Data processing framework for decision making

Jan Larsen

Intelligent Signal Processing Group

Department of Informatics and Mathematical Modelling

Technical University of Denmark

jl@imm.dtu.dk, www.imm.dtu.dk/~jl

DeFuse

scientific objectives

- Obtain general scientific knowledge about the advantages of deploying a combined approach
- Eliminate confounding factors through careful experimental design and specific scientific hypotheses
- Test the general scientific hypothesis is that there is little dependence between missed detections in successive runs of the same or different methods
- To accept the hypothesis under varying detection/clearance probability levels
- To lay the foundation for new practices for mine action, but it is not within scope of the pilot project

Objective of this talk

- To provide insight into some of the issues in data processing and detection systems
- To hint at possible solutions using statistical signal processing and machine learning methodologies
- To facilitate the discussion – the good solution requires a cross-disciplinary effort

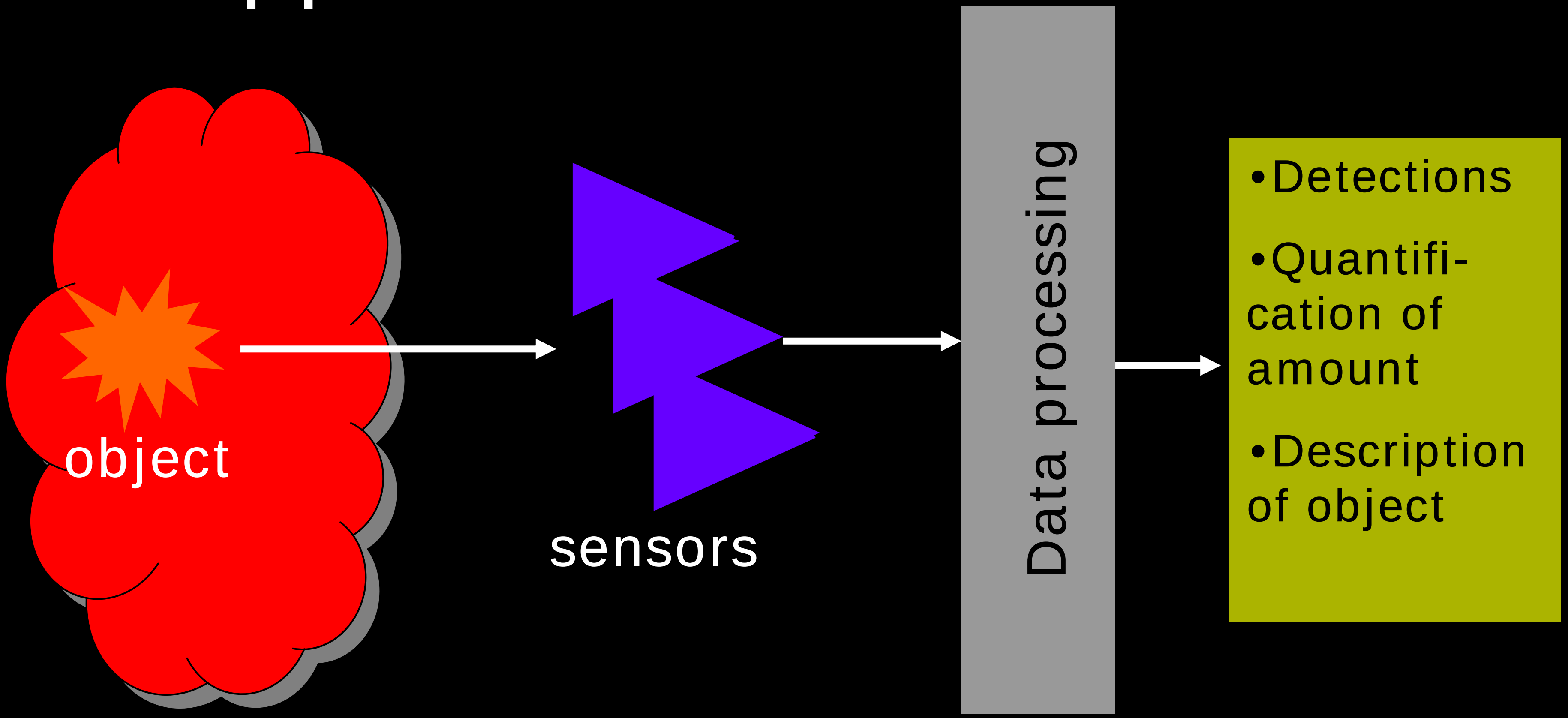
No math!

$$P(\theta \mid y) = \frac{P(y \mid \theta)p(\theta)}{P(y)}$$

Outline

- The data processing pipeline
- Methods for taking up the challenge: reliable detection
- Summary

Data pipeline



Sensing

- Sensing specific primary property of the object (e.g. odor component)
- Sensing a related property (e.g. reflected light)
- Sensing a mixture of properties – maybe only one is relevant
- Multiple sensors can sense different aspects

Sensing errors

- Various factors and other objects in the environment disturb the sensing
  - masking of related or primary property
  - other properties might be too strong
  - the environment is different from the environment in which the sensor was designed to work
- Errors in the sensors
  - Electrical noise
  - Drift
  - Degradation

Data processing

Data processing

- Extracting relevant features from sensor data
- Suppressing noise and error
- Segregation of relevant components from a mixture
- Integration of sensor data
- Prediction:
  - Presence of object
  - Classification of object type
  - Quantification of properties of the object (e.g. amount, size)
  - Description of object

Data processing errors

Data processing

- The sensed expression is too weak to make a reliable prediction of objects presence or quantification of an object property
- The processing device misinterprets the sensed expression
  - Maybe an unknown object in the environment
  - Not able to sufficiently suppress noise and errors
  - The processing can never done with 100% accuracy

Outline

- The data processing pipeline
- Methods for taking up the challenge: reliable detection
- Summary

How do we construct a reliable detector?

- Empirical method: systematic acquisition of knowledge which is used to build a mathematical model
- Specifying the relevant scenarios and performance measures – end user involvement is crucial!!!
- Cross-disciplinary R&D involving very competences

Mathematical models are prevalent: you need them to generate reliable results in a real use case

Knowledge acquisition

Physical modeling

- Study physical properties and mechanism of the environment and sensors
- Describe the knowledge as a mathematical model

Statistical modeling

Require real world related data

Use data to learn e.g. the relation between the sensor reading and the presence/absence of explosives

Why do we need statistical models?

> Scientist and engineers are born sceptical: they don't believe facts unless they see them often enough

- The process is influenced by many uncertain factors which makes classical physical modeling insufficient
- We can never achieve 100% accuracy – hence an estimate of the reliability is needed

There is no such thing as facts to spoil a good explanation!

- Pitfalls and misuse of statistical methods sometimes wrongly leads to the conclusion that they are of little practical use

Some data are in the tail of the distribution: generalization from few examples is not possible

The number of hazardous objects is very small

his live

Why do we need statistical models and machine learning?

- statistical modeling is the principled framework to handle uncertainty and complexity
- Statistic modeling usually focuses on identifying important parameters
- Estimation of performance and reliability is an integral part
- machine learning learns complex models from collections of data to ma optimal predictions in new situations

facts

prior information

consistent and robust information and decisions with associated risk estimates

Four examples of using statistical modeling

- Reliable detection
- Increasing detection rate by combining sensors
- Efficient MA as an hierarchical approach
- Segregation of mixed signals in order to reduce disturbances

Reliable detection of hazardous object – tossing a coin

$$Frequency = \frac{\text{no of heads}}{\text{no of tosses}}$$

$$probability = frequency \text{ when infinitely many tosses}$$

To achieve 99,6% detection probability

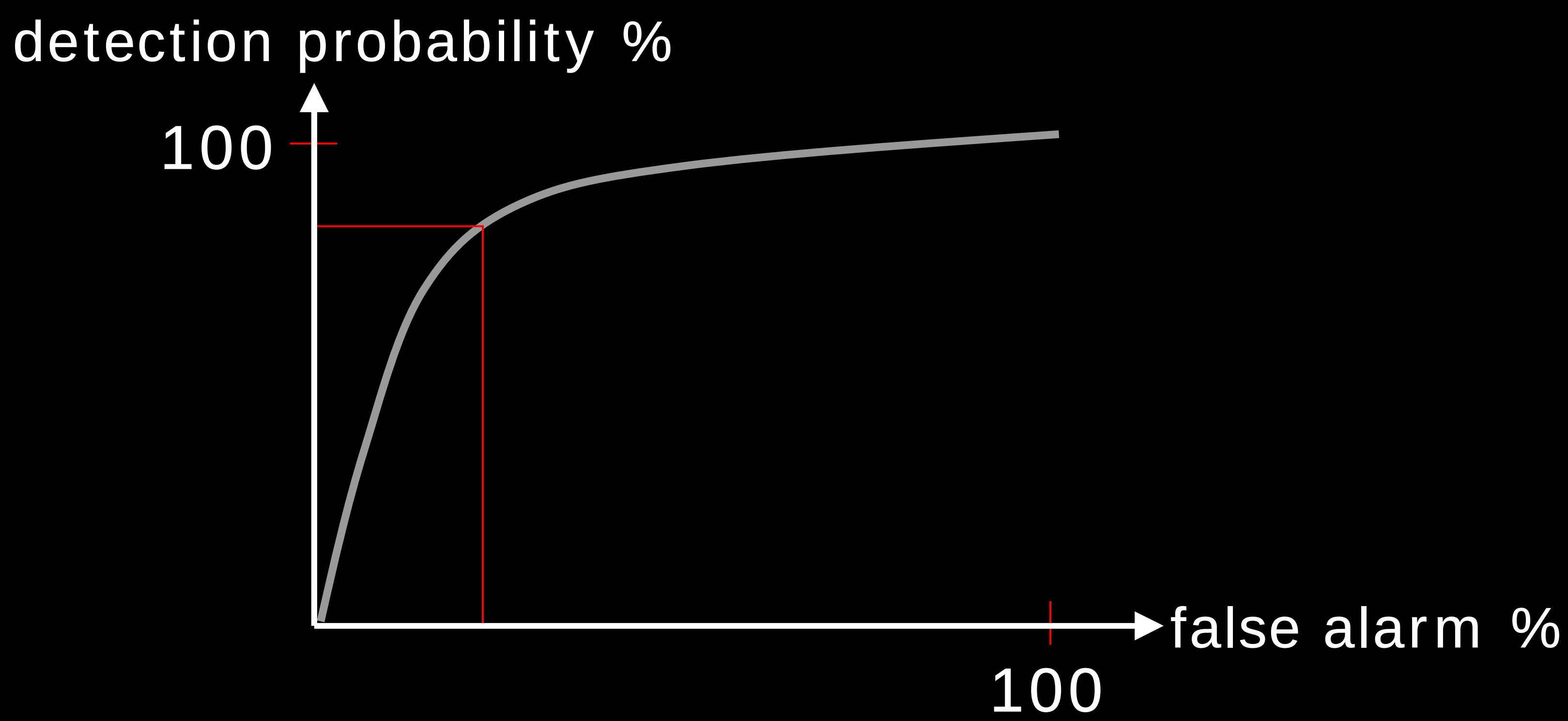
$$Frequency = \frac{996}{1000} = 99{,}6\%$$

$$Frequency = \frac{9960}{10000} = 99{,}60\%$$

One more (one less) count will change the frequency a lot!

You need 747 examples to be 95% sure that detection is better than 99,6% even if you detected all cases

Receiver operation characteristic (ROC)



Two types of errors in relation to ROC

Example: odor

•Sensing error:

has little explosi

content

•Decision error:

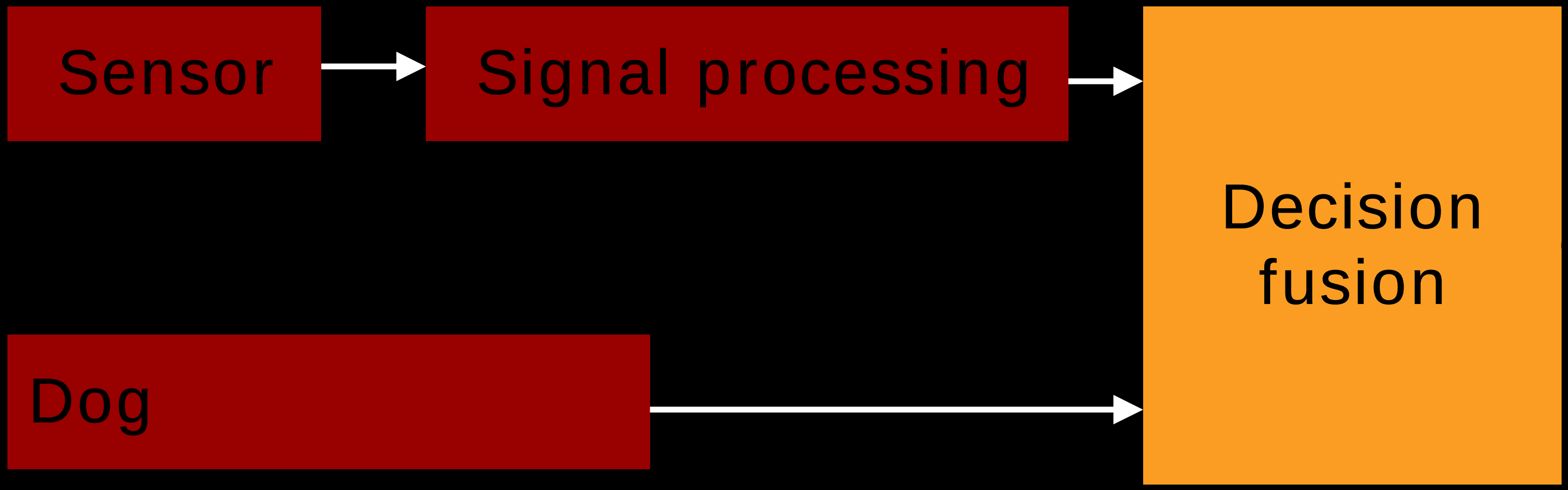
bee-wax was fou

Example: dog

•Sensing error: the TNT leakage from the object was too low

•Decision error: the dog handler misinterpreted the dogs indication

Late integration – decision fusion



Independent error assumption

- Combination leads to a possible exponential increase in detection performance

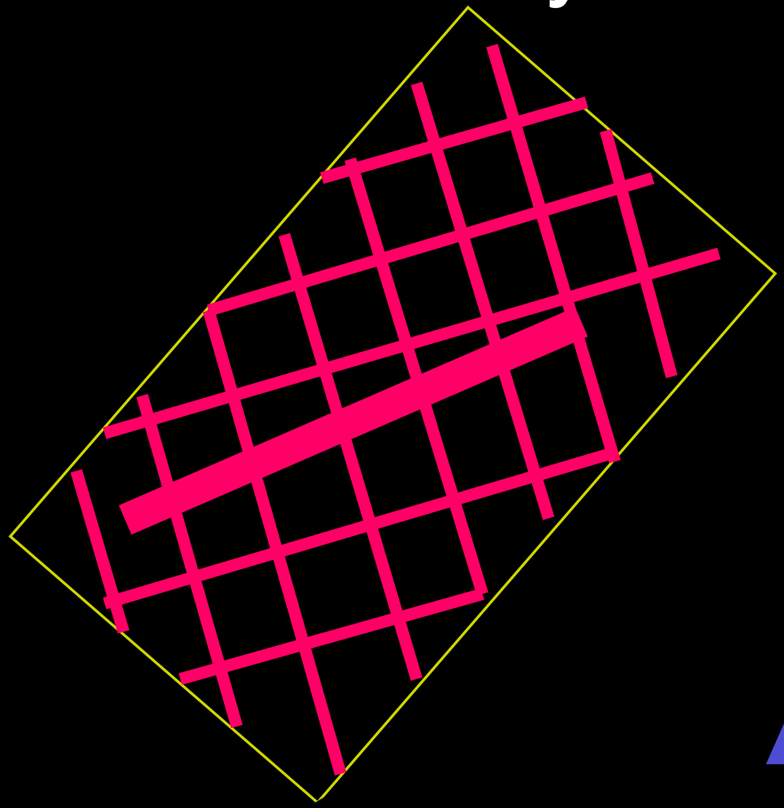
  System 1: 80%

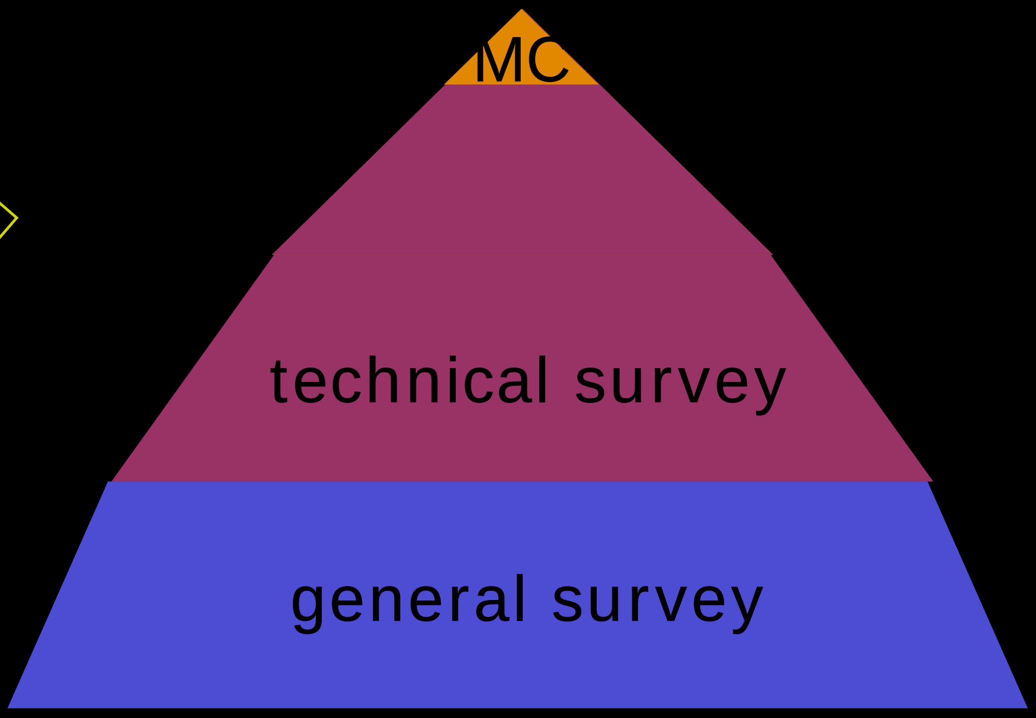
  System 2: 70%

  Combined system: 94%

- Combination leads to better robustness against changes in environmental conditions

Efficient MA by hierarchical approaches





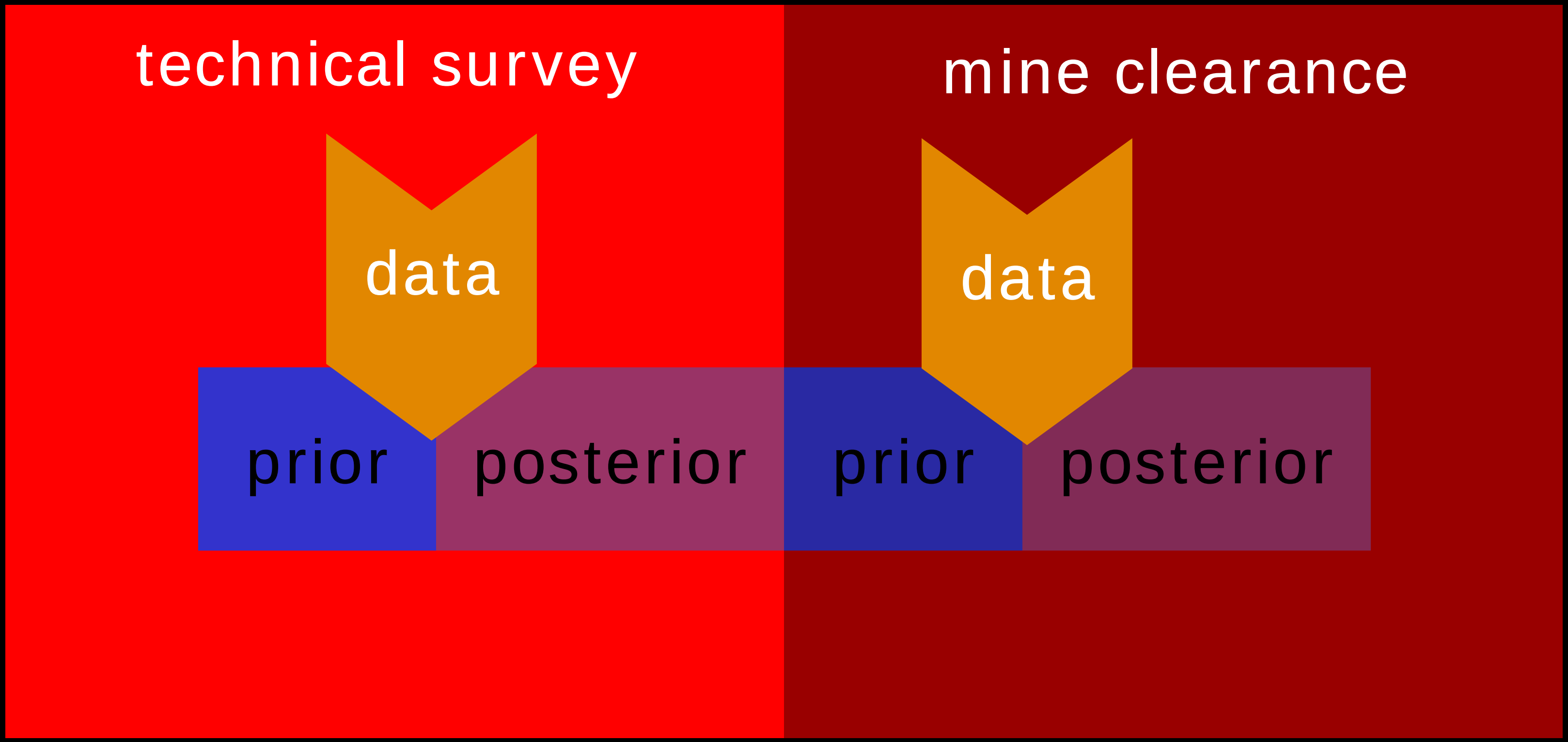
Ref: Håvard Bach, Paul Mackintosh

Danger maps

- The outcome of a hierarchical surveys
- Information about mine types, deployment patterns etc. should also be used
- Could be formulated/interpreted as a prior probability of mines

SMART system described in GICHD: Guidebook on Detection Technologies and Systems for Humanitarian Demining, 2006

Sequential information gathering



Statistical information aggregation

- e=1 indicates encounter of a mine in a box at a specific location
- probability of encounter $P(e=1)$ from current danger map
- d=1 indicates detection by the detection system
- probability of detection $P(d=1)$ from current accreditation

$$P(e=1 \wedge d=0) = P(e=1)(1-P(d=1))$$

$$P(\text{no mine}) = 1 - P(e=1 \wedge d=0)$$

Statistical information aggregation

Example: flail in a low danger area

$$P(e=1)=0.2,\ P(d=1)=0.8$$

$$P(\text{no mine})=1-P(e=1 \wedge d=0)=1-0.2*0.2=0.96$$

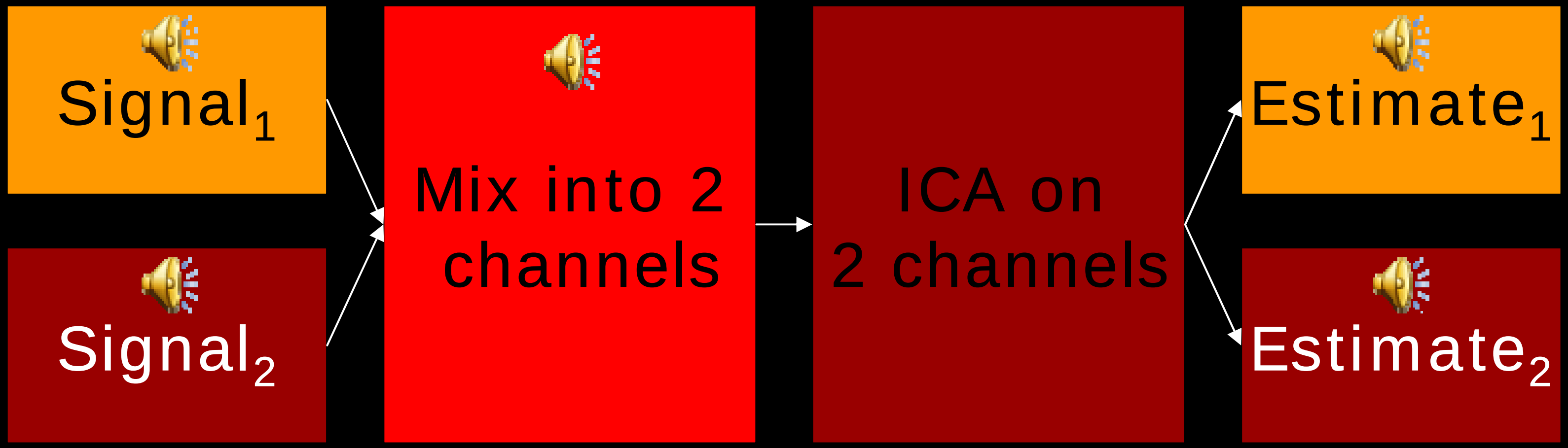
Example: manual raking in a high danger area

$$P(e=1)=1,\ P(d=1)=0.96$$

$$P(\text{no mine})=1-P(e=1 \wedge d=0)=1-1*0.04=0.96$$

Segregation of signals

- Independent Component Analysis of audio signals
  - Cocktail Party Problem
  - Two people talking together, recording two mixtures
  - Example: Molgedey and Schuster’s algorithm (1994)



Summary

- A cross-disciplinary effort is required to obtain sufficient knowledge about physical, operational and processing possibilities and constraints as well as clear definition of a measurable goal – the right tool for the right problem
- Statistical modeling is essential optimal handling of prior information, empirical evidence, robustness and uncertainty
- It is very hard to assess the necessary high performance which is required to have a tolerable risk of casualty
- The use of sequential information aggregation is promising for developing new hierarchical survey schemes (SOPs)
- Combination of methods is a promising avenue to overcome current problems