ER DEN KUNSTIGE INTELLIGENSEN TIL Å STOLE PÅ?

CAN ARTIFICIAL INTELLIGENCE BE TRUSTED?

Ole Jakob Mengshoel

ole.j.mengshoel@ntnu.no

Professor, Dept. of Computer Science, NTNU
Head, Norwegian Open AI Lab, NTNU
Adjunct Faculty, Dept. of Electrical and Computer Engineering, CMU

https://www.ntnu.edu/employees/ole.j.mengshoel
http://sv.cmu.edu/directory/faculty-and-researchers-directory/faculty-and-researchers/mengshoel.html
https://works.bepress.com/ole_mengshoel/
https://www.ntnu.edu/ailab

NOKIOS - a Norwegian Conference for eGovernment.
Clarion Hotel & Congress, 24 Oct. 2018, Trondheim, Norway

norwegian open ai lab

THE NORWEGIAN OPEN AI LAB

- To enable both **basic** and **applied research**
- To support a **wide variety of research areas**
- To perform **research at highest international level**
- To foster **cross-disciplinary collaboration**

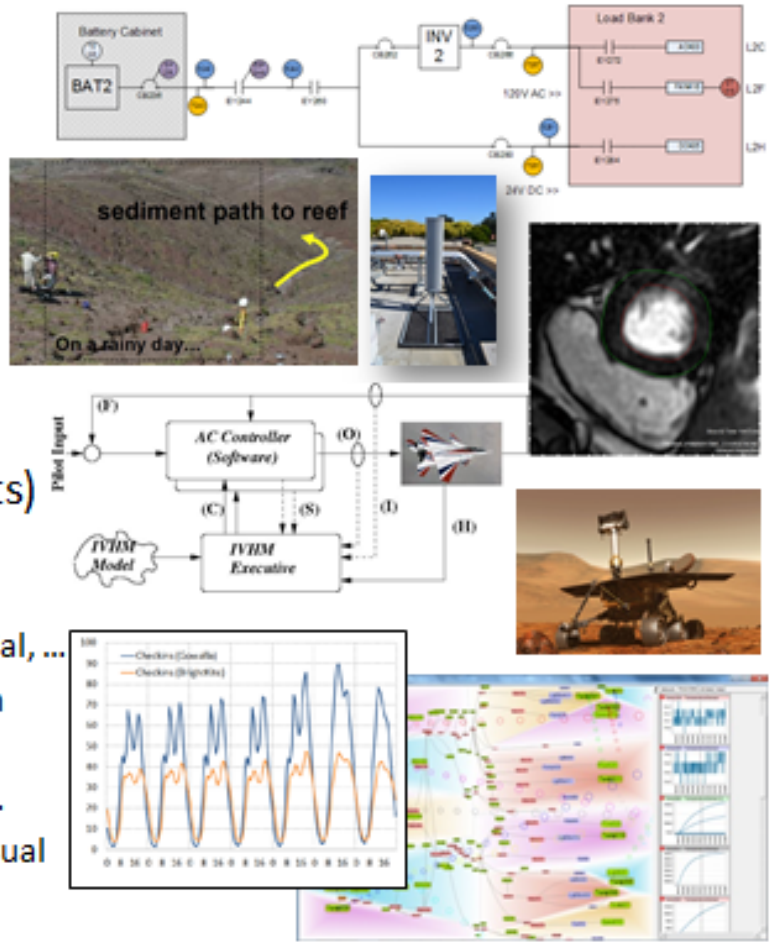
From the opening of the AI Lab at NTNU, Trondheim.

https://www.ntnu.edu/ailab

norwegian open ai lab

RESEARCH AND APPLICATION OVERVIEW: MENGSHOEL LAB

- Algorithms
  - Machine learning
  - Stochastic optimization
  - Inference, compilation, HW/SW…
- Analysis (and Models)
  - Probabilistic graphical models
  - Bayesian networks
  - Markov chains
- Applications (and Experiments)
  - System health: Power, software, aerospace, …
  - Networks: Computer, telecom, social, …
  - Mobility : Vehicles, devices, human activity, aerospace, traffic, …
  - Science: Earth sciences , medical, …
  - HCI: Recommendation systems, visual analytics, GUIs, …



https://works.bepress.com/ole_mengshoel/

norwegian open ai lab

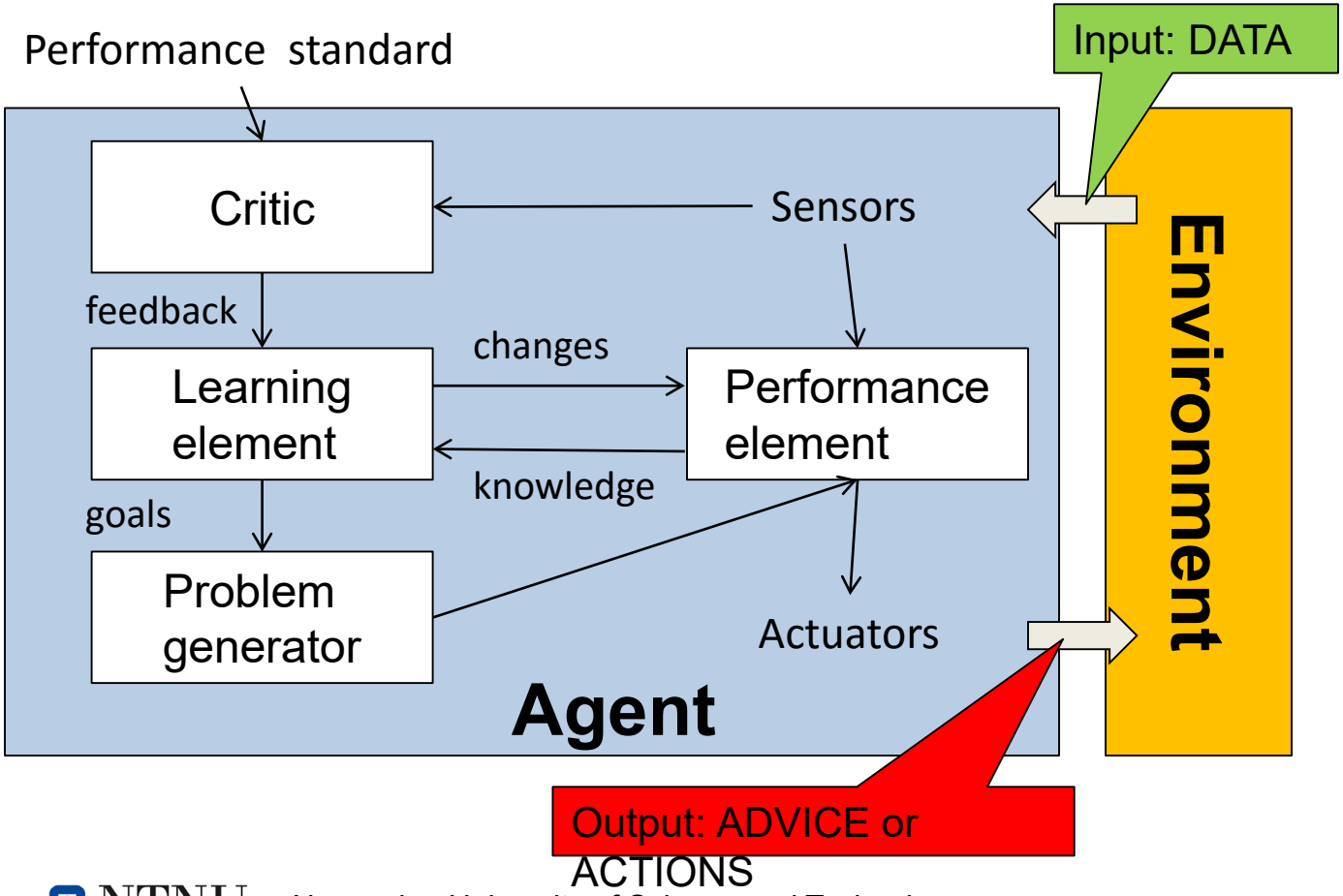
Artificial Intelligence (AI): When, What, How, and Why

Successes of Artificial Intelligence (AI)

- May 1997: Deep Blue was the first computer system to defeat a reigning world champion. It beat Kasparov 3½–2½ under standard chess tournament time controls.
- October 2005: Stanford Racing Team wins the DARPA Grand Challenge, a 212 km (132 mi) off-road course, near the California/Nevada state line.
- April 2006: Google introduces Translate, a services that translates text from one language into another. United Nations and European Parliament transcripts were used to gather linguistic data.
- February 2011: IBM's Watson computer system wins first place and $1 million in Jeopardy! against former winners Brad Rutter and Ken Jennings.
- October 2011: Apple introduced the iPhone 4S with Siri, an intelligent assistant with a voice recognition user interface.
- March 2016: AlphaGo, using Google's DeepMind AI, won its third Go match against Lee Sedol, one of Go's most dominant players.
- May 2016: Google Assistant, a virtual personal assistant, engages users in two-way conversations via voice and keyboard. It can search the Internet, schedule events and alarms, and show information from the user's Google account.

Artificial Intelligence: Rational Agent

For each possible percept sequence, a *rational agent* selects an action that is expected to maximize its performance measure given the percept sequence and the agent's knowledge [Russell and Norvig, 2009].



Note: Artificial intelligence is defined in terms of *broad problems* and *diverse methods*. (This is in contrast to many other fields, often defined in terms of narrow problems or a few methods.)

Machine Learning: From the Fringe to the Center of the AI Universe

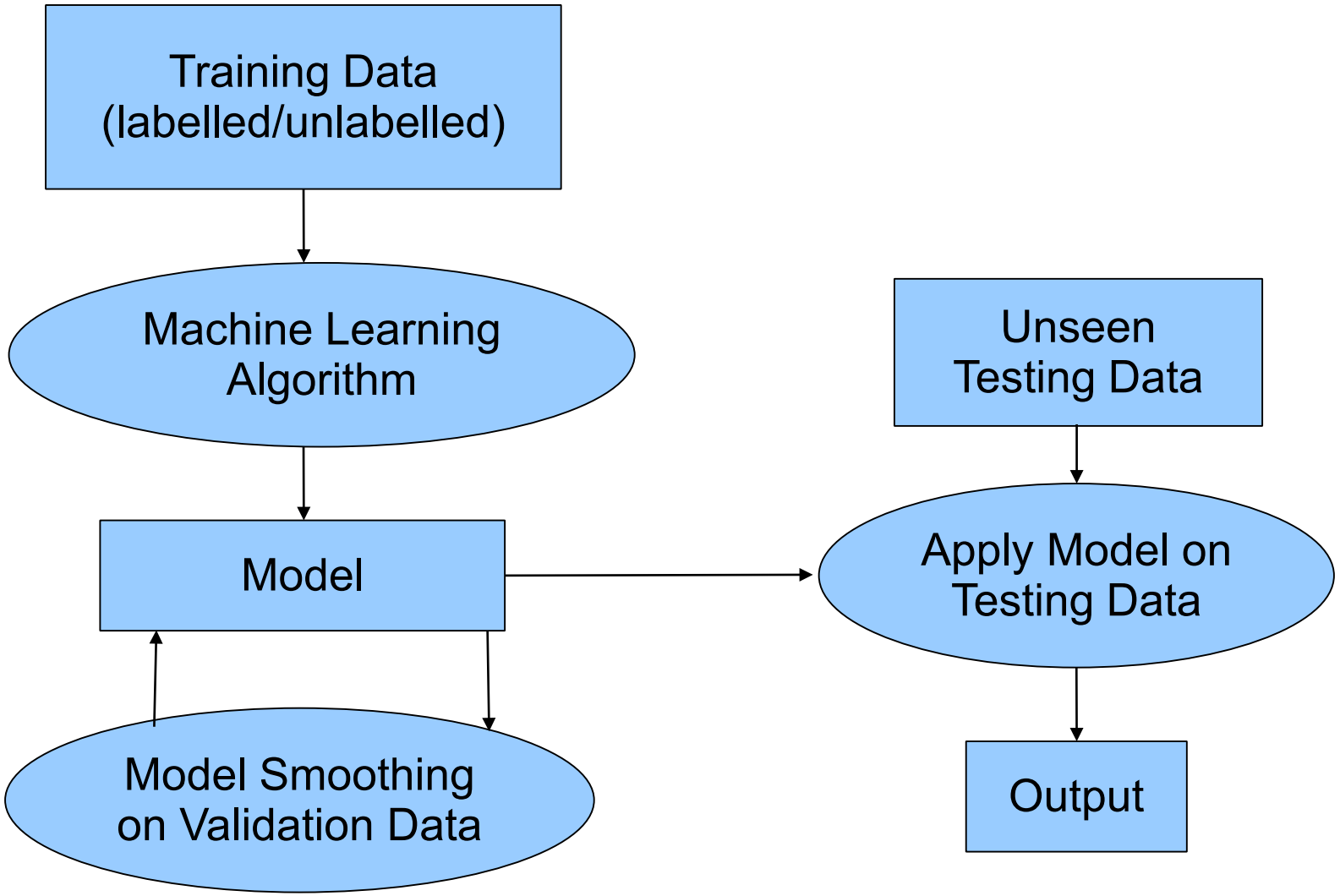
The central role of learning

Although researchers in artificial intelligence and psychology have long recognized the importance of learning, this topic has not always been the central focus of these fields. In the first years of AI, considerable attention was given to learning issues, but as pattern recognition and AI developed separate identities, learning research became associated with the former while the latter concentrated on problems of representation and performance. A similar phenomenon occurred in psychology. The behaviorist paradigm was almost exclusively concerned with learning phenomena, but as information processing psychology gained in popularity, psychologists turned their sights towards memory and performance phenomena and all but abandoned efforts to explain the learning process.

However, the past five years have seen a resurgence of interest in learning within both artificial intelligence and cognitive psychology. This has resulted partly from dissatisfaction with pure performance models of intelligence. One of the major insights of both fields has been that, except in the simplest domains, intelligent behavior requires significant knowledge of those domains. Although this insight has led to successful applied AI systems and to accurate psychological models of domain-specific performance, it has not led to systems or theories having any great degree of generality. By refocusing their efforts on learning, many researchers hope to discover more general principles of intelligence. In the case of psychology, such principles would lead to more encompassing theories of human behavior that move beyond particular domains. In the case of applied AI, general learning methods might let one automate the construction of knowledge-intensive systems, saving man-years of effort for each application area.

Pat Langley's Editorial in the "Machine Learning" journal's Inaugural Issue, 1986.

Machine Learning: Typical Approach



Why is Machine Learning Currently Achieving Success?

- Improvements in models, algorithms, and software:
  - Focus of my research and that of others in the AI Lab
- Improvements in hardware - computers are good at:
  - Fast computing of millions of simple operations
  - Huge memories: Cache, RAM, disk, ...
- Availability of large and interesting data sets:
  - Improvements in sensor technologies (including cameras, IoT, Internet, and Web)
  - Improvements in distribution technologies (including Internet and Web)
- No “real intelligence” needed:
  - We can start doing interesting machine learning work before human and animal intelligence is fully understood

Summary: Current machine learning methods take advantage of what computers are good at as well as the large data sets currently available.

Which Machine Learning Algorithm to Use?

"Tribes" in machine learning (and AI?) [Domingos, 2015]:

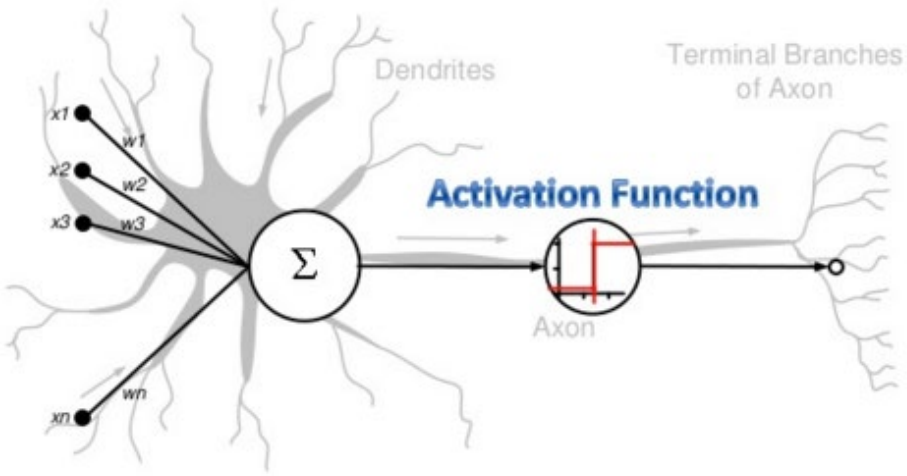
- **Evolutionaries**: use methods from evolution and genetics - evolutionary algorithms, genetic algorithms, and genetic programming [Darwin, 1859] [Holland, 1975] [Goldberg, 1989].
- **Bayesians**: learning as inference using - Bayes rule, Bayesian networks, and probabilistic graphical models [Duda & Hart, 1973] [Pearl, 1988] [Jelinek, 1997][Darwiche, 2009] [Koller & Friedman, 2009] [Blake, 2011].
- **Connectionists**: reverse engineer the brain – from neural networks to deep learning [Werbos, 1974] [Rumelhart & McClelland, 1986] [Bengio, 2009] [Goodfellow et al., 2016].
- **Symbolists**: intelligence as symbol manipulation[Newell & Simon, 1976] [Michalski et al., 1983] [Breiman et al., 1984] [Quinlan, 1992].
- **Analogizers**: learning by recognizing similarities [Boser et al., 1992] [Kolodner, 1993] [Cristianini & Shawe-Taylor, 2000].

Which algorithm(s) is (are) "best" depends on your project – data, goal, skills, resources, and so forth.

Tribe: Connectionists

Connectionists: Artificial Neural Networks (ANNs)

- Bio-inspiration: information processing in biological systems (brains).
- Mathematical abstraction of the biological processes.



Slide courtesy of: Andrew L Nelson

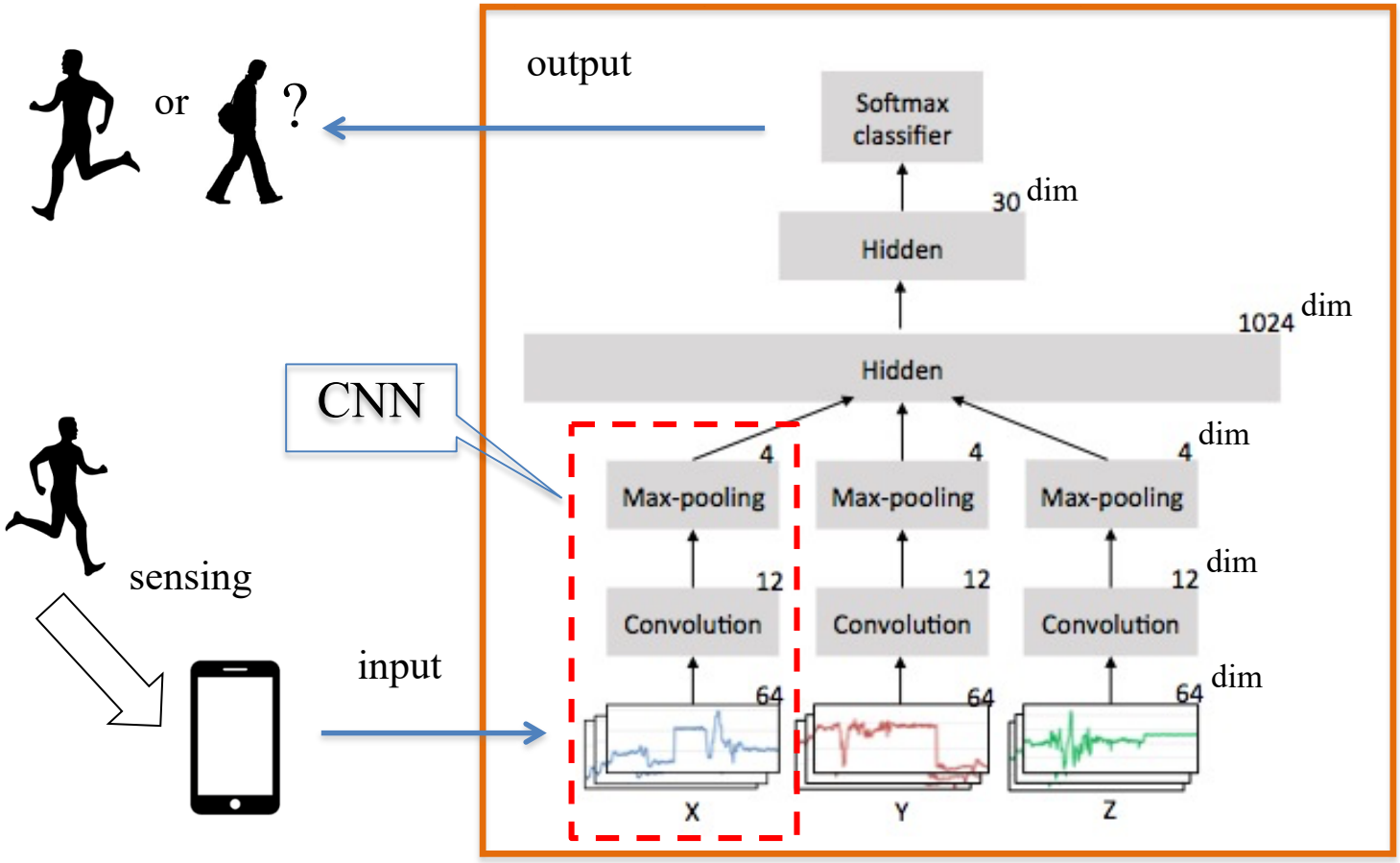
Generations of ANNs

- 1st generation (1958 – 1969):
  - Rosenblatt, F., The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain, Psychological review, 1958.
  - M. Minsky and S. A. Papert. Perceptrons, 1969.
  - Impossibility of representing linearly inseparable functions (e.g., XOR).
- 2nd generation (1980 – 2000):
  - Rumelhart, D. E., & McClelland, J. L., & the PDP Research Group, Parallel distributed processing: Explorations in the microstructure of cognition. volume I & II.
  - Not as feasible as other ML models (e.g., graphical models, SVM, boosting).
- 3rd generation (2006 – present):
  - Hinton, G. E. and Salakhutdinov, Reducing the dimensionality of data with neural networks, Science, 2006.
  - Lee, R. Grosse, R. Ranganath, and A.Y. Ng, Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations, ICML 2009.
  - Application success of autoencoders, RBMs, CNNs, RNNs, …

Deep Neural Networks (DNNs): The 3rd Generation of ANNs

- Deep neural networks have more layers than ANNs in previous generations
  - Typically 4-7 layers
- DNNs can take advantage of big data: With more data, DNN performance often improves
- GPUs accelerate the DNN network training process
  - Computational challenge of training
- New or "new" techniques:
  - Pre-training – find good locally optimal by getting good initial value
  - Dropout – prevent over-fitting
  - Momentum – find better local optima
  - Convolutional neural networks (CNNs) - connectivity pattern similar to the cortex of animals
  - …

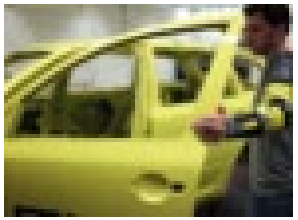
Convolutional Neural Networks (CNNs) for Human Activity Recognition (HAR)



M. Zeng, L. T. Nguyen, B. Yu, O. J. Mengshoel, J. Zhu, P. Wu, and J. Zhang. *Convolutional Neural Networks for human activity recognition using mobile sensors*, Proc. 6th International Conference on Mobile Computing, Applications and Services, Austin, TX, 2014, pp. 197-205.

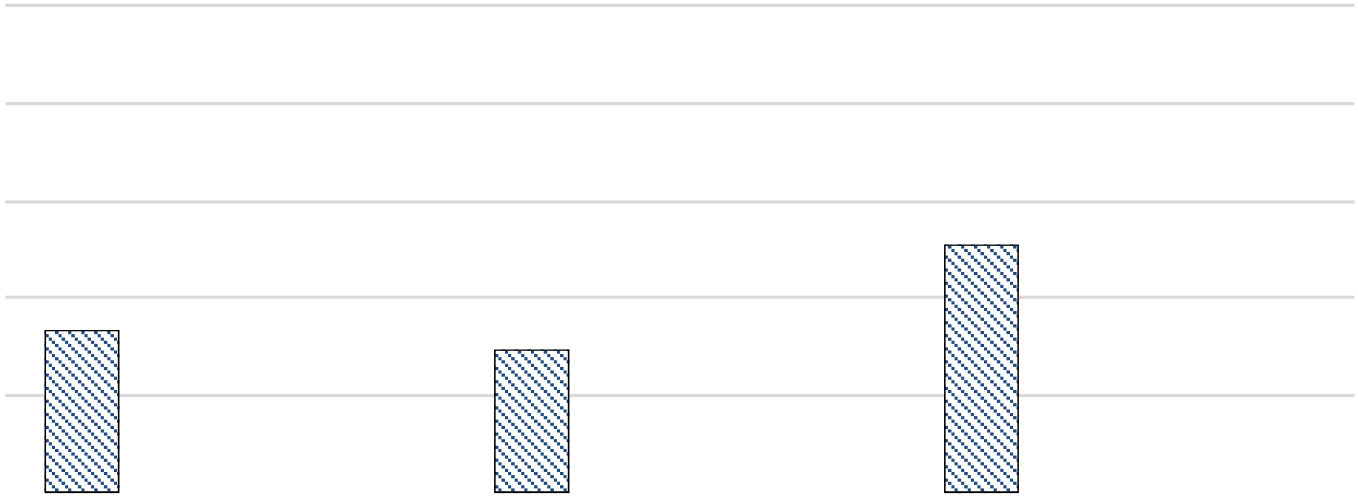
Experimental Result, CNNs: 3 HAR Datasets

1. Opportunity: activities in the kitchen
   - 10 activities (20 repetitions): open/close the fridge, drink while standing, clean the table, …
   - Body-worn sensors (19 in total, we used the right arm sensor)
   - 30,000+ instances, 64 dimension, 64Hz
2. Skoda: car assembly-line
   - 10 activities
   - Use 3D acceleration on the right arm
   - 20,000+ instances, 64 dimension, 96Hz



3. ActiTracker: primitive activities (walking, jogging, …)
   - 6 daily activities: jogging, walking, ascending stairs, …
   - Recorded from 36 users
   - 15,000+ instances, 64 dimension, 60Hz

Experimental Results, CNNs versus Baselines: Accuracy for 3 HAR Datasets



The CNN-based method, especially the CNN with partial weight sharing, performs better than other classifiers.

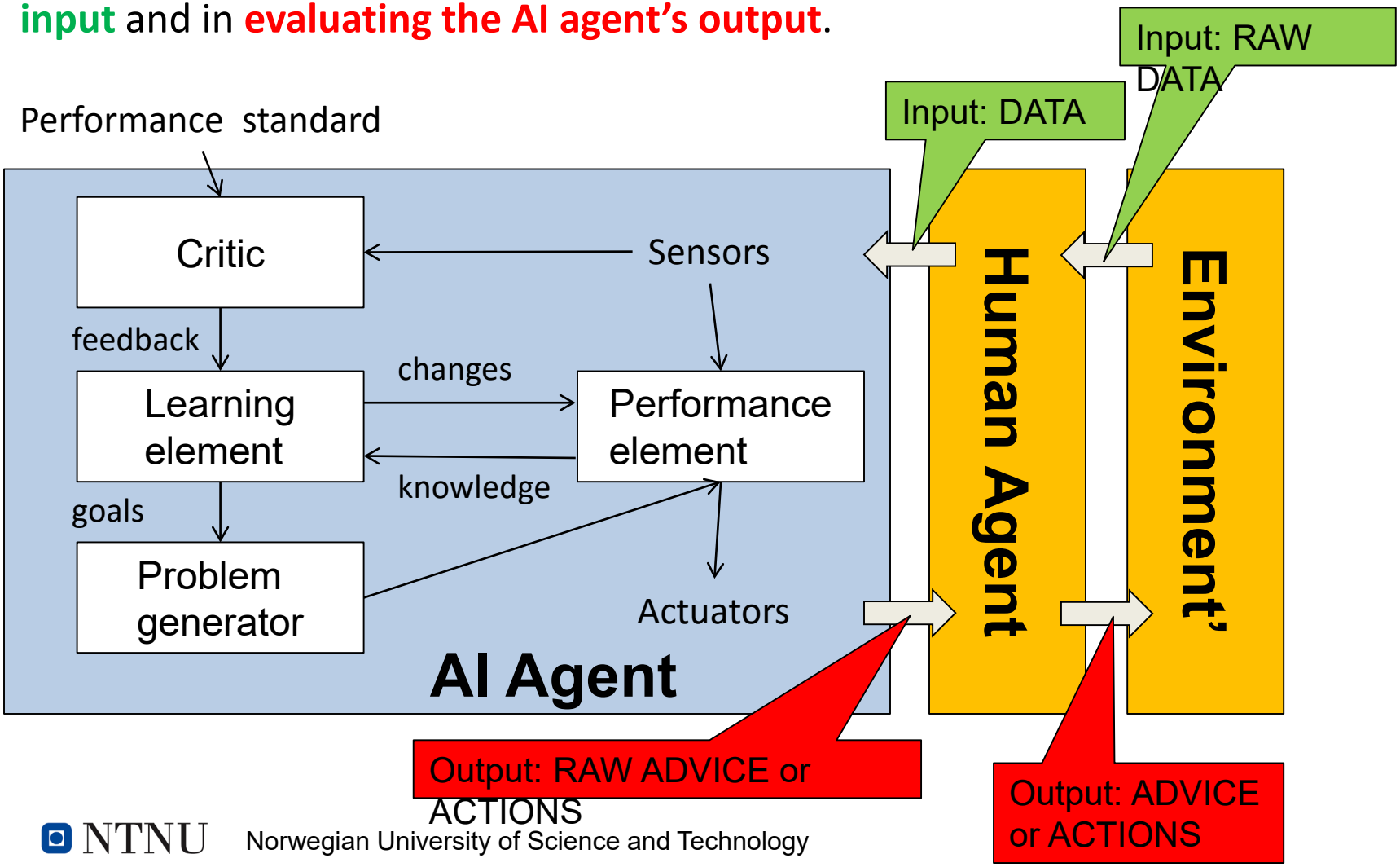
Pros and Cons of Connectionism

- Pros
  - Recent strong results in visual computing, speech recognition, natural language processing, human activity recognition, …
  - Extracts good features for classifiers – less need for feature engineering
  - Makes good use of big data
- Cons
  - Big data is typically needed for high-accuracy learning
  - The training time is typically long, powerful computing is needed
  - Interpretability and explainability are limited – the ANN is a “black box”
  - Creating a good architecture can be difficult

Rational Agent Architecture: Human in the Loop

Artificial Intelligence: Human Agent

The *rational AI agent* [Russell and Norvig, 2009] typically operates as a *decision support tool* for humans. Humans are involved both in **preparing the AI agent's input** and in **evaluating the AI agent's output**.



Preprocessing of Raw Data and Evaluation of Raw Decisions: Some Issues

(1) Data and preprocessing questions:

- Are the data representative or unbiased?
  - Recall iid assumption from statistics
- Is it Big Data? The 5 Vs of Big Data:
  - Veracity - Is the data noisy? How "clean" data?
  - Volume
  - Velocity
  - Variety
  - Value
- Are data skewed?
  - Rare or corner cases can be difficult
- Is it Small Data?
  - Learning from small, high-dimensional data is hard
- Are data complete?
- Is there data drift?
- Are there privacy or confidentiality concerns?
- Who owns the data?

(2) Decision (action or advice) and evaluation issues:

- The "garbage in, garbage out" rule still holds, likely more so than before
- Probabilistic (as opposed to deterministic) results – need for human in the loop and analytical thinking
- Less clear specifications and requirements due to learning component – most prominent in unsupervised machine learning
- Black box machine learning models are extremely hard to debug and fix when they fail

Preprocessing (of Raw Data) and Evaluation (of Raw Decisions)

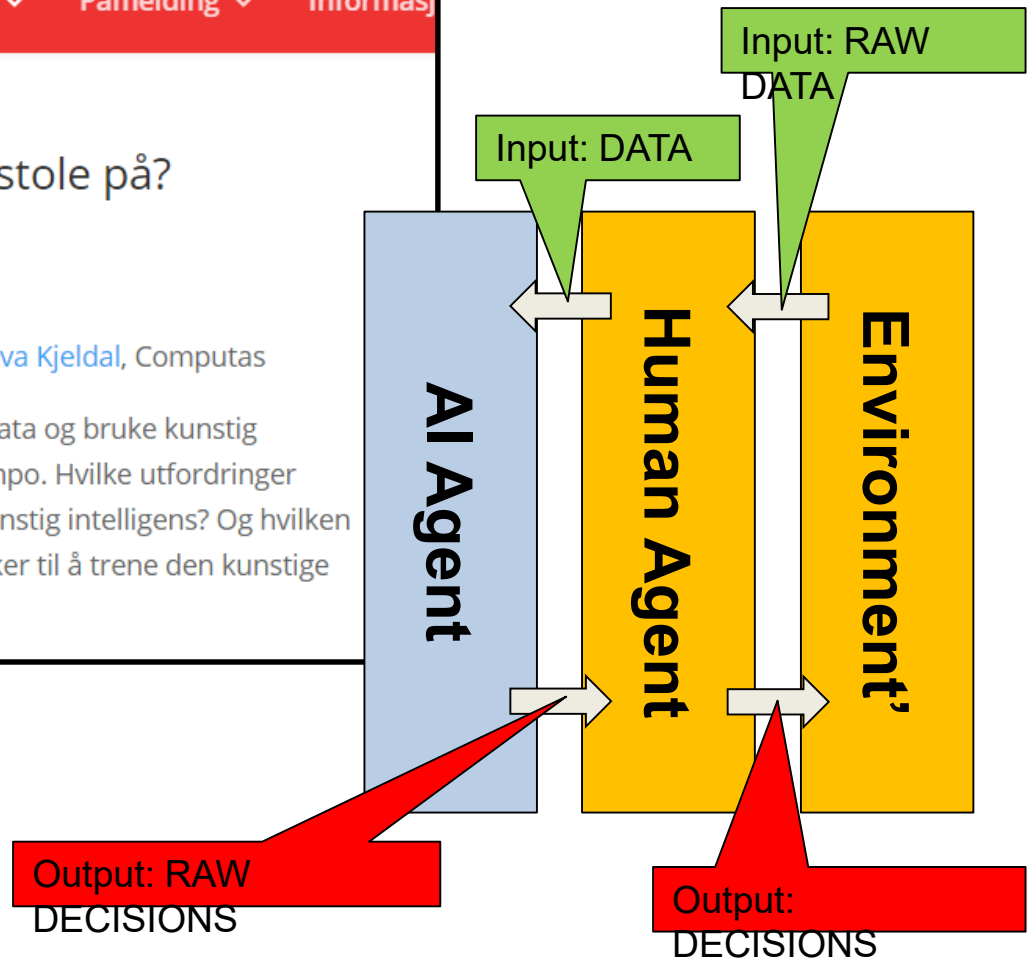
NOKIOS Hjem Program Påmelding Informasj

Sesjon 2B:
Er den kunstige intelligensen til å stole på?

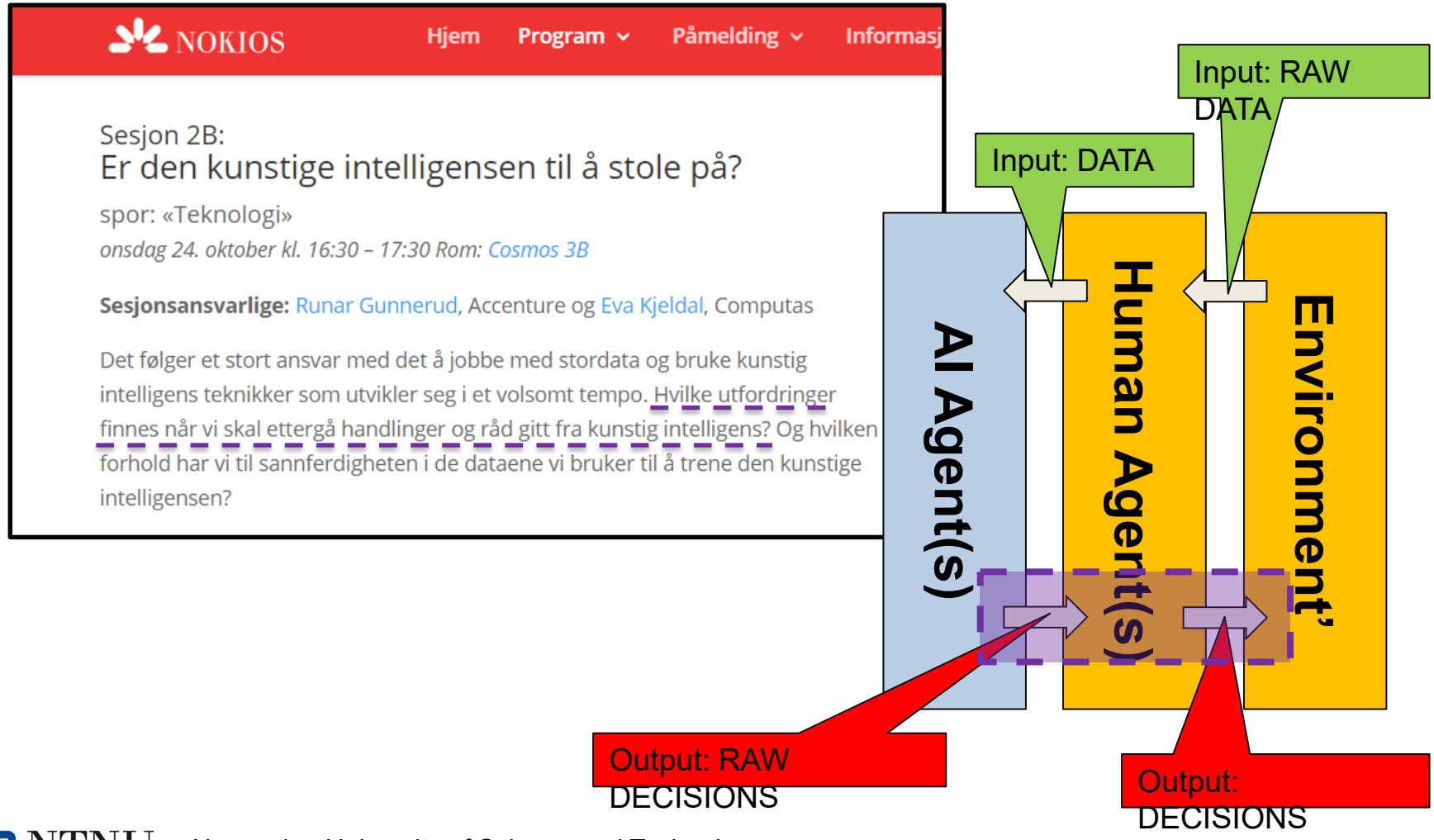
spor: «Teknologi»
onsdag 24. oktober kl. 16:30 – 17:30 Rom: Cosmos 3B

Sesjonsansvarlige: Runar Gunnerud, Accenture og Eva Kjeldal, Computas

Det følger et stort ansvar med det å jobbe med stordata og bruke kunstig intelligens teknikker som utvikler seg i et volsomt tempo. Hvilke utfordringer finnes når vi skal ettergå handlinger og råd gitt fra kunstig intelligens? Og hvilken forhold har vi til sannferdigheten i de dataene vi bruker til å trene den kunstige intelligensen?



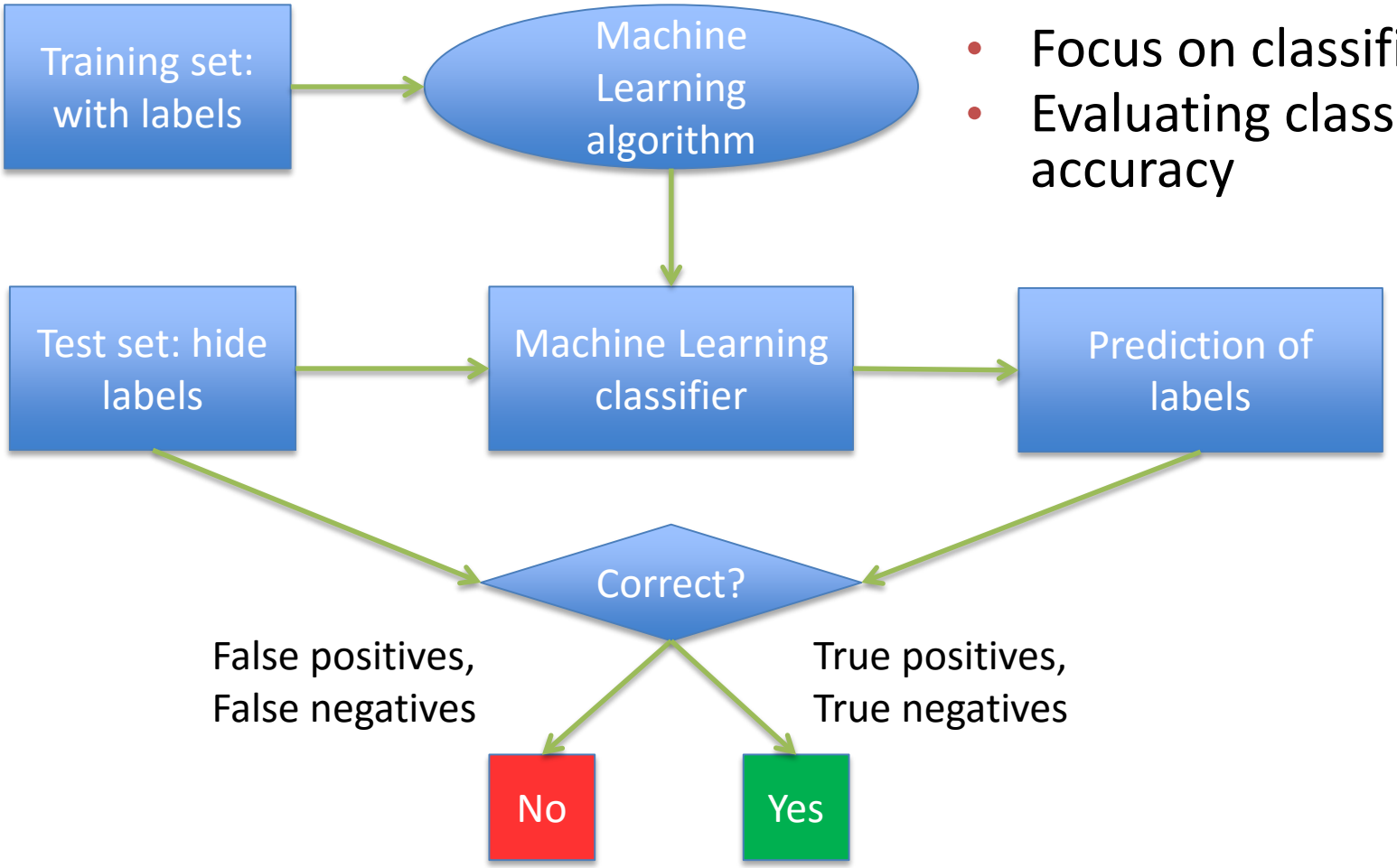
Evaluation of Decisions



Evaluation of AI Agent Decisions

- Ideally there should be:
  - A clear idea of what the AI agent is trying to achieve
  - A strong connection between AI agent decisions and the goal(s) of an organization (business)
- However, it can be difficult to:
  - Quantify the ultimate business goal(s)
    - We can try to use a surrogate in such cases
    - Need to decide the surrogate through careful analysis
    - In machine learning, the surrogate is often the ML model:
      - Created from a training data set + prior knowledge
      - Evaluated on a testing data set
    - Here: focus on classifier as the ML model
  - Guarantee that the decisions of the (learned) AI agent meet the goals - if they can be defined
    - Using an AI agent in a decision support role often makes sense

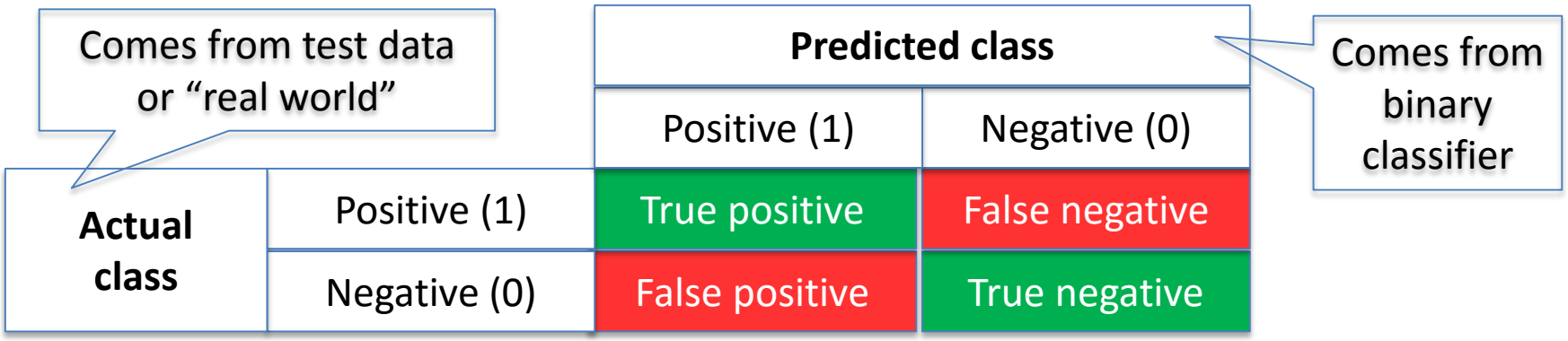
Evaluation Overview



- Focus on classifiers
- Evaluating classifiers: accuracy

Confusion Matrix for Binary Classification

- Assumption: binary (0/1, Yes/No, Positive/Negative) classification
- Positives and negatives - in machine learning terminology:
  - **Negatives** are the uninteresting outcomes
  - **Positives** are the outcomes of interest (sometimes few)
- Confusion Matrix :
  - An *n* x *n* matrix for a classification problem with *n* classes
  - For binary classification: *2 x 2* confusion matrix
  - Main diagonal (green) contains the correct outputs of the classifier



Goal: (1) minimize FPs and FNs while maximizing TPs and TNs.

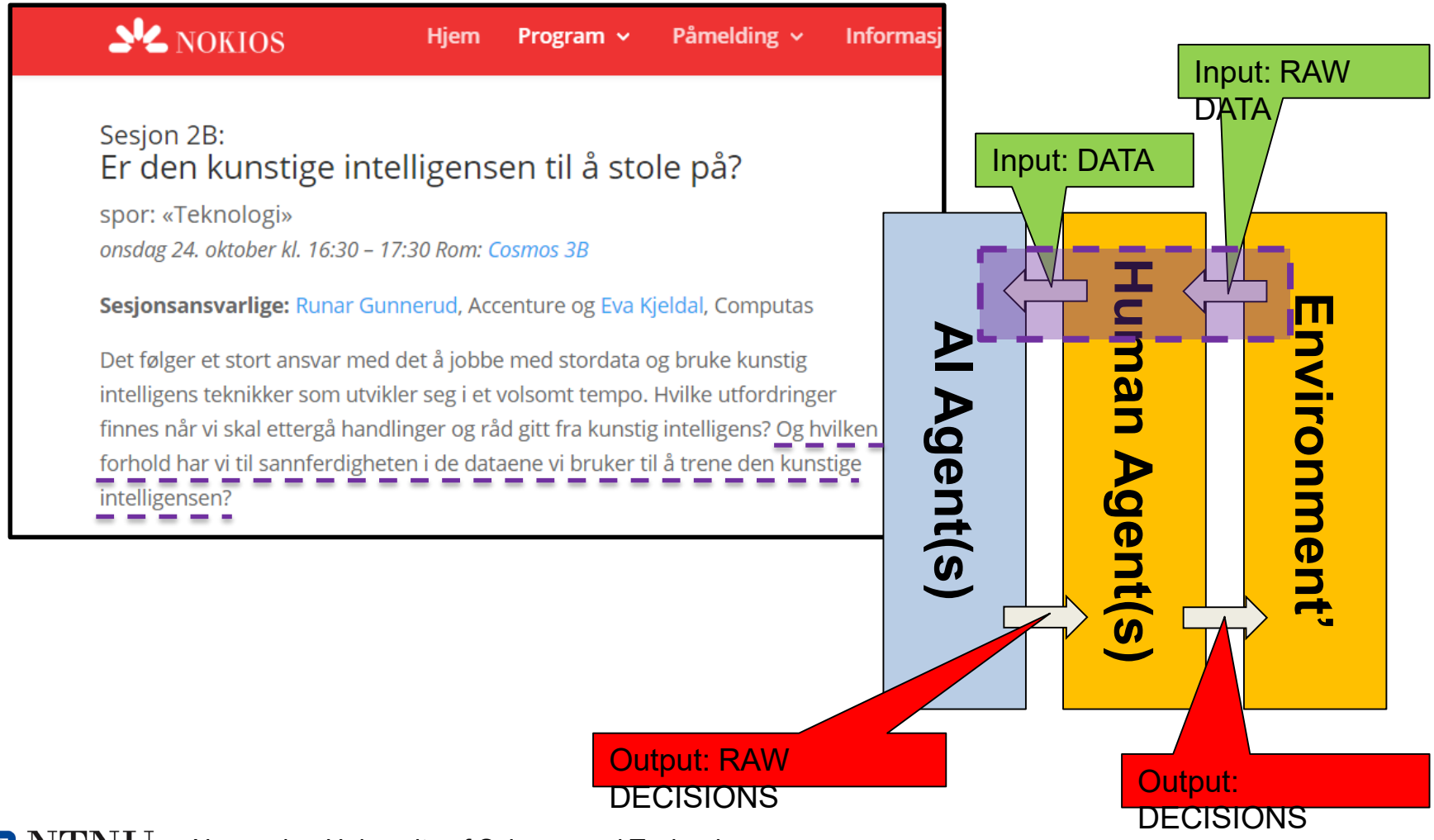
Evaluation Metrics: Accuracy and Error

- Define (based on the confusion matrix):
  - TPs: Number of true positives
  - TNs: Number of true negatives
  - FPs: Number of false positives
  - FNs: Number of false negatives
- Metrics:
  - Accuracy *a* - proportion of *correct* decisions
  - Error *e* - proportion of *incorrect* decisions

$$a = \frac{TPs+TNs}{TPs+TNs+FPs+FNs} = 1 - e$$

- A typical goal of machine learning is to maximize accuracy *a* and minimize error rate *e*

Preprocessing of Data

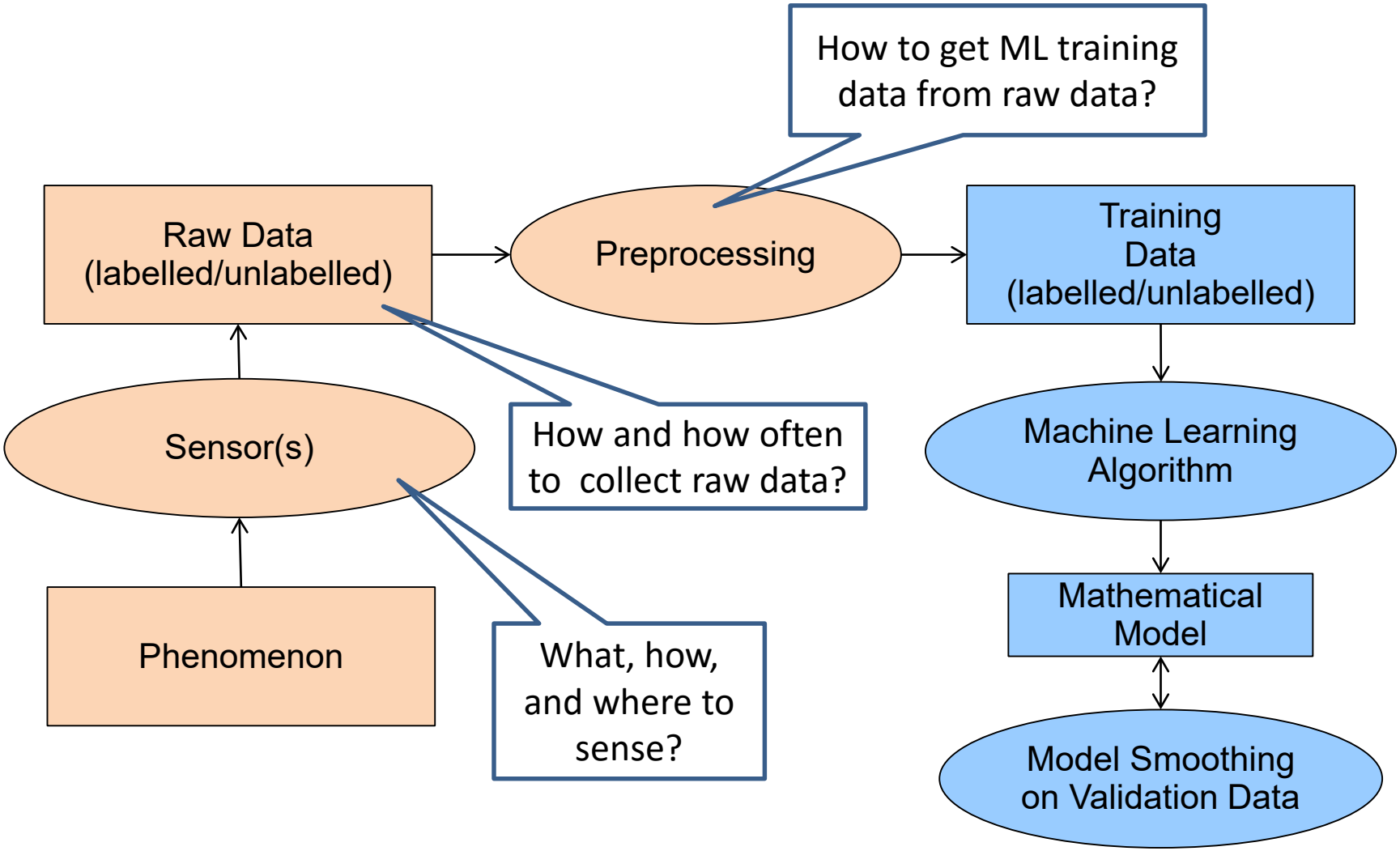


Data Preprocessing

- Metadata: Information about data set and its attributes
- Statistics: Mean, std. dev., outliers, clusters, correlation,…
- Missing values and data cleansing
- Normalization: satisfy statistical and/or visualization constraints
- Continuous versus discrete:
  - Segmentation and discretization: continuous to discrete
  - Nominal to ordinal mapping: discrete to continuous
- Sampling and sub-setting
- Dimension reduction: reduce to smaller number of dimensions
- Aggregation and summarization
- Smoothing and filtering: signal processing techniques
- …

If data preprocessing is performed, it is often important to (1) clearly indicate so and (2) provide drill-down capability to the raw data.

From Raw Data to Training Data



Feature Engineering

- Scenarios for features:
  - Many, independent, predictive features: Easy learning
  - Few, dependent, non-predictive features: Hard learning
- Applied machine learning project:
  - Much (most?) time might be spent on feature engineering
- Feature engineering is typically application-specific:
  - Feature construction is often semi-automatic or manual
  - Approaches to features selection:
    - Filter: First feature selection, then machine learning
    - Wrapper: Iterate between feature selection and machine learning
- Holy grail of ML:
  - Automated construction of features
  - Today:
    - Traditional ML: Generate (feature construction) and test (feature selection)
    - Deep ML: Progress has and is being made on feature construction

Accuracy and Unbalanced Classes

- Accuracy, and closely related metrics, are good starting points for evaluation of ML decisions
- But: is accuracy sufficient to evaluate a model?
- In problems where data for one class is rare (including not observed at all), using only accuracy can give poor results:
  - E.g.: credit card transactions
    - 100 transactions: 98 legitimate, 2 fraudulent (actual)
    - Classifier classifies all transactions as legitimate (predicted)
    - Accuracy a = 98/100 = 98%
  - E.g.: diagnosis of infants for cerebral palsy (CP)
    - 1000 live births: 2 positive, 998 negative[1] (actual)
    - Classifier classifies all births as negative (predicted)
    - Accuracy a = 998/1000 = 99.8%
  - Are these good classifiers?

[1]"After validation, 1784 children born 1996–2007 in Norway were confirmed to have CP, with a corrected prevalence of 2.5 (95% CI: 2.4–2.7) per 1000 live births." S. Hollung, G. Andersen, R. Wiik, I. Bakken, and T. Vik. *What is the prevalence of cerebral palsy in Norway? Developmental Medicine & Child Neurology,* Volume 57, Issue S5, October 2015.

Concluding Remarks

In Conclusion

- Question: Can artificial intelligence be trusted?
- Answer: Yes and no.
- Rational:
  - Artificial intelligence and machine is different from traditional computer science (for example *sorting*)
  - There are right and wrong ways to sort numbers, and a programmer can write a provably correct sorting program
  - In contrast, machine learning is typically used when no obvious program can easily be written – the program is learned from data
  - A further complications comes with machine learning – outputs (advice or actions) are typically uncertain
    - "Doing machine learning means to always say I'm sorry" (Prof. D. Wilkins)
  - Trust in AI Agents needs to be built, using verification and validation methods, similar to other engineering artifacts (aerospace vehicles, buildings, cars, ships, …)
    - Thought as your plane takes off: "Did anyone prove (mathematically) that it will fly?" (Prof. D. Goldberg)

ANY COMMENTS OR QUESTIONS?

Ole Jakob Mengshoel

ole.j.mengshoel@ntnu.no

Professor, Dept. of Computer Science, NTNU
Head, Norwegian Open AI Lab, NTNU
Adjunct Faculty, Dept. of Electrical and Computer Engineering, CMU

https://www.ntnu.edu/employees/ole.j.mengshoel
https://www.ntnu.edu/ailab
http://sv.cmu.edu/directory/faculty-and-researchers-directory/faculty-and-researchers/mengshoel.html
https://works.bepress.com/ole_mengshoel/