Artificial Intelligence: Status and Opportunities or What AI Can Do for Science

Kerstin Kleese van Dam
Brookhaven National Laboratory

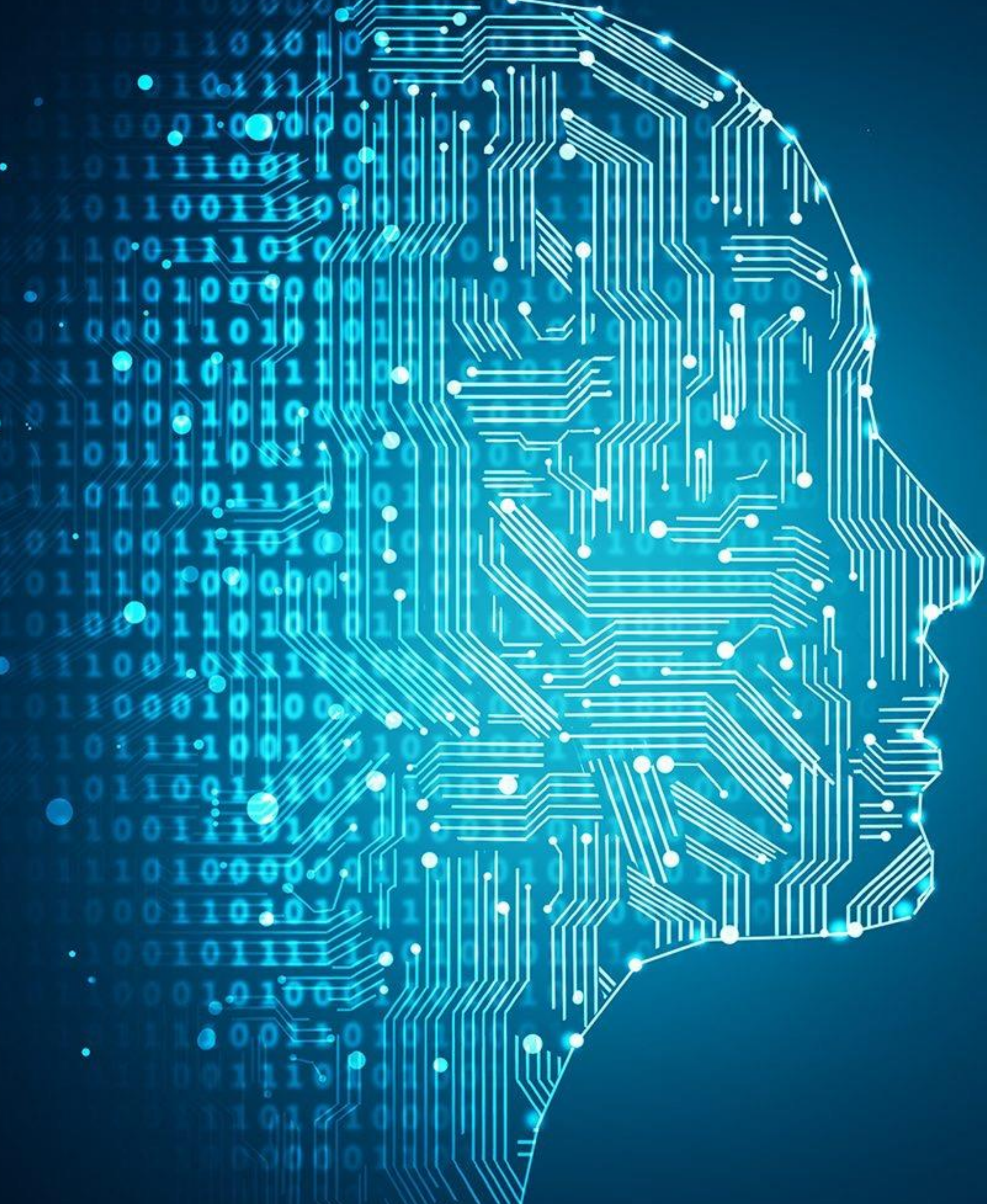
Why is Everyone talking about AI Today?

- AI has been around for a long time, in 1951 a machine called Ferranti Mark 1 used the algorithm to master Checkers.

- Big Data - Drives demand for AI, but is also the lifeblood of modern AI technologies
- Big Machines - Training a good algorithm used to take a very long time, new hardware and bigger systems, make it now a question of hours.

Identifying these are Dogs is Machine Learning

Autonomously driving this Car is AI



Machine Learning Methods

Neural Networks are one of the most popular machine learning methods today.

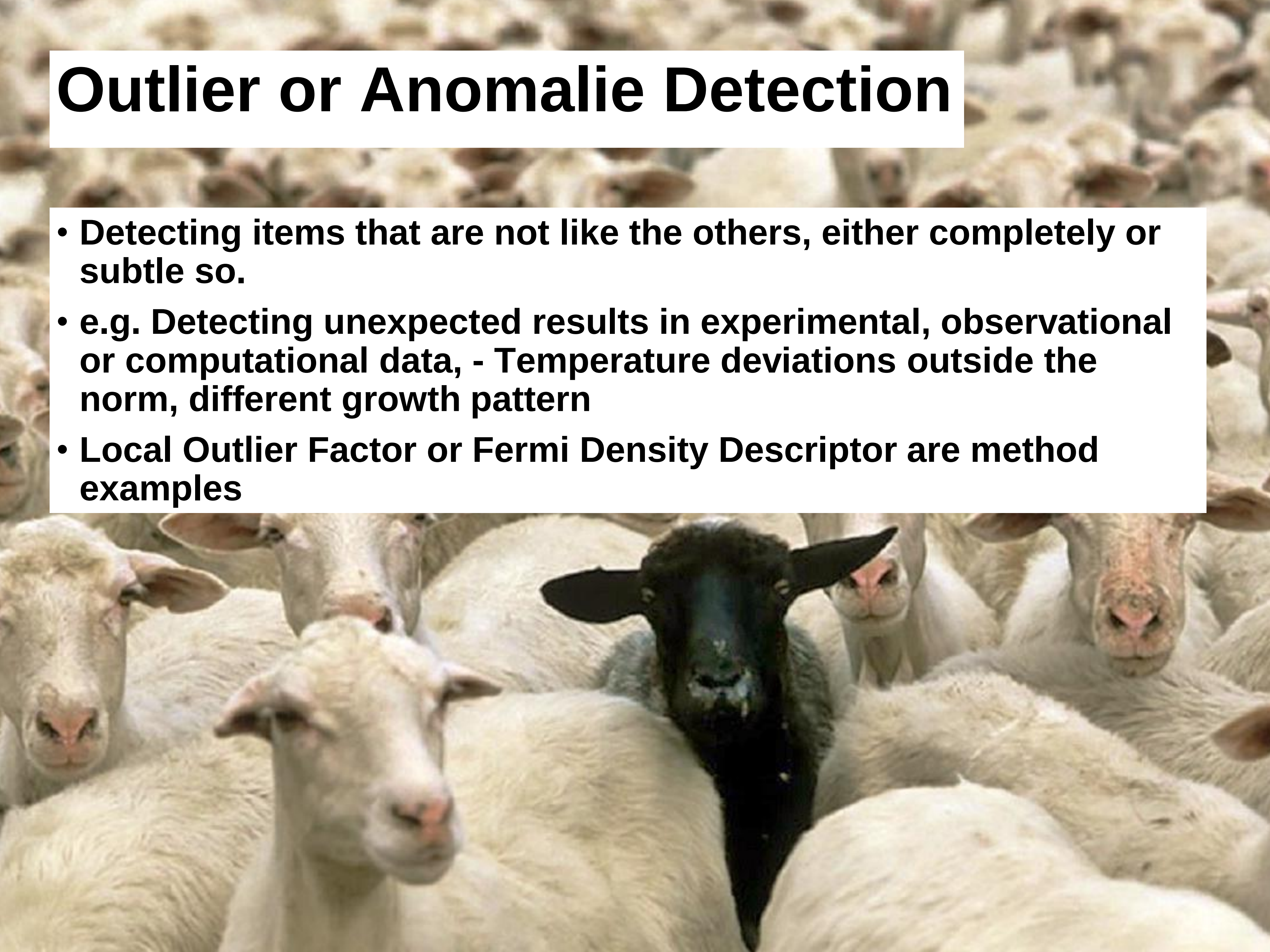
They come in many forms - Deep Learning, Convolutional NN, Graph NN for time series data etc.

Cryo-EM Image Analysis
Restore compressed climate modeling output, 2018 Gordon Bell Prize - Identification of Extreme Weather Patterns

Classification, Pattern and Feature Detection

Outlier or Anomalie Detection

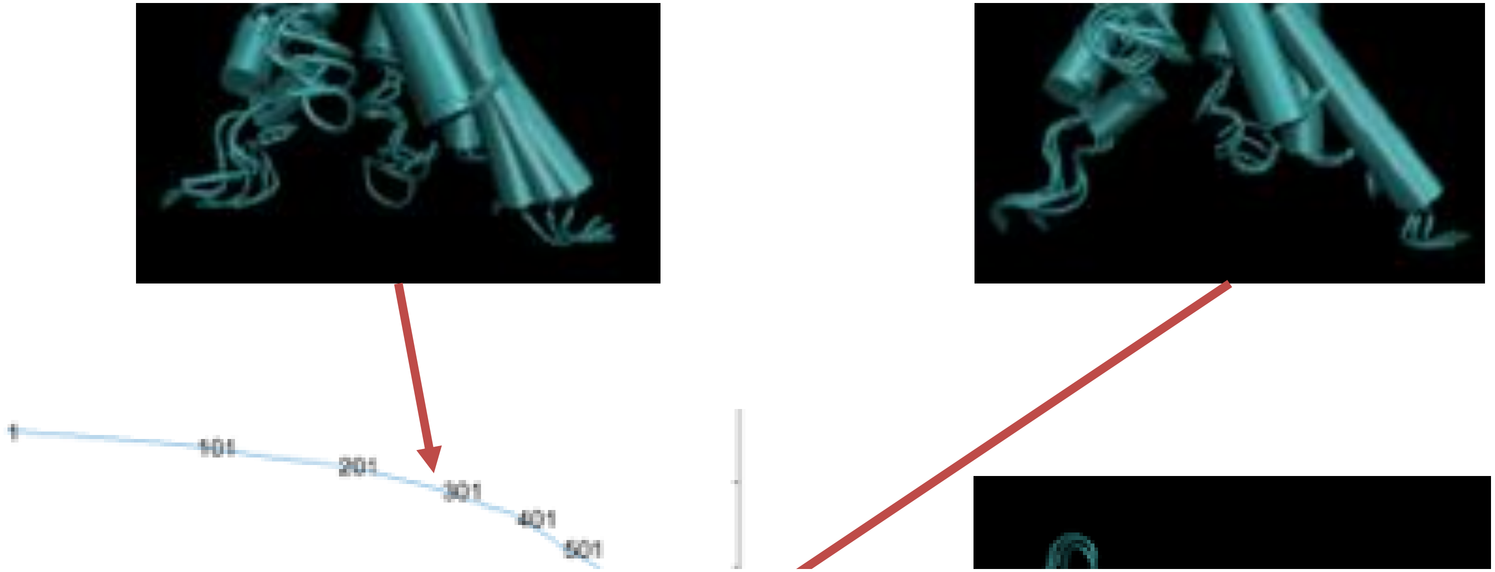
- Detecting items that are not like the others, either completely or subtle so.
- e.g. Detecting unexpected results in experimental, observational or computational data, - Temperature deviations outside the norm, different growth pattern
- Local Outlier Factor or Fermi Density Descriptor are method examples



Causal Analysis

- **Identify the reason why an event may have occurred or what relationships may have led to the event.**
- **e.g. Gene regulatory network inferred by robust prior-knowledge-driven causal analysis. Analysis of the *Saccharomyces* genome revealed interesting regulatory relationships.**
- **e.g. Lasso or Conditional Granger Causal Inference with two step prior incorporation are example methods**

Dimension Reduction



- You have complex data with many features, and want to reduce the number of those features, while retaining as much information as possible.
- e.g. reducing the output volume of climate simulations without loosing key features, identifying phase changes in MD calculations
- Linear Dimensionality Reduction (LDA) or Manifold Learning algorithms such as Diverse Power Iteration Embedding (DPIE) are example methods.

Prediction / Surrogate Models

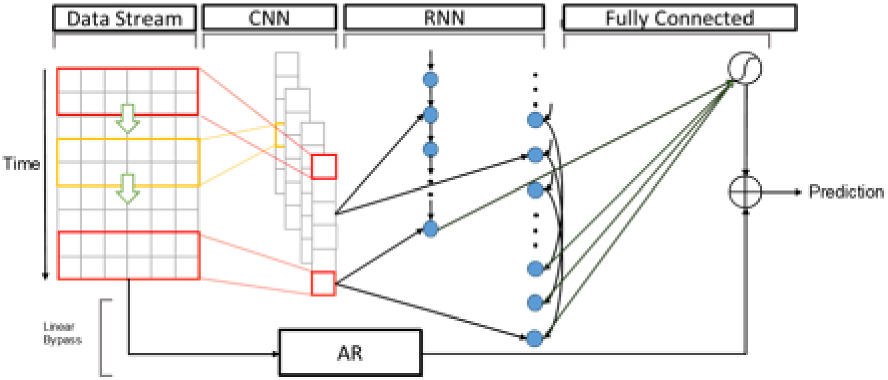
- Machine Learning Methods can also be used to make prediction of future states
- e.g. There have been a number of recent projects that created surrogate models by using machine learning. Though not perfect they outperform classical approaches and are faster to execute
- Autoencoder, Convolutional Neural Networks, and CNN Autoencoder are example methods.

Science is Complex and so are the ML solutions

Spatio-Temporal Learning for Solar Energy Forecasting

ML Methods Combined in LSTnet:

- Jointly modeled short- and long-term time dependency by combining deep learning (DL) and autoregressive (AR) model
- CNN (Convolutional Neural Network) captures local dependency patterns whereas RNN (Recurrent Neural Network) captures long-term dependency patterns
- Automatically adapting the mixtures weights of the AR and CNN/RNN components based on input



Natural Language Processing

- **Extract scientifically relevant information from text, tables and graphics in scholarly publications, reports etc.**

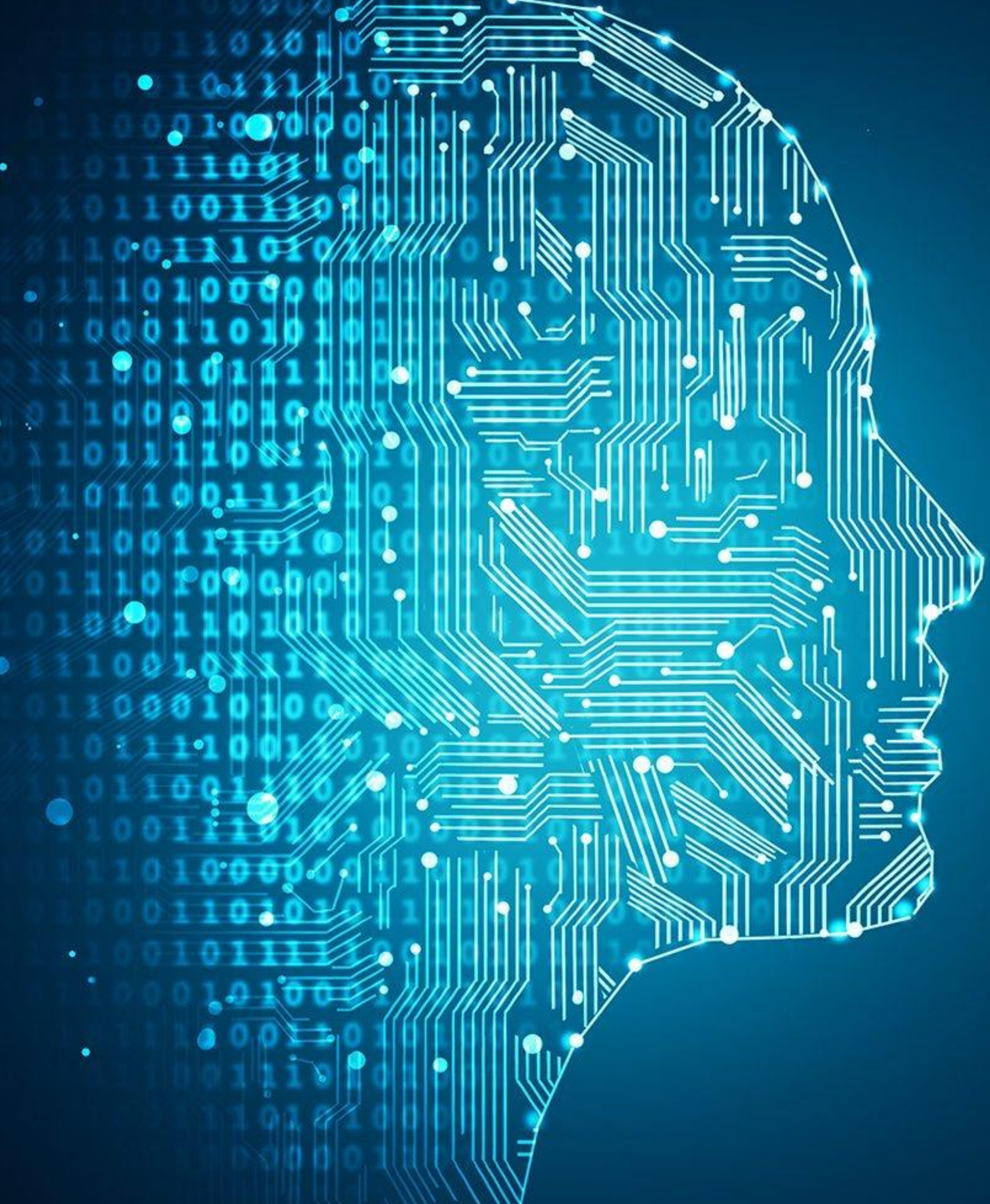
- **e.g. Identify mentions of a specific protein, co-occurrences, and interaction predictions - build knowledge graph with the results to help interpret your results**

- **Faster-RCNN with Contextual Information, Word2Vec are example methods**

Batch versus Streaming

- Traditionally Machine Learning was applied to a large collection of Static Data
- All data is available at the start of the analysis
- There is time to consider all data before producing a result or making a prediction
- Example: Long range pattern detection in climate data across many climate runs

- Data is arriving while the algorithm is working
- Results are expected before the last data arrives
- In some cases you only get to look at each data point once (YOLO)
- Example: Real time data analysis and reduction of high throughput biological experiments

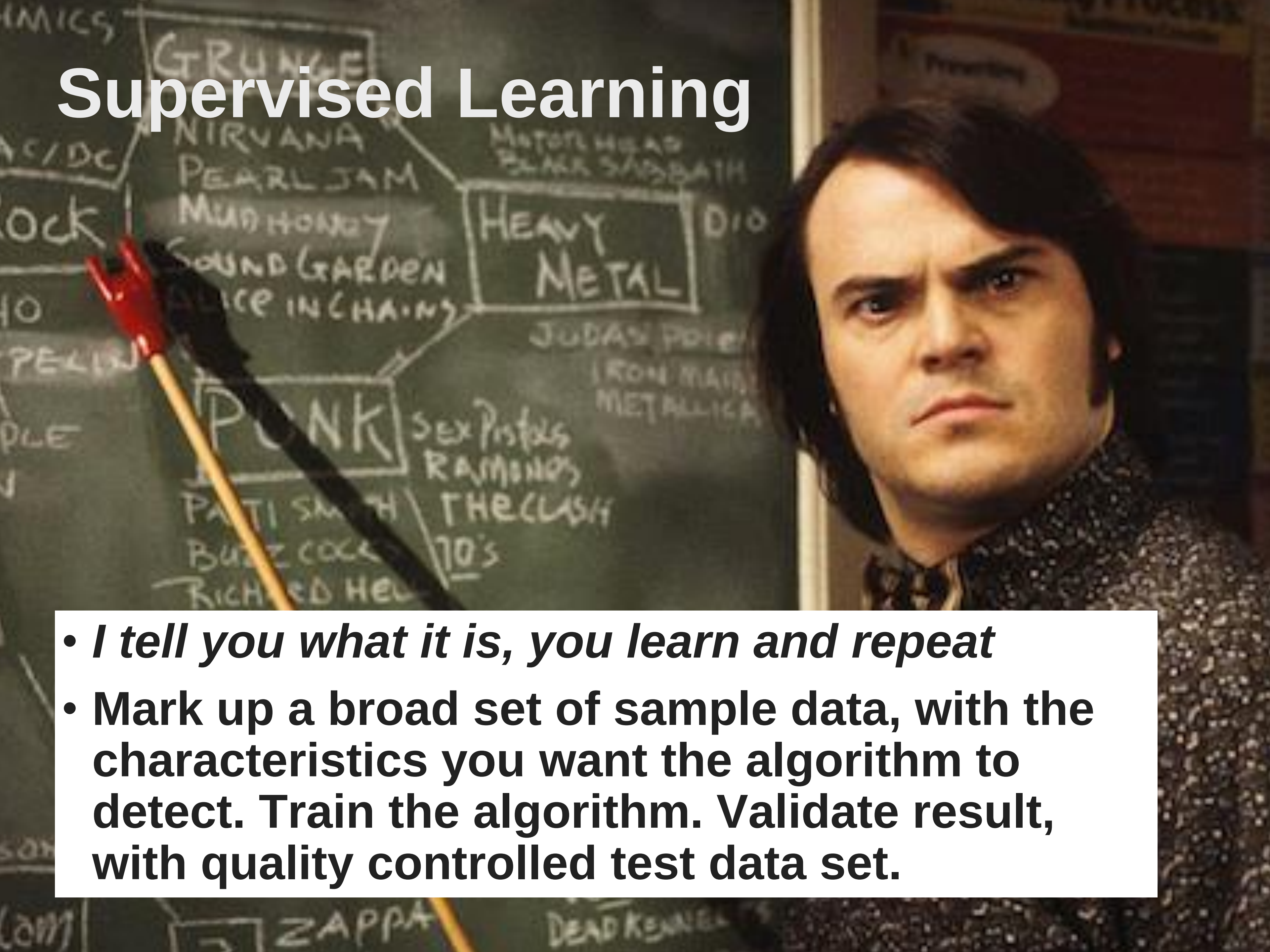


How can machines learn?

Training Data

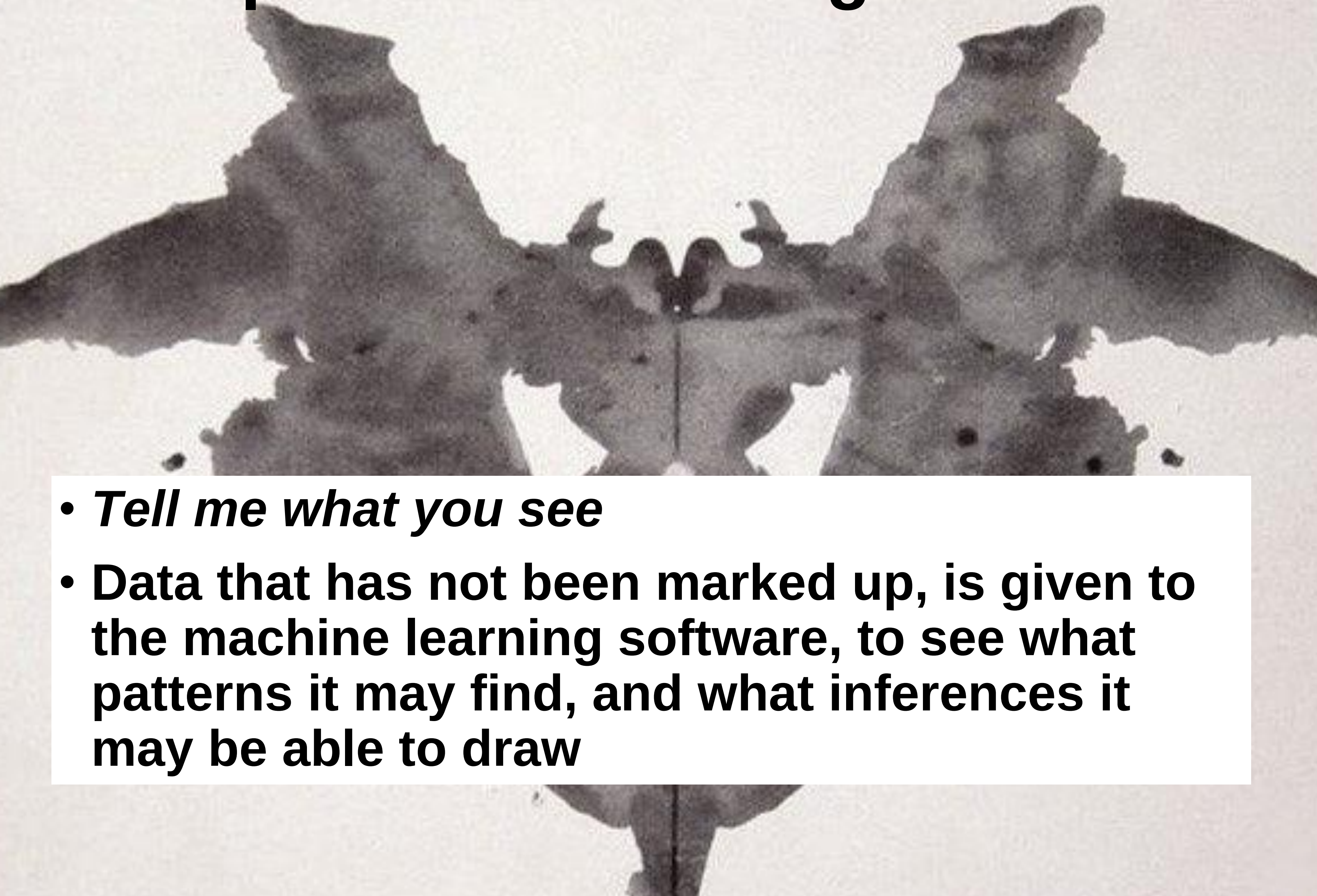
- *Ground Truth* - there is always a need for labelled data
- *Coverage* - when selecting training data its important to select a broad range, that covers all possible cases equally well
- *Bias* - Beware of introducing bias into your training data - checks needed!

Supervised Learning



- *I tell you what it is, you learn and repeat*
- **Mark up a broad set of sample data, with the characteristics you want the algorithm to detect. Train the algorithm. Validate result, with quality controlled test data set.**

Unsupervised Learning



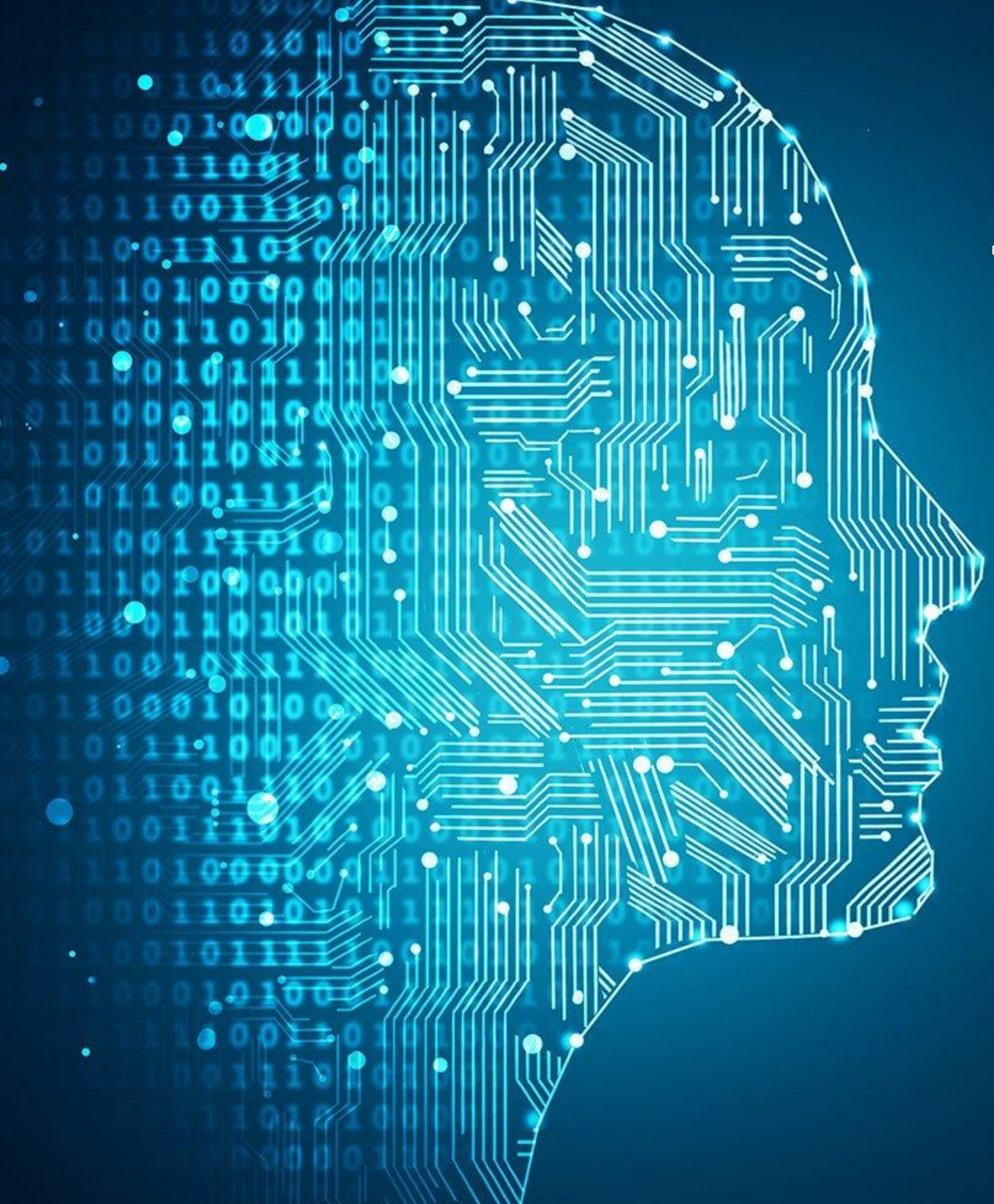
- ***Tell me what you see***
- **Data that has not been marked up, is given to the machine learning software, to see what patterns it may find, and what inferences it may be able to draw**

Reinforcement Learning

- ***Tell me what you think and I tell you if you are right***
- **Provide suitable actions to reward specific behavior in the algorithm, in certain situations, to encourage a desired path to solution or reaction.**

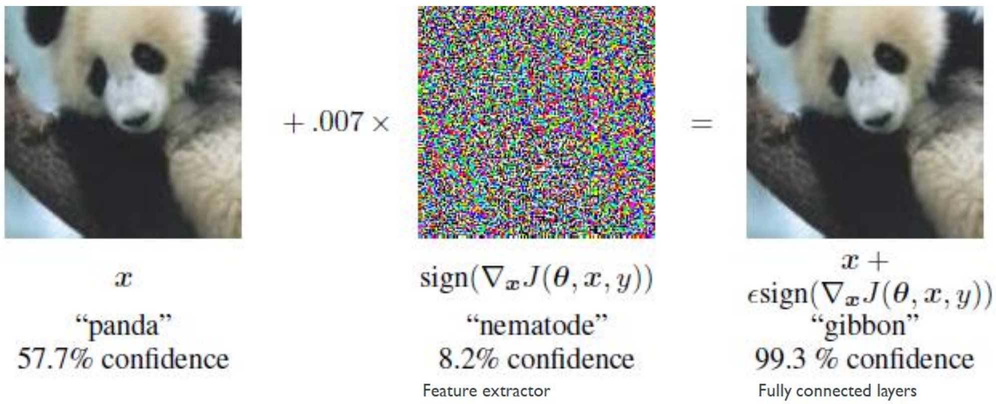
How to train for Rare Events?

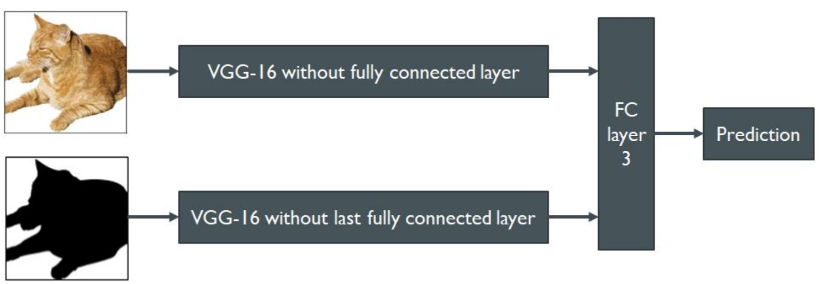
- Unfortunately we do not always have sufficient data - rare cases - what can we do?
  - *Transfer Learning* - We train on data that is similar first and than refine with the rare data
  - *Reinforcement Learning* - train with a small labeled data set and the add capability to the model by giving active user feedback
  - *Numerical Simulation* - Where models exist, these can be used to create labelled data that can act as ground truth
  - *Create Synthetic Data* - In more complex cases you can use machine learning to generate the synthetic data - e.g. via Generative Adversarial Networks (GANs)



Can you trust machines learning?

Robustness of Algorithms





Not all algorithms are robust. Sometimes even a small perturbation in the data can cause misclassifications.

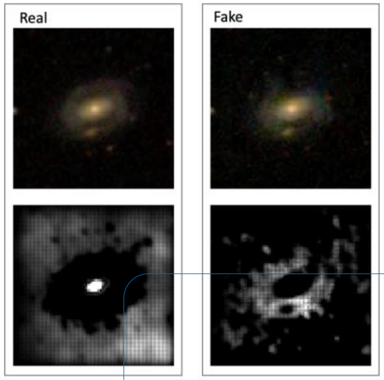
New Approaches taking global information into account are needed.

Accuracy	Clean images	AE	Merge from 3rd FC layer	Train two branches separately then finetune	Only train saliency maps branch
$\varepsilon = 0.04$, run 5 iterations	98.23%	79.68%	91.12%	82.50%	85.40%

Explainability

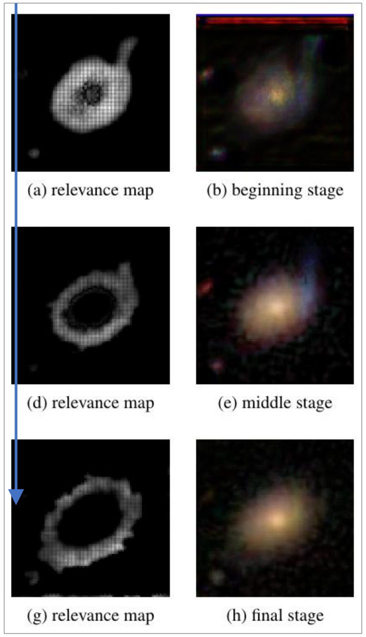
Interpreting Galaxy Deblender GAN

- **Machine Learning is often seen as a black box.**
- **In normal algorithms every decision, every rule has to be programmed and can therefore be verified by mathematical or physical models.**
- **New verification methods needed for ML. E.g. Layer-wise Relevance Propagation (LRP) method for GAN interpretation by defining a complete relevance map - visualization of results through heat map**



Different attention areas identifie
different classification decision

See where your ML model focusses its learning



ML Reproducibility through Provenance

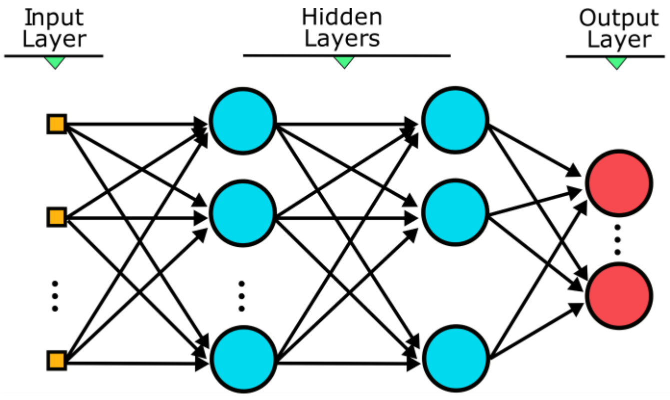
Task: Test Reproducibility of common ML Methods

Three regression models evaluated:

- Gradient boosted trees (GBT), an ensembles a set of decision trees.
- Multilayer perceptron (MLP), fully connected multiple layer neural networks.
- 1D convolutional neural networks (1D-CNN).

2 Frameworks: Tensorflow and PyTorch

Result: Different platforms, cannot reproduce exactly the same model, while different versions of the same platform show good reproducibility.

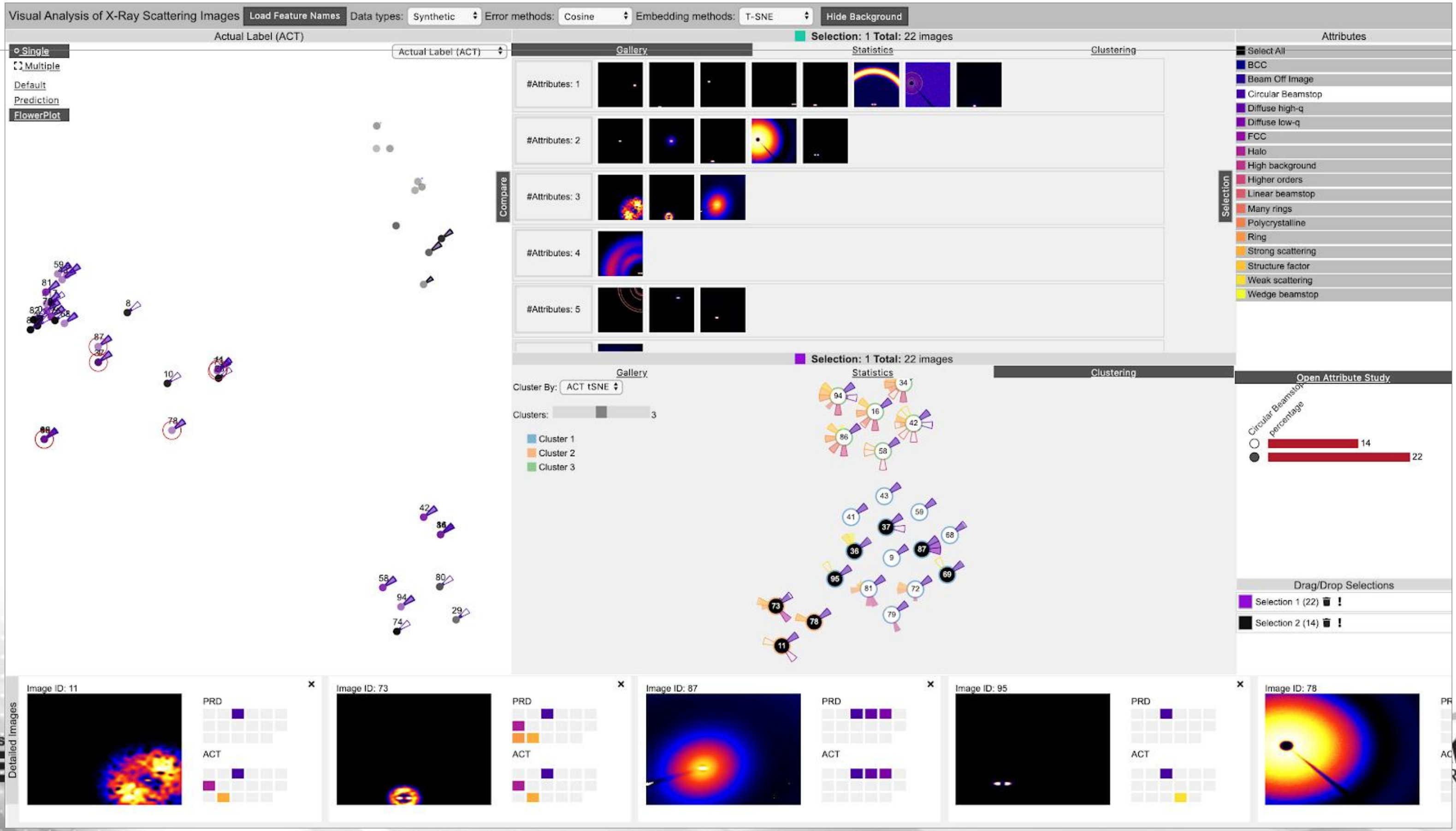


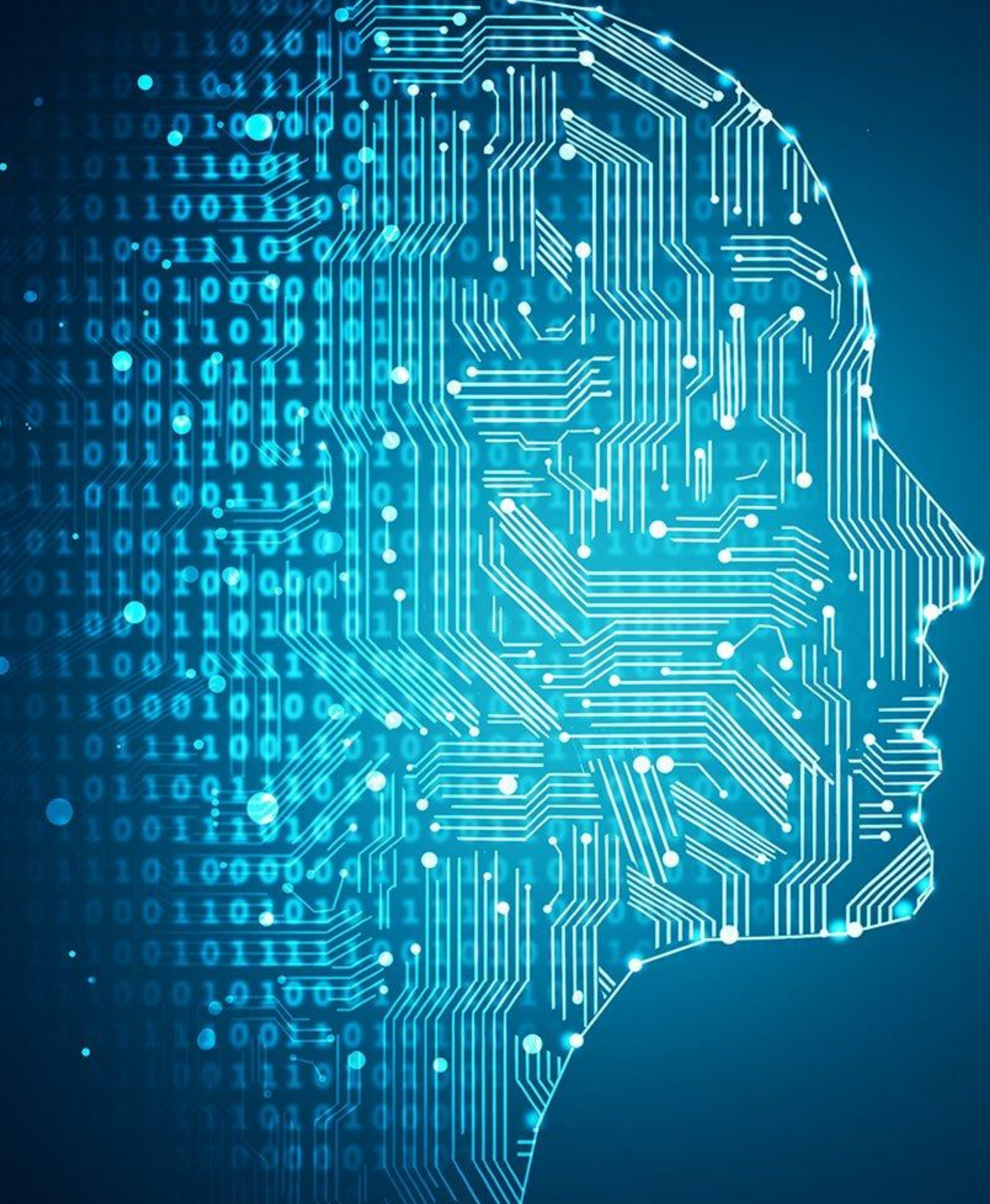
A typical MLP model.

		Tensorflow		PyTorch	
		1.9.0	1.14.0	1.2.0	1.13.0-dev
Tensorflow	1.9.0	Y	Y	N	N
	1.14.0	Y	Y	N	N
PyTorch	1.2.0	N	N	Y	Y
	1.3.0-dev	N	N	Y	Y

Integration of Domain Knowledge

- Integration of domain knowledge into the ML models is vital to achieve the needed accuracy and help explainability
- e.g. See below Show multi-label classification, Co-existence and Co-relation analysis of the labels

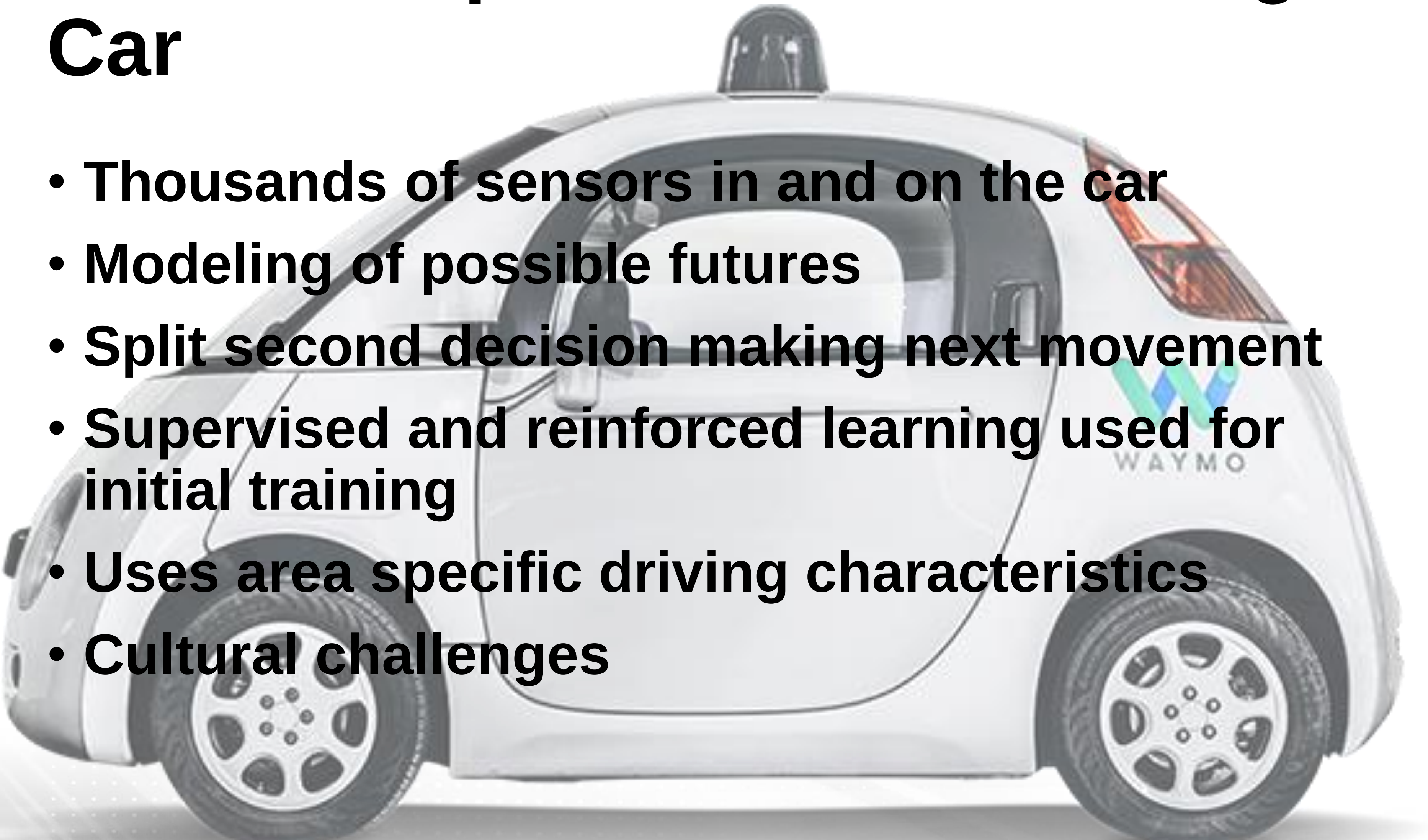




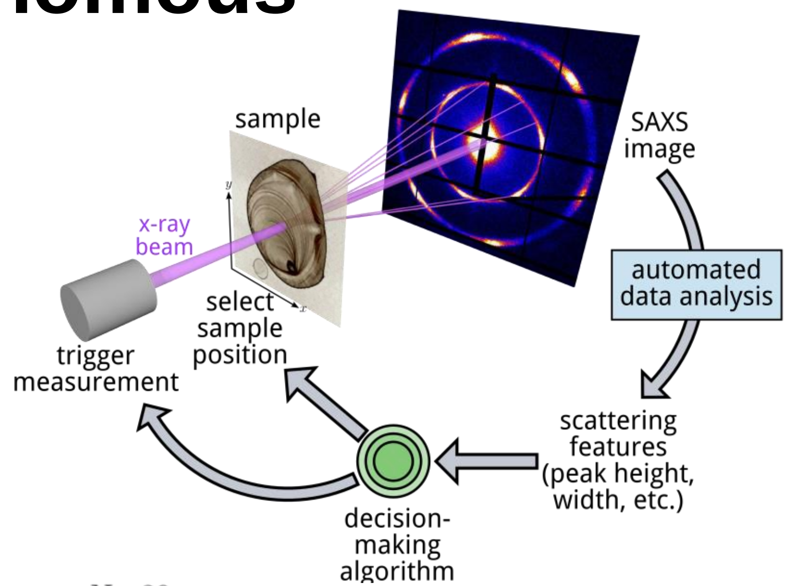
Where can you find AI today?

Human Inspired AI - Self-Driving Car

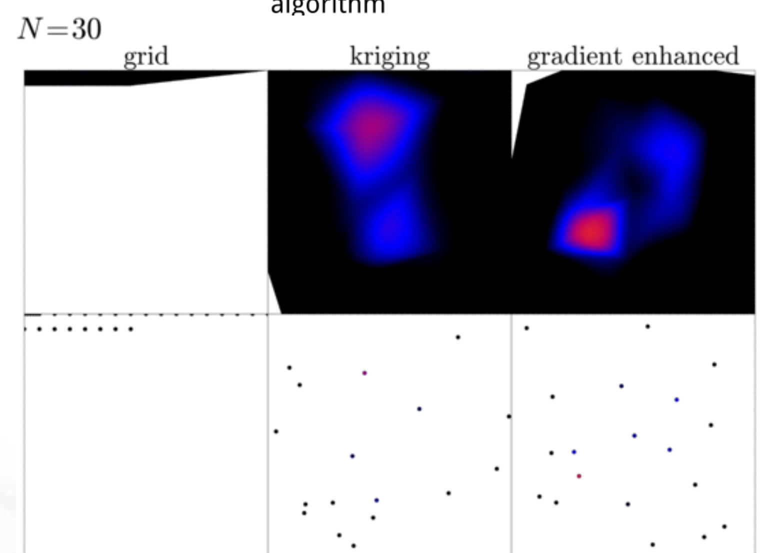
- **Thousands of sensors in and on the car**
- **Modeling of possible futures**
- **Split second decision making next movement**
- **Supervised and reinforced learning used for initial training**
- **Uses area specific driving characteristics**
- **Cultural challenges**



Science Example - Autonomous Experiments



- **Goal based steering of an experiment, a complex simulation or complex workflow.**
- **Real time data analysis on data as it arrives, drives optimization algorithms that determine the next best step, based o the given scientific goal.**
- **Takes into account domain knowledge and uncertainty**
- **First simpler, highly specialized examples exist today**
- **Future: not only autonomous experiments, but a personalized science infrastructure that is aware of your goals and needs and shapes resources accordingly**



What does it take to build an AI application?

- Identify emerging phenomena in high velocity streaming data **- Streaming Statistics, Data Mining, Machine Learning**
- Determine what is of interest and impact, generate candidate explanations **– Streaming Deductive Reasoning, Computational Models**
- Human-Computer collaboration to jointly adjust data collection, reasoning and insights **– Science of Interaction, Cognitive Depletion Detection, Hypothesis Exchange, Adaptive Algorithms and Workflows**
- Evaluate the impact of possible decisions **- On Demand Prediction, ML Surrogate Models**
- Document which decisions were taken during the analysis process to explain the results **- Provenance, Explainability, Reproducibility**

Imitating Human Creativity

- New AI solutions today are starting to mimic human creativity - so far they are largely imitating specific artists, rather than creating on their own.
- Examples are:
  - Painting - A new Rembrandt, through analysis of several hundred original works of the master, studying techniques, features and composition
  - Music - Several solutions exist to compose music that sounds like e.g. the Beatles
  - Acting - AI solutions have written comedy sketches

Science Example - ??

- Not aware of scientific example today
- Imagine AI autonomously learning and replicating physical, chemical and biological processes
- Imagine AI rewriting computational models based on inter-comparison of different model evaluations
- Further out - your personal assistant - combine Einstein over your shoulder with the ability to do the work of 100 grad students

Dark Side

- **Too Clever:**
- AI algorithms have learned how to play games successfully, they have also been observed to explore bugs in the code or to cheat to win.
- **Fraud:**
- As we are learning how to perfectly mimic other humans (e.g. Rembrandt), how do we guard against imitations in our computational and experimental research.

We are only at the beginning of exploring what AI can do for us. There are many possible applications, but we must proceed with care, being aware of the limitations and pitfalls.