Base Rate Fallacy

Prof. Ravi Sandhu
Executive Director and Endowed Chair

Lecture 13

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

Base-Rate Fallacy

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

R: Test **R**esult
is positive

	S	¬S
R	R ∧ S True positive	R ∧ ¬S False positive
¬R	¬R ∧ S False negative	¬R ∧ ¬S True negative

Base-Rate Fallacy

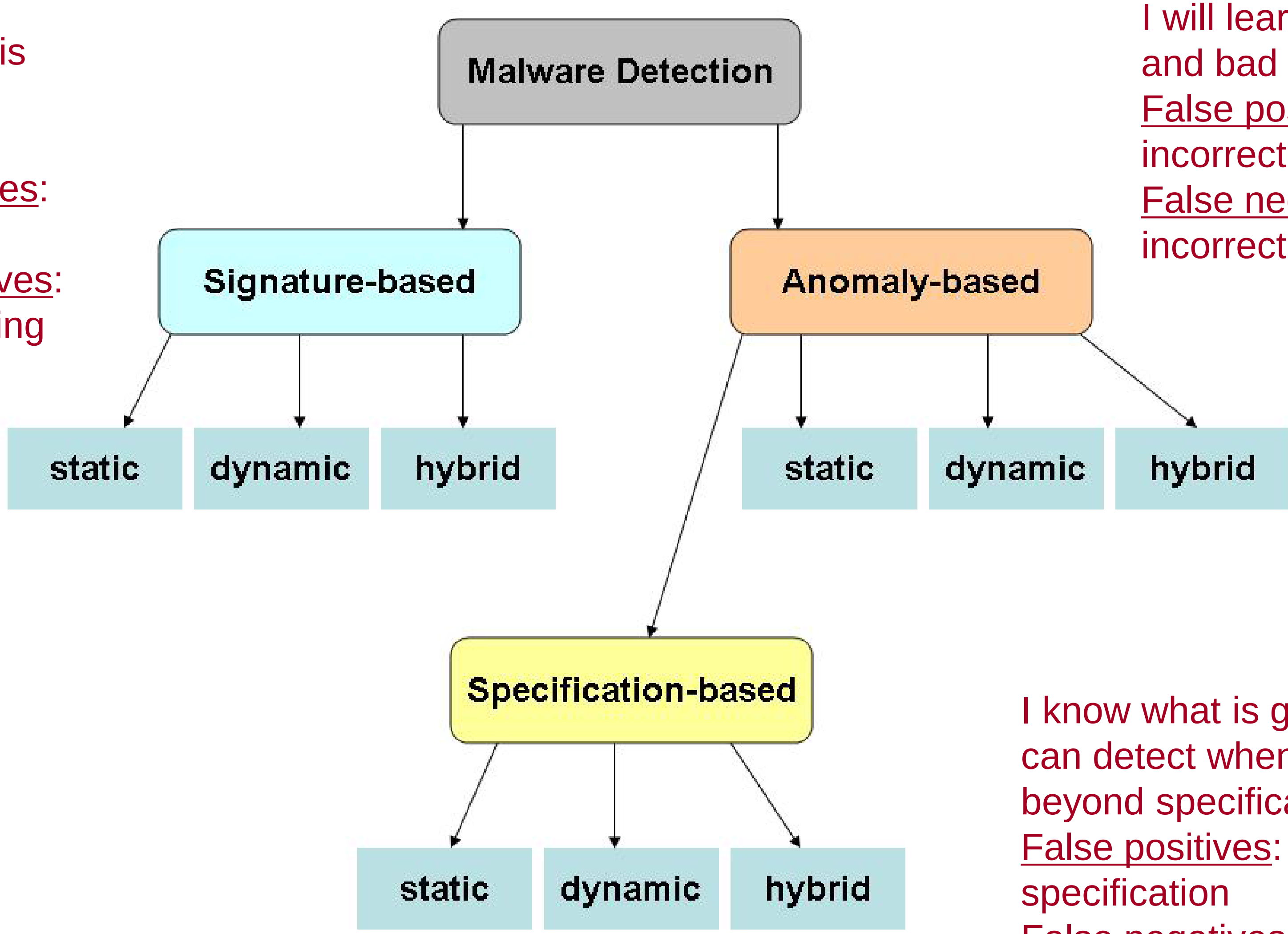
S: Patient is **S**ick
(has the disease)
System is under attack

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

	S	¬S
R	$R \wedge S$ True positive	$R \wedge \neg S$ False positive
¬R	$\neg R \wedge S$ False negative	$\neg R \wedge \neg S$ 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.

Base-Rate Fallacy

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

R: Test **R**esult
is positive

	S	¬S
R	R ∧ S True positive	R ∧ ¬S False positive
¬R	¬R ∧ S False negative	¬R ∧ ¬S True negative

Base-Rate Fallacy

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

R: Test **R**esult
is positive

	S	¬S
R	$R \wedge S$ True positive $P(R\|S) = 0.99$	$R \wedge \neg S$ False positive $P(R\|\neg S) = 0.01$
¬R	$\neg R \wedge S$ False negative $P(\neg R\|S) = 0.01$	$\neg R \wedge \neg S$ True negative $P(\neg R\|\neg S) = 0.99$

These probabilities can be empirically estimated

Estimating P(R|S) etc

2000 sick

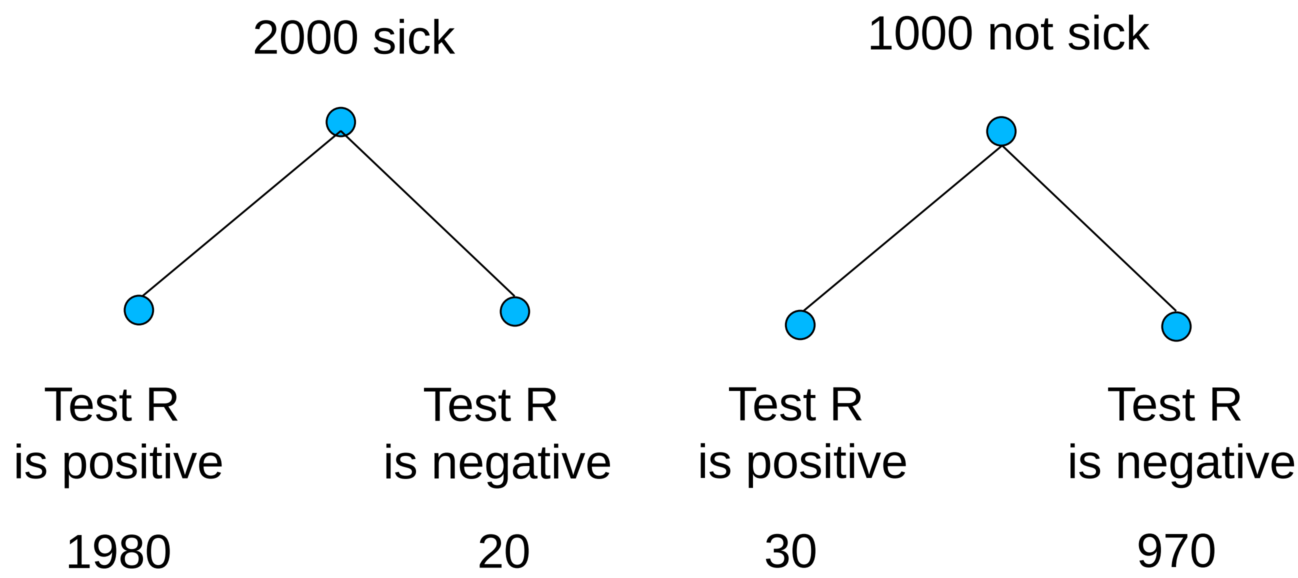
1000 not sick

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

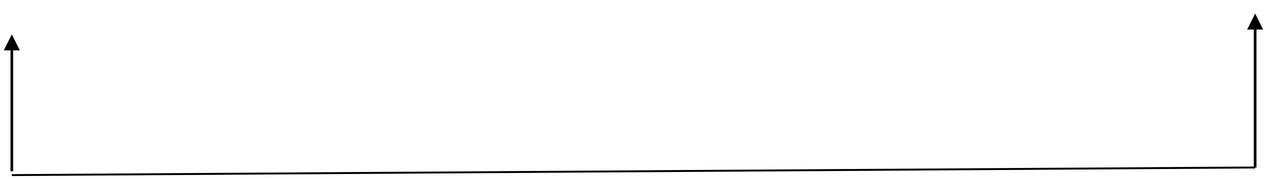
estimate P(R|S) = 0.99 P(¬R|S) = 0.01 P(R|¬S) = 0.01 P(¬R|¬S) = 0.99

Coincidentally equal

Estimating P(R|S) etc



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 True positive P(R|S) = 0.99	R ∧ ¬S False positive P(R|¬S) = 0.03
¬R	¬R ∧ S False negative P(¬R|S) = 0.01	¬R ∧ ¬S True negative 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 True positive P(R|S) = 0.99	R ∧ ¬S False positive P(R|¬S) = 0.01
¬R	¬R ∧ S False negative P(¬R|S) = 0.01	¬R ∧ ¬S True negative 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 True positive $P(S \mid R) = ??$	R ∧ ¬S False positive $P(\neg S \mid R) = ??$
¬R	¬R ∧ S False negative $P(S \mid \neg R) = ??$	¬R ∧ ¬S True negative $P(\neg S \mid \neg R) = ??$

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 True positive $P(R \mid S) = 0.99$	R ∧ ¬S False positive $P(R \mid \neg S) = 0.01$
¬R	¬R ∧ S False negative $P(\neg R \mid S) = 0.01$	¬R ∧ ¬S True negative $P(\neg R \mid \neg S) = 0.99$

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 ∧ S True positive P(S|R) = 0.009804	R ∧ ¬S False positive P(¬S|R) = 0.990196
¬R	¬R ∧ S False negative P(S|¬R) = 0.000001	¬R ∧ ¬S True negative P(¬S|¬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

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

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

		S	¬S
		100	999,900
R: Test **R**esult is positive	R	R ∧ S True positive	R ∧ ¬S False positive
	¬R	¬R ∧ S False negative	¬R ∧ ¬S 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

R: Test **R**esult is positive

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

	S 100	¬S 999,900
R	R ∧ S True positive 99	R ∧ ¬S False positive 9,999
¬R	¬R ∧ S False negative 1	¬R ∧ ¬S True negative 989,901