

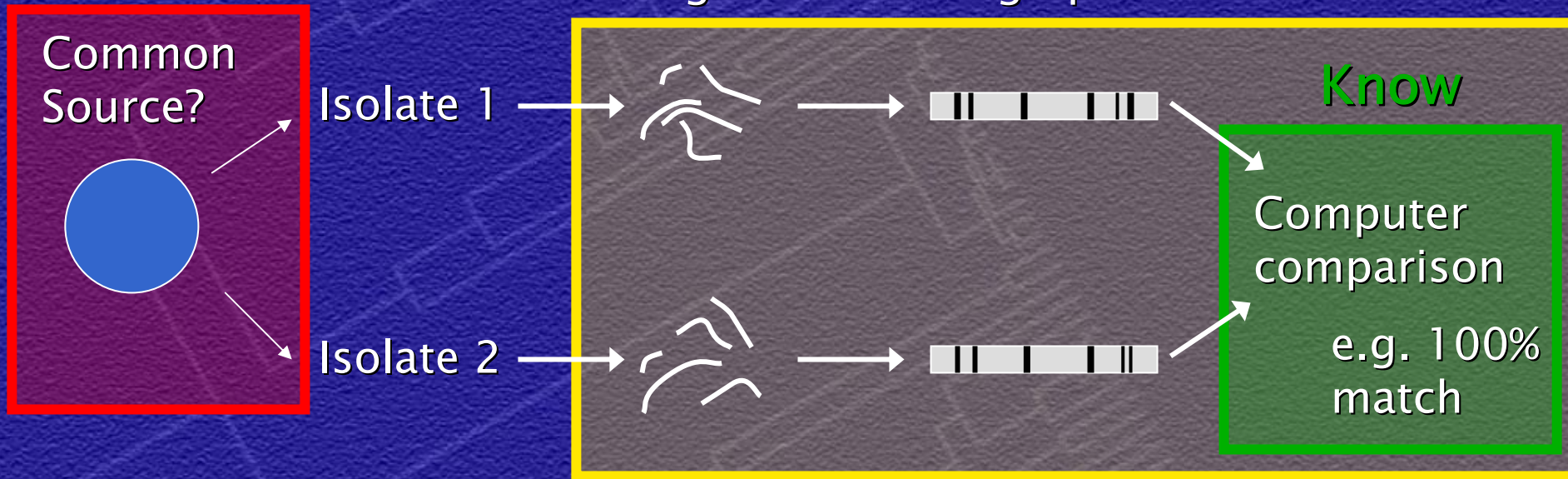
Some elementary statistics to address:

# The significance of computer-reported PFGE matches

Paul Vauterin  
Applied Maths

- Significance of two matches
- Significance of clusters in time

## Want to know



**Hypothesis:** DNA fragments are the same

**Test:** computerized comparison of fingerprints, using a certain tolerance for matching

**Criterion:** e.g. 100% similarity

**How reliable?**

|           |                     |                       |                          |
|-----------|---------------------|-----------------------|--------------------------|
|           |                     | ↓ Test ↓              |                          |
|           |                     | Computer match        |                          |
|           |                     | Yes                   | No                       |
| Reality ↓ | Same fragments<br>↓ | Yes<br>Correct<br>(I) | No<br>False<br>negatives |
|           | No                  | False<br>positives    | Correct<br>(II)          |

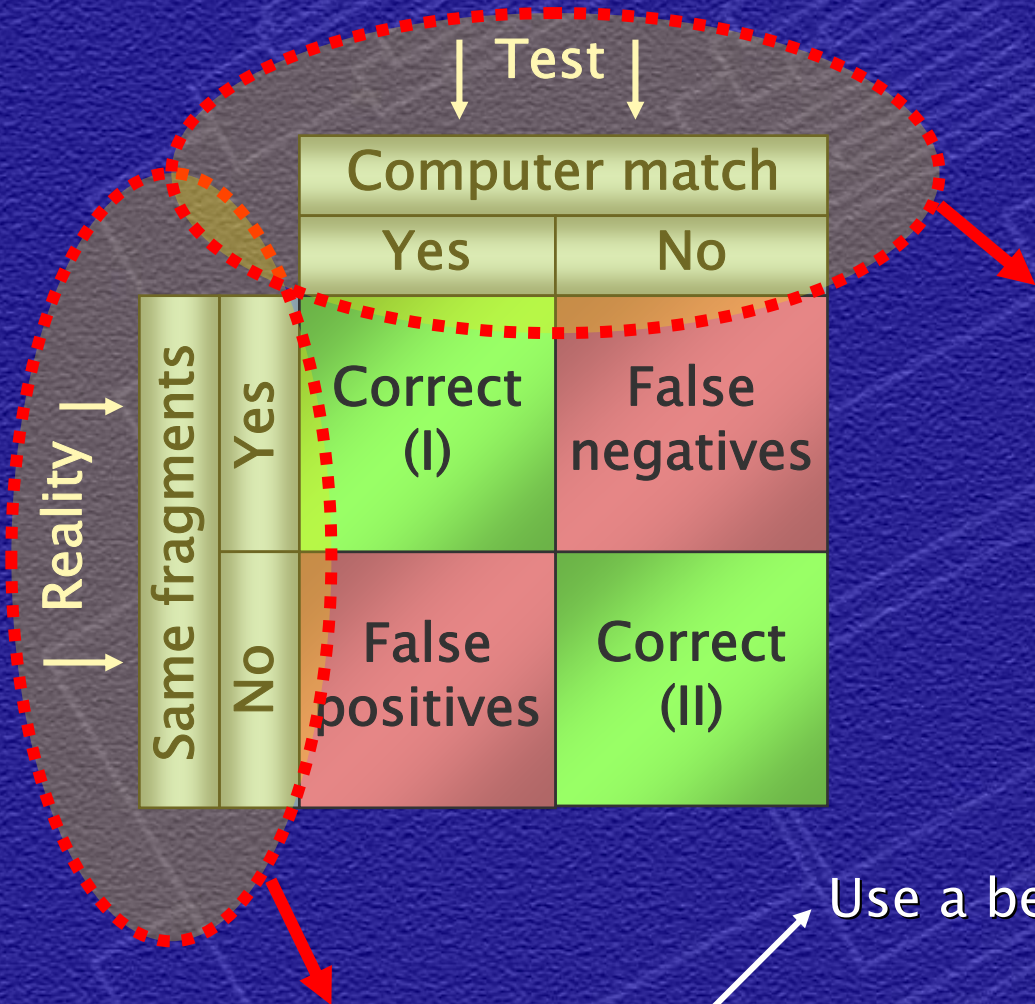
False negatives (FN) and false positives (FP) will always exist. The goal is to keep them as low as possible

→ optimize PFGE & processing (normalization, tolerance, ...)

$$\text{False negatives rate (FNR)} = \frac{\text{False negatives}}{\text{False negatives} + \text{Correct (I)}}$$

$$\text{False positives rate (FPR)} = \frac{\text{False positives}}{\text{False positives} + \text{Correct (II)}}$$

# How to determine?



Straightforward  
(computer matches)

Hard (this is what we  
want to know!)

Use a benchmark set (known isolates)

Use pattern names

(= tests the ability of the computer  
to reproduce visual assignments)

## Example:

Ecoli O157 H7, XbaI, 400 fingerprints,  $\leq 1$  band different  
FP / FN as a function of matching tolerance



Tolerance = maximum shift allowed  
for a pair of bands to be matching

In data set of 400 fingerprints:

- Take all possible pairs
- Determine band matching
- Compare to pattern names

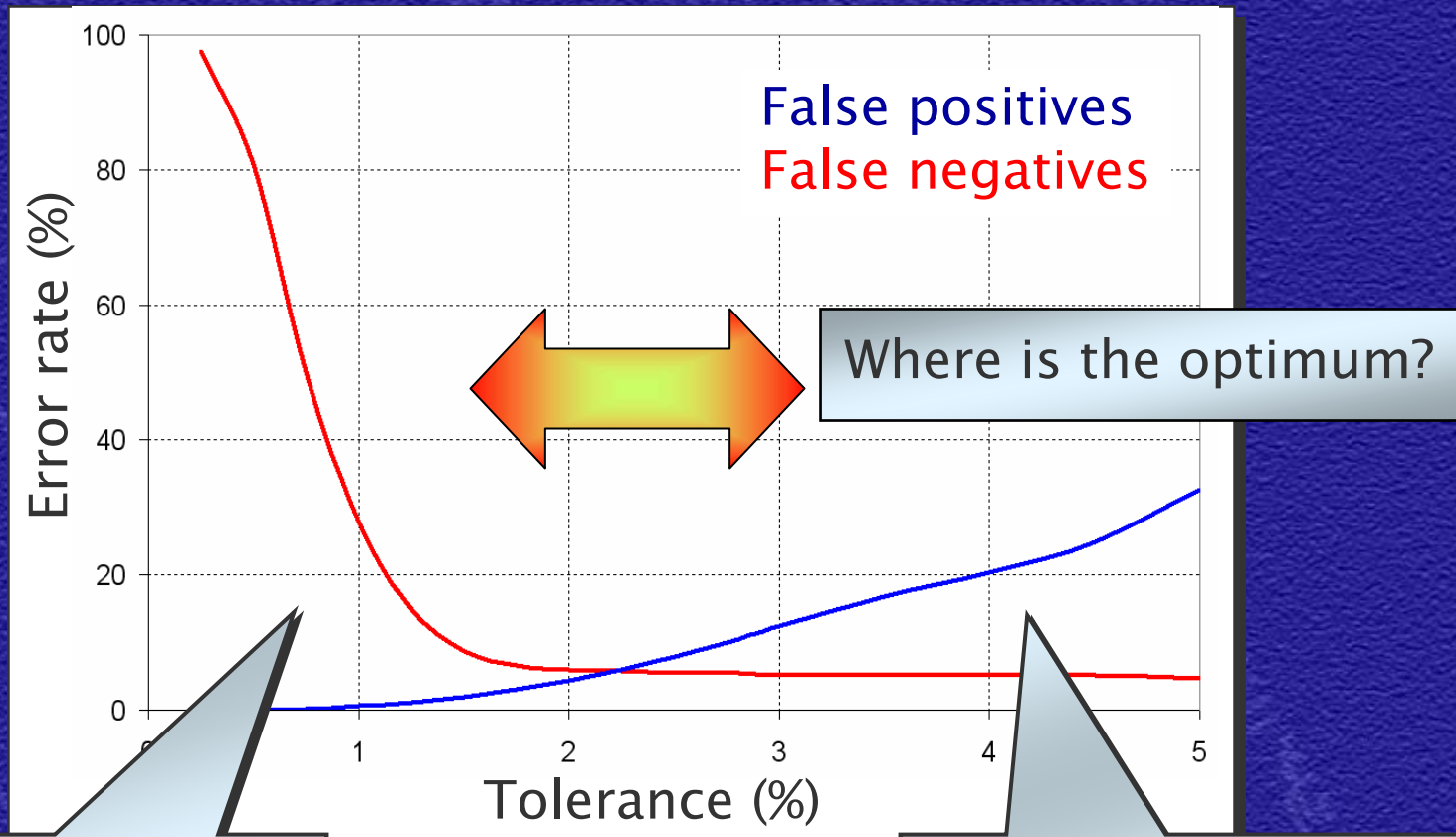
With varying tolerance!

|                   |     | Computer match     |                    |
|-------------------|-----|--------------------|--------------------|
|                   |     | Yes                | No                 |
| Same pattern name | Yes | Correct<br>(I)     | False<br>negatives |
|                   | No  | False<br>positives | Correct<br>(II)    |

## Example:

Ecoli O157 H7, Xbal, 400 fingerprints,  $\leq 1$  band different

As a function of matching tolerance

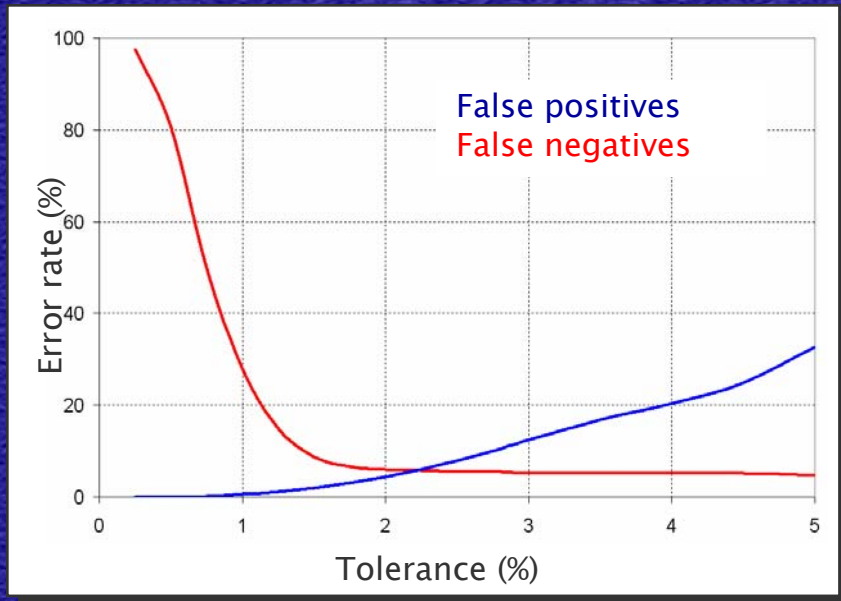


Too strict:

- few false positives
- many false negatives

Too relaxed:

- many false positives
- few false negatives



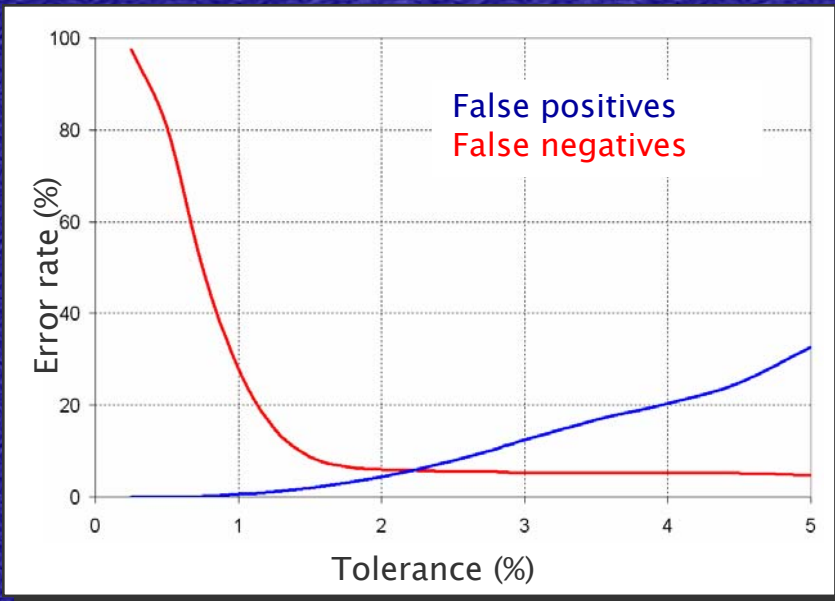
Sweet point =  
find a balance between two curses

Important observation:  
the consequences of both types of errors are not the same !!

**False positive** : an incorrect candidate match appears in the list  
→ Noticed & corrected down the road

**False negative** : an correct match is not shown  
→ May never be picked up (=lost case)

*... Compare it to a spam email filter...*



Sweet point =  
find a balance between two curses

-> introduce the notion of a “cost of mistake”

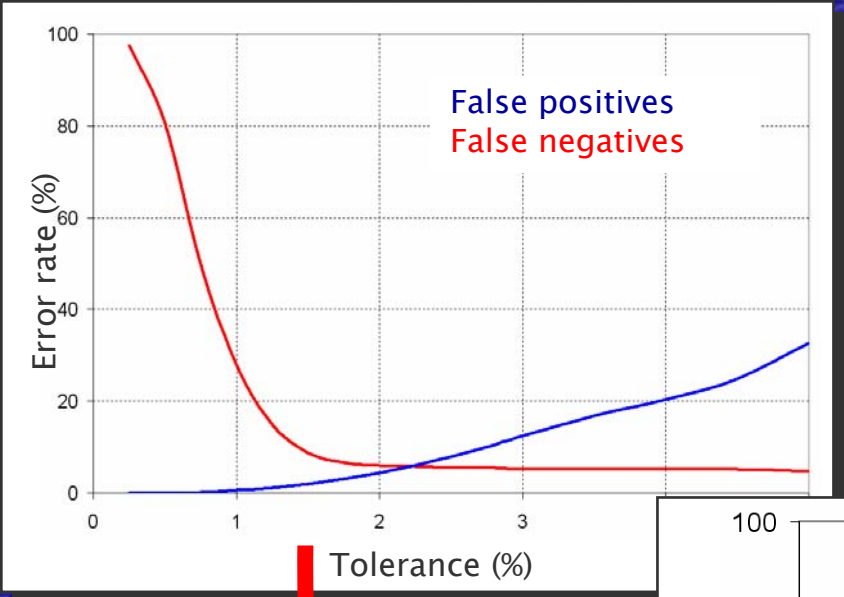
$$\text{Overall cost of mistakes} = \text{FNR} \times \text{CostFN} + \text{FPR} \times \text{CostFP}$$

FNR = False Negatives Rate

CostFN = Cost of a false negative

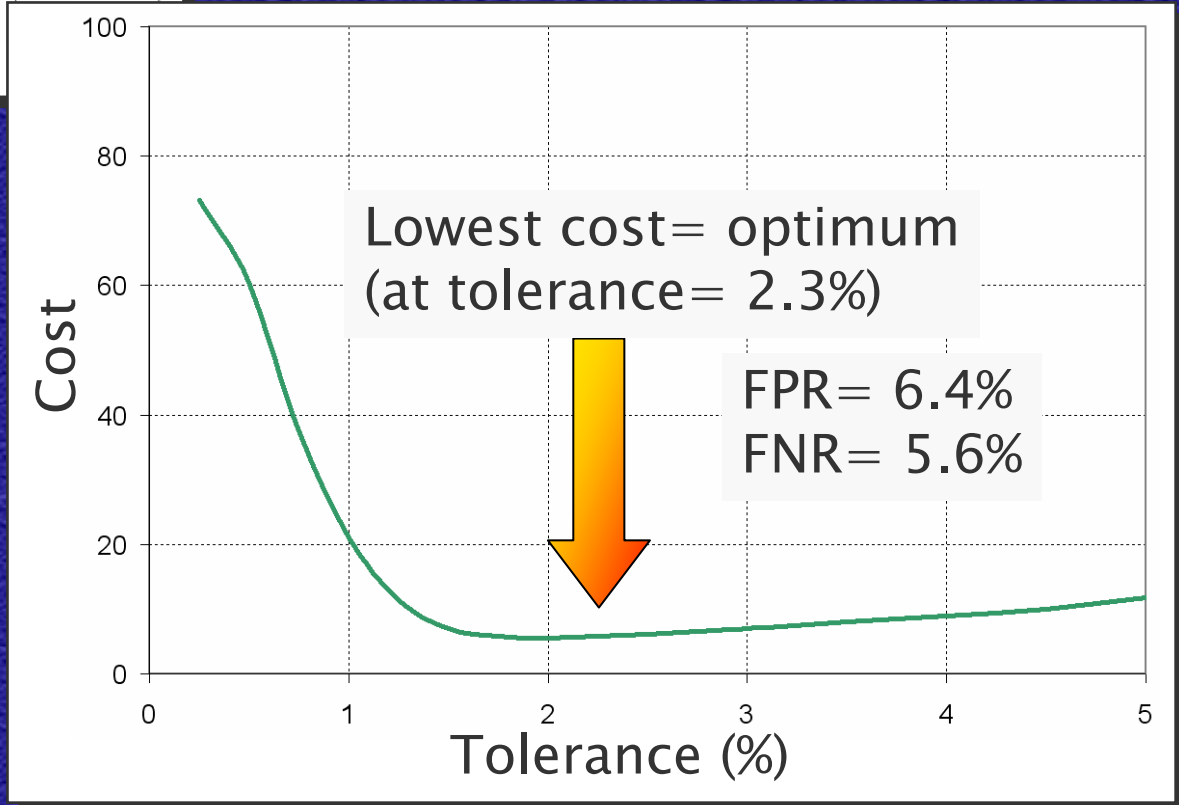
FPR = False Positives Rate

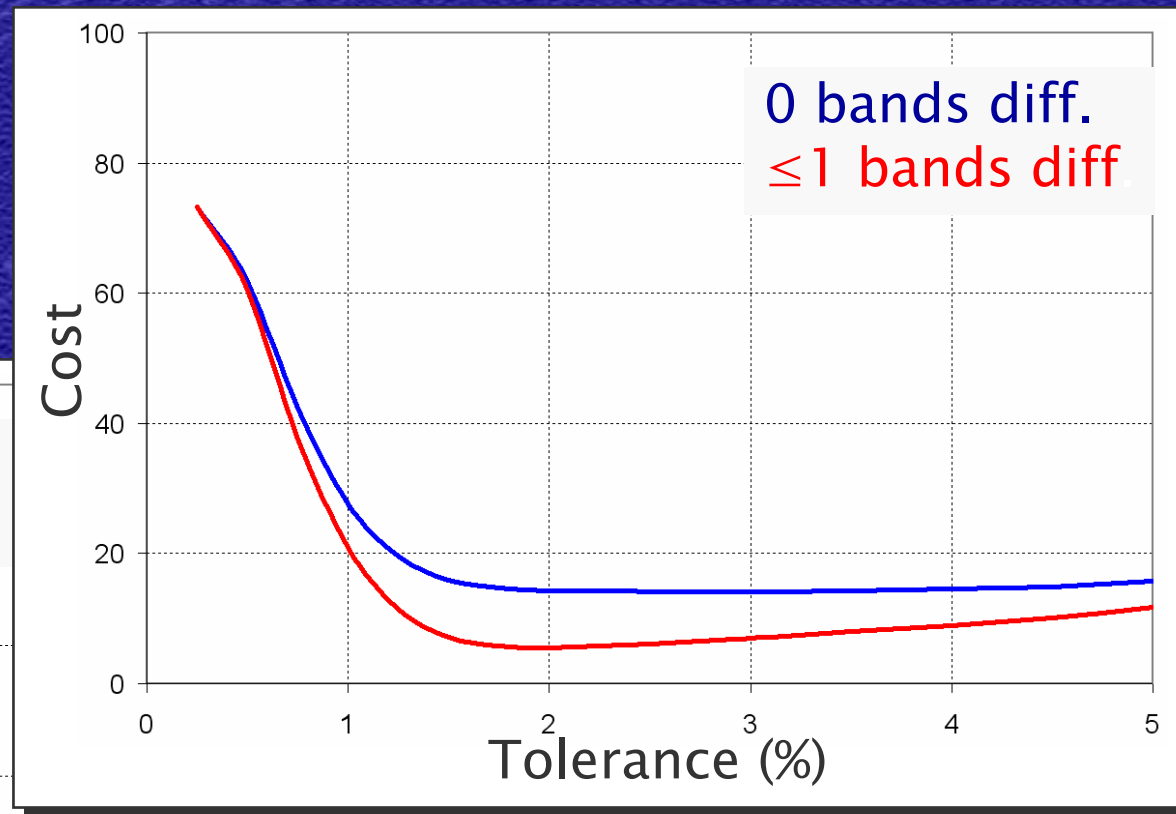
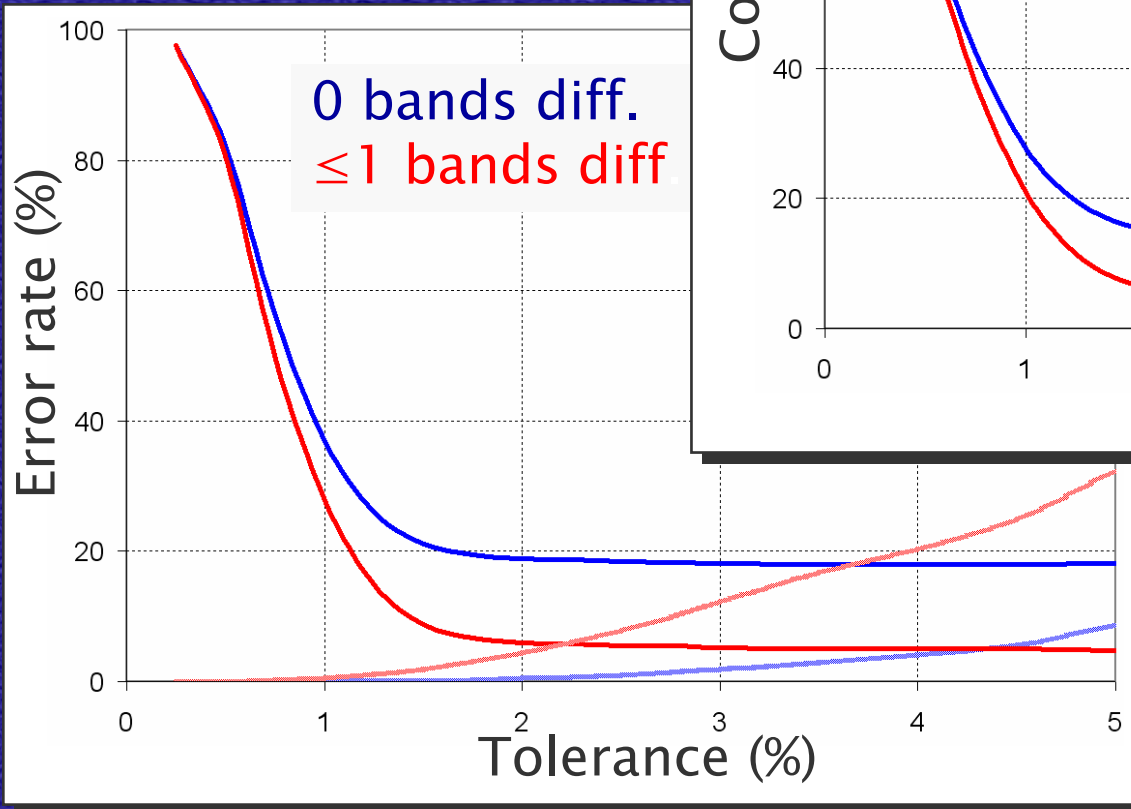
CostFP = Cost of a false positive



Sweet point =  
find a balance between two curses

$Cost_{FN} = 3 \times Cost_{FP}$





- $\leq 1$  bands difference better than 0 bands difference
- High tolerance is better ( $> 2\%$ )

## Further work needs to be done:

- Larger data set
- Serotype dependency?
- Use a set of known repeated isolates as benchmark (proficiency testing data set?)

# Significance of clusters in time

For example: pattern EXHX01.0086 has been found 4x in a period of 12 days

Is this normal, or exceptional?

- calculate “p-value” of this set of coincidences  
= probability that this occurs, given the normal background rate of this pattern (= nothing special going on)
- low p-value is a hint that something might be going on (unlikely to be explained by background rate)

## Relevant ingredients:

- 'Natural' frequency of the pattern (is it rare? is it common?)
- Number of coincidences
- Time period of these coincidences

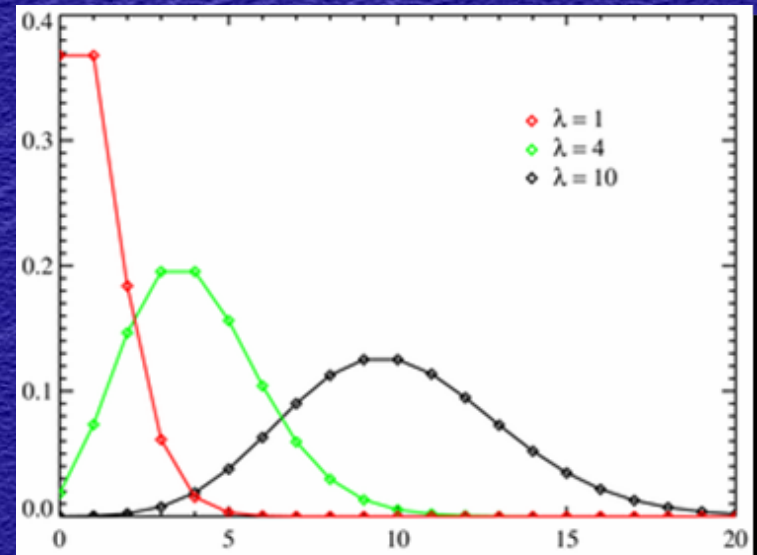


## Probability theory:

Probability to find a certain number of events over a certain time span, given the average event rate

= Poisson distribution

$$f(k, \lambda) = \frac{e^{-\lambda} \lambda^k}{k!}$$



$K$  events, during a period of  $Q$  days,  
given a pattern frequency of  $F$  patterns per year

Theory:

$$P(\text{\#events} \geq K)$$

$$P(\text{\#events} \geq 1)$$

P(...) from cumulative Poisson distribution:

$$P(\text{\#events} \geq K) = 1 - \sum_{i=0}^{K-1} f(\lambda, i)$$

$$f(k, \lambda) = \frac{e^{-\lambda} \lambda^k}{k!}$$

with

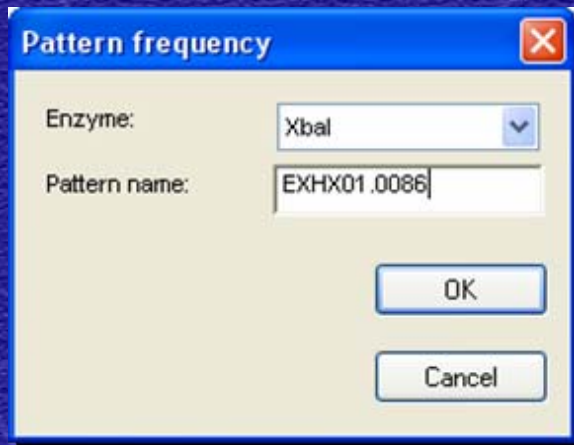
$$\lambda = \frac{Q \times F}{365.25}$$

**Reality = as easy as running a script in BN !**

# How to use

Pattern EXHX01.0086 has been found 4x in a period of 12 days

- Run script “getpatternfreq” (“getpatternfreq\_srv” for server)



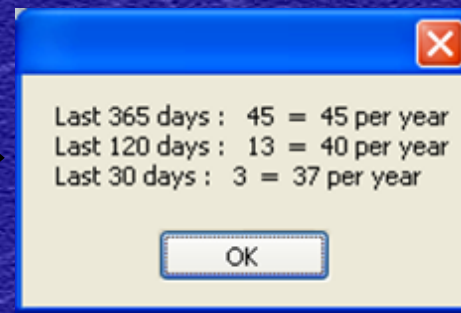
Pattern frequency

Enzyme: Xbal

Pattern name: EXHX01.0086

OK

Cancel

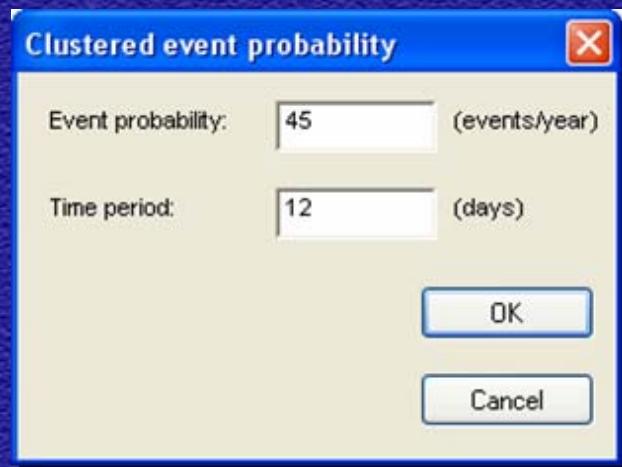


Last 365 days : 45 = 45 per year  
Last 120 days : 13 = 40 per year  
Last 30 days : 3 = 37 per year

OK

(based on IsolationDate)

- Run script “poissonstat”



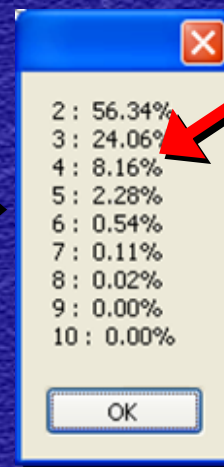
Clustered event probability

Event probability: 45 (events/year)

Time period: 12 (days)

OK

Cancel



2 : 56.34%  
3 : 24.06%  
4 : 8.16%  
5 : 2.28%  
6 : 0.54%  
7 : 0.11%  
8 : 0.02%  
9 : 0.00%  
10 : 0.00%

OK

4 events:  
p-value = 8.16%  
= ‘a bit’ significant?

## Some warnings:

- For a new pattern, a certain 'background' rate is assumed (once every 6 months)
- Batched processing may lump dates together  
→ always take a reasonable minimum period
- Simple model → Take these numbers with a grain of salt

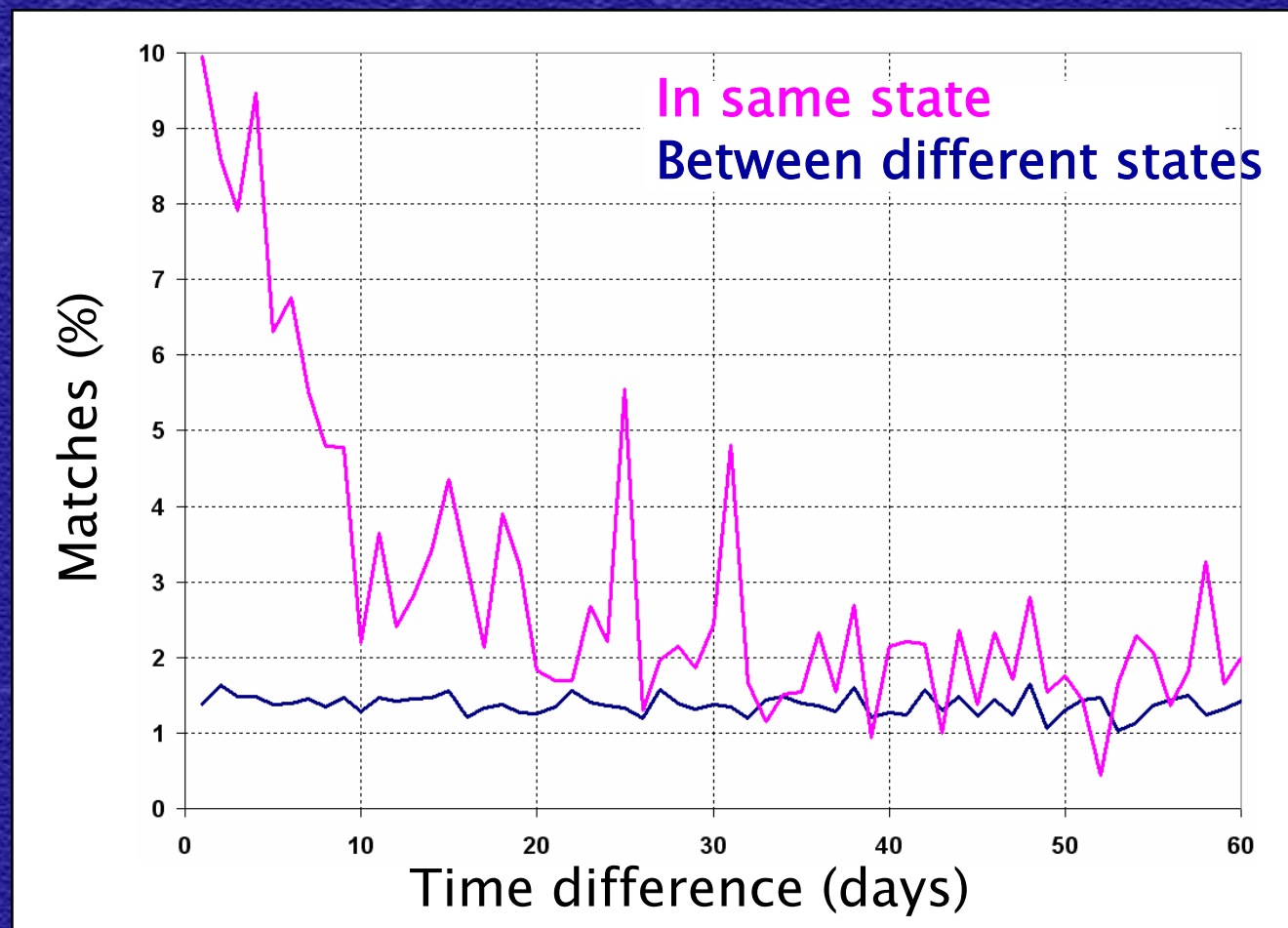
**!!! NEVER USE AS PROOF OF OUTBREAK !!!**

(too far away from actual evidence)

but could be a red flag that something might be going on

A little extra to conclude...

- Take all pairs of matching fingerprints in a data set
- Plot # of matches as a function of time between them



# Conclusions

- **False Positives / False Negatives analysis**

- = powerful tool to investigate the quality of a technique

- = useful to optimize certain process parameters

- Tolerance for PFGE

- Choice of loci for MLVA ...

- **Poisson Statistics on time series**

- = can be used to get some idea about

- the 'oddness' of a set of coincidences over a time period

- **Database = goldmine ready for exploitation**

- some simple graphs can already tell a lot!