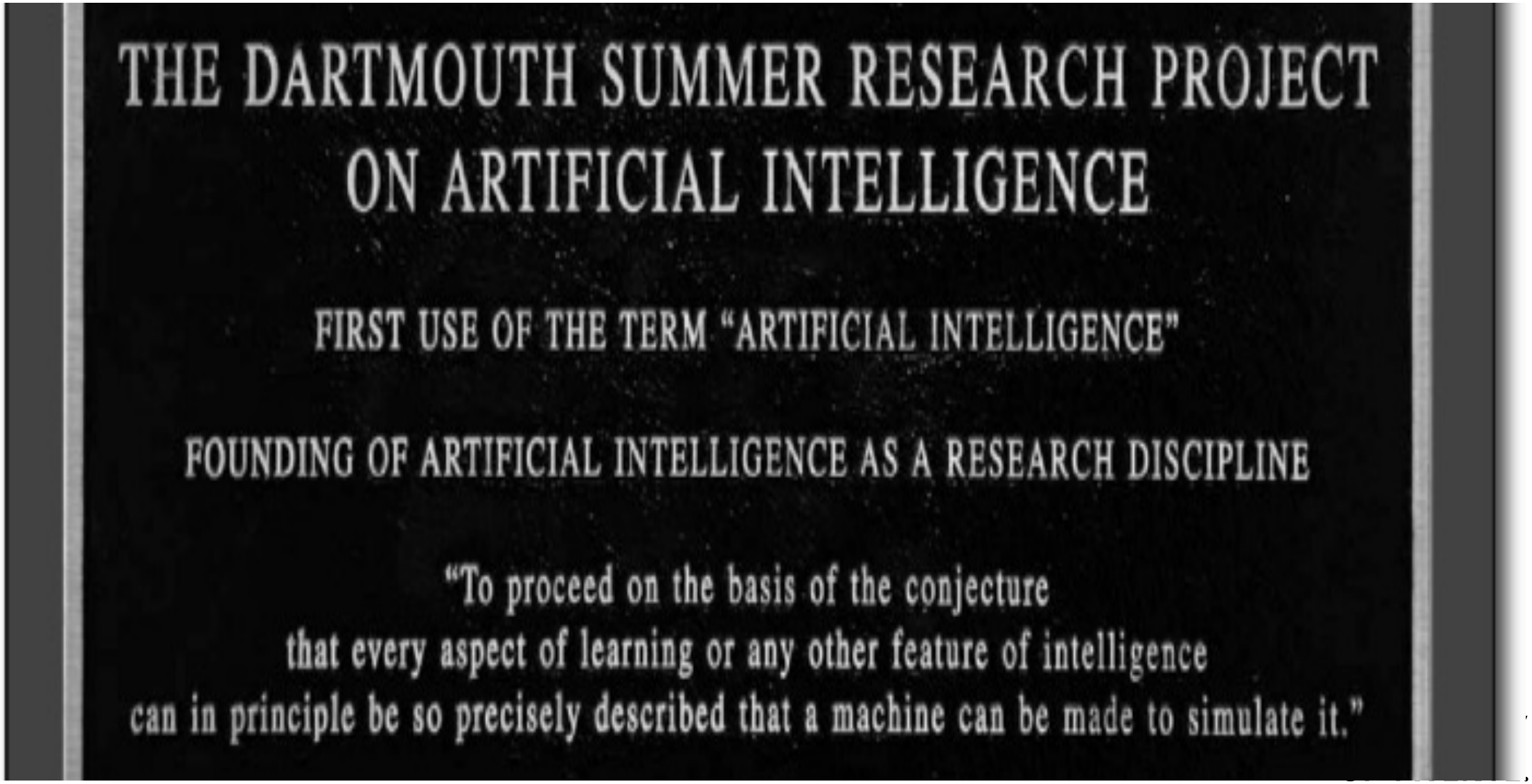
UNIVERSITY OF TWENTE.

MOD6/AI & CYBER SECURITY: INTRODUCTION

PART I: HISTORY OF AI

M. BIRNA VAN RIEMSDIJK

DARTMOUTH, 1956



THE GOLDEN YEARS, 1956-1974: SHAKEY

Helen Chan Wolf

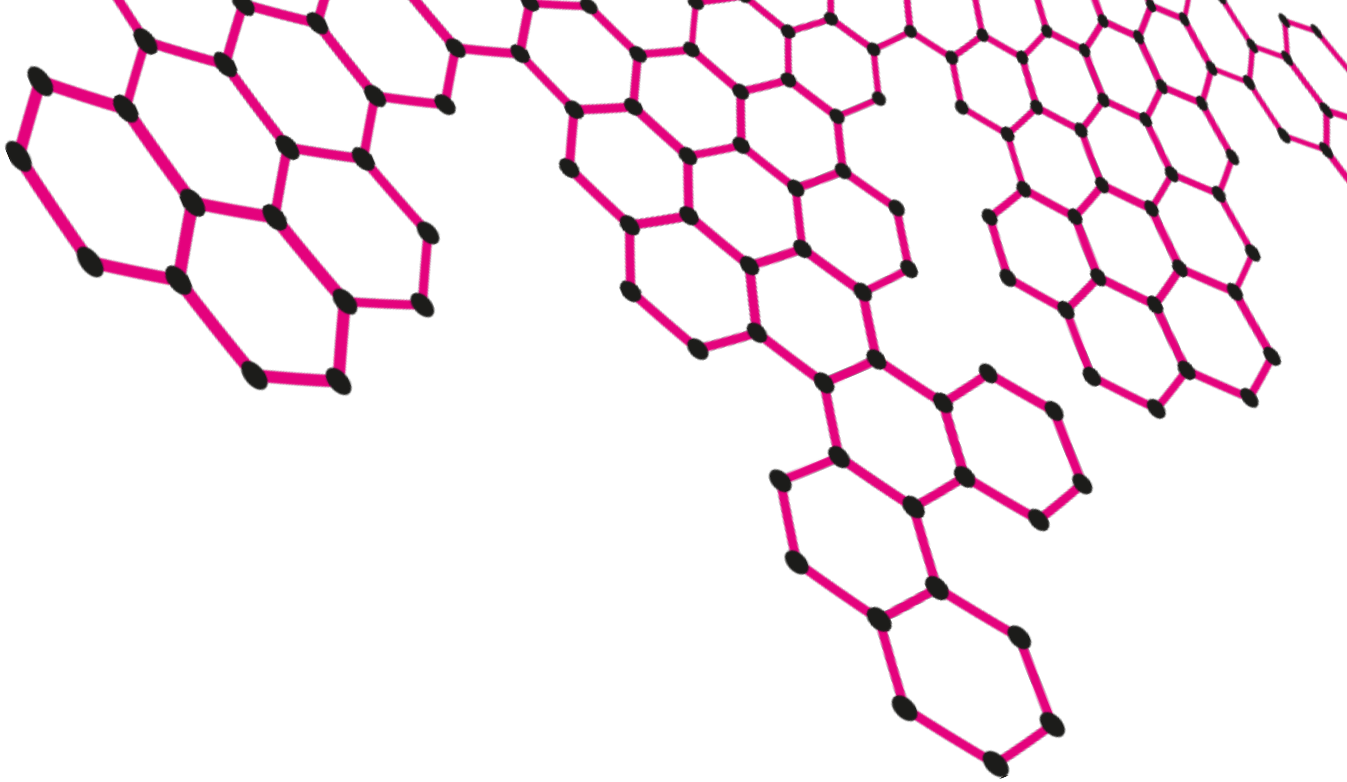
UNIVERSITY OF TWENTE.

AI WINTERS AND SPRING

- Inflated expectations of what AI might be capable of
- Problems:
  - modelling commonsense knowledge
  - intractability
  - frame problem
- 1st AI winter 1974-1980
- 1980-1987: Rise of expert systems
  - rule-based systems for modelling the knowledge of experts in a machine
  - but brittle, difficult to update
- 1987-1993: 2nd AI winter
- 1993-2011
  - increase in computing power
  - intelligent agents
  - decision theory
- 2011 – now
  - big data, machine learning

AI PENDULUM

- Oscillation of inflated expectations with high levels of funding, and realization of limitations followed by AI winters
- Oscillation of attention for different types of AI techniques: symbolic vs. subsymbolic/data-oriented
- Recent years
  - data-oriented approaches at the basis of recent AI successes
  - but: also realization of limitations and risks
- The future
  - bringing together different AI subdisciplines to use their complementary strengths and work towards well-functioning human-centered AI?



MOD6/AI & Cyber security: Introduction

PART II: WHAT IS AI?

M. BIRNA VAN RIEMSDIJK

12 NOVEMBER 2021

TYPES OF DEFINITIONS

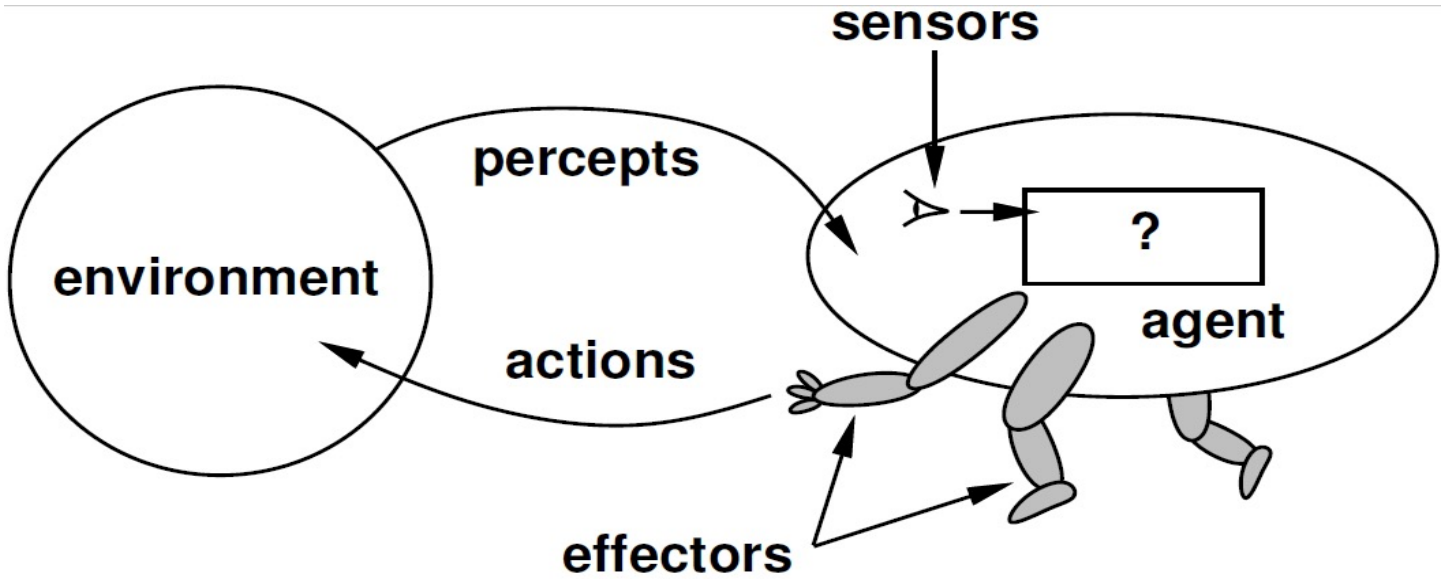
1. AI as a system (agent) interacting with the environment

2. AI as a collection of computational techniques for creating “intelligent” systems (this course)

3. AI as a multidisciplinary research field (HCI/value-sensitive design course)

AI SYSTEM: AGENTS

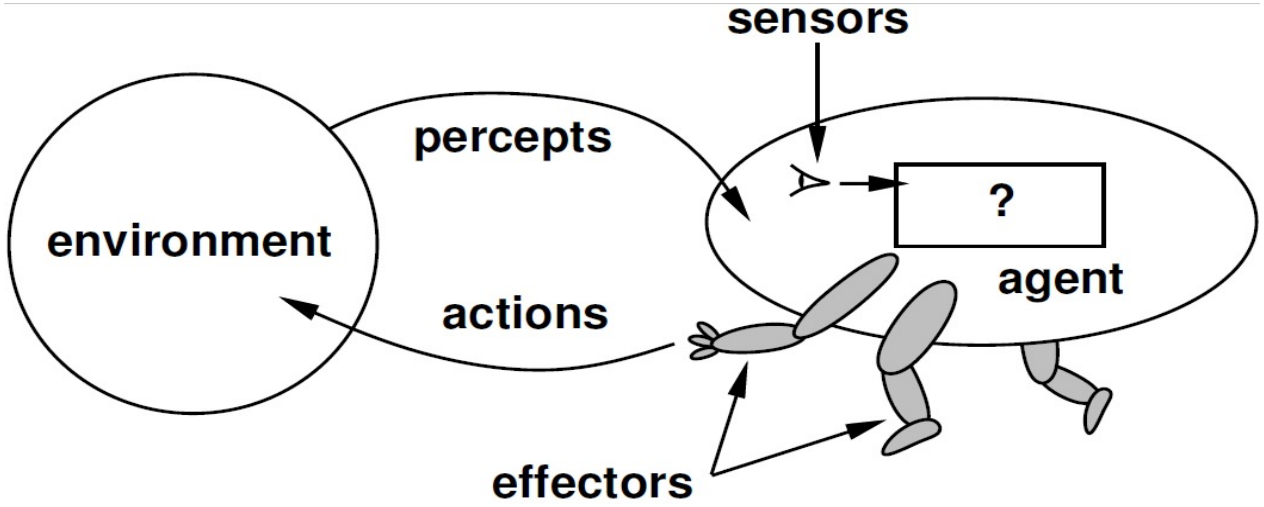
- Definition according to Russel and Norvig:
  An **agent** is anything that can be viewed as percieving its **environment** through **sensors** and acting upon this environment through **actuators.**



UNIVERSITY OF TWENTE.

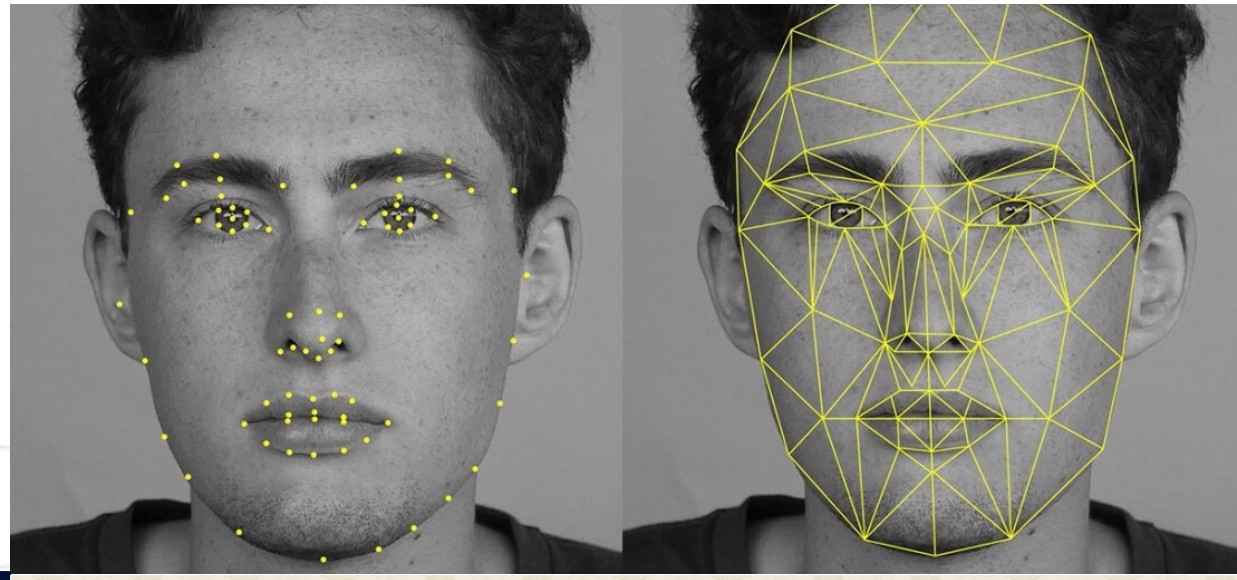
“INTELLIGENT” AGENT: RATIONALITY

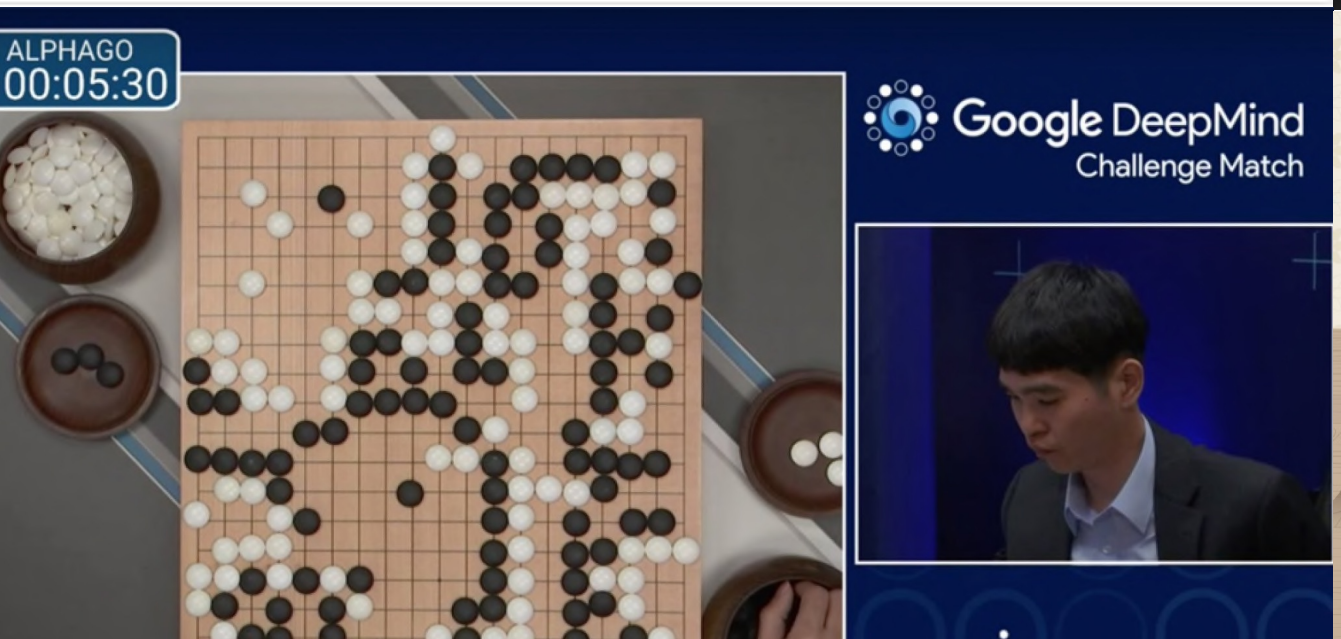
- Definition according to Russel and Norvig:
  A rational agent is an agent that selects actions in order to maximize its performance measure, given evidence provided by the percepts and any built-in knowledge
- Action selection [?] = Reasoning/information processing & decision making



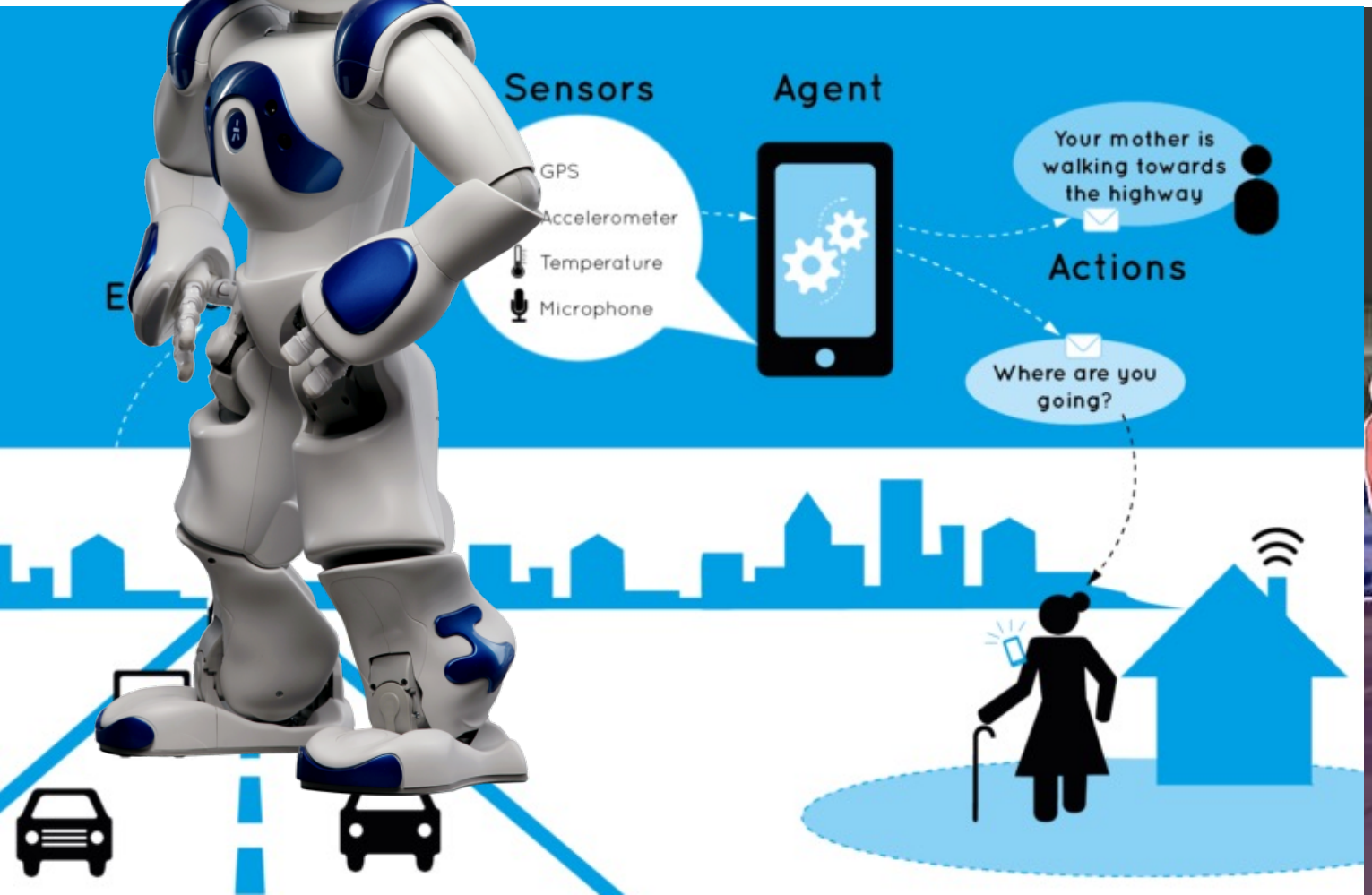
SOFTWARE-BASED AI SYSTEMS

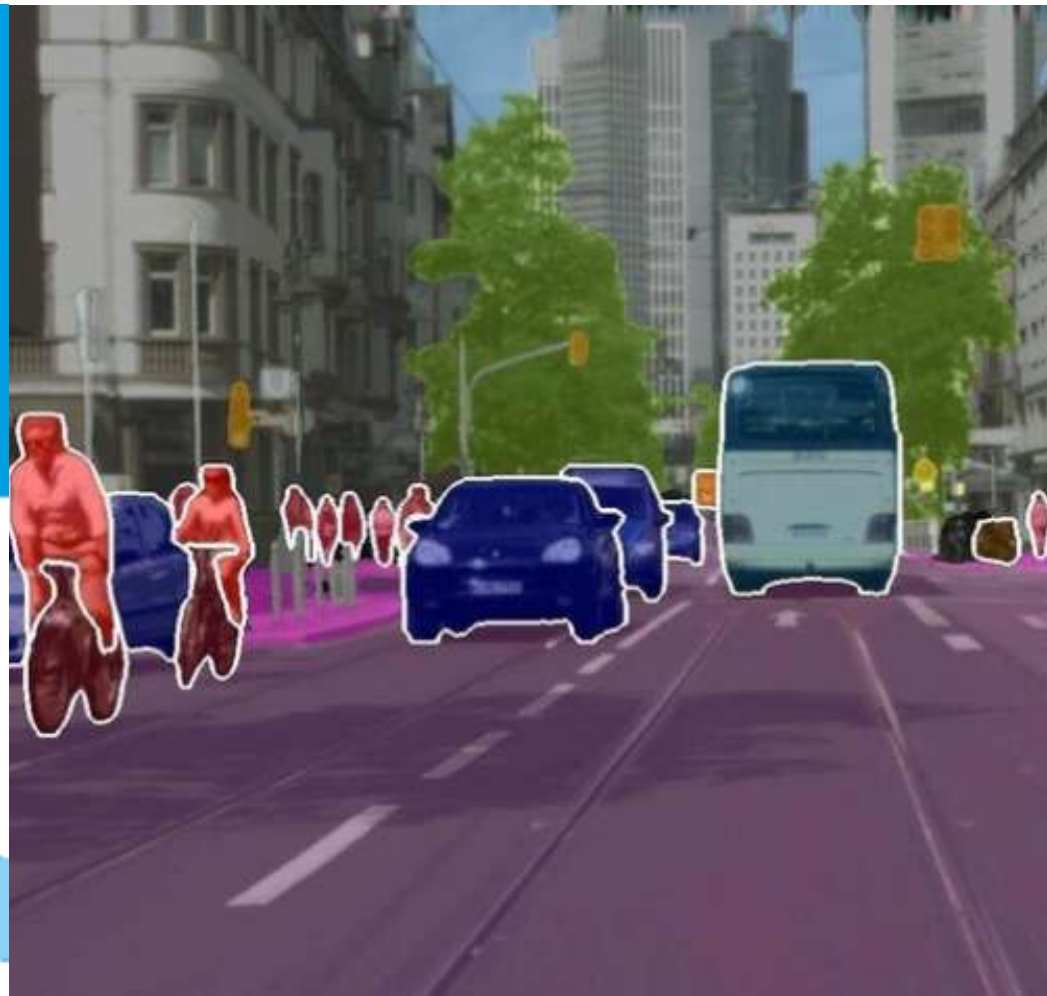
Google





AI SYSTEMS EMBEDDED IN HARDWARE





AI SYSTEM

- See also report “A definition of AI: Main capabilities and scientific disciplines” by the EU High-Level Expert Group on Artificial Intelligence, 2019.

AI AS FAMILY OF COMPUTATIONAL TECHNIQUES FOR INTELLIGENT SYSTEMS

Machine Reasoning:
logic-based approaches

Week 1-3

Optimisation:
algorithmic approaches

Week 2

Machine Learning:
data-oriented approaches

Week 4-6

Towards combining knowledge-based and data-oriented approaches! Cf.:

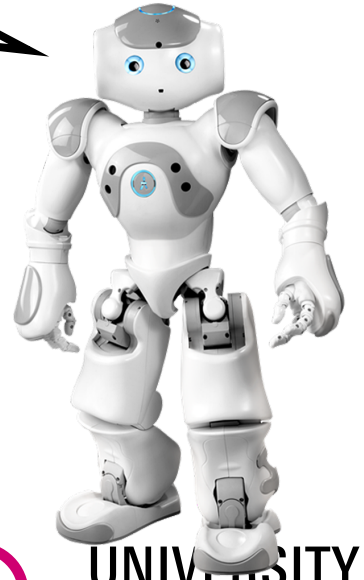
CLAIRE

MACHINE REASONING (1)

Explicit knowledge representation + inferences (= reasoning) to derive new knowledge

What do you usually have for breakfast?

- *Normally I have a slice of bread to eat.*
- *I always have a warm drink and a juice.*
- *If I have an early meeting, I eat yoghurt.*
- *I don't drink coffee with yoghurt.*



MACHINE REASONING (2)

INFORMAL DESCRIPTION	FORMAL REPRESENTATION
1. Normally I have a slice of bread to eat.	1. $bread$
2. I always have a warm drink and a juice.	2. $(tea \vee coffee) \wedge juice$
3. If I have an early meeting, I eat yoghurt.	3. $earlyMeeting \rightarrow yoghurt$
4. I don't drink coffee with my yoghurt.	4. $yoghurt \rightarrow \neg coffee$

It is Thursday. On Thursday there is a weekly meeting which starts at 8am. What do I have for breakfast on this day?

$earlyMeeting$
$yoghurt\ (3), \neg coffee\ (4), tea\ (2), juice\ (2)$
$bread\ (1)$

MACHINE REASONING (3)

ADVANTAGES

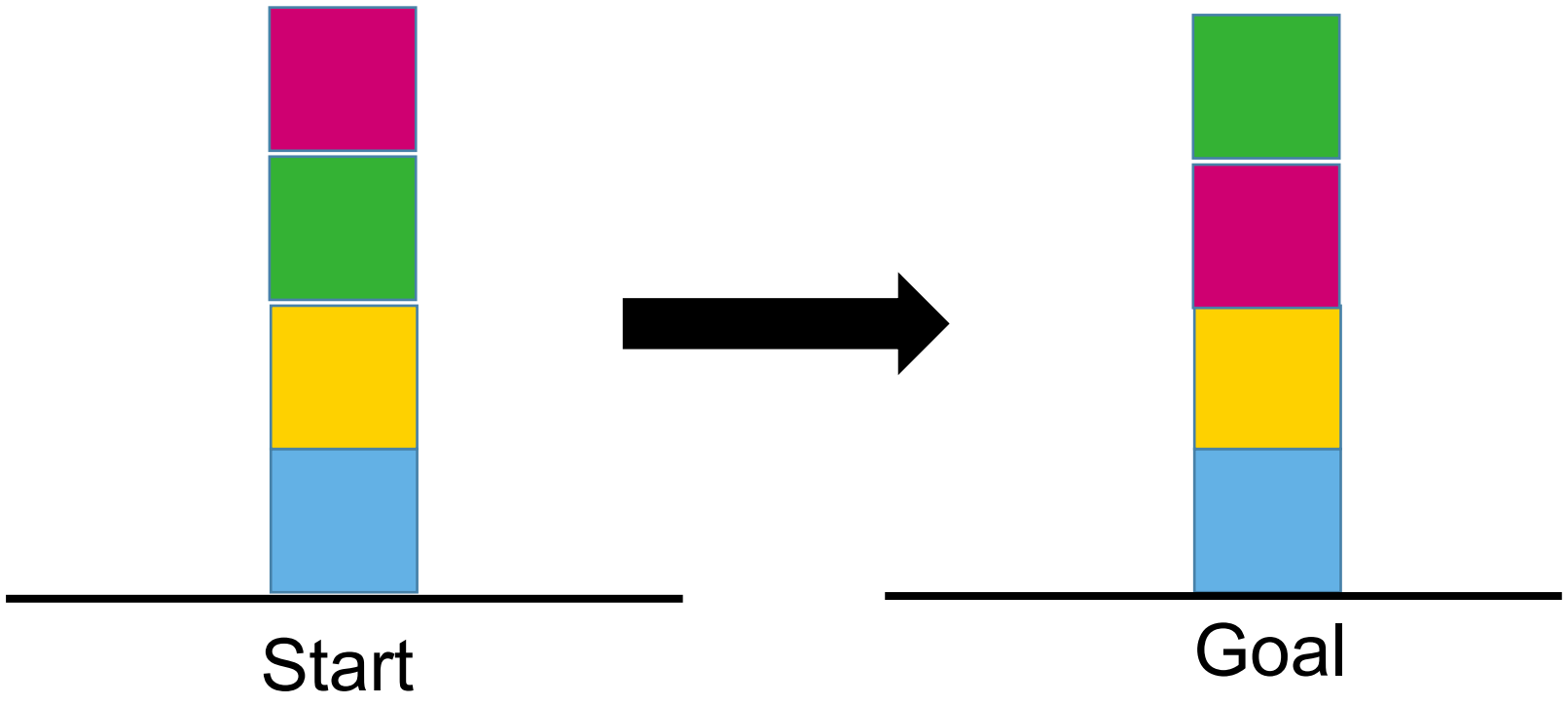
- Precise specification
  - *What do you mean by early meeting?*
  - Usable by a computer: computational logic
- Potentially more explainable and open to correction
  - *Why did I get bread for breakfast?*
  - *I don't want bread when I have yoghurt*

DISADVANTAGES

- Restricted expressiveness
- Some knowledge hard to capture explicitly
  - *Describe what a coffee looks like*
- How to obtain all relevant knowledge?
  - *What is a warm drink? What do I put on my bread? What to do if I am out of juice?*
  - knowledge engineering bottleneck
  - common sense knowledge: *I only have one warm drink for breakfast*

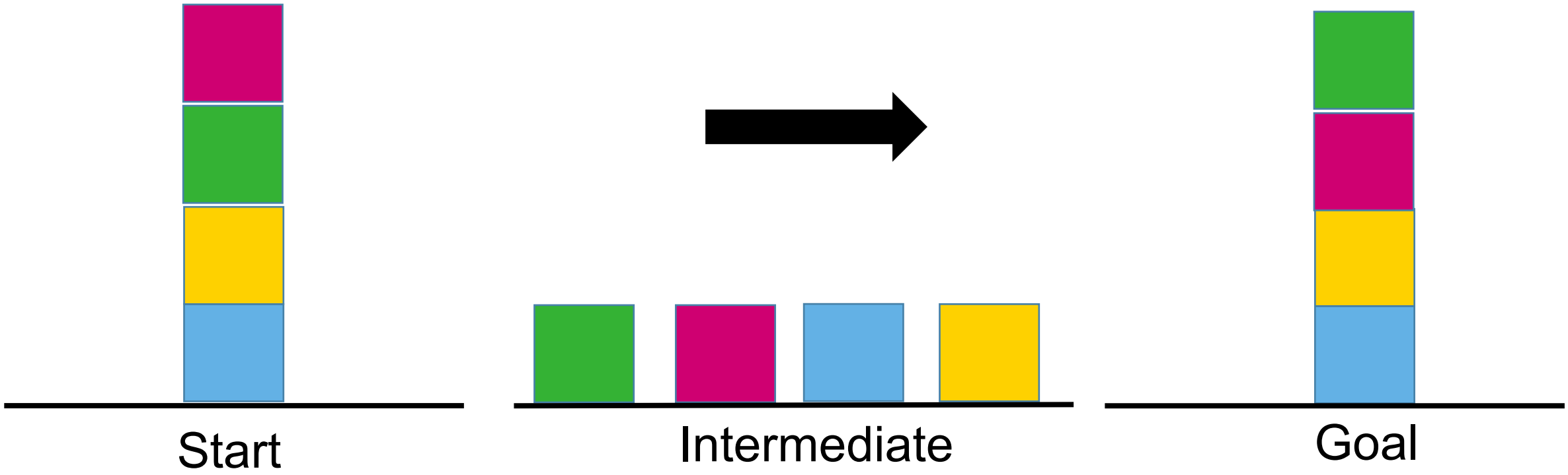
OPTIMISATION (1)

algorithms for finding the best solution according to some criterion of optimality, e.g., number of steps, execution time,…

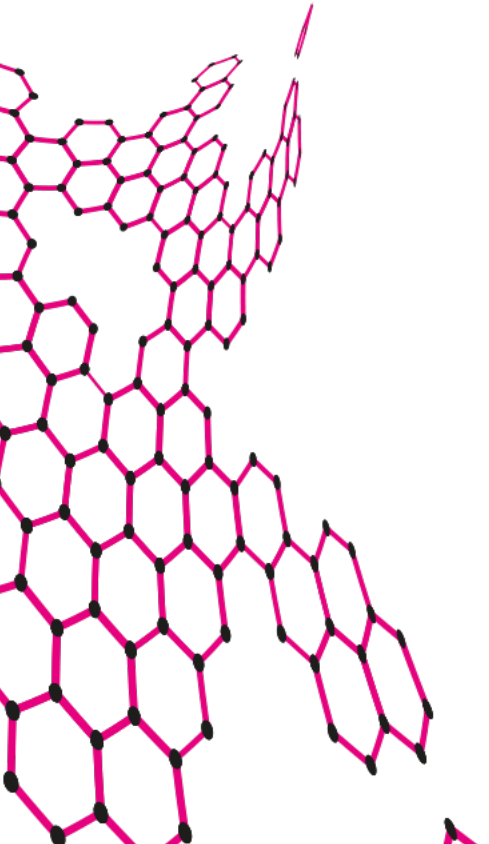


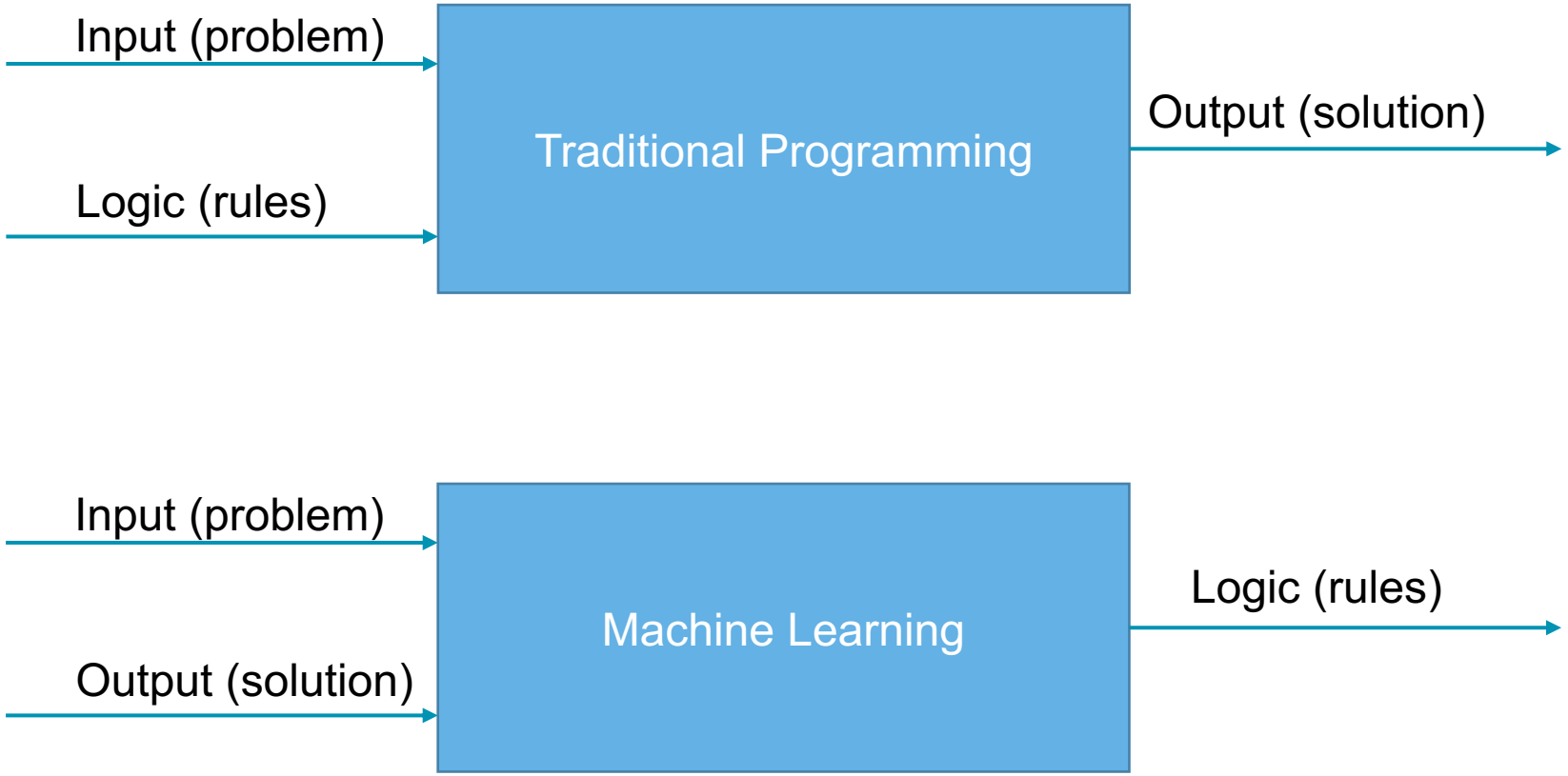
OPTIMISATION (2)

algorithms for finding the best solution according to some criterion of optimality, e.g., number of steps, execution time,…



MACHINE LEARNING (1)



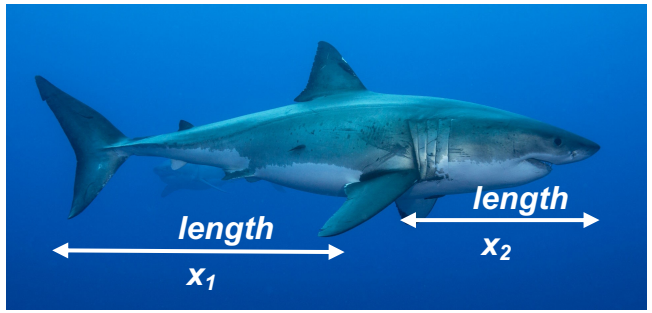


A paradigm to infer knowledge from input/output pairs → supervised learning

UNIVERSITY OF TWENTE.

MACHINE LEARNING(2)
Example: classification

data: observations, e.g. vectors of num. values



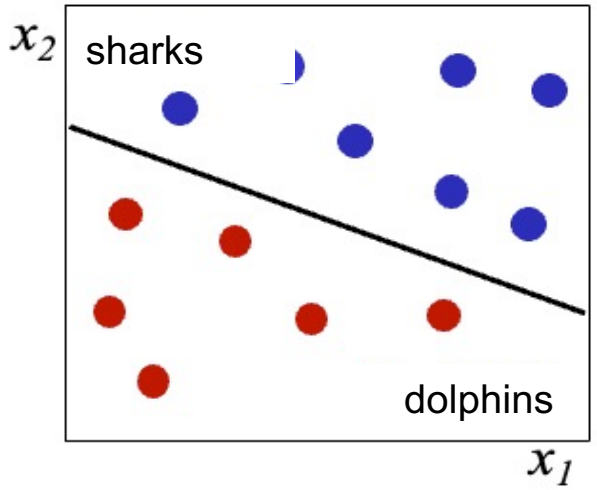
example data set

classification tasks: assign data to a **category**

model: linear separation

$\alpha x_1 + \beta x_2 > \gamma$ "shark"

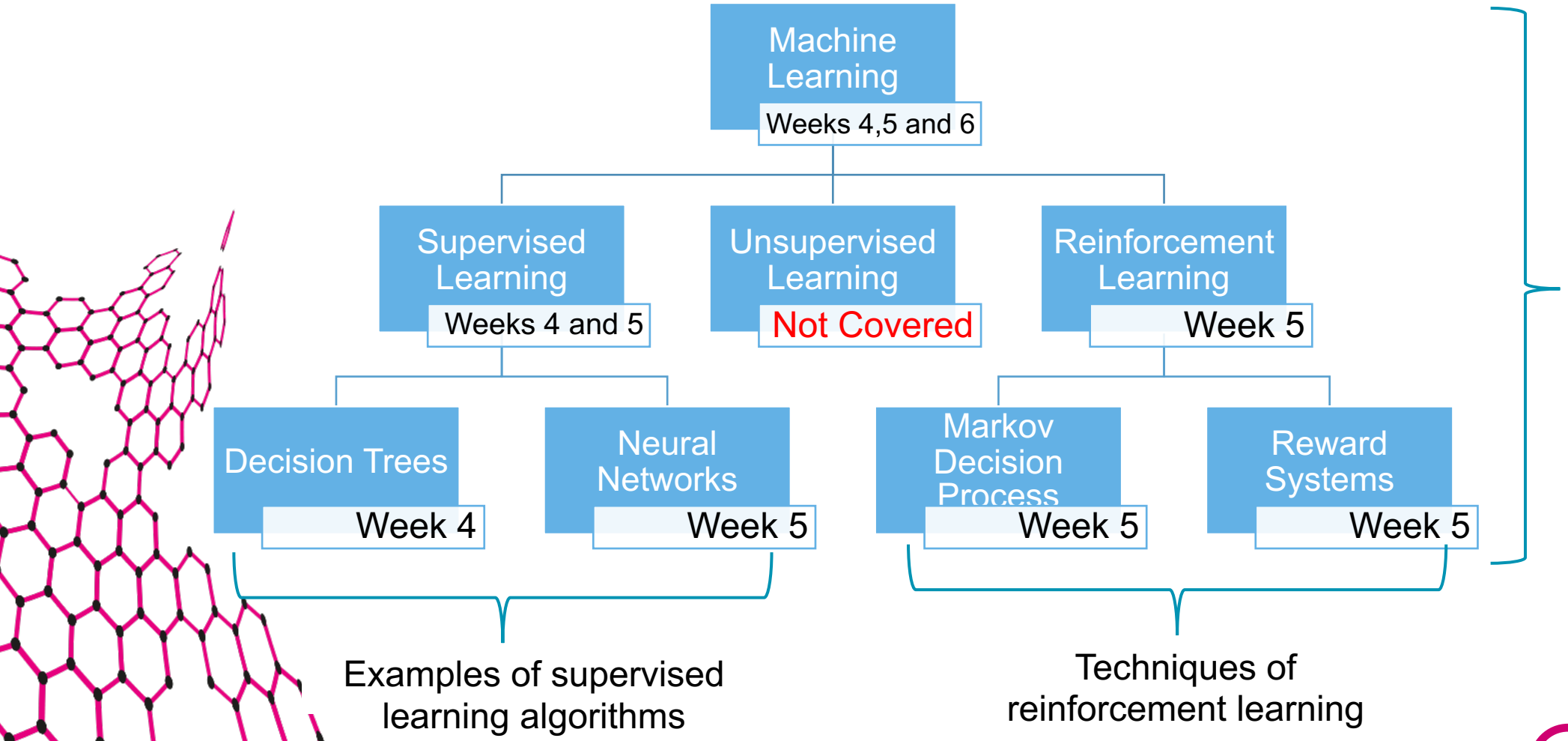
else "dolphin"



training:

⇨ optimize parameters α, β, γ

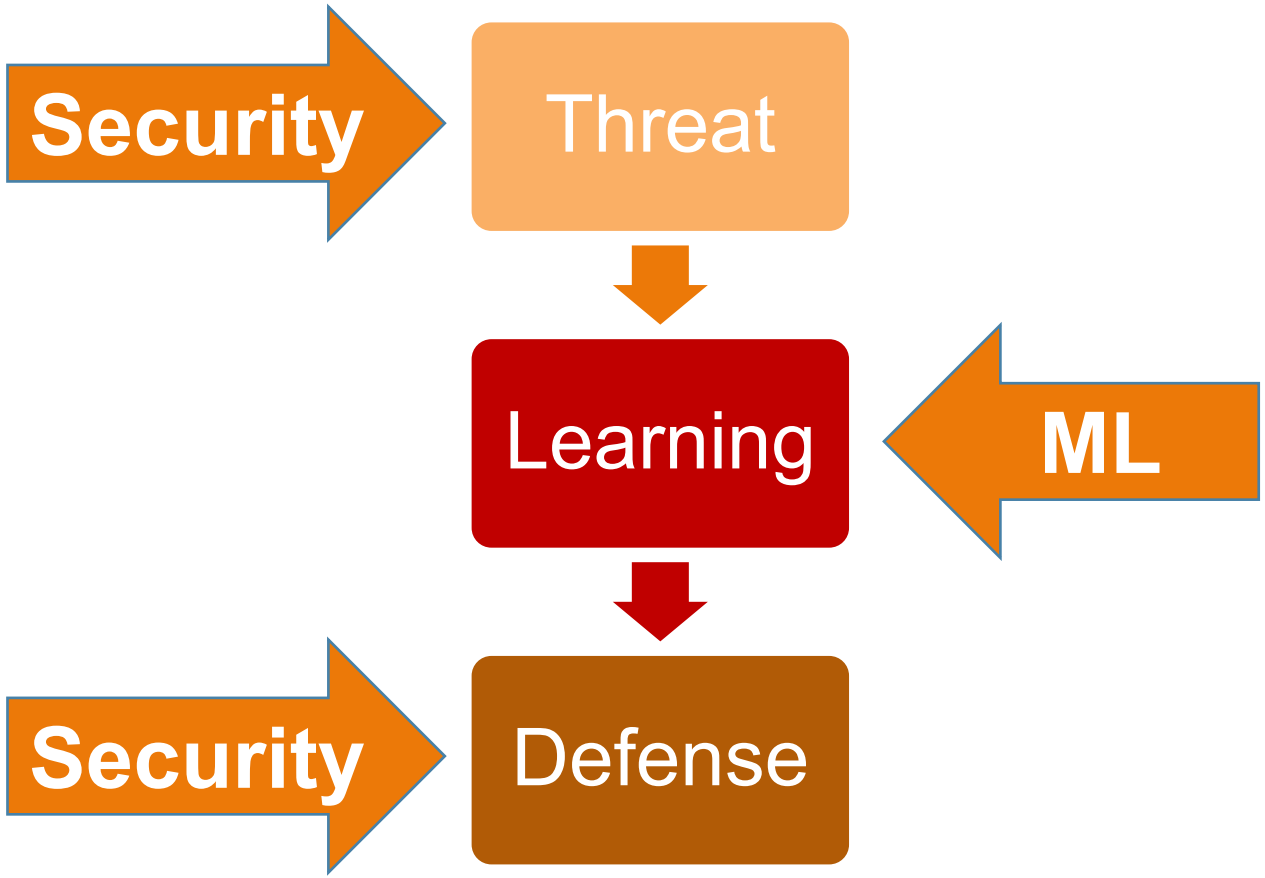
MACHINE LEARNING (3)



UNIVERSITY OF TWENTE.

SECURITY & MACHINE LEARNING

- Automate & generalize identification of new threats
- Domain specific challenges
  - Robustness against evasion attacks (adversarial scenario)
  - Robustness against evolving attacks



(SUPERVISED) MACHINE LEARNING

ADVANTAGES

- no need to model knowledge explicitly
  - Relation between input and output not known beforehand
  - Addresses knowledge engineering bottleneck
  - Commonsense knowledge
- discover new relations between input and output
  - But: correlation is not causation!

CHALLENGES

- how to handle noise, e.g. wrong labels in example data
- biases in data → biased algorithms
- optimization of generalization ability
- validation: can we test and predict the performance with respect to new unseen data?
  - Explainability & scrutability

... more will be discussed in Week 4!

AI AS A MULTIDISCIPLINARY FIELD

- Viewing AI as a socio-technical system which involves technology, people, and social and organizational context
- Combining expertise from many fields:
  - computer science
  - human-computer interaction
  - cognitive science
  - (social) psychology
  - law
  - ethics & philosophy
  - science and technology studies (STS)
  - gender studies
  - political science
  - neuroscience
  - embedded systems
  - robotics
  - economics
  - linguistics
  - ….