Testing AI Systems: Creating Awareness

Peter Collewijn
TestNet, Netherlands

Hannie van Kooten
TestNet, Netherlands

EuroSTAR 2021 ONLINE SEPT. 28-30 | Software Testing & Quality Conference

#EuroSTARConf

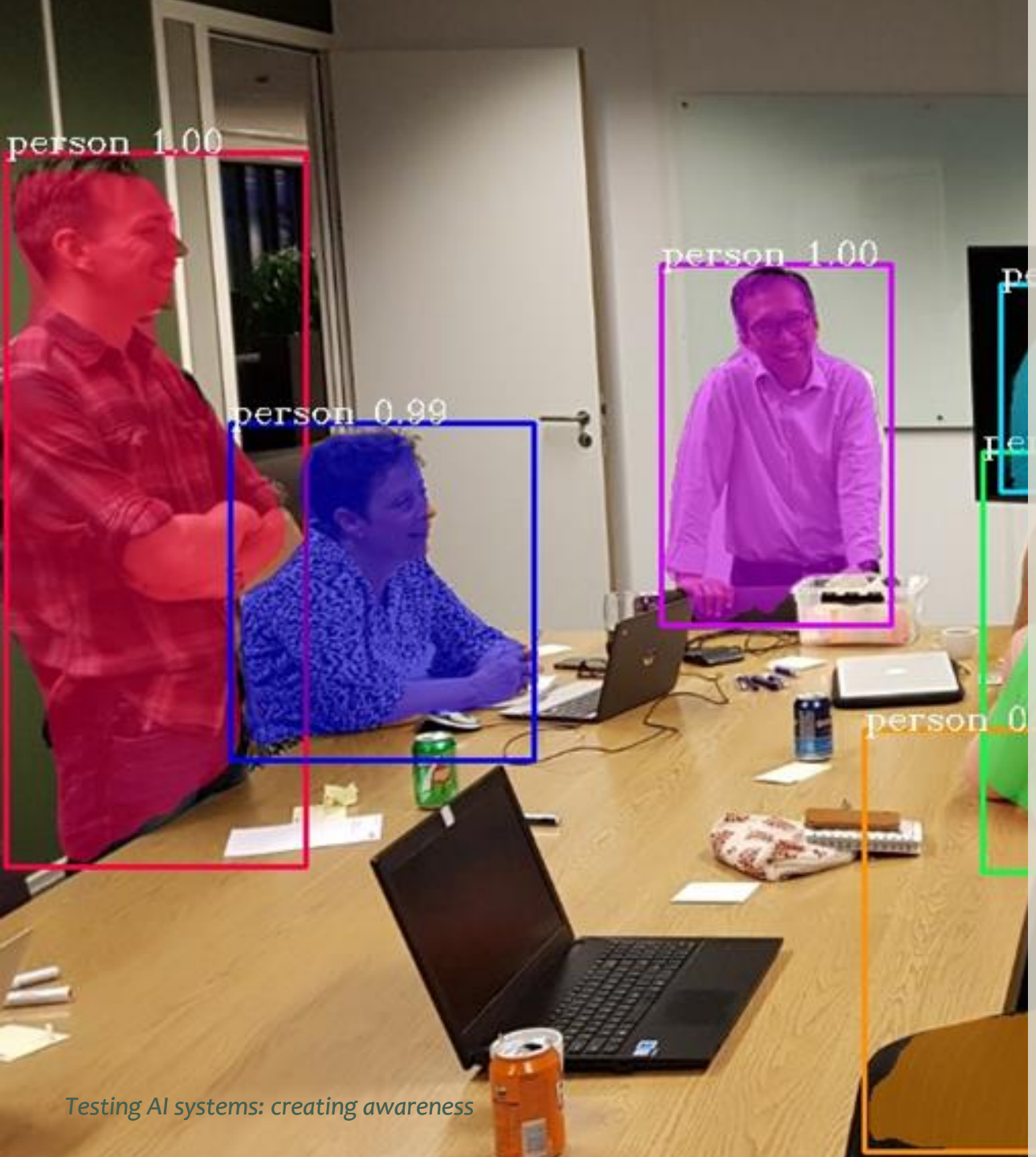
TestNet

- Independent
- Non-profit
- Close to 2.000 members

- Founded in 1997

- Professionalize testing in the IT world
- Experience and expertise exchange
- Stimulating research

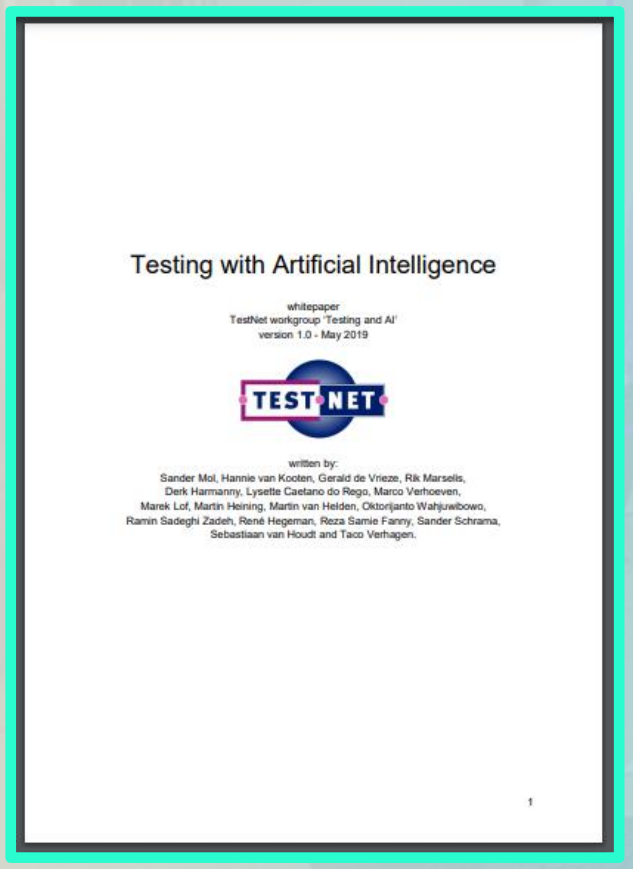
- https://www.testnet.org



Working group testing & ai

- Bram van den Reijen
- Gerline van Lieburg
- Hannie van Kooten
- Johannes Sim
- Marco Verhoeven
- Mariëlle van der Sluys
- Martin van Helden
- Peter Collewijn
- Richard van Emmerik
- Rik Marselis
- Sander Mol

Publications


Testing with Artificial Intelligence

whitepaper
TestNet workgroup 'Testing and AI'
version 1.0 - May 2019

TEST NET

written by:
Sander Mol, Hannie van Kooten, Gerald de Vrieze, Rik Marselis,
Derk Harmanny, Lysette Caetano do Rego, Marco Verhoeven,
Marek Lof, Martin Heining, Martin van Helden, Oktorijanto Wahjuwibowo,
Ramin Sadeghi Zadeh, René Hegeman, Reza Samie Fanny, Sander Schrama,
Sebastiaan van Houdt and Taco Verhagen.

1

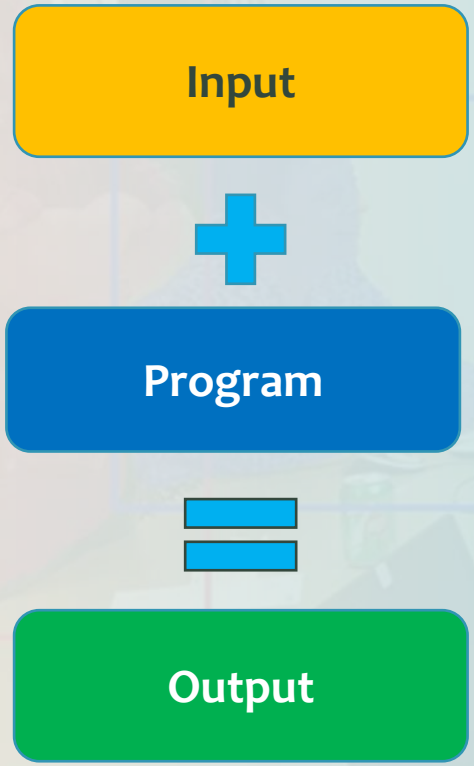
2019

2021

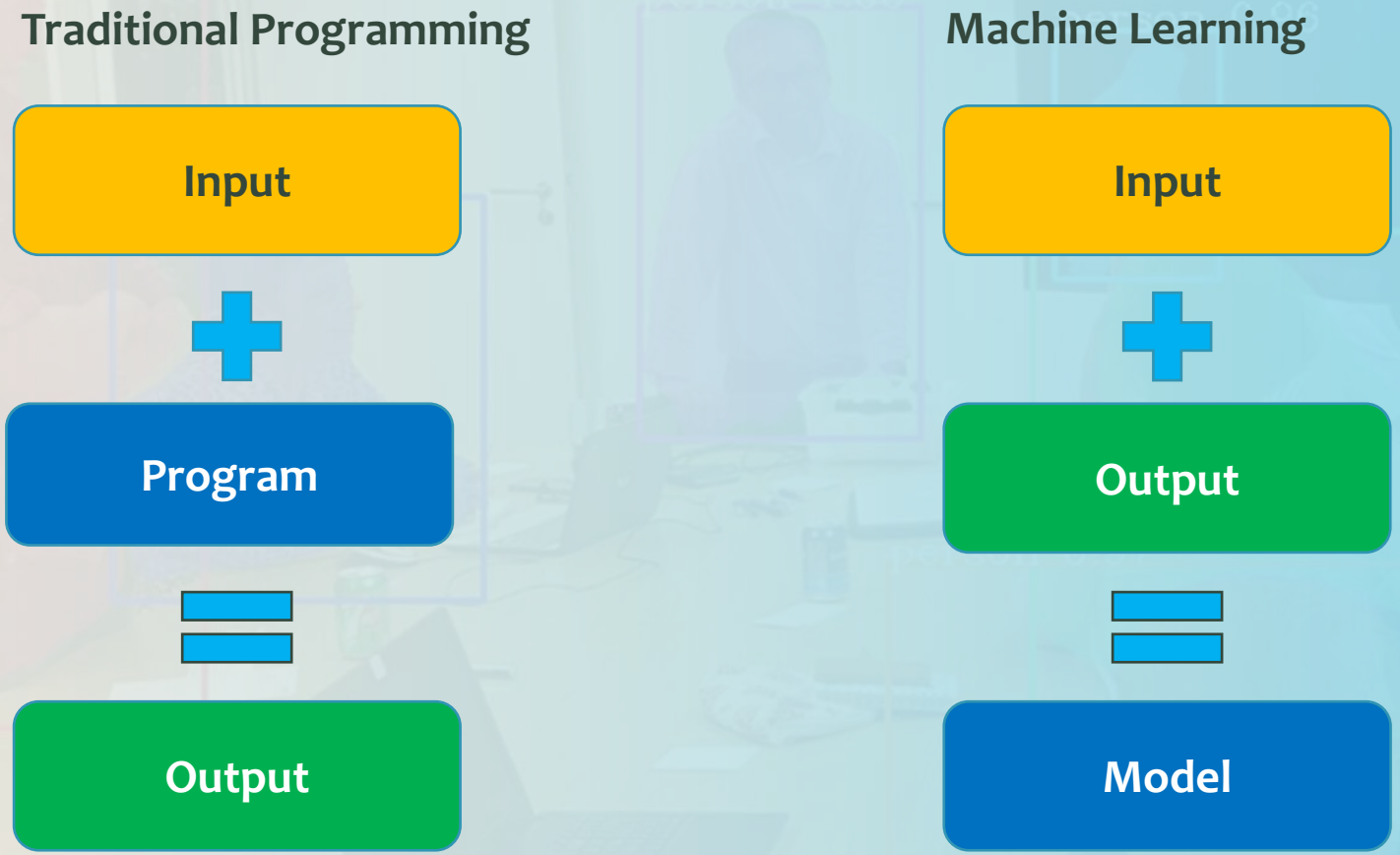
Testing AI systems: creating awareness

Traditional Programming – Machine Learning

Traditional Programming

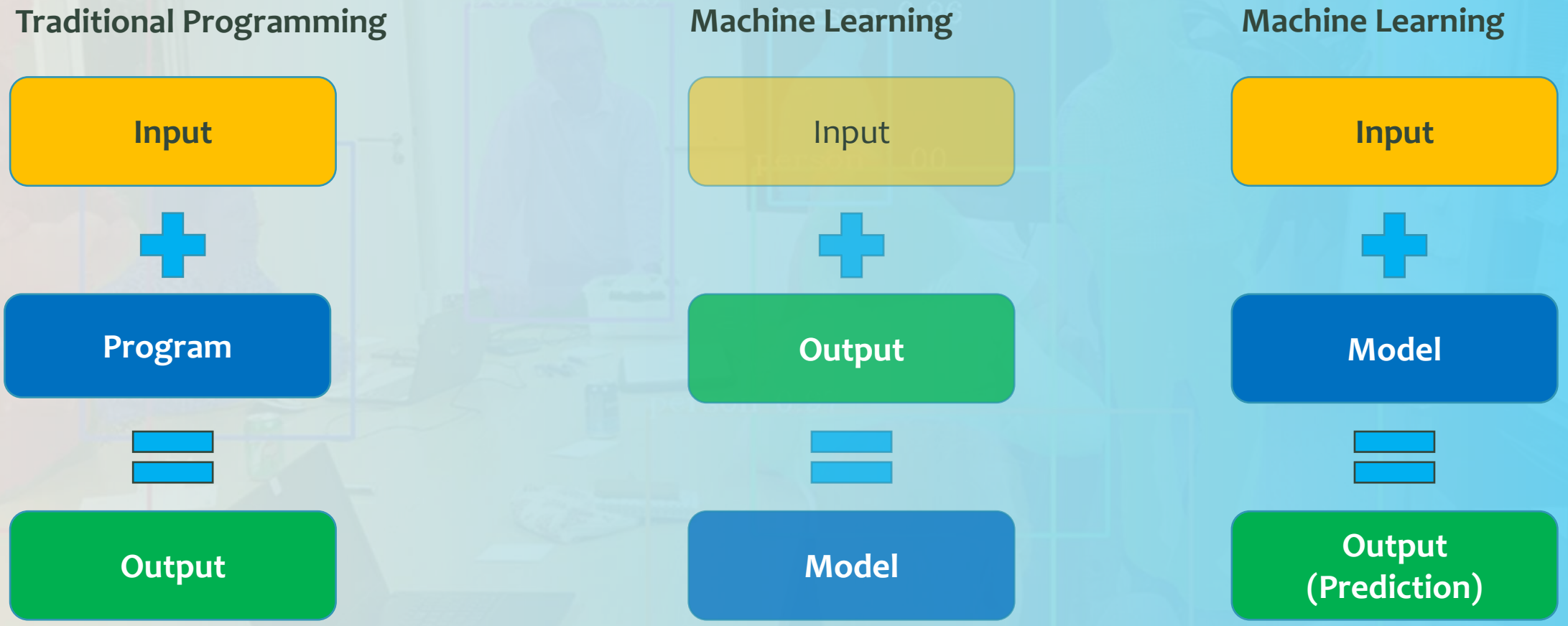


Traditional Programming – Machine Learning



Learning phase

Traditional Programming – Machine Learning

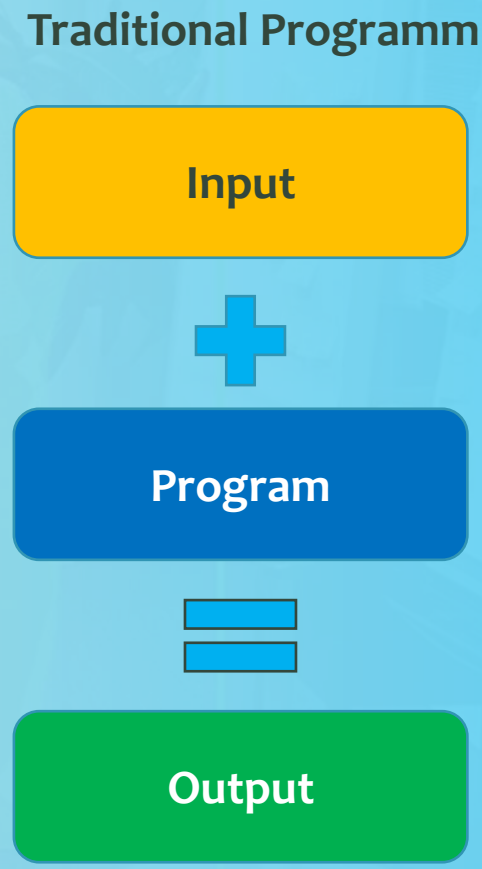


Traditional Programming – Machine Learning

```
# Input
age = 18

# Rule
if age < 18:
    ageClass = 0
else:
    ageClass = 1

# Result
if ageClass == 0:
    print("This person is a Child")
else:
    print("This person is an Adult")
```

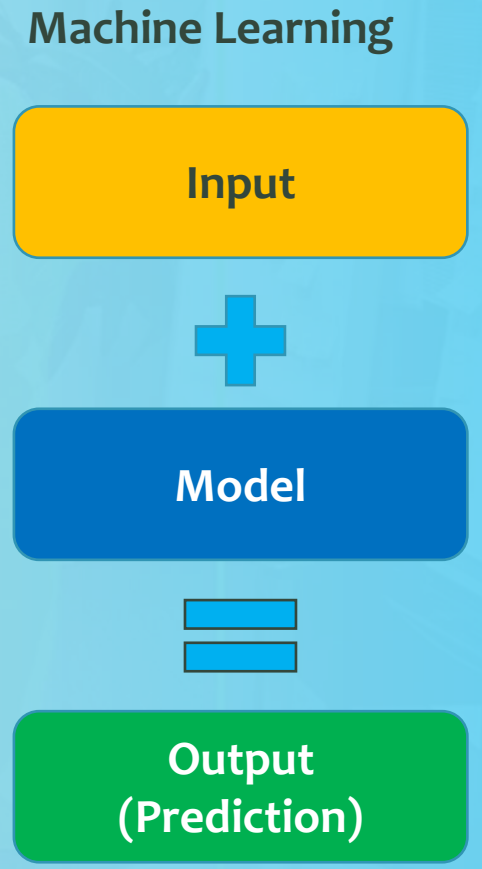


Traditional Programming – Machine Learning

```
# Input
age = 18

# Model - Function
ageClass = 1 / (1 + 2.7**-(-15.6 + age * 0.9))

# Result
if ageClass < 0.5:
    print("This person is a Child")
else:
    print("This person is an Adult")
```



The Model

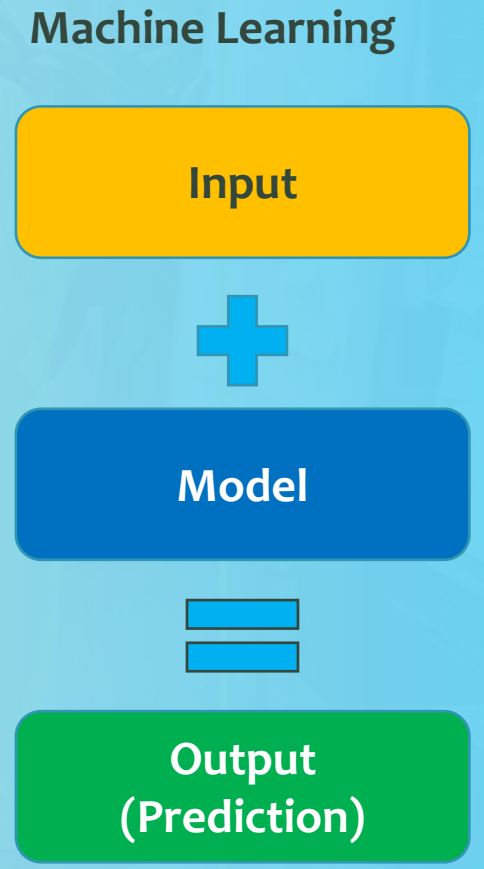
```
# Model - Function
ageClass = 1 / (1 + 2.7**-(-15.6 + age * 0.9))
```

$$ageClass = \frac{1}{1 + 2.7^{-(-15.6 + age * 0.9)}}$$

Points to remember:

"A model is a kind of Function or Formula“

"It is (almost) impossible to understand a model"



Questions

1. This was a presentation about Artificial Intelligence why are you talking about Machine Learning?

2. Machine Learning looks difficult and is hard to understand, why should we use it?

Question 1

1. This was a presentation about Artificial Intelligence why are you talking about Machine Learning?

For this presentation:

Artificial Intelligence == Machine Learning

Question 2

2. Machine Learning looks difficult and is hard to understand, why should we use it?

   When we don't know the rules!

Question 2

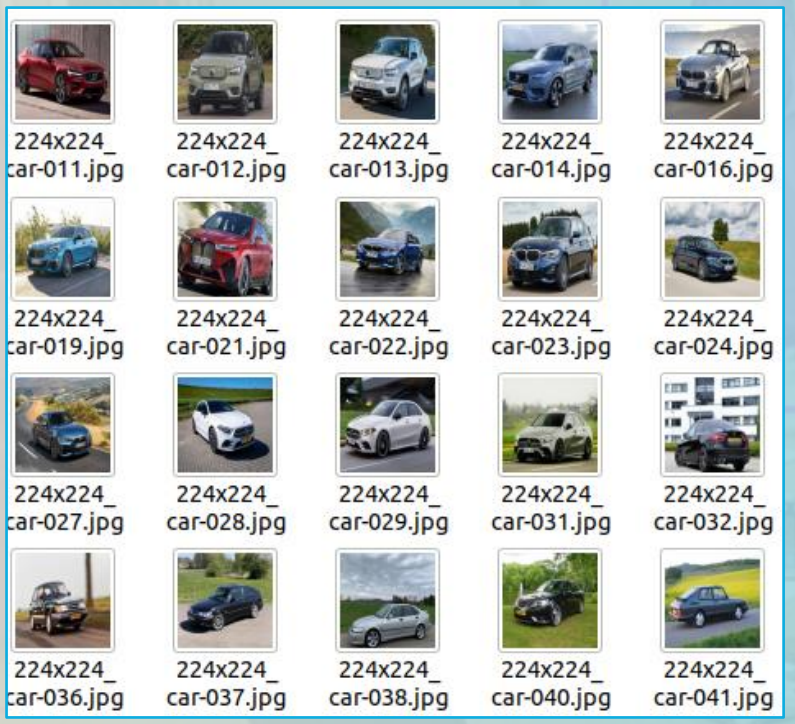
2. Machine Learning looks difficult and is hard to understand, why should we use it?

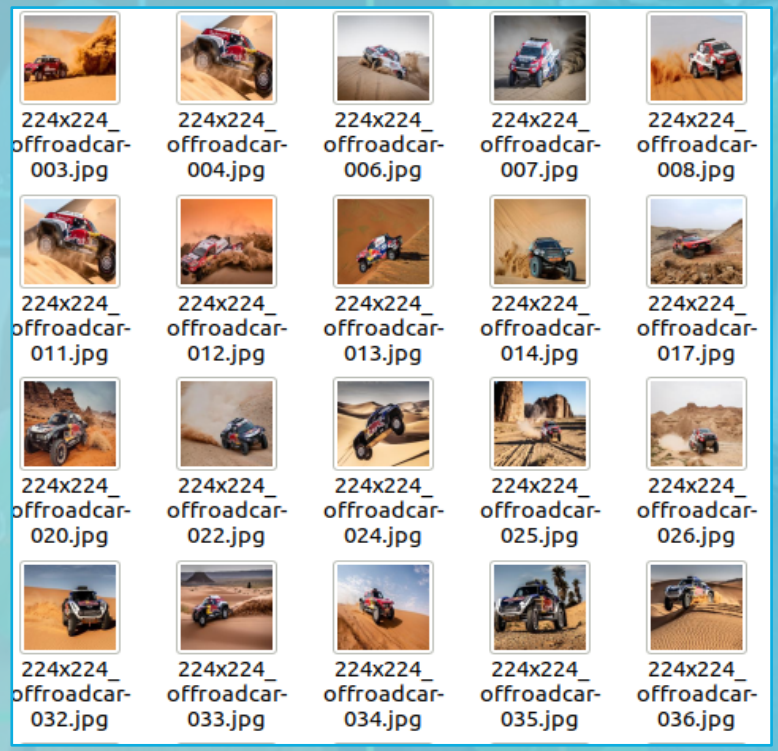
When we don't know the rules!

We don't know the rules to distinguish a Cat from a Dog!

Streetcar vs Paris Dakar car

Streetcar vs Paris Dakar car





Data: 100 images of street cars and 100 of Paris-Dakar cars

Streetcar vs Paris Dakar car

Classify image: **1. 0-Car: 99.98%**
2. 1-OffRoadCar: 0.02%

Classify image **1. 1-OffRoadCar: 99.96%**
2. 0-Car: 0.04%

General Risks of AI

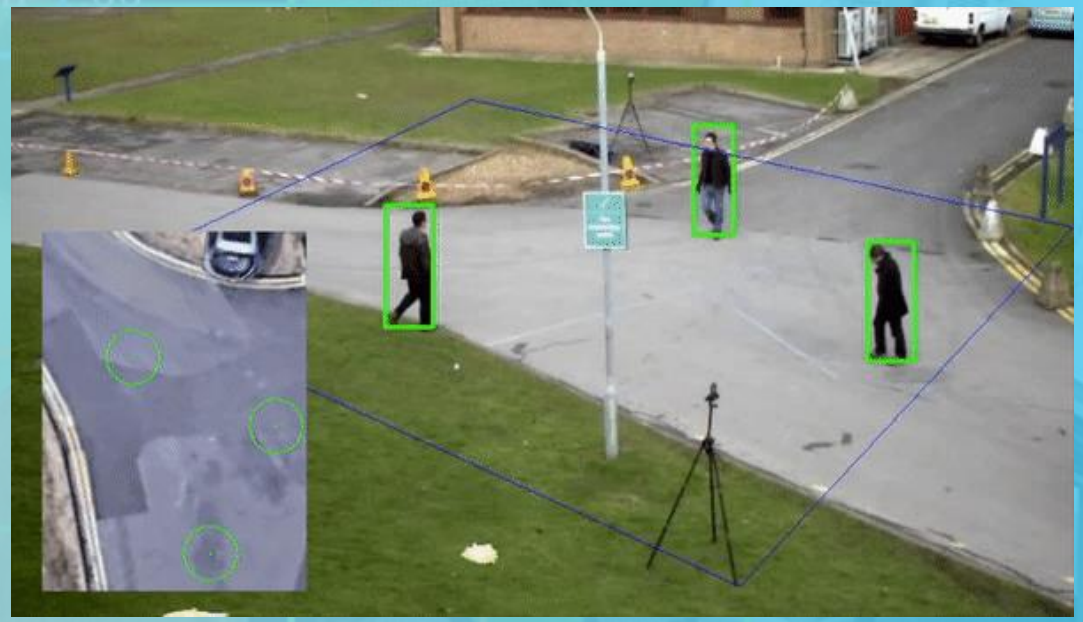
We have defined 5 main risks:

1. Uncertain outcomes
2. Dependency on data
3. Limited explainability
4. Changing reality or need
5. General fear of AI

Different appearances of AI

We matched the risks for the following appearances:

- Pattern recognition in datasets
- Image Recognition
- Sequence Recognition
- Regression
- Text generation
- Speech Generation
- Image Generation

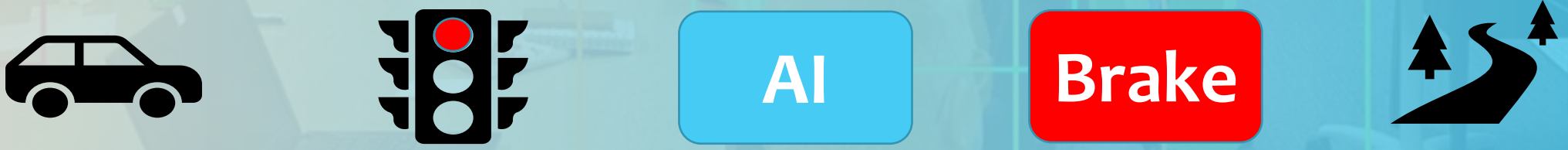


Different degrees of autonomy

Another view: the risks based on the degree of autonomy

- Person uses AI and makes decisions

- Machines take over control



Resume

Difference between machine learning and traditional algorithms

General risks of AI

Appearances of AI

Different degrees of autonomy

Ethics in relation to AI

In the past, ethics mainly concerned human actions.
As technology becomes more autonomous, new ethical issues arise.

- Decisions made by AI applications are hard to trace
- Degree of autonomy
- Fear for AI

What 'the good thing' is and how to 'do the right thing'

Ethics in relation to AI

- **Transparency**
- **Fairness**
- **Explainability**

What 'the good thing' is and how to 'do the right thing'

Ethics guidelines and regulations

- **Robust AI**
- **Lawful AI**
- **Ethical AI**

What 'the good thing' is and how to 'do the right thing'

Quality attributes

- Functionality
  - Completeness
  - Correctness
- Performance
- Compatibility
- Usability
- Reliability
- Security
- Maintainability
  - Modularity
  - Reusability
  - Analyzability
  - Modifiability
  - Testability
- Portability

ISO 25010

Quality attributes

- Functionality
  - Completeness
  - Correctness
- Performance
- Compatibility
- Usability
- Reliability
- Security
- Maintainability
  - Modularity
  - Reusability
  - Analyzability
  - Modifiability
  - Testability
- Portability

ISO 25010

DIN SPEC 92001-1

Functionality &
Performance
Robustness
Comprehensibility

Quality attributes

- Functionality
  - Completeness
  - Correctness
- Performance
- Compatibility
- Usability
- Reliability
- Security
- Maintainability
  - Modularity
  - Reusability
  - Analyzability
  - Modifiability
  - Testability
- Portability

ISO 25010

Book
Testing in the digital age

Intelligent behavior
Ability to learn
Improvisation
Transparency of choices
Collaboration
Natural interaction

Morality
Ethics
Privacy
Human friendliness

Personality
Mood
Empathy
Humor
Charisms

DIN SPEC 92001-1
Functionality &
Performance
Robustness
Comprehensibility

Quality attributes

- Functionality
  - Completeness
  - Correctness
- Performance
- Compatibility
- Usability
- Reliability
- Security
- Maintainability
  - Modularity
  - Reusability
  - Analyzability
  - Modifiability
  - Testability
- Portability

ISO 25010

ISO/CEN 5059
Ability to learn
Ability to generalize
Trustworthiness
Robustness
Controllability
Explainability
Accountability
Respect for democracy, justice and the rule of law
Responsibility
Privacy
Fairness and non-discrimination
Transparency
Reinforcement of existing bias
Consistency
Free from bias

Book
Testing in the digital age

Intelligent behavior
Ability to learn
Improvisation
Transparency of choices
Collaboration
Natural interaction

Morality
Ethics
Privacy
Human friendliness

Personality
Mood
Empathy
Humor
Charisms

DIN SPEC 92001-1
Functionality &
Performance
Robustness
Comprehensibility

Quality attributes

- Functionality
  - Completeness
  - Correctness
- Performance
- Compatibility
- Usability
- Reliability
- Security
- Maintainability
  - Modularity
  - Reusability
  - Analyzability
  - Modifiability
  - Testability
- Portability

ISO 25010

ISO/CEN 5059

Ability to learn
Ability to generalize
Trustworthiness
Robustness
Controllability
Explainability
Accountability
Respect for democracy, justice and the rule of law
Responsibility
Privacy
Fairness and non-discrimination
Transparency
Reinforcement of existing bias
Consistency
Free from bias

Book
Testing in the digital age

Intelligent behavior

Ability to learn
Improvisation
Transparency of choices
Collaboration
Natural interaction

Morality

Ethics
Privacy
Human friendliness

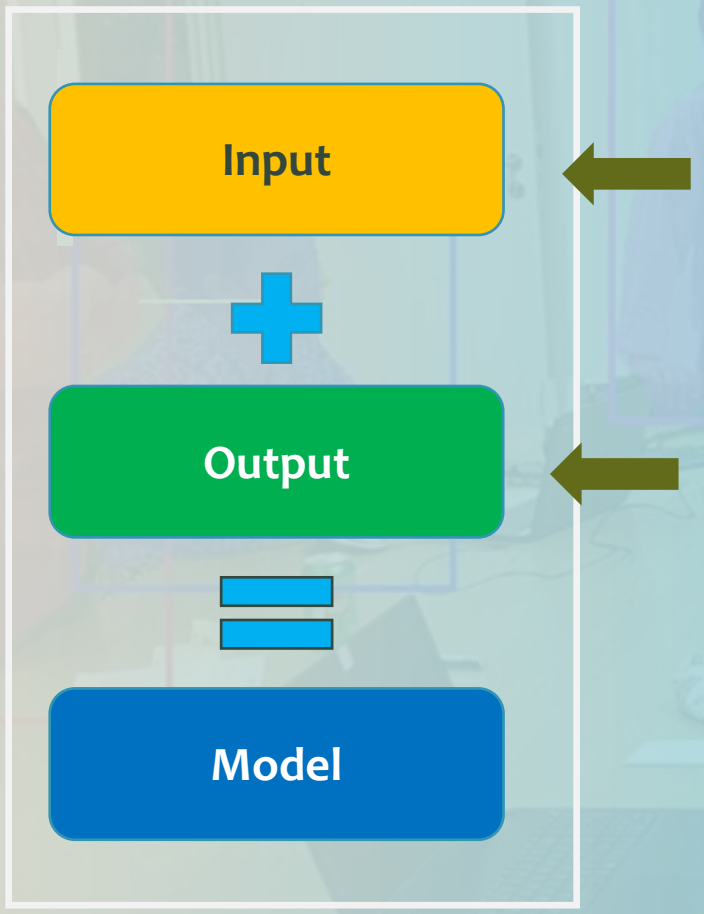
Personality

Mood
Empathy
Humor
Charisms

DIN SPEC 92001-1

Functionality &
Performance
Robustness
Comprehensibility

Data testing



Is the data complete? And if not, how are ‘unknowns’ handled? With default values?

Are formats, like dates and numbers with periods and commas, the same in all the data?

Are the sources of the data known?

Is the data recent enough?

Can older data be re-used?

Is the data correct for the purpose?

Free from biases?

.....

Testing the functionality of the model

Sometimes a bit different

- A/B testing
- Equivalence partitioning
- Boundary value analysis
- Metamorphic testing
- User story testing / Use case testing
- Expert panel testing
- Exploratory testing
- Testing with personas

Personas

Example

chatbot

Each persona represents a group of users

Boundary testing

Streetcar vs Paris-Dakar Car - Model

Classification report

	precision	recall	f1-score	support
Car	1.00	0.91	0.95	23
OffRoadCar	0.91	1.00	0.95	21
accuracy			0.95	44
macro avg	0.96	0.96	0.95	44
weighted avg	0.96	0.95	0.95	44

accuracy of 95% on the Test Dataset

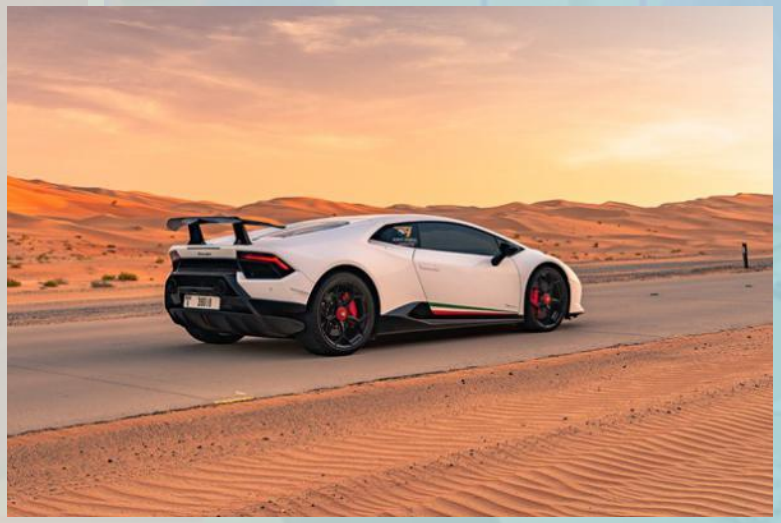
Classify image ...

1. 1-OffRoadCar: 99.96%
2. 0-Car: 0.04%

Image classification

Boundary testing

Car model: street or Paris-Dakar



Amount of light in a picture can be a boundary:

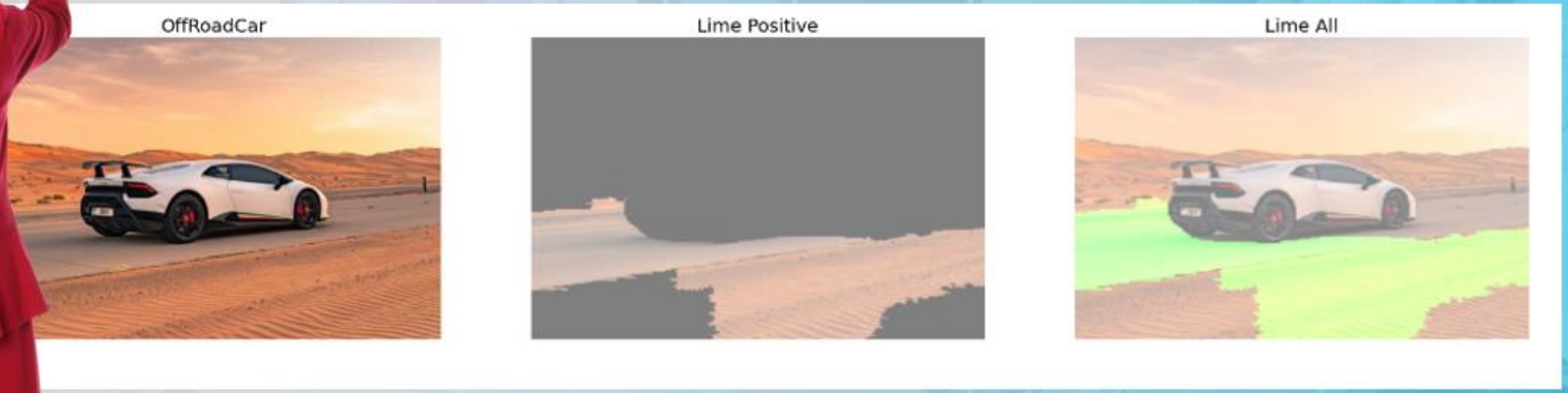
- **Sunset**
- **Low light conditions**
- **More red light**

Classify image …

1. **1-OffRoadCar: 99.99%**
2. 0-Car: 0.01%

But ..

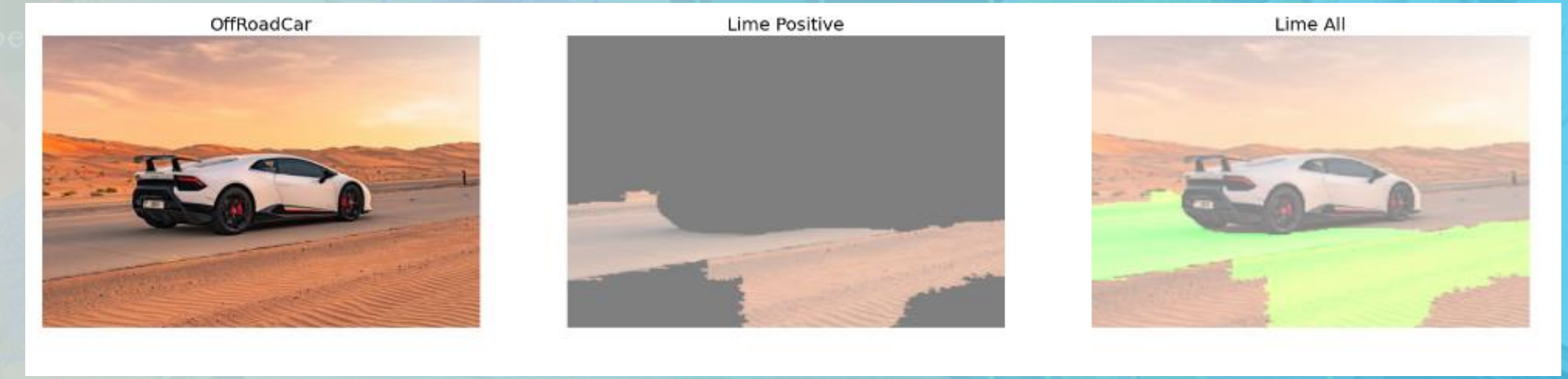
Lime technique to see what is selected

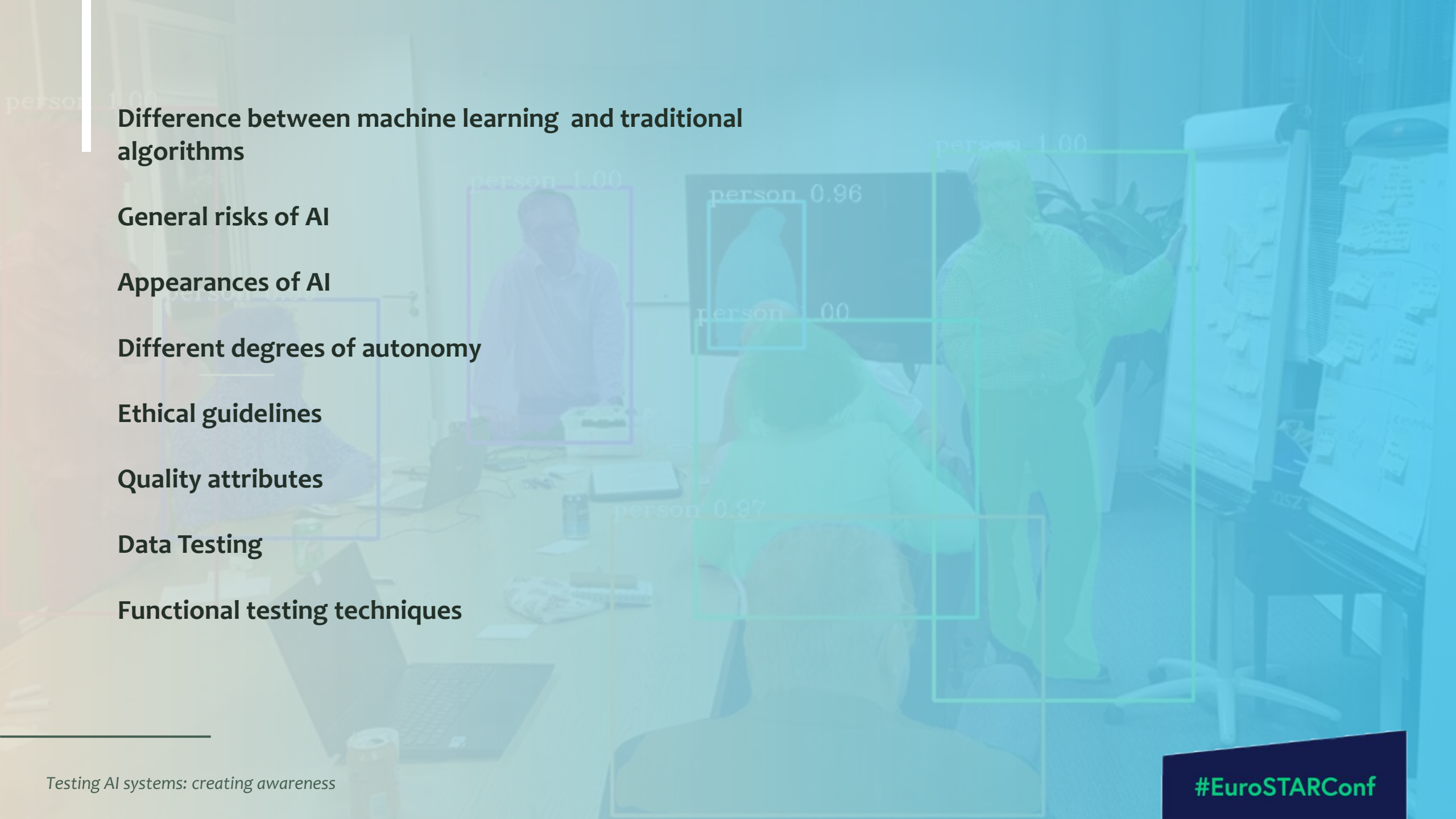


Model accuracy = 95%

Prediction of a street-car vs a Paris-Dakar car?

It is a sand detector!





Difference between machine learning and traditional algorithms

General risks of AI

Appearances of AI

Different degrees of autonomy

Ethical guidelines

Quality attributes

Data Testing

Functional testing techniques

Difference between machine learning and traditional algorithms

General risks of AI

Appearances of AI

Different degrees of autonomy

Ethical guidelines

Quality attributes

Data Testing

Functional testing techniques

We believe that testing AI systems is important, not only to test functionality, but also for ethical aspects.

European guidelines

- Traditional Testing is not enough!
- Start early and throughout the lifecycle to ensure intended behavior and consistency
- Verify, validate, and monitor the data processing and model of the system as a whole and at every stage
- A diverse group of people should design and implement the system

Goals

YOU!

Testing of AI

To give you enough information and confidence to recognize the risks associated with an AI implementation and enable you to shape the AI testing process with your own knowledge and skills.

White paper & working group

YOU!

Starting point for new knowledge

Read it and help us improve it

Share your experience

We like to meet other groups to share knowledge about ‘Testing and AI’

Contact details

Working group Testing and AI

Sander Mol
Peter Collewijn
Hannie van Kooten

ai.workgroup.testnet@gmail.com

www.testnet.org/testnet/p000610/werkgroep/werkgroep-testen-en-ai

www.testnet.org

Thank You!

workgroup. www.testnet.org/ testnet/p000610/werkgroep/werkgroep-testen-en-ai

Website www.testnet.org

Email. ai.workgroup.testnet@gmail.com

LinkedIn. www.linkedin.com/in/peter-collewijn-nl

LinkedIn. www.linkedin.com/in/hannie-van-kooten-nl