Certificeret software og explainable AI

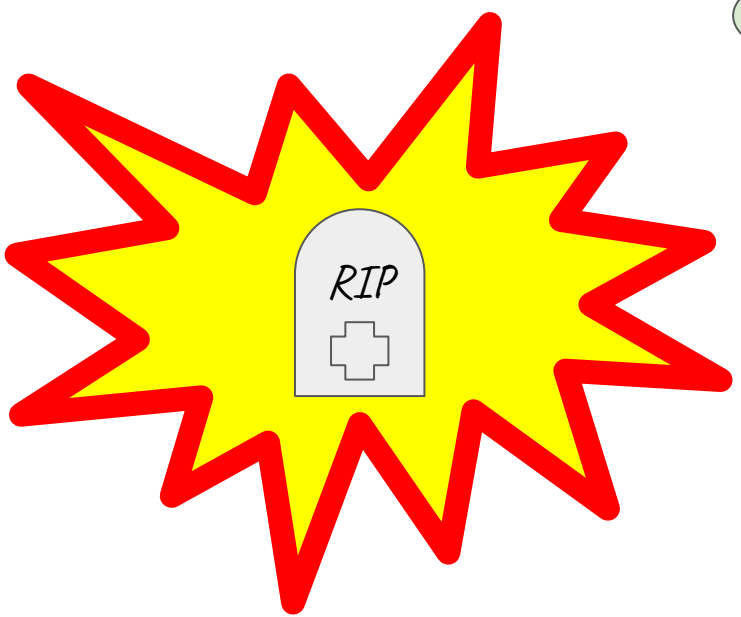
Luís Cruz-Filipe
lcf@imada.sdu.dk

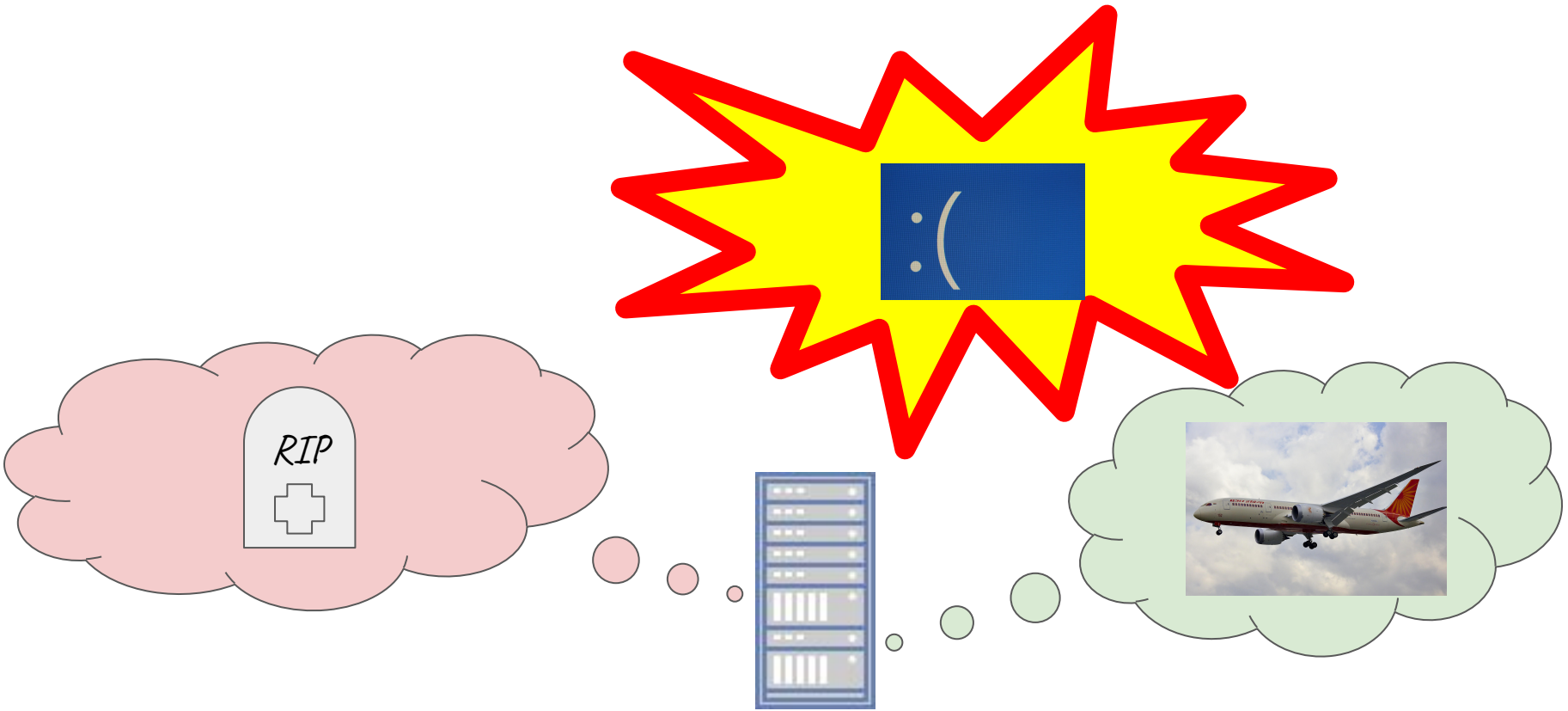
Billeder: Saverio Giallorenzo
www.unsplash.com
www.pixabay.com

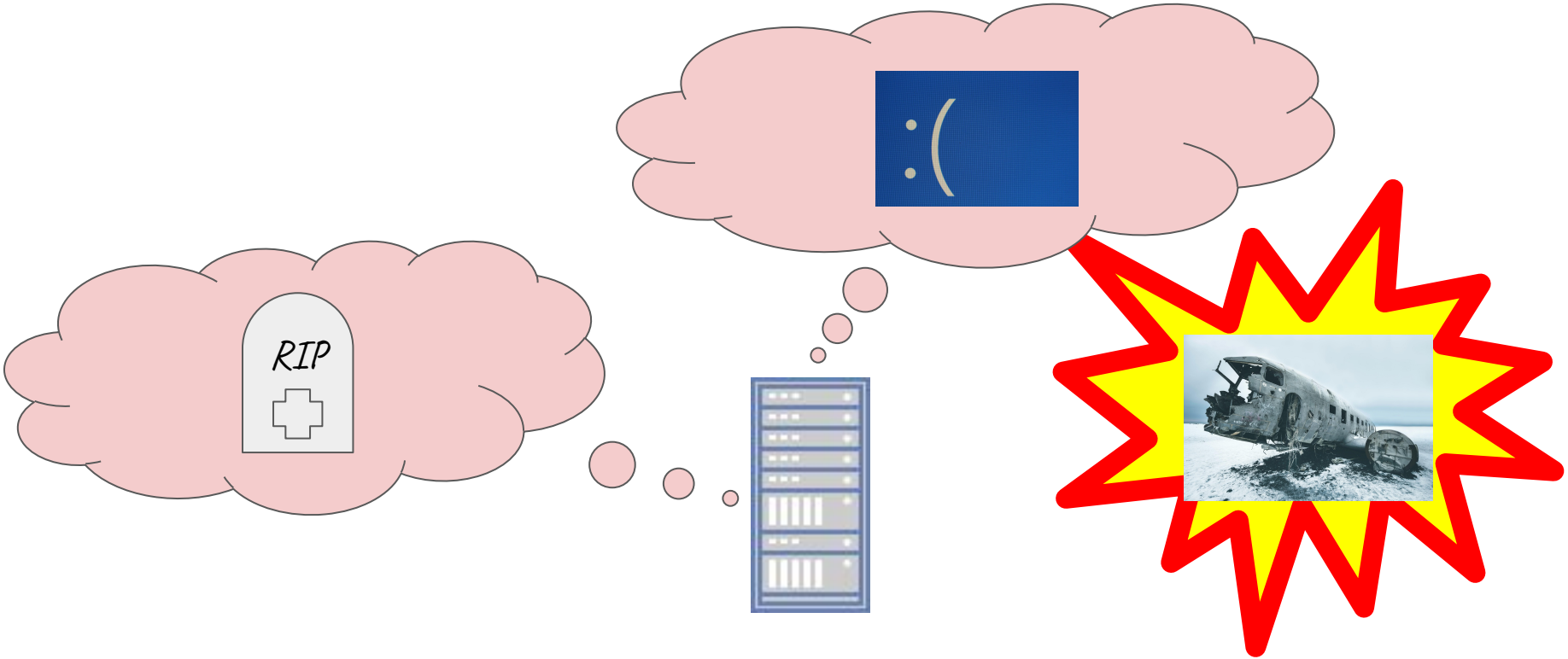
SDU
DEPARTMENT OF MATHEMATICS
AND COMPUTER SCIENCE

Vi stoler mere og mere på computere i vores hverdag…

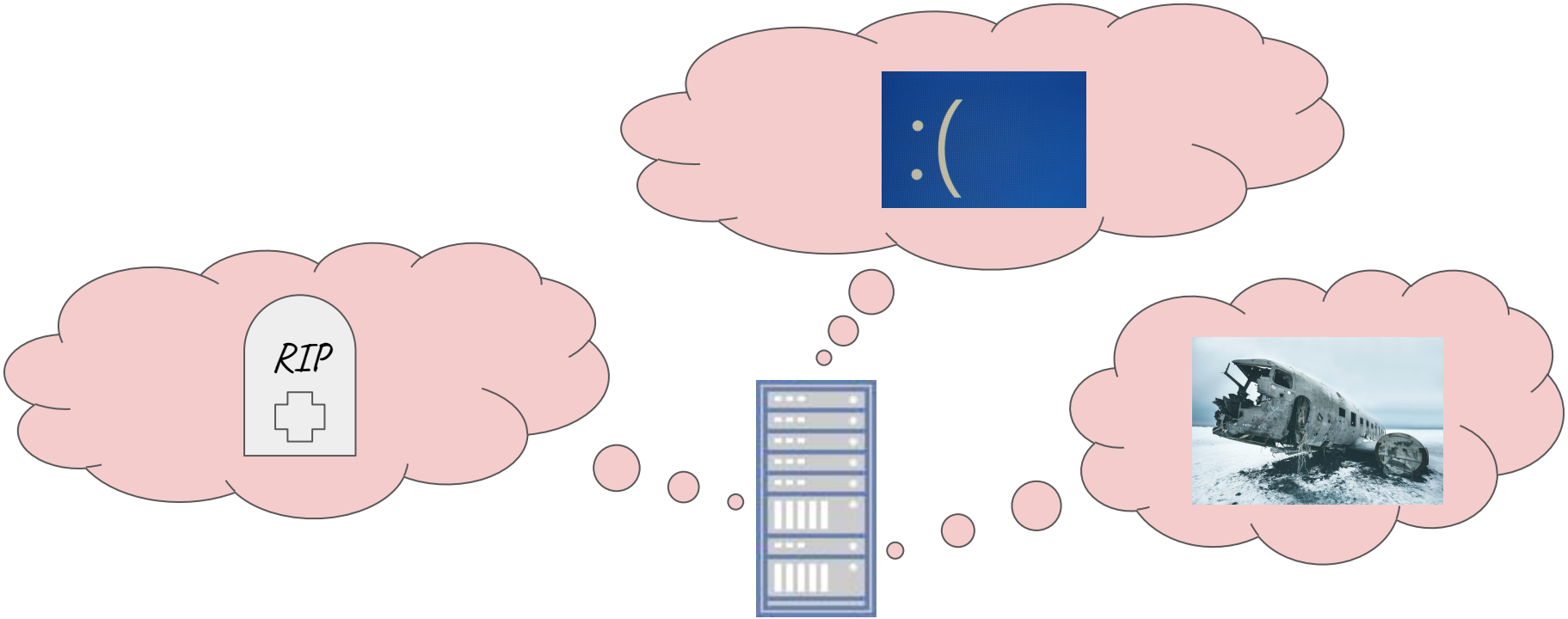
…men hvordan kan vi vide, at det er sikkert at gøre det?


RIP


RIP

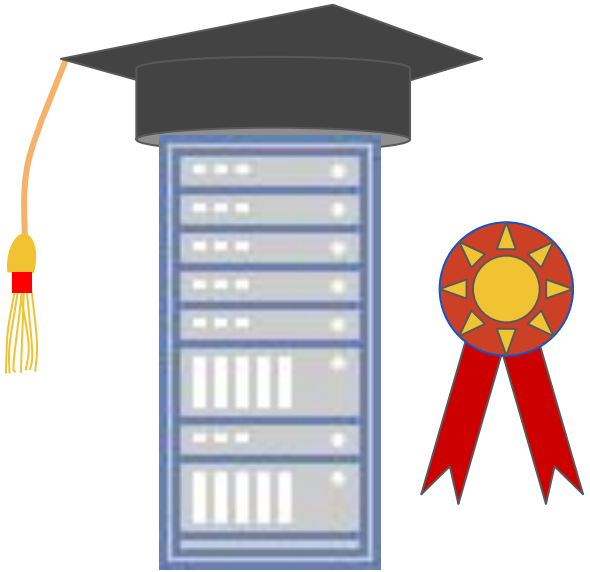

RIP
:(

Computerfejl kan koste penge og liv…



…og “et eller andet gik galt” er ikke nogen god forklaring.

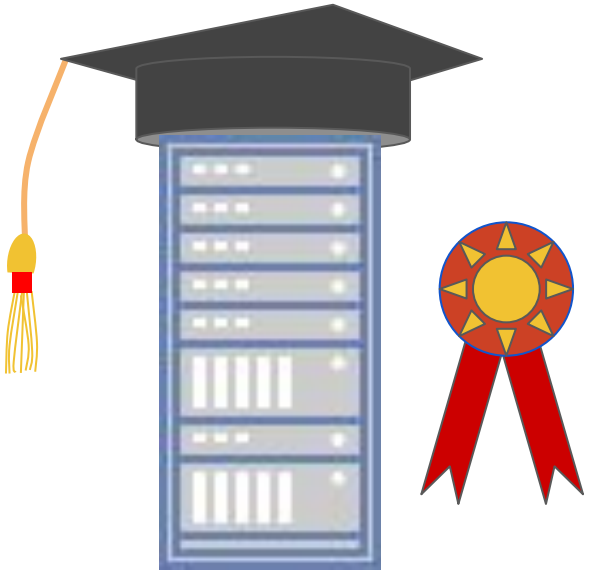
Certificerede programmer



❌ Testning

- Tests kan finde fejl, men de kan ikke bevise, at der ikke findes nogen
- Tests kan kun dække de scenarier, som testeren har tænkt på

Certificerede programmer



❌ Testning

✔️ Matematiske metoder

- Correct-by-design
  Programmeringssprog, der garanterer sikkerhedsegenskaber for *alle* programmer
- Validering
  Matematiske teknikker, der beviser sikkerhedsegenskaber for *bestemte* programmer

Maskinelæringsalgoritmer bruges mere og mere:

- Kredit ansøgning
- Jobsøgning
- Beregning af straf til dømte
- Tildeling af sociale ydelser
- Coronavirusscreening

Målet er, at blive mere effektiv, ved at bruge computere

Maskinelæringsalgoritmer lærer også fordomme:

- Kredit ansøgning ✔
- Jobsøgning ✔
- Beregning af straf til dømte ✔
- Tildeling af sociale ydelser ✔
- Coronavirusscreening ✔

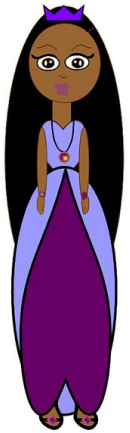
Maskinelæringsalgoritmer lærer også fordomme:

- Kredit ansøgning ✅
- Jobsøgning ❌
- Beregning af straf til dømte ✅
- Tildeling af sociale ydelser ✅
- Coronavirusscreening ✅

Maskinelæringsalgoritmer lærer også fordomme:

- Kredit ansøgning
- Jobsøgning
- Beregning af straf til dømte
- Tildeling af sociale ydelser
- Coronavirusscreening

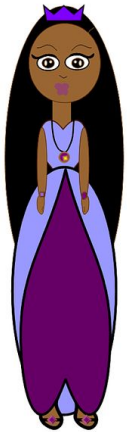
❌❌❌❌❌



Maskinelæringsalgoritmer lærer også fordomme:

- Kredit ansøgning
- Jobsøgning
- Beregning af straf til dømte
- Tildeling af sociale ydelser
- Coronavirusscreening



…selvom programmører kæmper aktivt imod dem!

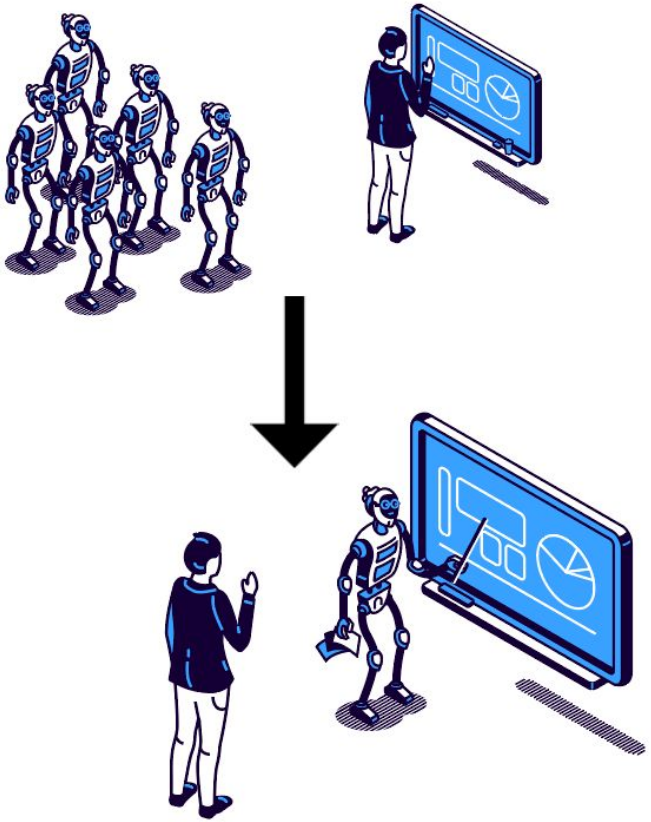
Maskinelæringsalgoritmer lærer også fordomme:

- Kredit ansøgning
- Jobsøgning
- Beregning af straf til dømte
- Tildeling af sociale ydelser
- Coronavirusscreening

Man kan forestille sig, hvad det betyder for fx:

- Risikovurdering
- Asylansøgninger
- Indvandringskontrol

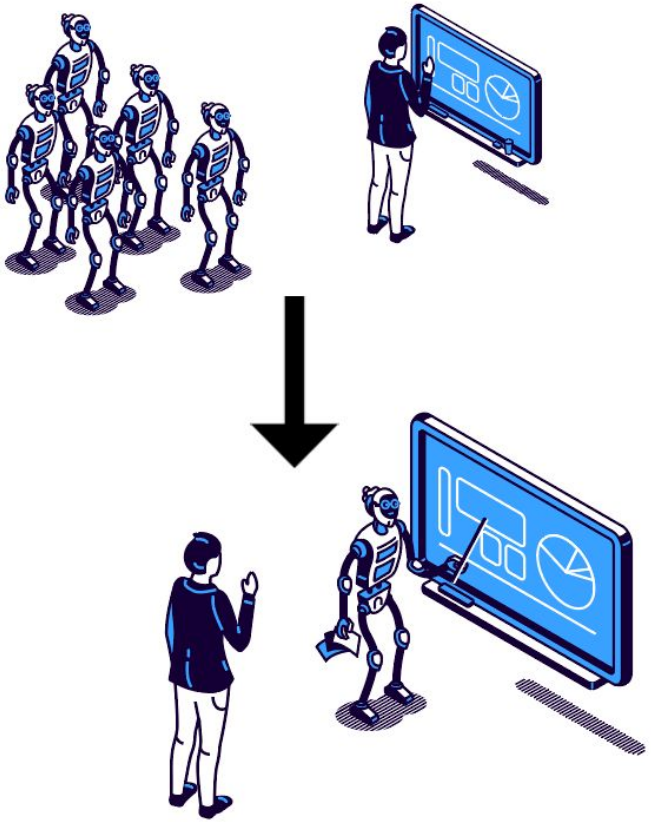
Explainable AI



❌ Ingen “black-box” systemer

- Ingen forklaring på svar
- Computere finder selv ud af, hvad der er “rigtigt”
- Systemer kan ikke ændres, når der sker fejl

Explainable AI

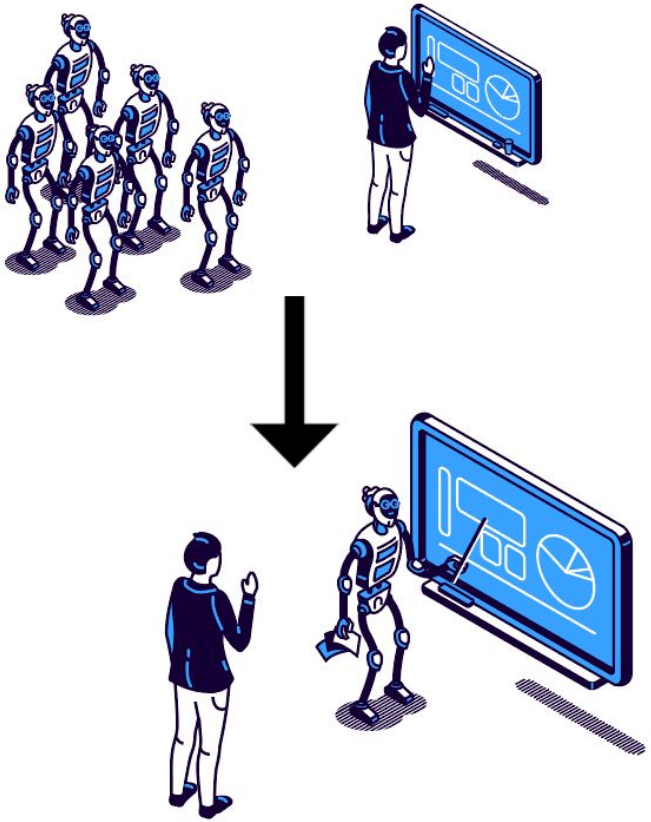


❌ Ingen “black-box” systemer

✔️ Resultater, der kan forklares

- Computere oplæres som voksne
- Computere kan forklare deres svar
- Forklaringer kan analyseres

Explainable AI



❌ Ingen “black-box” systemer

✔️ Resultater, der kan forklares

✔️ Resultater, der kan bruges

- tydelige antagelser
- fejl kan bruges til at forbedre systemet

Vores track record

SDU
DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE

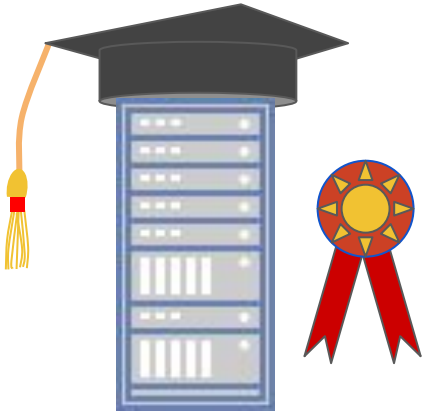
NIST
National Institute of Standards and Technology
U.S. Department of Commerce

- Udvikling af “best practices” i cybersikkerhed for microservices

- Guld- og sølvmedaljer i internationale konkurrencer i kunstig intelligens og optimering

- Deltagelse i ekspertpaneler i internationale konferencer

Opsummering



- **Certificerede metoder** udleverer programmer, der ikke kan tage fejl

- **Explainable AI** udleverer svar, der kan undersøges, forstås og muligvis rettes