Bayes' rule

Example: Fraud Detection in Credit Card Transactions

Imagine a bank is using an algorithm to detect potential credit card fraud.

- Let *F represent a fraudulent transaction*
- *Let T represent a transaction flagged as fraudulent*

We know the so called ***Sensitivity*** $P(T|F)$: The probability that the algorithm flags a transaction as fraudulent given that it is actually fraudulent.

However, what the bank (and its customers) really want to know is:

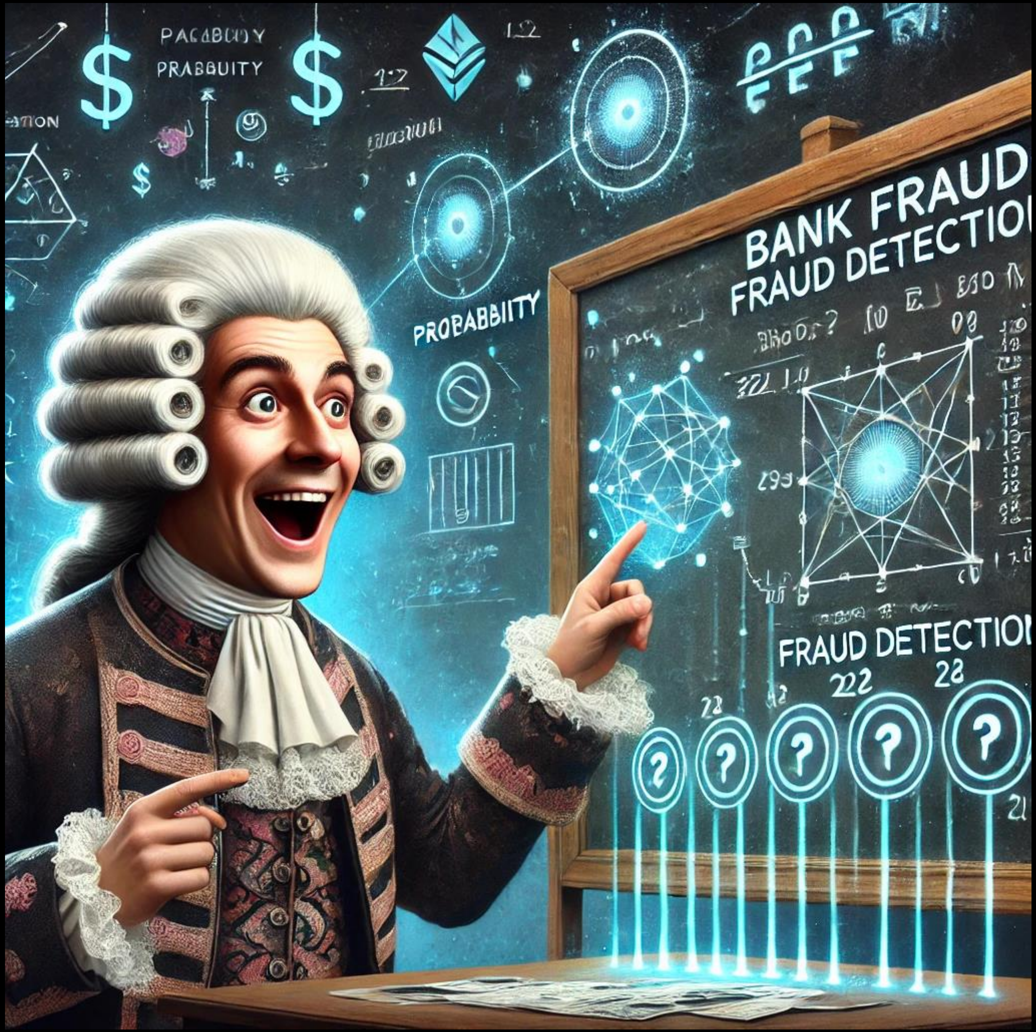
What is the probability that a transaction is actually fraudulent, given that the algorithm has flagged it as potentially fraudulent?

$$P(F|T)$$

NHH TECH3

$P(B|A)$ known, but want to know $P(A|B)$?

Then we need Bayes' rule!



NHH TECH3

Bayes rule

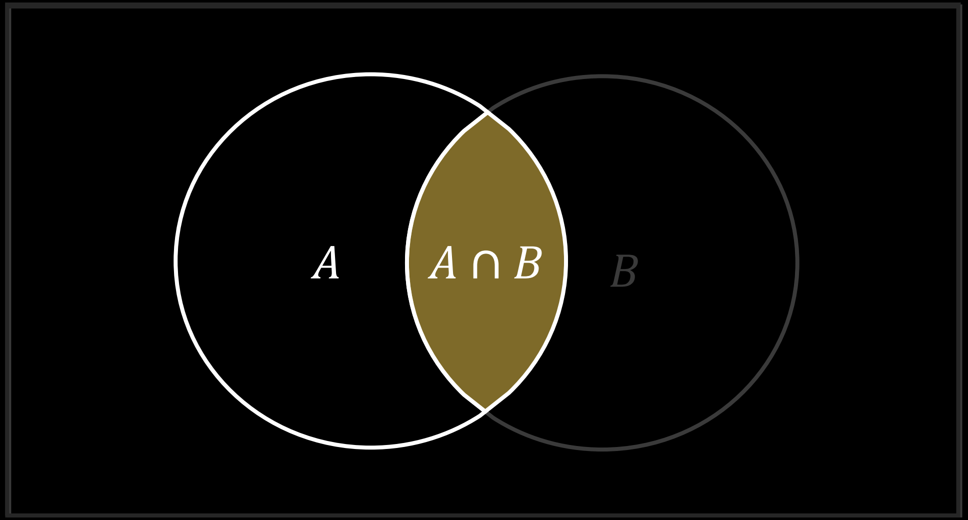
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

NHH
TECH3

PROOF

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$



NHH TECH3

A TRICKY DENOMINATOR

$$P(A|B) = \frac{P(B|A)P(A)}{\boxed{P(B)}}$$

NHH
TECH3

3 versions of Law of total probability

Assume $A_1, A_2, \dots, A_k$ are disjoint events that divide up the whole sample space so that their probabilities add to exactly 1. Then, if B is any other event

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + \dots + P(A_k \cap B)$$

$$= P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \dots + P(B|A_k)P(A_k)$$

Special case: A and A^c are examples of disjoint events dividing up the whole sample space:

$$P(B) = P(B|A)P(A) + P(B|A^c)P(A^c)$$

NHH TECH3

3 versions of Bayes' rule

$$P(B) = P(B|A)P(A) + P(B|A^c)P(A^c)$$
$$= P(B|A_1)P(A_1) + \cdots + P(B|A_k)P(A_k)$$

1. $P(A|B) = \dfrac{P(B|A)P(A)}{P(B)}$

2. $P(A|B) = \dfrac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^C)P(A^C)}$

3. $P(A_j|B) = \dfrac{P(B|A_j)P(A_j)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \ldots + P(B|A_k)P(A_k)}$

NHH TECH3

Example: Fraud Detection in Credit Card Transactions

We know:

- $P(T|F) = 0.90$ *(sensitivity)*
- $P(F) = 0.01$ *(base rate of fraud)*
- $P(T|F^c) = 0.05$ *(false positive rate)*
- $P(F^c) = 0.99$ *(base rate of legitimate transactions)*

We can then use the following version of Bayes rule:

$$P(F|T) = \frac{P(T|F)P(F)}{P(T|F)P(F) + P(T|F^c)P(F^c)} = \frac{\cdot}{\cdot + \cdot} \approx 0.15$$

NHH TECH3

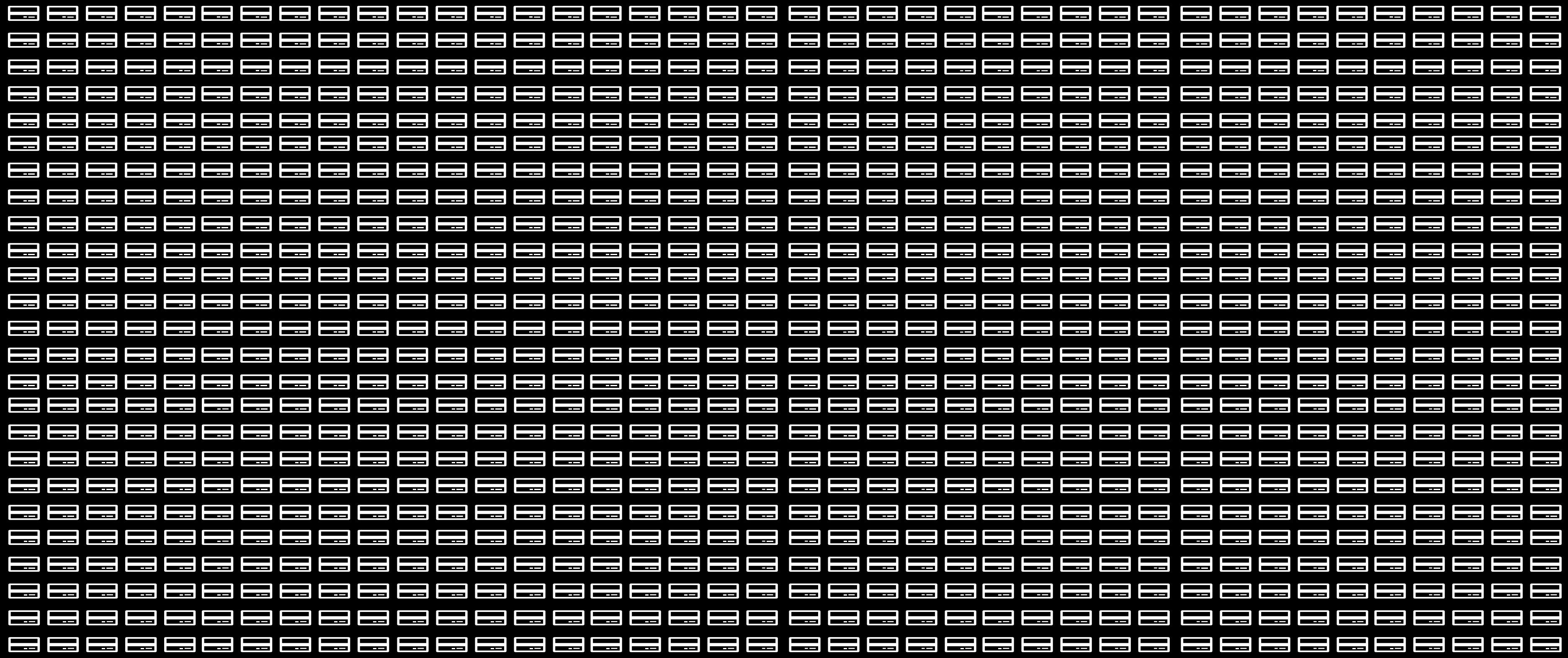
A TYPICAL MISTAKE

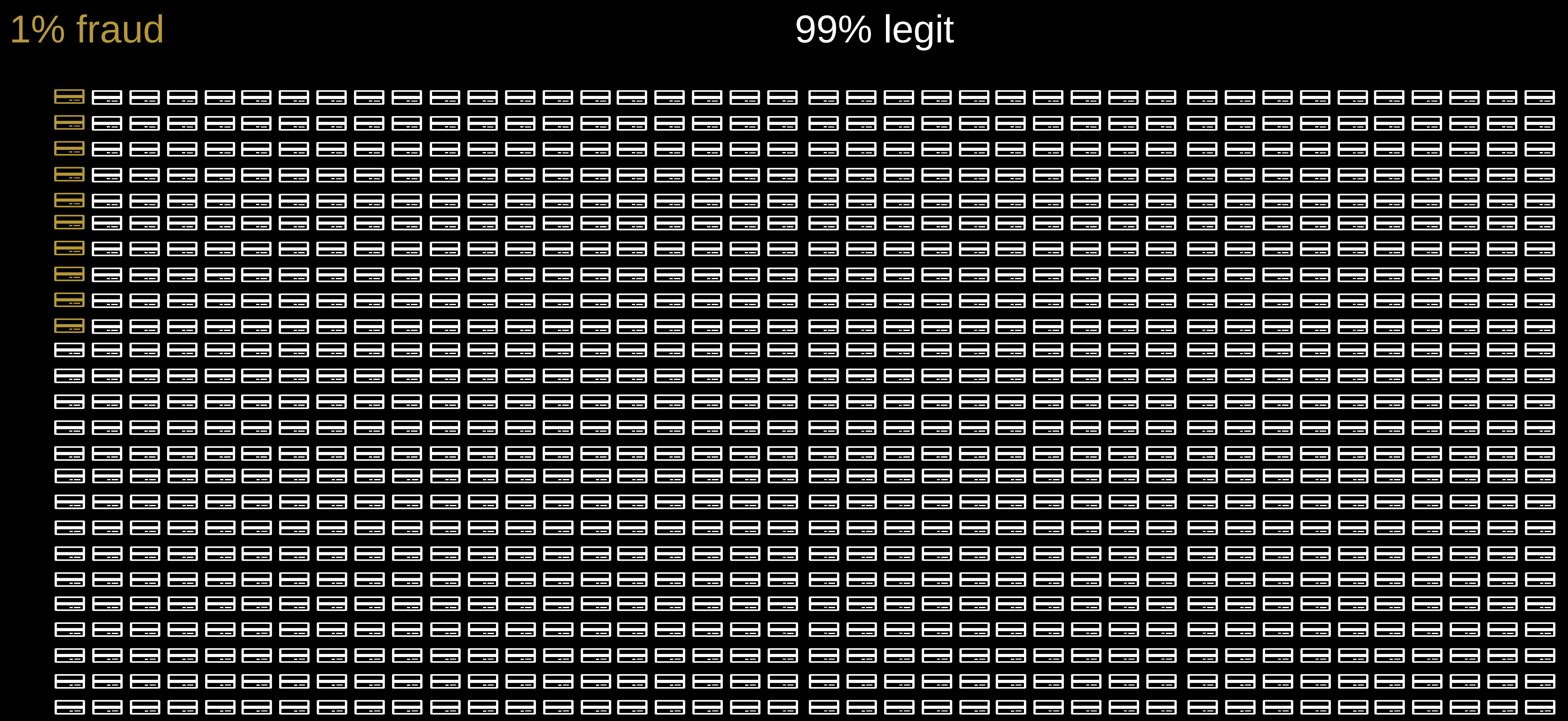
Base rate fallacy: It is easy to overestimate the likelihood of fraud when a transaction is flagged because we focus on the high sensitivity and low false positives of the algorithm, neglecting the very low base rate of fraud P(F).

NHH TECH3

Base rate fallacy

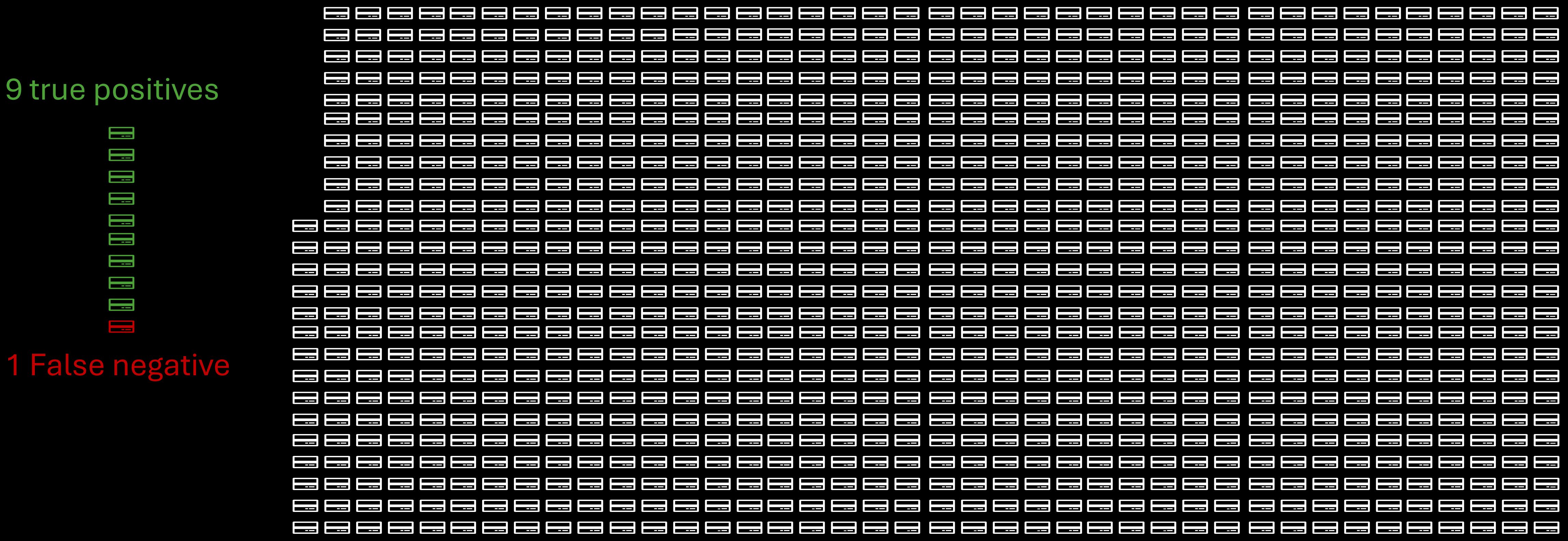
Sample of 1000 transactions:





- $P(T|F) = 0.90$ *(sensitivity)*

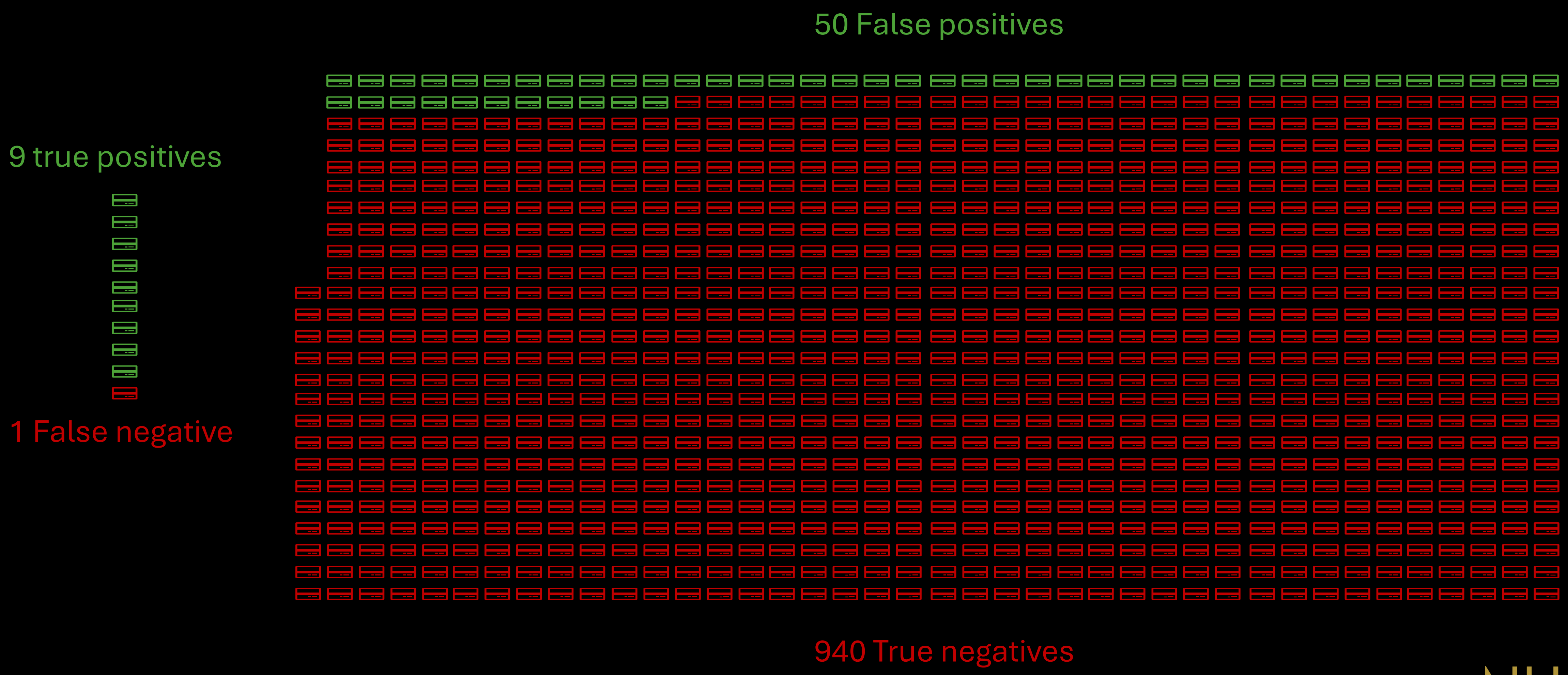
NHH
TECH3



- $P(T|F^c) = 0.05$ *(false positive rate)*

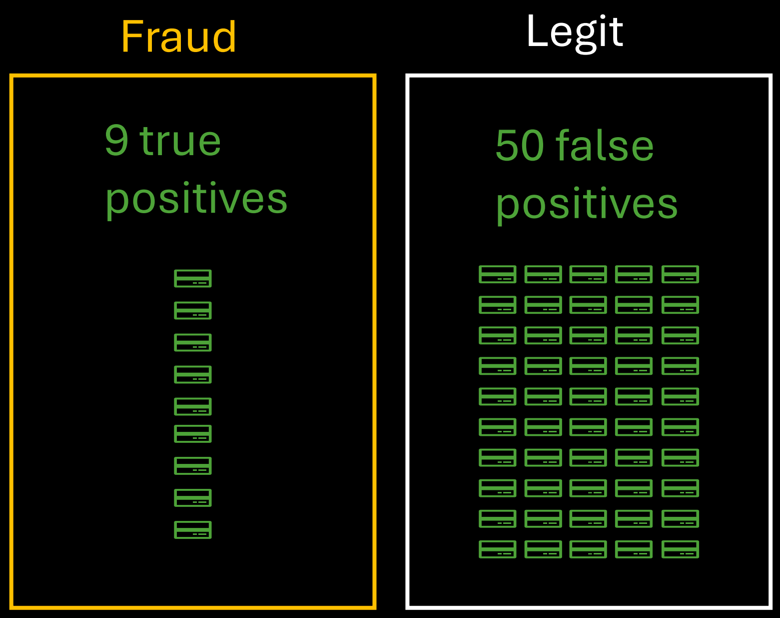
- $P(T|F) = 0.90$ *(sensitivity)*

NHH TECH3


50 False positives
9 true positives
1 False negative
940 True negatives

NHH
TECH3

NHH
TECH3



$$P(F|T) \approx \frac{9}{9+50} \approx 0.15$$

NHH

TECH3

Sondre Hølleland
Geir Drage Berentsen