg R|\neg S) = 0.97$

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 | $R \wedge S$<br>True positive<br>$P(R \mid S) = 0.99$ | $R \wedge \neg S$<br>False positive<br>$P(R \mid \neg S) = 0.03$ |
| ¬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>$P(\neg R \mid \neg S) = 0.97$ |

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

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

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

# 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

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

Rows must total 1

Columns must total between 0 and 2

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

# 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

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

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

R: Test **R**esult
is positive

| | 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 |

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

# 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