UHH Vorlesung „Hochleistungsrechnen“

Machine Learning in Climate Science on Super Computers

Introduction and Research

Christopher Kadow

German Climate Computing Center

Machine Learning Mailing List: machinelearning@lists.dkrz.de

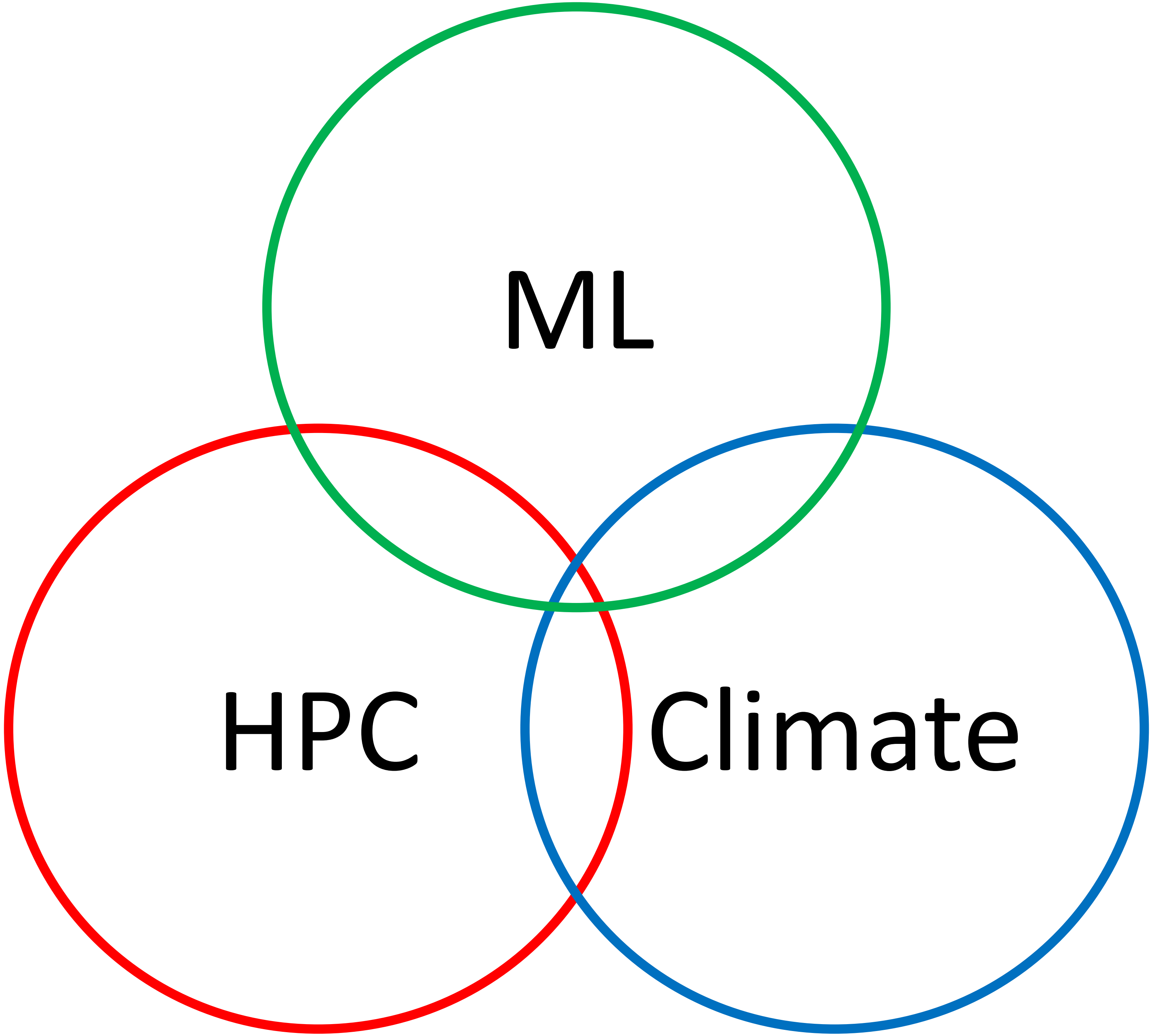
HZG ML Seminar Tuesdays 2-weekly http://m-dml.org/seminar.html

Machine Learning in Climate Science MSc Class in *Ocean and Climate Physics*

Class @ UHH

BSc/MSc Studies

Jan, 5th 2021


ML
HPC
Climate

German Climate Computing Center

Mission

DKRZ – Partner for Climate Science.
Maximum Compute Performance.
Sophisticated Data Management.
Competent Service.

Vision

DKRZ reliably unlocks the potential of the accelerating technological progress for climate research

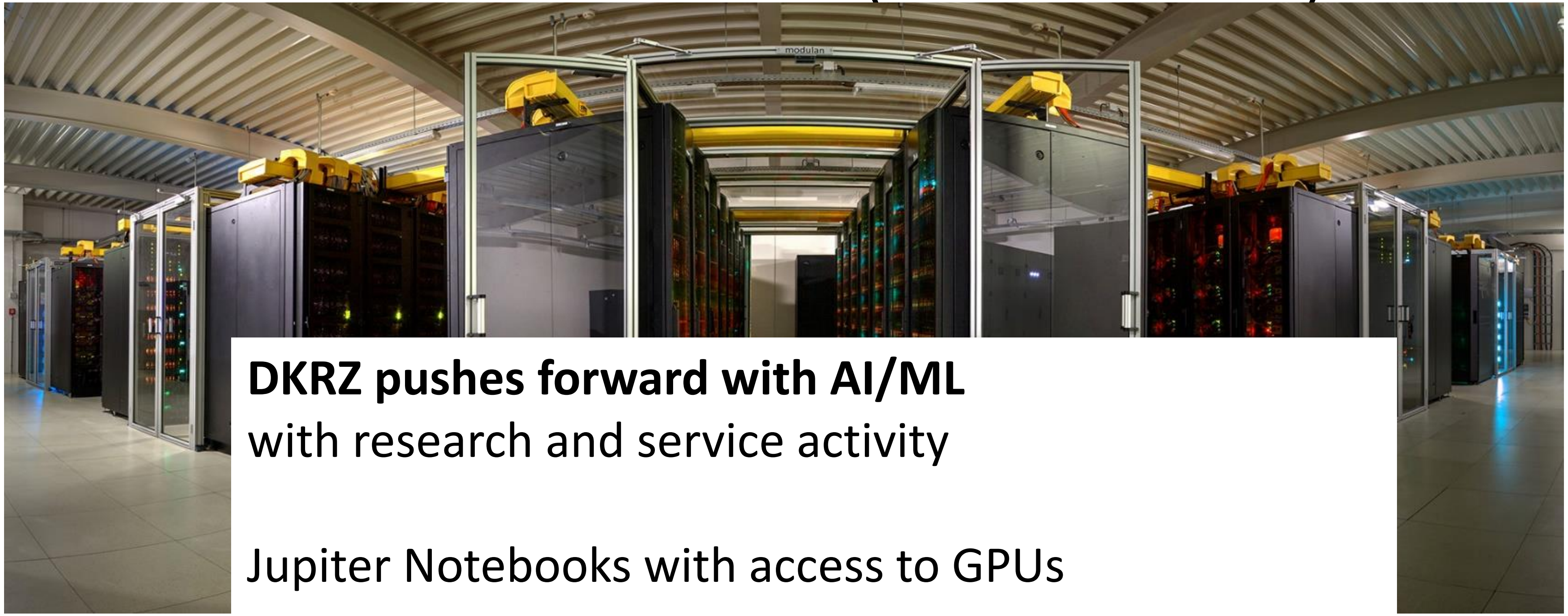
Partner

Climate institutions play an important in climate science

- Max Planck Institute for Meteorology (MPI-M),
- Climate Service Center Germany (GERICS),
- University of Hamburg and Climate Campus,
- Helmholtz Center for Coastal Research
- and more...

DKRZ
DEUTSCHES KLIMARECHENZENTRUM

German Climate Computing Center
HLRE-3 – Mistral (2015-2021)



DKRZ pushes forward with AI/ML
with research and service activity

Jupiter Notebooks with access to GPUs

However, focus is not on ML needed technology (yet)

bullx DLC 720, 3,500+ nodes, 100,000+ cores, Haswell/Broadwell, 3.6 PFLOPS
240 TB main memory, 54 PB disk storage, 450 GB/s mem-disk rate, FDR network
21 nodes for visualization
hot liquid cooling with high efficiency

DKRZ Machine Learning Research Group

Christopher Kadow, Martin Bergemann, Etor Lucio, Mahesh Ramadoss

Climate Informatics and Technologies

Artificial Intelligence
Machine Learning
Data Mining
Deep Learning

Software Development
Data Analytics
Evaluation and Validation
HPC (CPU/GPU/TPU)

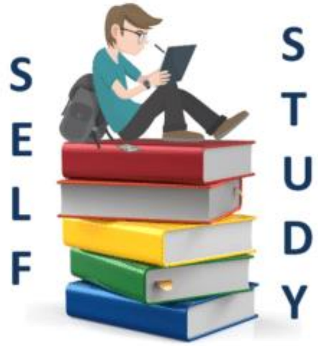
- Interface between AI/ML and Climate Science
- AI/ML for DKRZ HPC Infrastructure
- Knowledge Transfer and Method Research for Climate Community
- Utilization of cutting-egde AI/ML Technologies for Climate Scientists

SPOILER ALERT
Master Thesis

Climate Science & Machine Learning

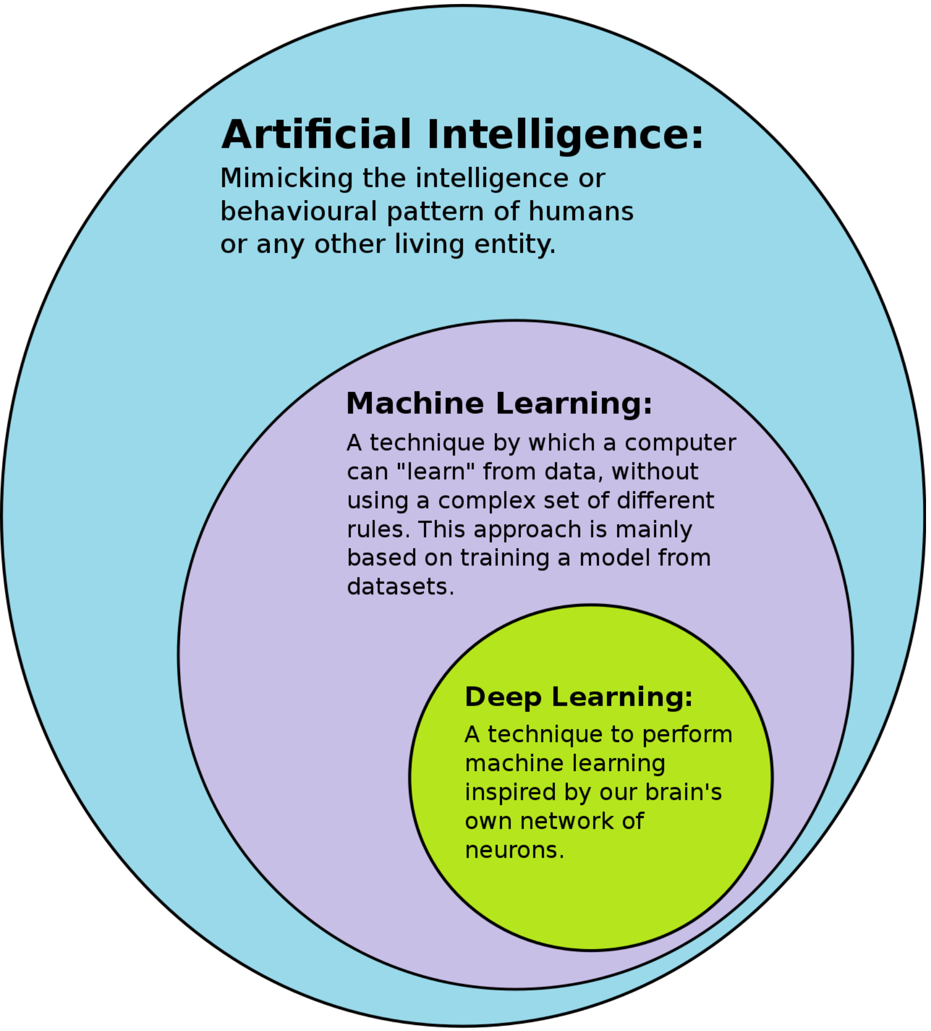
Agenda *Introduction and Research*

- General
  - DL, ML, AI? WTF? Literature?
  - What is a Neural Network?
- Methods & Networks
  - Supervised Learning, Unsupervised Learning, Reinforcement Learning
  - Convolutional Neural Network, Recurrent Neural Network, Generative Adversal Network
- Hardware & Software
  - PCs, HPCs, Clouds
  - Tools, Frameworks, First Steps
- AI reconstructs missing Climate Information
  - A Research Journey
  - Transfer Learning
  - What is next?

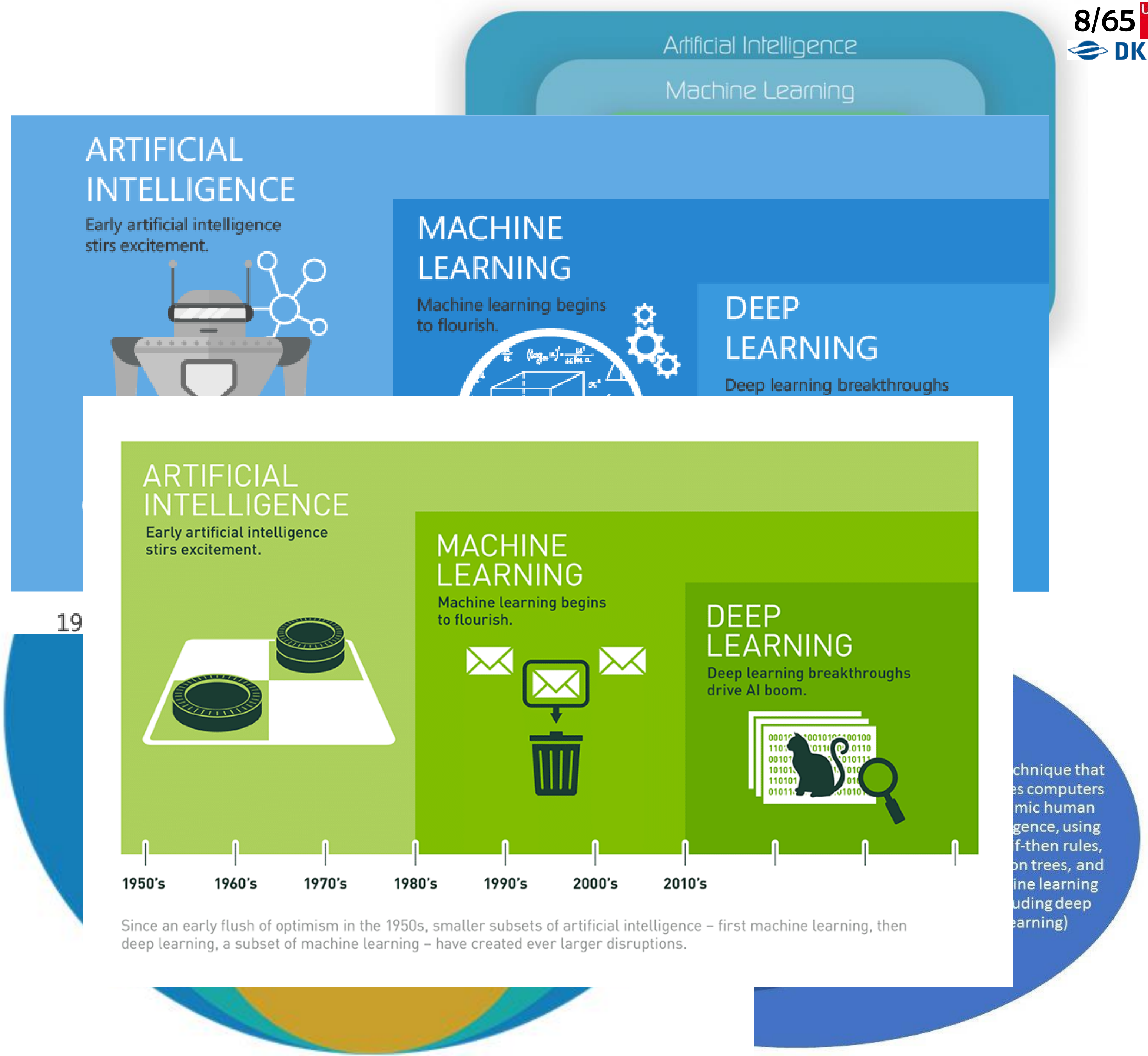


General

General



Wikipedia.com



General

Data Science

- Need of entire analytics universe
- Branch that deals with data
- Different operations related to data i.e.
  - Data Gathering
  - Data Cleaning
  - Data Subsetting
  - Data Manipulation
  - Data Insights [Data Mining]

Machine Learning

- Combination of Machine and Data Science
- Machines utilize Data Science techniques to learn about the data hence called as Machine Learning
- Model Building, Model Evaluation and Validation
- 3 Types:
  - Unsupervised Learning
  - Reinforcement Learning
  - Supervised Learning
- Most popular tools are Python, R and SAS

Deep Learning

- Specific branch of Machine Learning that deals with different flavours of Neural Network
- Examples
  - Simple Neural Network
  - Convolutional Neural Network
  - Recurrent Neural Network
  - Long Short Term Memory
- Mainly utilized in..
  - Object detection in Image and Video
  - Speech Recognition
  - Natural Language Processing and Understandings

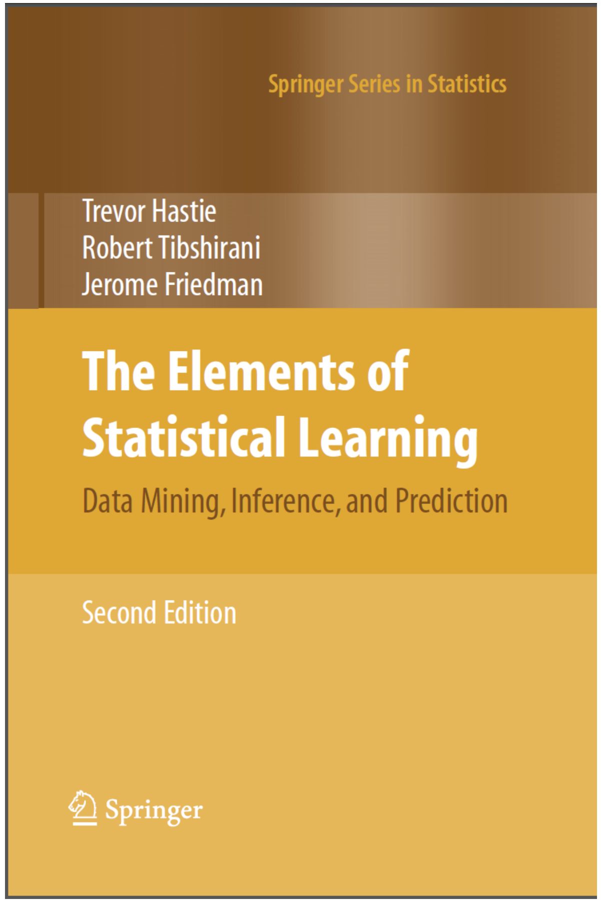
Artificial Intelligence

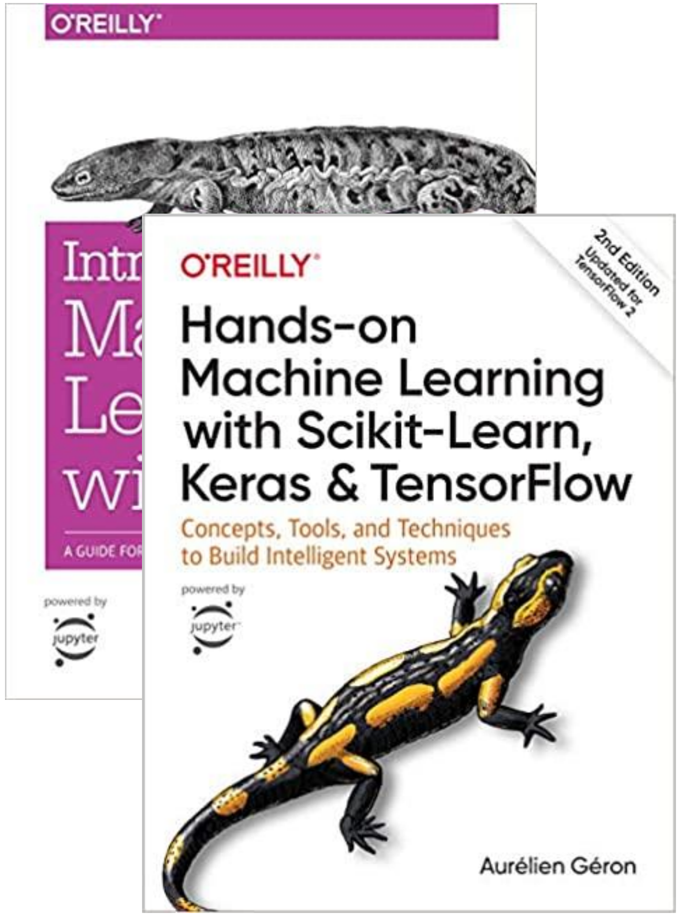
- Big Umbrella
- Empowering machines to take decisions on their own
- As the name suggest imparting humans' natural intelligence in machines
- Thus machines have ability to understand and react according to the situation

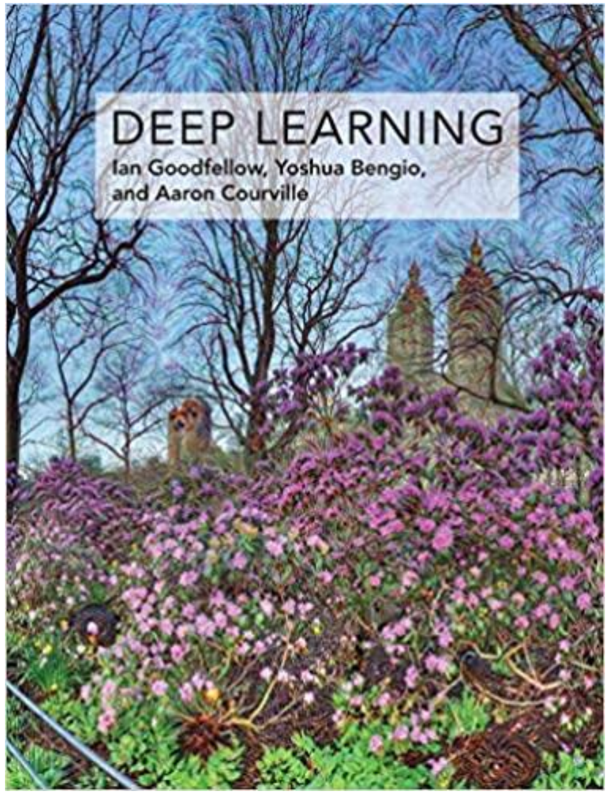
https://mc.ai

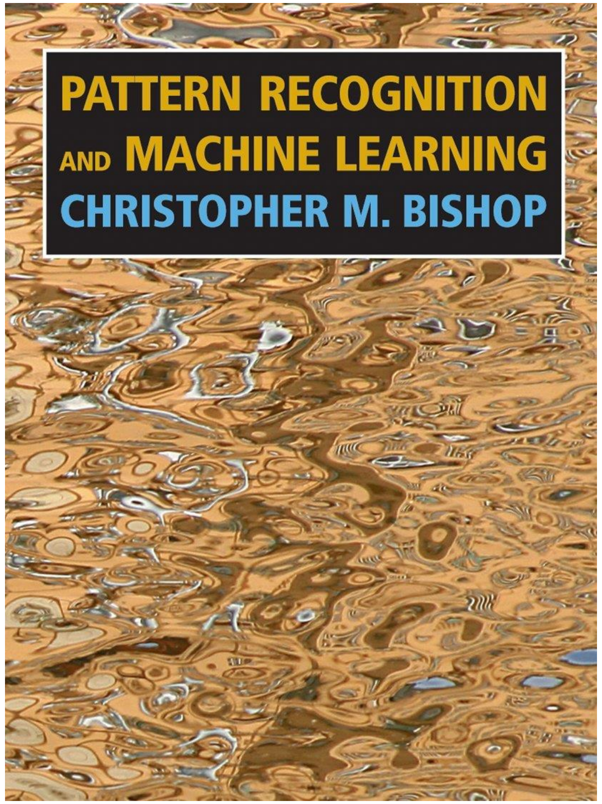
Literature

Books









PDF free online

PDF free online

PDF free online

General

German

PDF free online

PDF and Printed

PDF

Literature

DKRZ

MACHINE LEARNING

Climate Informatics: Accelerating Discovering in Climate Science with Machine Learning

The goal of climate informatics, an emerging discipline, is to inspire collaboration between climate scientists and data scientists, in order to develop tools to analyze complex and ever-growing amounts of observed and simulated climate data, and thereby bridge the gap between data and understanding. Here, recent climate informatics work is discussed, along with some of the field's remaining challenges.

The impacts of present and potential future climate change pose important scientific and societal challenges. Scientists have observed changes in temperature, sea ice, and sea level, and attributed those changes to human activity. It is an urgent international priority to improve our understanding of the climate system—a system characterized by complex phenomena that are difficult to observe and even more difficult to simulate. Despite the increasing availability of computational resources, current analytical tools have been outpaced by the ever-growing amounts of observed climate data from satellites, environmental sensors, and climate-model simulations. Computational approaches will therefore be indispensable for these analysis challenges. The goal of the fledgling research discipline, *climate informatics*, is to inspire collaboration between climate scientists and data scientists (machine learning, statistics, and data mining researchers), and thus bridge the gap between data and understanding. Research on climate informatics will accelerate discovery and answer pressing questions in climate science.

Machine learning is an active research area at the interface of computer science and statistics. The goal of machine learning research is to develop *algorithms*, automated techniques, to detect patterns in data. Such algorithms are critical to a range of technologies including Web search, recommendation systems, personalized Internet advertising, computer vision, and natural language processing. Machine learning also benefits the natural sciences, such as biology; the interdisciplinary bioinformatics field has facilitated many discoveries in genomics and proteomics. The impact of machine learning on climate science has the potential to be similarly profound.

Here, we focus specifically on challenges in climate modeling; however, there are myriad collaborations possible at the intersection of these two fields. Recent work reveals that collaborations with climate scientists also generate interesting new problems for machine learning.[1] To broaden the discussion, we propose challenge problems for climate informatics, some

1521-9615/13/$31.00 © 2013 IEEE
Copublished by the IEEE CS and the AIP

Claire Monteleoni
George Washington University
Gavin A. Schmidt
NASA Goddard Institute for Space Studies
Scott McQuade
George Washington University

Monteleoni, C., G.A. Schmidt, and S. McQuade, 2013: Climate informatics: Accelerating discovering in climate science with machine learning. *Comput. Sci. Eng.*, **15**, 32-41, doi:10.1109/MCSE.2013.50.

PERSPECTIVE

https://doi.org/10.1038/s41586-019-0912-1

Deep learning and process understanding for data-driven Earth system science

Markus Reichstein[1,2*], Gustau Camps-Valls[3], Bjorn Stevens[4], Martin Jung[1], Joachim Denzler[2,5], Nuno Carvalhais[1,6] & Prabhat[7]

Machine learning approaches are increasingly used to extract patterns and insights from the ever-increasing stream of geospatial data, but current approaches may not be optimal when system behaviour is dominated by spatial or temporal context. Here, rather than amending classical machine learning, we argue that these contextual cues should be used as part of deep learning (an approach that is able to extract spatio-temporal features automatically) to gain further process understanding of Earth system science problems, improving the predictive ability of seasonal forecasting and modelling of long-range spatial connections across multiple timescales, for example. The next step will be a hybrid modelling approach, coupling physical process models with the versatility of data-driven machine learning.

Humans have always striven to predict and understand the world, and the ability to make better predictions has given competitive advantages in diverse contexts (such as weather, diseases or financial markets). Yet the tools for prediction have substantially changed over time, from ancient Greek philosophical reasoning to non-scientific medieval methods such as soothsaying, towards modern scientific discourse, which has come to include hypothesis testing, theory development and computer modelling underpinned by statistical and physical relationships, that is, laws[1]. A success story in the geosciences is weather prediction, which has greatly improved through the integration of better theory, increased computational power, and established observational systems, which allow for the assimilation of large amounts of data into the modelling system[2]. Nevertheless, we can accurately predict the evolution of the weather on a timescale of days, not months. Seasonal meteorological predictions, forecasting extreme events such as flooding or fire, and long-term climate projections are still major challenges. This is especially true for predicting dynamics in the biosphere, which is dominated by biologically mediated processes such as growth or reproduction, and is strongly controlled by seemingly stochastic disturbances such as fires and landslides. Such predictive problems have not seen much progress in the past few decades[3].

At the same time, a deluge of Earth system data has become available, with storage volumes already well beyond dozens of petabytes and rapidly increasing transmission rates exceeding hundreds of terabytes per day[4]. These data come from a plethora of sensors measuring states, fluxes and intensive or time/space-integrated variables, representing fifteen or more orders of temporal and spatial magnitude. They include remote sensing from a few metres to hundreds of kilometres above Earth as well as in situ observations (increasingly from autonomous sensors) at and below the surface and in the atmosphere, many of which are further being complemented by citizen science observations. Model simulation output adds to this deluge; the CMIP-5 dataset of the Climate Model Intercomparison Project, used extensively for scientific groundwork towards periodic climate assessments, is over 3 petabytes in size, and the next generation, CMIP-6, is estimated to reach up to 30 petabytes[5]. The data from models share many of the challenges and statistical properties of observational data, including many forms of uncertainty. In summary, Earth system data are exemplary of all four of the 'four Vs' of 'big data': volume, velocity, variety and veracity (see Fig. 1). One key challenge is to extract interpretable information and knowledge from this big data, possibly almost in real time and integrating between disciplines.

Taken together, our ability to collect and create data far outpaces our ability to sensibly assimilate it, let alone understand it. Predictive ability in the last few decades has not increased apace with data availability. To get the most out of the explosive growth and diversity of Earth system data, we face two major tasks in the coming years: (1) extracting knowledge from the data deluge, and (2) deriving models that learn much more from data than traditional data assimilation approaches can, while still respecting our evolving understanding of nature's laws.

The combination of unprecedented data sources, increased computational power, and the recent advances in statistical modelling and machine learning offer exciting new opportunities for expanding our knowledge about the Earth system from data. In particular, many tools are available from the fields of machine learning and artificial intelligence, but they need to be further developed and adapted to geo-scientific analysis. Earth system science offers new opportunities, challenges and methodological demands, in particular for recent research lines focusing on spatio-temporal context and uncertainties (Box 1; see https://developers.google.com/machine-learning/glossary/ and http://www.wildml.com/deep-learning-glossary/ for more complete glossaries).

In the following sections we review the development of machine learning in the geoscientific context, and highlight how deep learning—that is, the automatic extraction of abstract (spatio-temporal) features—has the potential to overcome many of the limitations that have, until now, hindered a more wide-spread adoption of machine learning. We further lay out the most promising but also challenging approaches in combining machine learning with physical modelling.

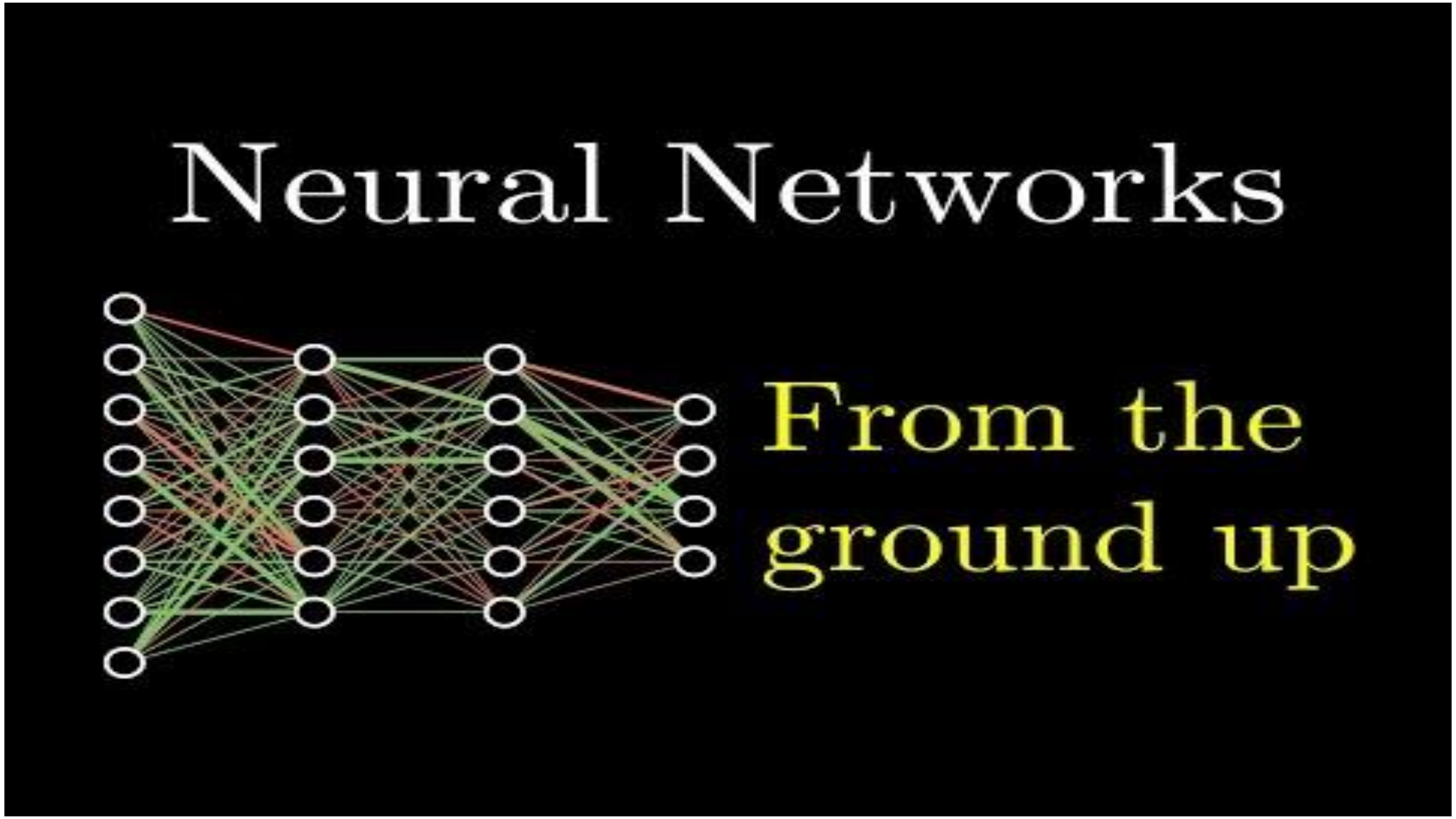
State-of-the-art geoscientific machine learning

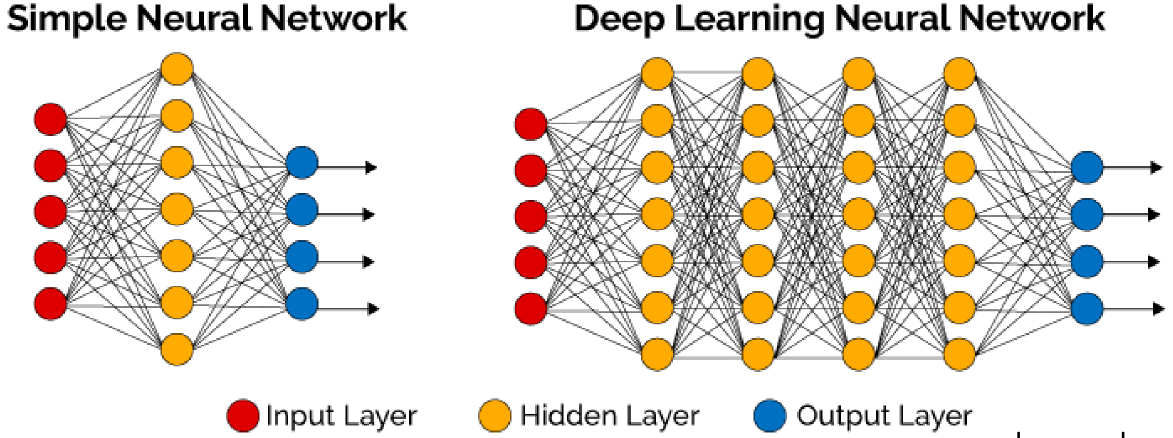
Machine learning is now a successful part of several research-driven and operational geoscientific processing schemes, addressing the atmosphere, the land surface and the ocean, and has co-evolved with data availability over the past decade. Early landmarks in classification of land cover and clouds emerged almost 30 years ago through the coincidence of high-resolution satellite data and the first revival of neural networks[6,7]. Most major machine learning methodological

[1]Department of Biogeochemical Integration, Max Planck Institute for Biogeochemistry, Jena, Germany. [2]Michael-Stifel-Center Jena for Data-driven and Simulation Science, Jena, Germany. [3]Image Processing Laboratory (IPL), University of València, Valencia, Spain. [4]Max Planck Institute for Meteorology, Hamburg, Germany. [5]Computer Vision Group, Computer Science, Friedrich Schiller University, Jena, Germany. [6]CENSE, Departamento de Ciências e Engenharia do Ambiente, Faculdade de Ciências e Tecnologia, Universidade NOVA de Lisboa, Lisbon, Portugal. [7]National Energy Research Supercomputing Center, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. *e-mail: mreichstein@bgc-jena.mpg.de

Reichstein, M., Camps-Valls, G., **Stevens, B**. *et al.* Deep learning and process understanding for data-driven Earth system science. *Nature* **566,** 195–204 (2019). https://doi.org/10.1038/s41586-019-0912-1

General





becomehuman.ai

Linear model	$f(x) = softmax(W_1 x)$
Neural network	$f(x) = softmax(W_2(g(W_1 x)))$
Deep neural network	$f(x) = softmax(W_3(g(W_2(g(W_1 x)))))$

6 Stages of Neural Network Learning

1. Initialization—initial weights are applied to all the neurons.

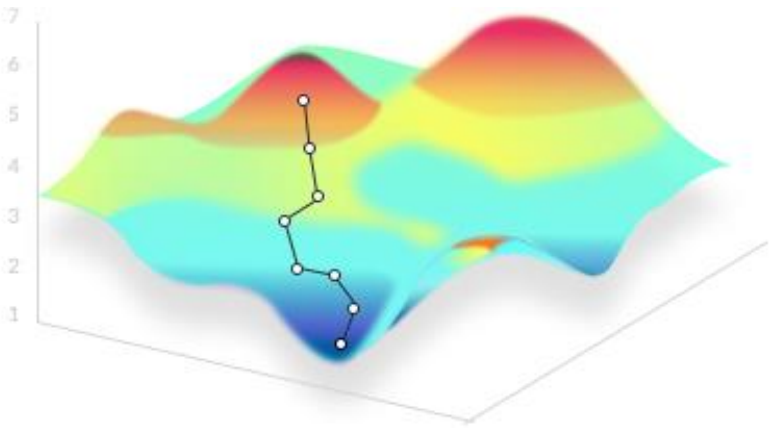
2. Forward propagation—the inputs from a training set are passed through the neural network and an output is computed.

3. Error function—because we are working with a training set, the correct output is known. An error function is defined, which captures the delta between the correct output and the actual output of the model, given the current model weights.

4. **Backpropagation**—the objective of backpropagation is to change the weights for the neurons, in order to bring the error function to a minimum.

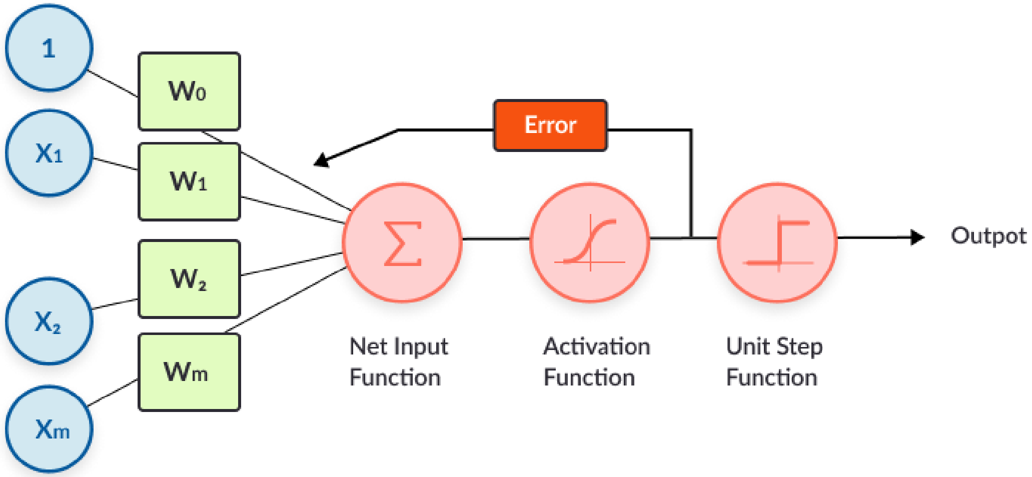
5. Weight update—weights are changed to the optimal values according to the results of the backpropagation algorithm.

6. Iterate until convergence—because the weights are updated a small delta step at a time, several iterations are required in order for the network to learn. After each iteration, the gradient descent force updates the weights towards less and less global loss function.



Backward pass

- Backprop: efficient method to calculate gradients
- Gradient descent: nudge parameters a bit in the opposite direction

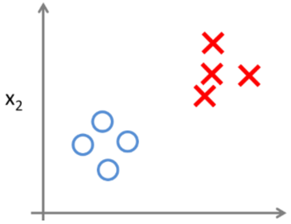


From missinglink.ai

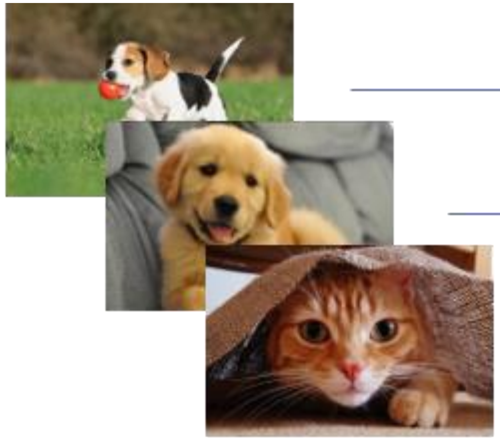
Methods

Methods

Machine Learning → Supervised Learning

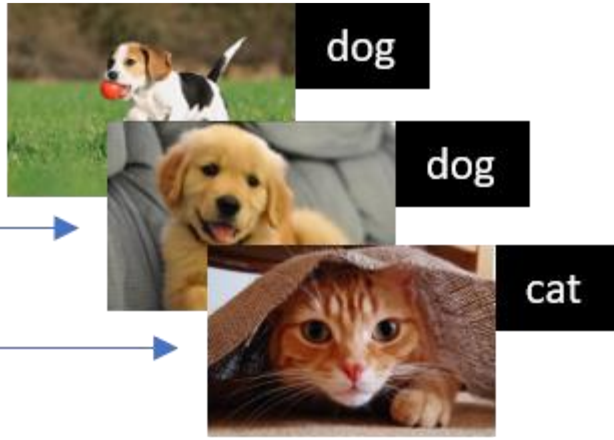


Example:



Supervised ML algorithm

allagora.wordpress.com



Research:

Climate Prediction

Input

Training

Observations

Labeling

Deep Learning

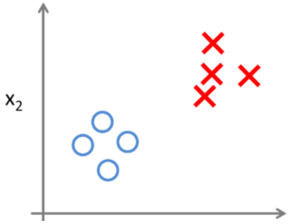
Output

Climate Prediction

Methods

Machine Learning

Supervised Learning



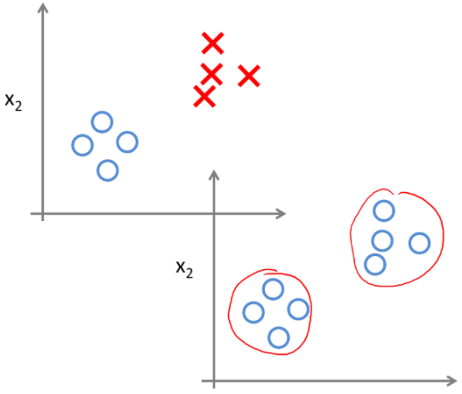
TWO MINUTE PAPERS

YouTube

Methods

Machine Learning → Supervised Learning

Machine Learning → Unsupervised Learning



Example:

allagora.wordpress.com

Research:

Climate Observation $\xrightarrow{\text{Input}}$ Deep Learning → Climate Observation

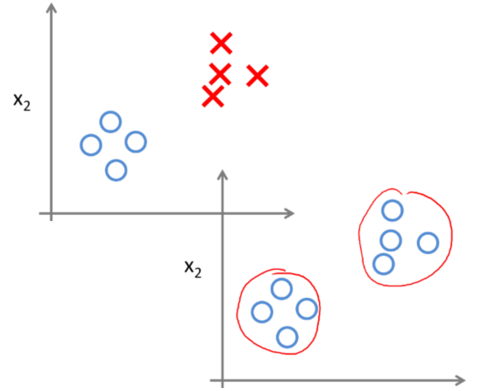
Deep Learning → Climate Observation

Methods

Machine Learning

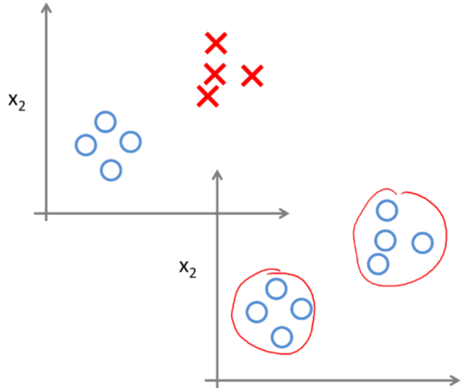
Supervised Learning

Unsupervised Learning



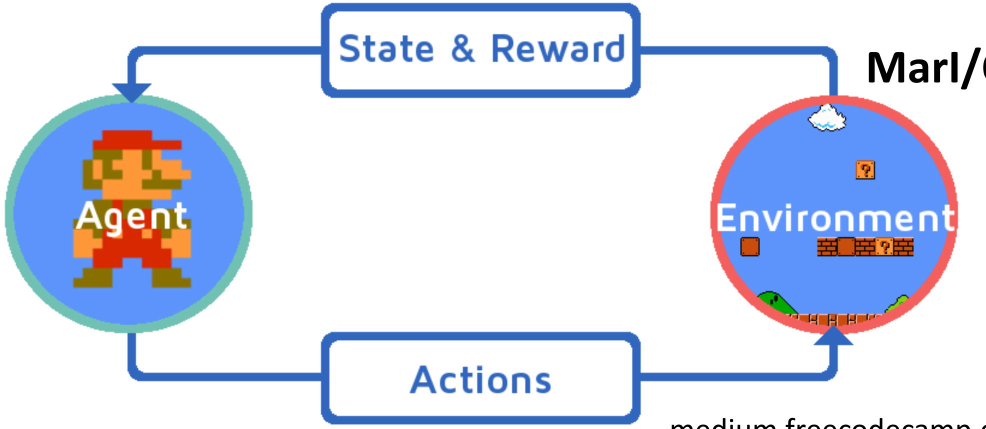
Methods

Machine Learning → Supervised Learning, Unsupervised Learning, Reinforcement Learning



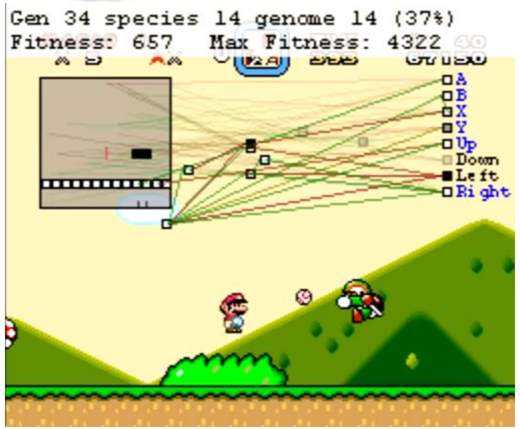
APE2010.de; www.coursera.org

Example:



MarI/O

medium.freecodecamp.org



www.youtube.com/watch?v=qv6UVOQ0F44

Research:

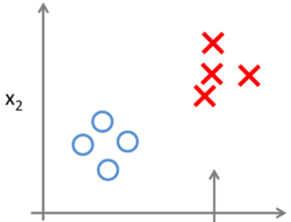
Climate Prediction —Input→ Deep Learning —Output→ Climate Prediction

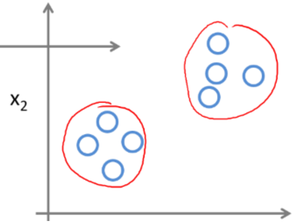
Global Mean Temperature Correlation ←Reward→ Deep Learning

Methods

Machine Learning

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

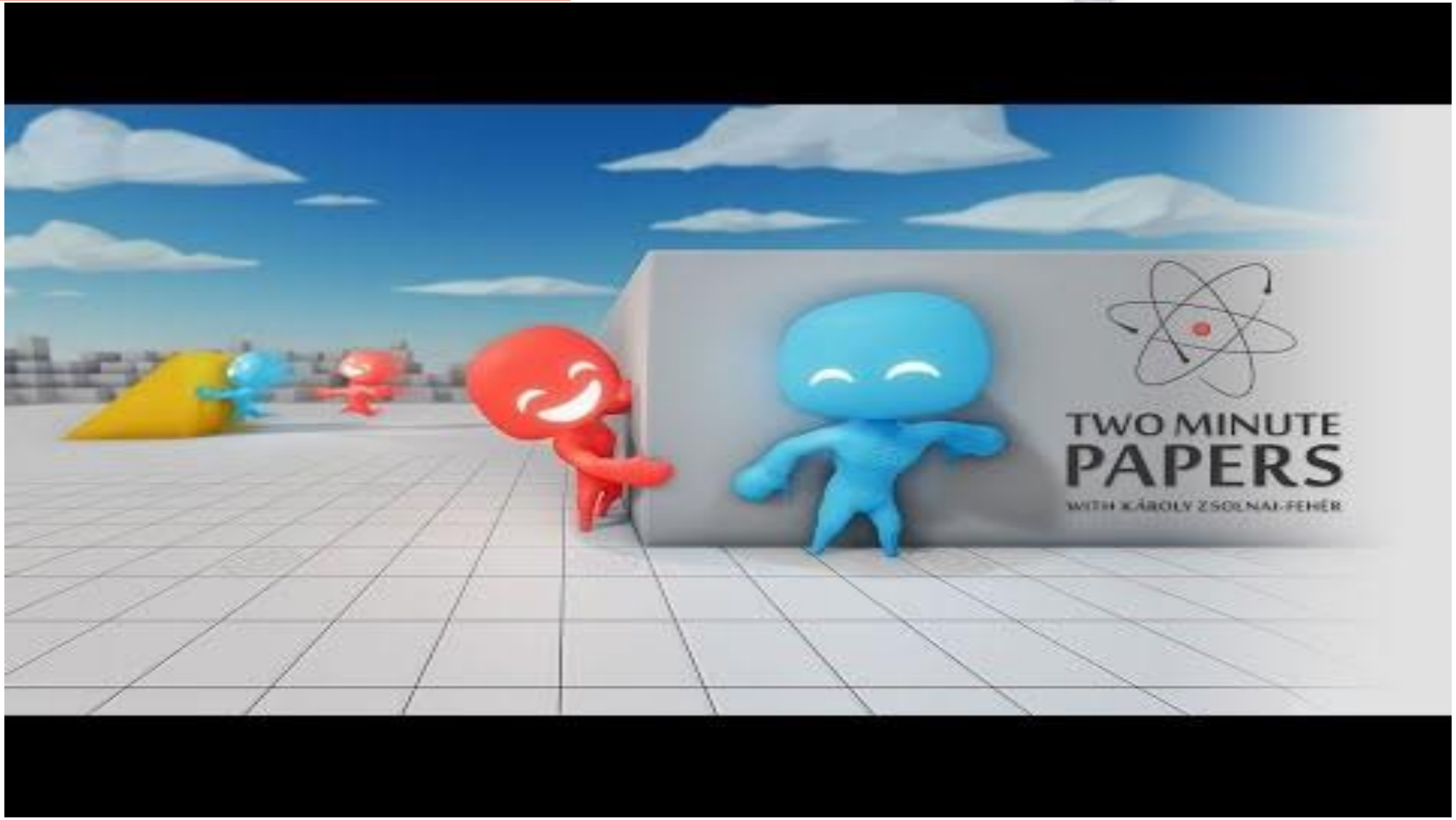




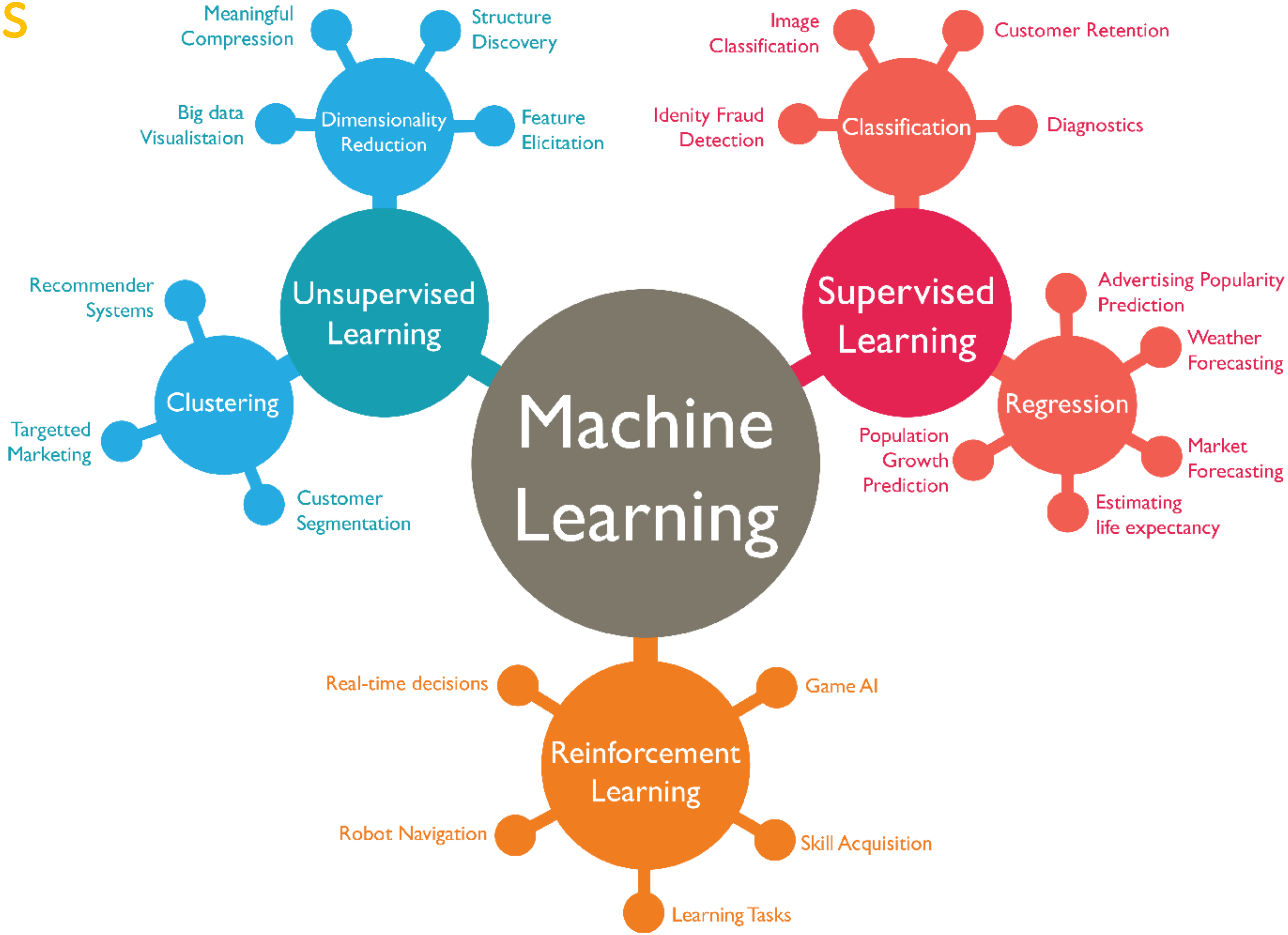
APE2010.de; www.coursera.org

TWO MINUTE PAPERS

YouTube



Methods

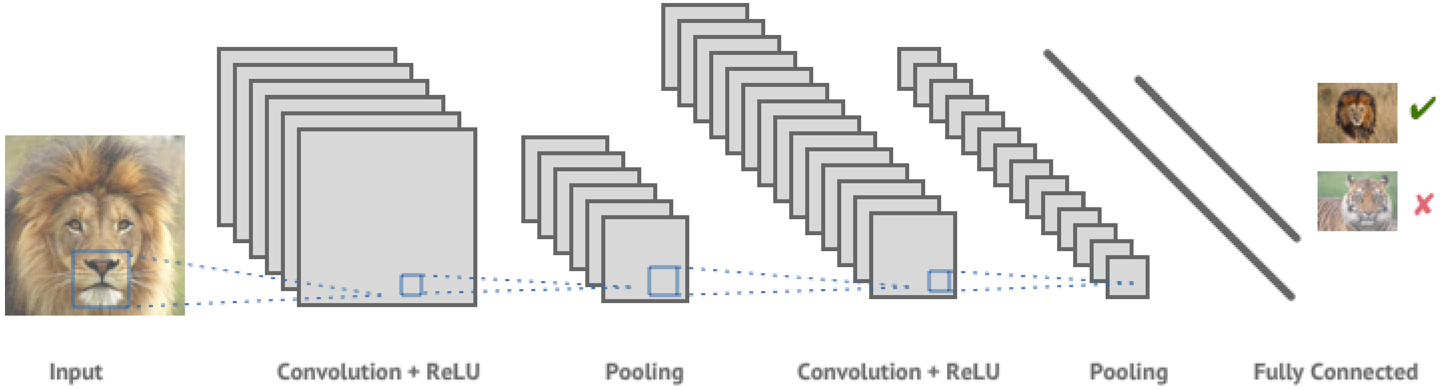

Meaningful Compression
Structure Discovery
Big data Visualistaion
Dimensionality Reduction
Feature Elicitation
Image Classification
Customer Retention
Idenity Fraud Detection
Classification
Diagnostics
Unsupervised Learning
Supervised Learning
Recommender Systems
Clustering
Targetted Marketing
Customer Segmentation
Machine Learning
Advertising Popularity Prediction
Weather Forecasting
Regression
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Real-time decisions
Game AI
Reinforcement Learning
Robot Navigation
Skill Acquisition
Learning Tasks

Networks

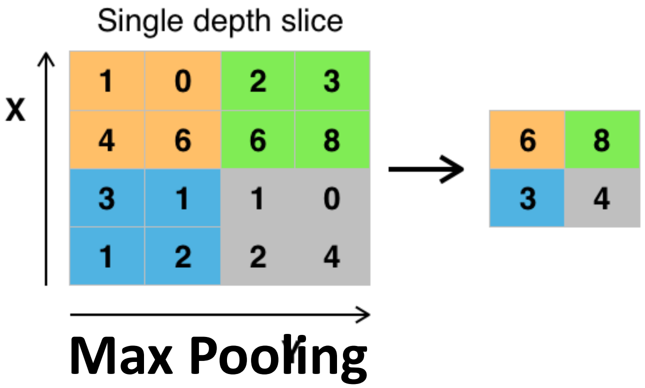
Convolutional Neural Network (CNN, or ConvNet)

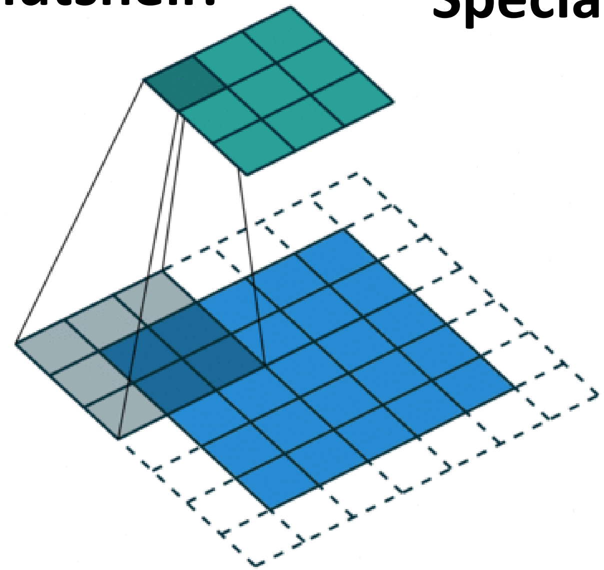
Good for: Classification, Supervised Learning, Image Recognition

A **Convolutional Neural Network (ConvNet/CNN)** is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. The pre-processing required in a CNN is much lower as compared to other classification algorithms.

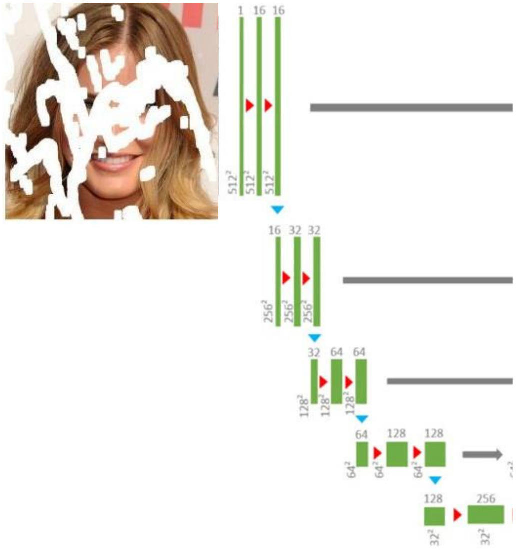


How does this work in a nutshell?





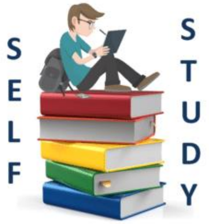
Special: U-Net



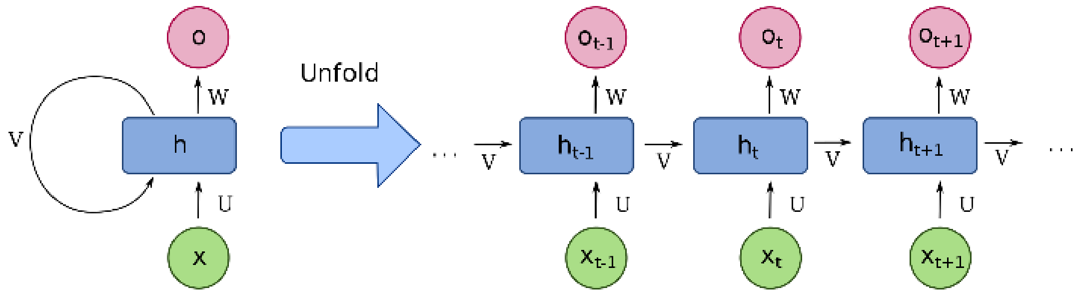
MIT Introduction to Deep Learning: CNN ~40min
https://www.youtube.com/watch?v=iaSUYvmCekI

Recurrent Neural Network (RNN)

Good for: make use of sequential information, have a "memory" which captures info about what has been calculated so far.



A **recurrent neural network** (**RNN**) is a class of artificial neural networks where connections between nodes form a directed graph along a temporal sequence. This allows it to exhibit temporal dynamic behavior. Derived from feedforward neural networks, RNNs can use their internal state (memory) to process variable length sequences of inputs.[

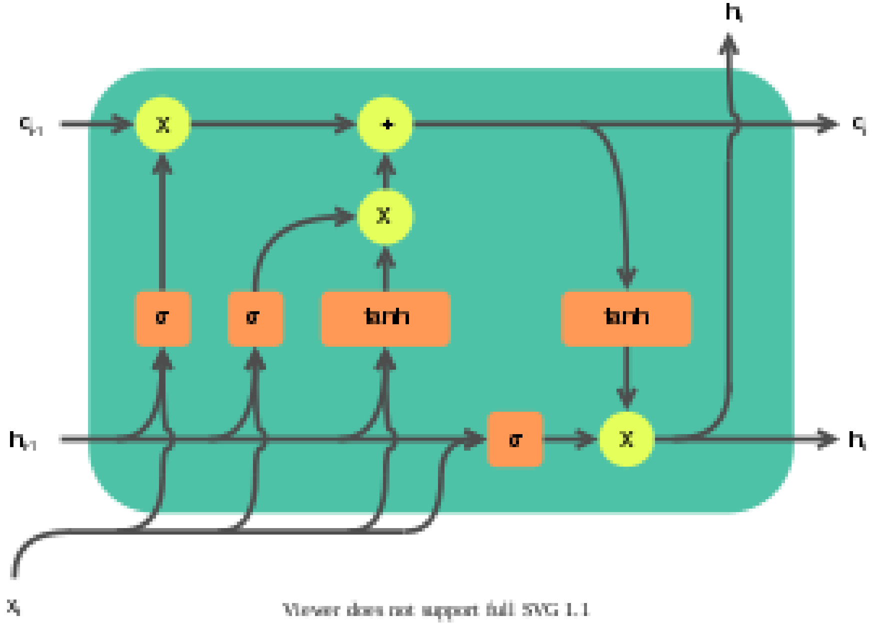


How does this work in a nutshell?

RNN is a generalization of feed-forward neural network that has an internal memory. RNNs are designed to recognize a data's sequential characteristics and use patterns to predict the next likely scenario.

Special: LSTM

Long short-term memory, has feedbacks, can process sequences of data. It has gates, which decide about information to be stored as memory.



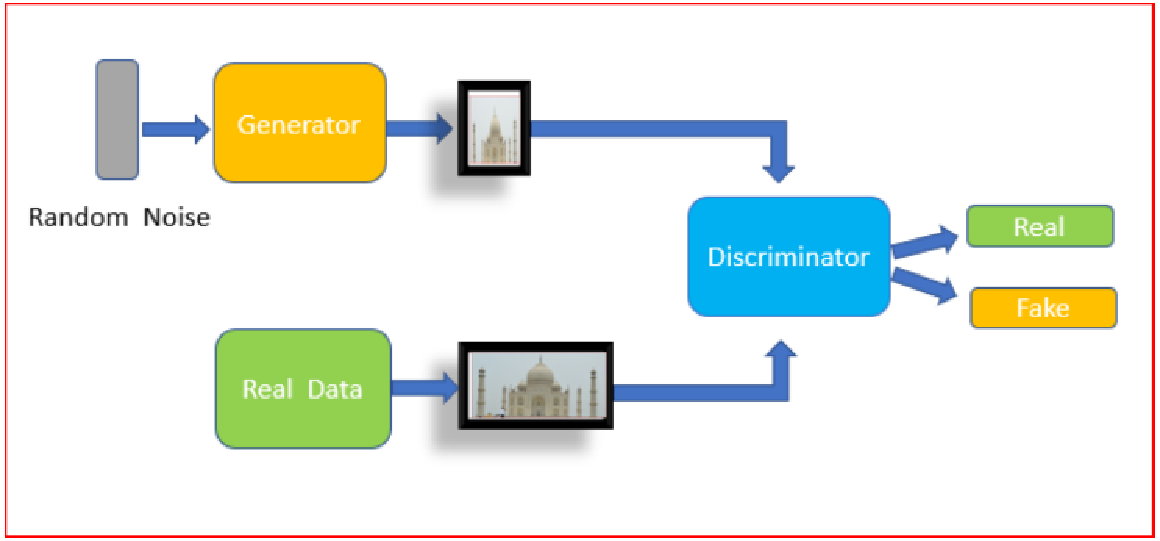
MIT Introduction to Deep Learning: RNN ~40min
https://www.youtube.com/watch?v=SEnXr6v2ifU

Issues?
Gradient vanishing
Training is difficult/Failure to converge
Cannot process very long sequences

Generative Adversarial Networks (GAN)

Good for: imitation of data, structures, pictures, systems

A **generative adversarial network** (**GAN**) is a class of machine learning frameworks. The *generative* network generates candidates while the *discriminative* network evaluates them. The contest operates in terms of data distributions. GANs often suffer from a "mode collapse" where they fail to generalize properly, missing entire modes from the input data.



How does this work in a nutshell?

Two neural networks contest with each other in a gam, in the form of a zero-sum game, where one agent's gain is another agent's loss. Can produce realistic fake fotos of humans.

Special: Conditional GAN

The conditional generative adversarial network, or cGAN for short, is a type of GAN that involves the conditional generation of images by a generator model

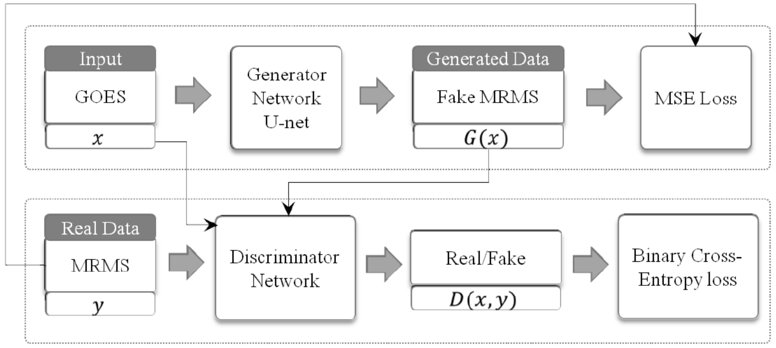
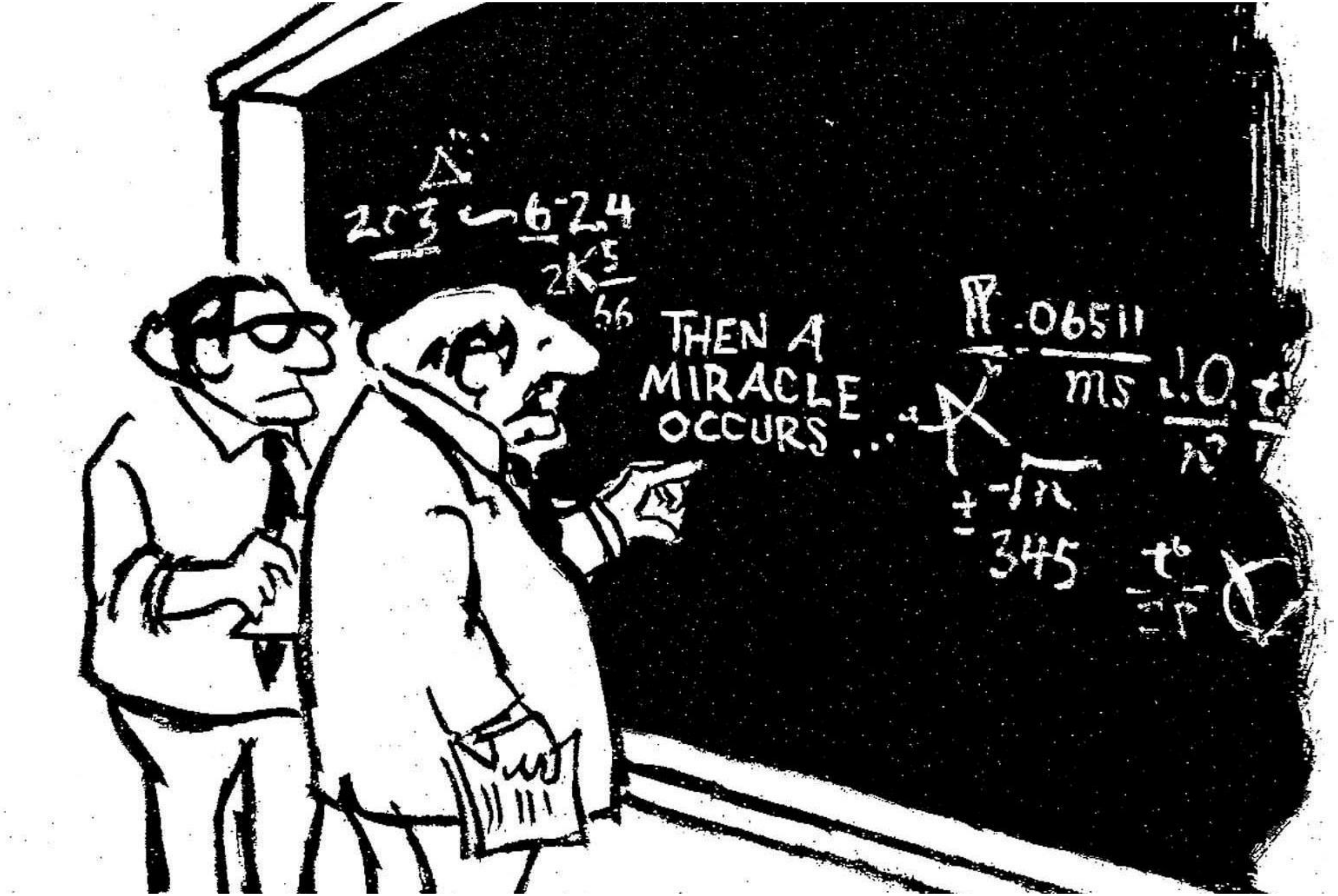


Figure 3. Schematic conditional Generative Adversarial Network Structure.

MIT Introduction to Deep Learning: GAN ~40min
https://www.youtube.com/watch?v=rZufA635dq4

Issues?
Hyperparameter tuning can be tricky and time consuming.
What do you do with „fake“ data?

How2open the Black Box? Where are limits, where physics?



„I think you should be more explicit in step two“

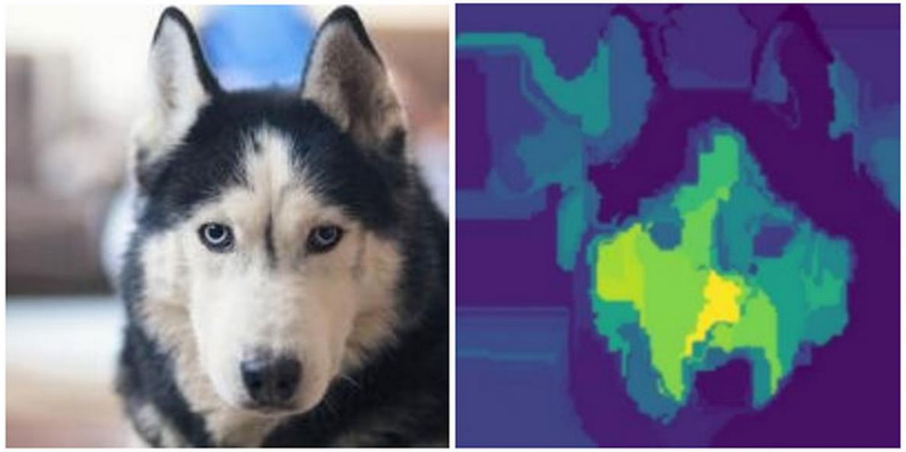
Networks How2open the Black Box? Where are limits, where physics?

1. Scientific Background & Evaluation!

- Build upon weather and climate validation, verification, and evaluation from centuries of research.
- Climate data needs climate data tests.
- We do probably something we already did before, like e.g. forecasts.
- Scientific setups need to make sure to make things right for the right reasons.

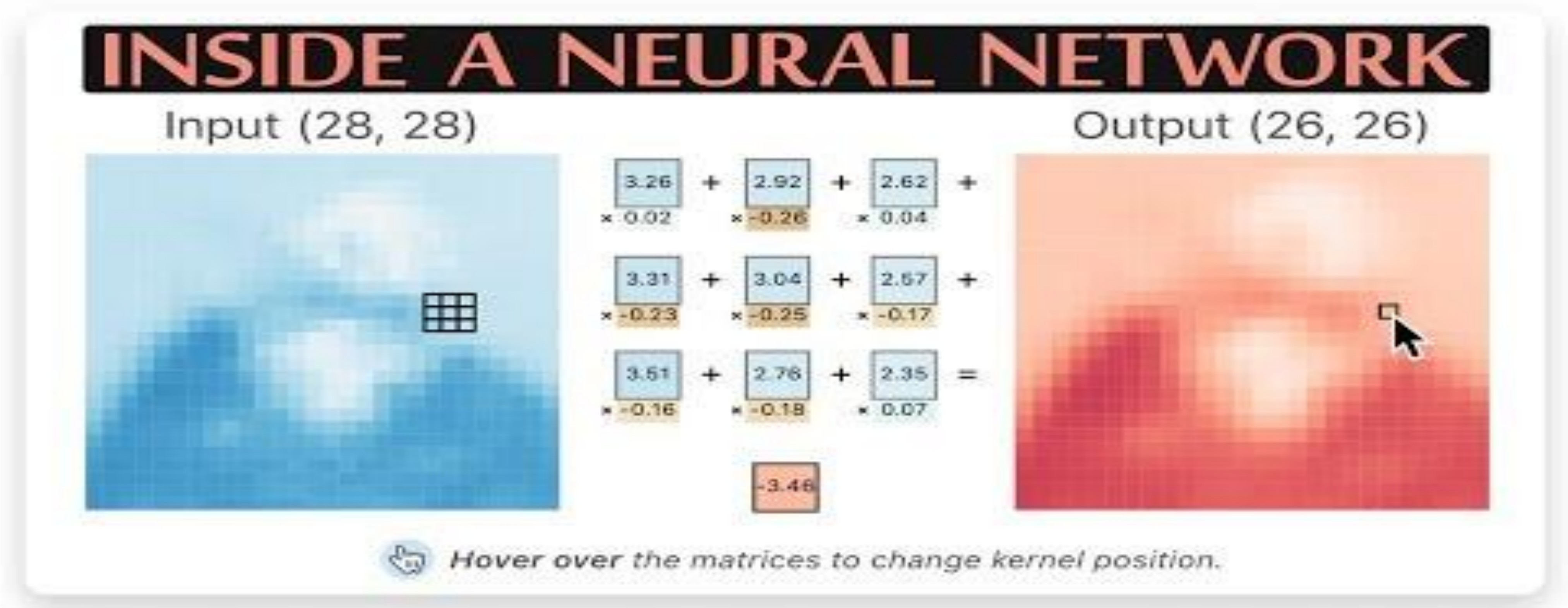
2. Explainable AI

A lot efforts in the ML community to make everything explainable.
Important research for climate science are for example „Heat Maps“, showing where in a 2D field the outcome (Dog/Cat) is mostly based on:



https://cloud.google.com https://arxiv.org/abs/1906.02825

3. Look Inside

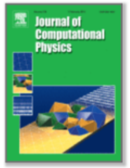


Also see: https://www.youtube.com/watch?v=rGOy9rqGX1k

ELSEVIER

Journal of Computational Physics

Volume 378, 1 February 2019, Pages 686-707



Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations

M. Raissi [a], P. Perdikaris [b], G.E. Karniadakis [a]

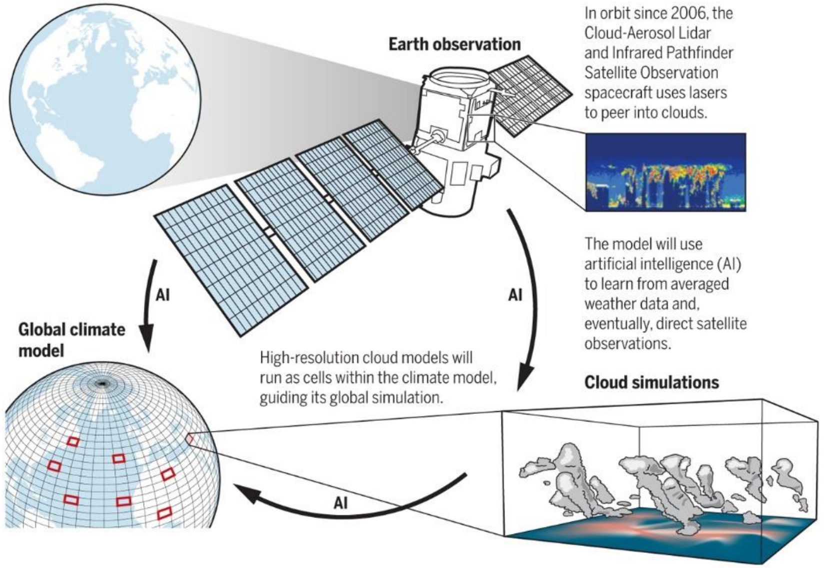
Show more

Share Cite

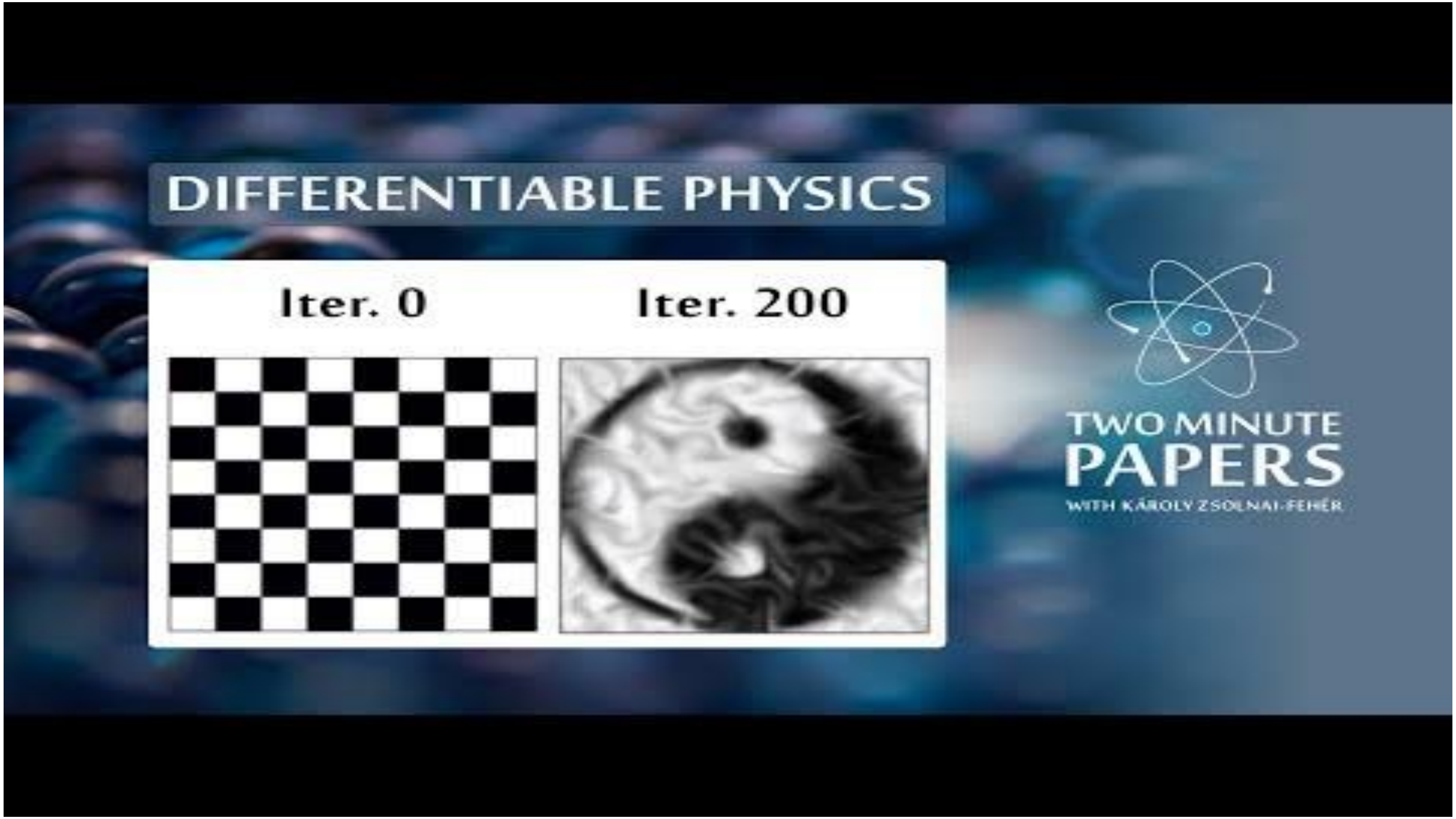
The Earth Machine

Learning the climate

A new data-driven climate model will use satellite observations and high-resolution simulations to learn how best to render its clouds. Similar methods will also be applied to other, small-scale phenomena, such as sea ice and ocean eddies.



Much more groups work on that!


DIFFERENTIABLE PHYSICS
Iter. 0
Iter. 200
TWO MINUTE PAPERS
WITH KÁROLY ZSOLNAI-FEHÉR

Hardware & Software

Hardware & Software

Great News: Deep Learning, Machine Learning, Artificial Intelligence is possible on **CPU, GPU and TPU**

Hardware & Software

Great News: Deep Learning, Machine Learning, Artificial Intelligence is possible on **CPU, GPU and TPU**

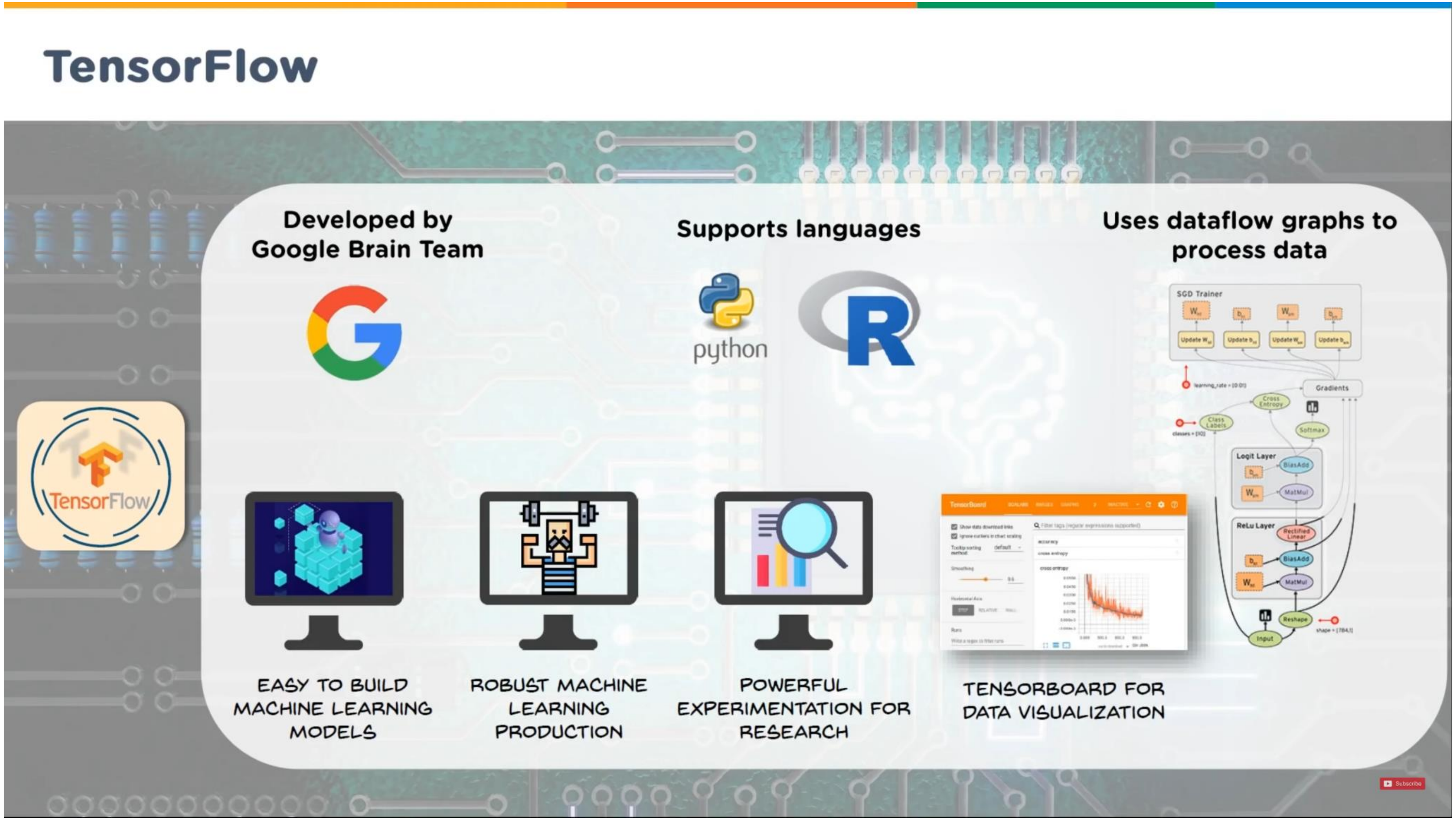
You can use bigger or high performance computer like DKRZ or UHH.

Or also you could use Amazon Web Services (AWS) or Google Cloud.

Hardware & Software

https://www.tensorflow.org/

TensorFlow



https://www.simplilearn.com/

 https://ludwig-ai.github.io/ludwig-docs/

For absolute programming beginners

The core design principles Ludwig:

- **No coding required**: no coding skills are required to train a model and use it for obtaining predictions.

- **Generality**: a new data type-based approach to deep learning model design that makes the tool usable across many different use cases.

- **Flexibility**: experienced users have extensive control over model building and training, while newcomers will find it easy to use.

- **Extensibility**: easy to add new model architecture and new feature data types.

- **Understandability**: deep learning model internals are often considered black boxes, but we provide standard visualizations to understand their performance and compare their predictions.

Hardware & Software

PyTorch

PyTorch is an open source machine learning library based on the Torch library used for applications such as computer vision and natural language processing, primarily developed by Facebook's AI Research lab (FAIR).

WELCOME TO PYTORCH TUTORIALS

New to PyTorch?

The 60 min blitz is the most common starting point and provides a broad view on how to use PyTorch. It covers the basics all the way to constructing deep neural networks.

Start 60-min blitz >

PyTorch Recipes

Bite-size, ready-to-deploy PyTorch code examples.

Explore Recipes >

PyTorch Build	Stable (1.7.1)	Preview (Nightly)			
Your OS	Linux	Mac	Windows		
Package	Conda	Pip	LibTorch	Source	
Language	Python	C++ / Java			
CUDA	9.2	10.1	10.2	11.0	None

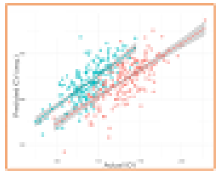
Run this Command:

```
NOTE: Python 3.9 users will need to add '-c=conda-forge' for installation
conda install pytorch torchvision torchaudio cudatoolkit=10.2 -c pytorch
```

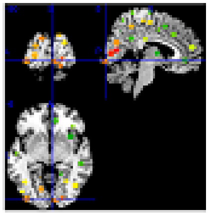
- Alibaba Cloud >
- Amazon Web Services >
- Google Cloud Platform >
- Microsoft Azure >

Hardware & Software

Applied Machine Learning
Dr. Kaustubh Patil



Prediction: Brain-age and size; Clinical status



Biological insights of clinical relevance and evolutionary origins

HIGH PRODUCTIVITY DATA PROCESSING RESEARCH GROUP
DR. -ING. GABRIELE CAVALLARO

What is important for ML on HPCs?

"Reproducibility and Data Management"

julearn — just ml it!

This library provides users with the possibility of testing ML models directly from pandas dataframes, while keeping the flexibility of using scikit-learn's models.
https://juaml.github.io/julearn/main/index.html

DataLad

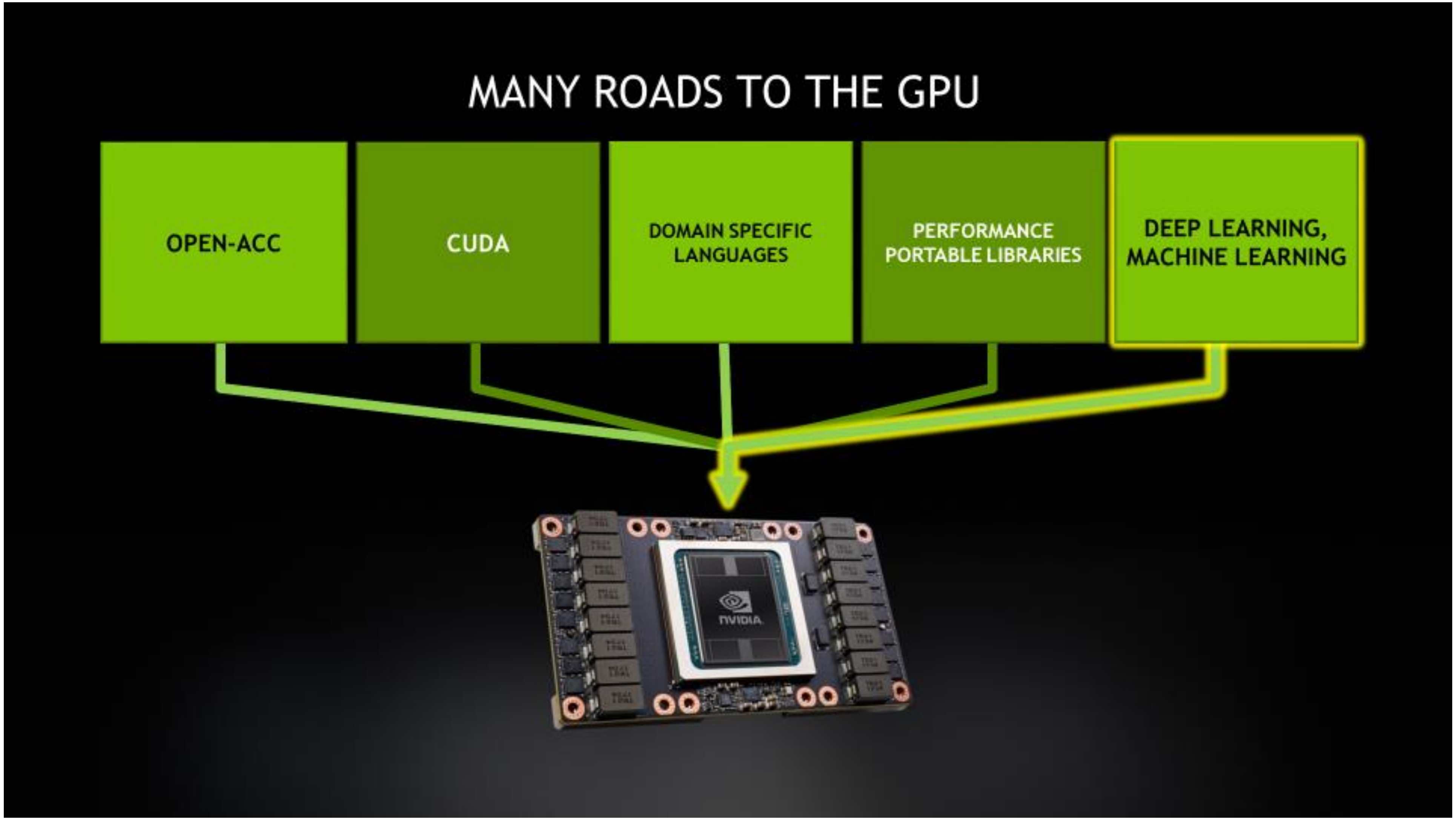
Providing a data portal and a versioning system for everyone, DataLad lets you have your data and control it too.
https://www.datalad.org

	JUWELS-Cluster	JUWELS-Booster
Performance	12 petaflops (12 quadrillion computing operations per second)	73 petaflops (73 quadrillion computing operations per second)
Compute nodes	2511 CPU nodes + 56 GPU nodes	936 GPU nodes
Processors	total of 5134 CPUs (Intel Xeon Skylake) + total of 224 GPUs (NVIDIA V100)	total of 1872 CPUs (AMD EPYC Rome) + total of 3744 GPUs (NVIDIA A100)
Cores	122,768 CPU cores + 71.680 FP64 CUDA cores (GPUs in total)	44,928 CPU cores + 12,939,264 FP64 CUDA cores (GPUs in total)
Main Memory	total of 264 TB	total of 479 TB + total of 150 TB High Bandwidth Memory
Network	100 Gb/s (Mellanox InfiniBand EDR)	200 Gb/s (NVIDIA Mellanox HDR InfiniBand)

ML@HPC at INM-7
Example scenario: brain-age prediction

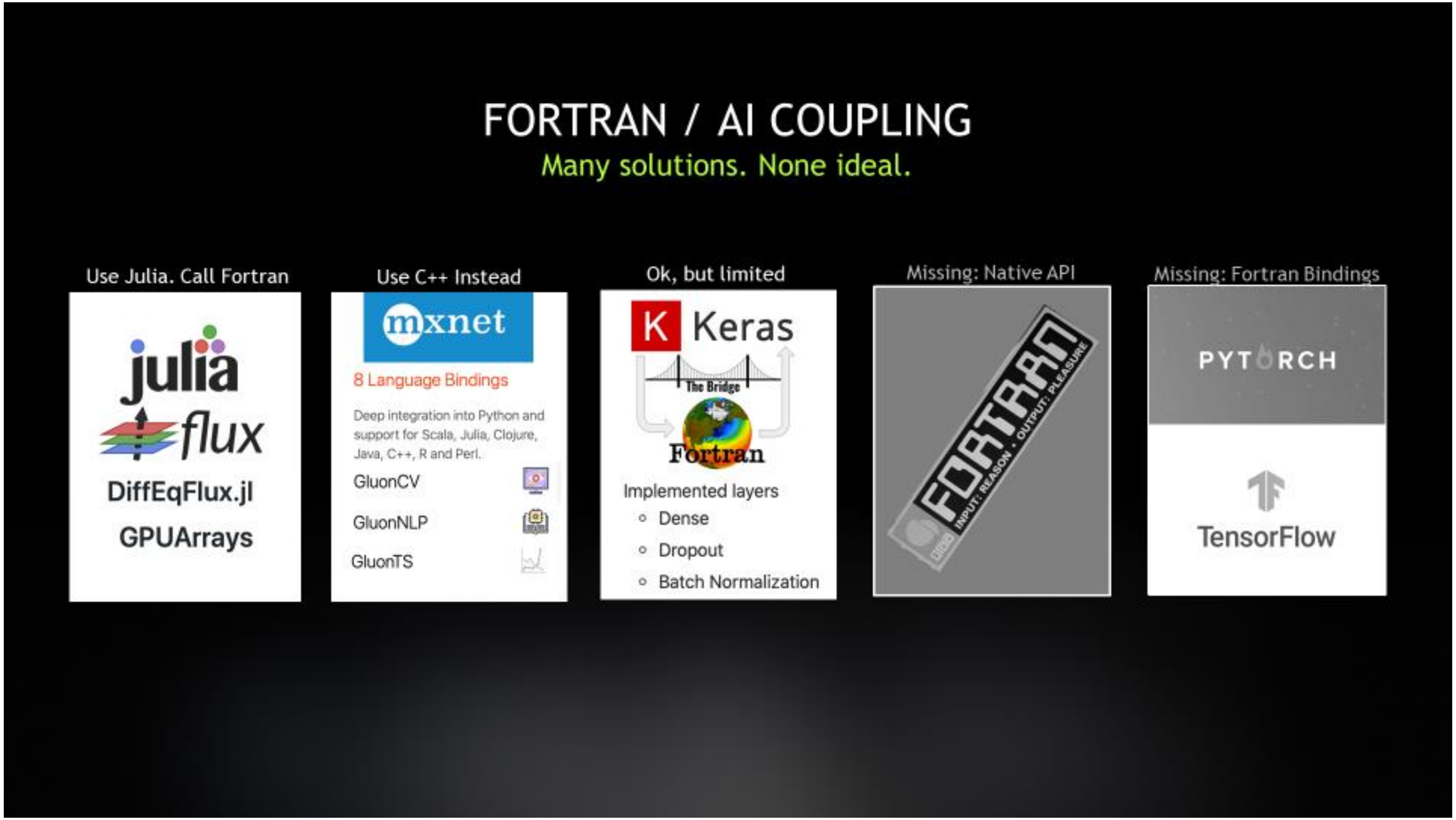
- Problem: Predict chronological age using structural MRI image
- Importance: Large difference in actual and predicted age indicates atypical ageing
- Data: UK biobank with > 40k subjects
- Use of HPC: in all stages of the ML pipeline
  - Data management: dynamic using DataLad (all data does not fit in user folder)
  - Preprocessing (CAT12): ~1hr/subject (parallelized subject-wise on JURECA)
  - Feature extraction: Gray matter volume from thousands of brain regions
  - Learning:
    - Traditional methods: SVM
    - Deep learning: multi-GPU using PyTorch

Hardware & Software



Courtesy by David M. Hall - NVIDIA

Hardware & Software



Courtesy by David M. Hall - NVIDIA

AI reconstructs Climate

Image Inpainting - Restoration

Human Intelligence

Paintings

”Ground Truth“ ”Broken“ ”Restoration“

Artificial Intelligence

Photos

”Ground Truth“ ”Broken“ ”Restoration“

Sanctuary of Mercy church in Borja, Spain
https://en.wikipedia.org

Image Inpainting with Deep Learning
https://medium.com Tarun Bonu

Literature

Bertalmio, M., Sapiro, G. Caselles, V. & Ballester, C. **Image inpainting**. In *Proc. ACM Conf. Comp. Graphics (SIGGRAPH)* (eds Brown, J. R. & Akeley, K.) 417–424 (ACM/Addison-Wesley, 2000)

Image Inpainting

Marcelo Bertalmio and Guillermo Sapiro*
Electrical and Computer Engineering, University of Minnesota

Vicent Caselles and Coloma Ballester
Escola Superior Politecnica, Universitat Pompeu Fabra

Abstract

Inpainting, the technique of modifying an image in an undetectable form, is as ancient as art itself. The goals and applications of inpainting are numerous, from the restoration of damaged paintings and photographs to the removal/replacement of selected objects. In this paper, we introduce a novel algorithm for digital inpainting of still images that attempts to replicate the basic techniques used by professional restorators. After the user selects the regions to be restored, the algorithm automatically fills-in these regions with information surrounding them. The fill-in is done in such a way that isophote lines arriving at the regions' boundaries are completed inside. In contrast with previous approaches, the technique here introduced does not require the user to specify where the novel information comes from. This is automatically done (and in a fast way), thereby allowing to simultaneously fill-in numerous regions containing completely different structures and surrounding backgrounds. In addition, no limitations are imposed on the topology of the region to be inpainted. Applications of this technique include the restoration of old photographs and damaged film; removal of superimposed text like dates, subtitles, or publicity; and the removal of entire objects from the image like microphones or wires in special effects.

CR Categories: I.3.3 [Computer Graphics]: Picture/Image Generation—; I.3.4 [Computer Graphics]: Graphics Utilities—; I.4.4 [Image Processing and Computer Vision]: Restoration—; I.4.9 [Image Processing and Computer Vision]: Applications—;

(e.g., removal of stamped date and red-eye from photographs, the infamous "airbrushing" of political enemies [3]).

Digital techniques are starting to be a widespread way of performing inpainting, ranging from attempts to fully automatic detection and removal of scratches in film [4, 5], all the way to software tools that allow a sophisticated but mostly manual process [6].

In this article we introduce a novel algorithm for automatic digital inpainting, being its main motivation to replicate the basic techniques used by professional restorators. At this point, the only user interaction required by the algorithm here introduced is to mark the regions to be inpainted. Although a number of techniques exist for the semi-automatic detection of image defects (mainly in films), addressing this is out of the scope of this paper. Moreover, since the inpainting algorithm here presented can be used not just to restore damaged photographs but also to remove undesired objects and writings on the image, the regions to be inpainted must be marked by the user, since they depend on his/her subjective selection. Here we are concerned on how to "fill-in" the regions to be inpainted, once they have been selected.[1] Marked regions are automatically filled with the structure of their surrounding, in a form that will be explained later in this paper.

2 Related work and our contribution

We should first note that classical image denoising algorithms do not apply to image inpainting. In common image enhancement applications, the pixels contain both information about the real data

Elharrouss, O., Almaadeed, N., Al-Maadeed, S. & Akbari, Y. Image inpainting: a review. *Neural Process. Lett.* **51**, 2007–2028 (2019).

Neural Processing Letters (2020) 51:2007–2028
https://doi.org/10.1007/s11063-019-10163-0

Check for updates

Image Inpainting: A Review

Omar Elharrouss[1] · Noor Almaadeed[1] · Somaya Al-Maadeed[1] · Younes Akbari[1]

Published online: 6 December 2019

Abstract

Although image inpainting, or the art of repairing the old and deteriorated images, has been around for many years, it has recently gained even more popularity, because of the recent development in image processing techniques. With the improvement of image processing tools and the flexibility of digital image editing, automatic image inpainting has found important applications in computer vision and has also become an important and challenging topic of research in image processing. This paper reviews the existing image inpainting approaches, that were classified into three subcategories, sequential-based, CNN-based, and GAN-based methods. In addition, for each category, a list of methods for different types of distortion on images are presented. Furthermore, the paper also presents available datasets. Last but not least, we present the results of real evaluations of the three categories of image inpainting methods performed on the used datasets, for different types of image distortion. We also present the evaluations metrics and discuss the performance of these methods in terms of these metrics. This overview can be used as a reference for image inpainting researchers, and it can also facilitate the comparison of the methods as well as the datasets used. The main contribution of this paper is the presentation of the three categories of image inpainting methods along with a list of available datasets that the researchers can use to evaluate their proposed methodology against.

Keyword Image inpainting · Objects removal · Image repairing · CNN · GAN

Liu et al. 2018 Image Inpainting for **Irregular Holes** Using Partial Convolutions

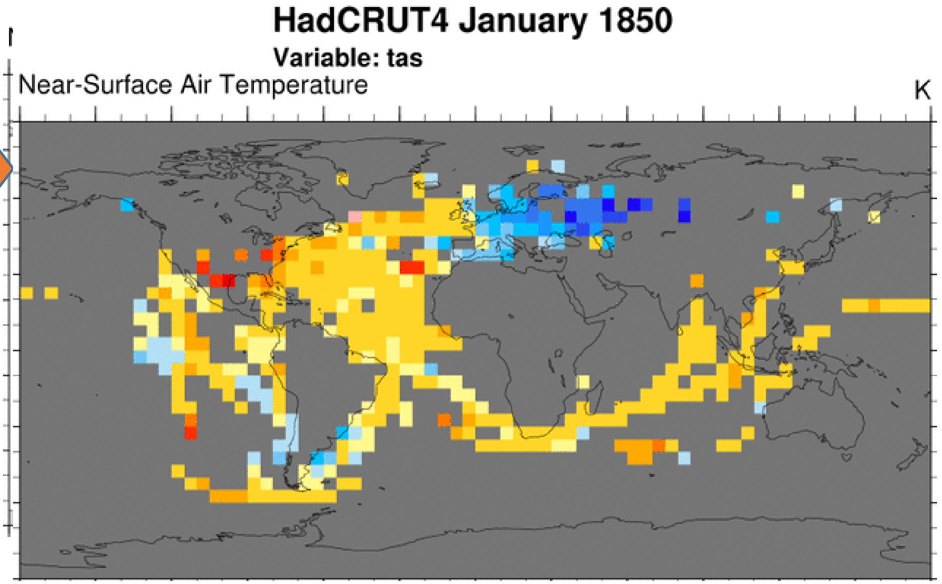
Transfer Learning

Youtube comment:

Google: Let's make AI that teaches itself to walk.
Facebook: Let's make AI that develop their own language.
Nvidia: Let's make Healing brush tool from Photoshop...

Climate Science? BUT HOW?

Observations Annual Mean HadCRUT4
Station, Ship, Boye based data



GRAY -> Missing Values

Transfer Learning

Observations – HadCRUT4	**Observations - What is this?** perception and **recording** of data via the use of scientific **instruments**	**HadCRUT4 - What is this?** It contains newly digitised measurement data, both over **land and sea**, new sea-surface temperature bias adjustments and a more comprehensive error model for describing uncertainties in sea-surface **temperature measurements**	**How does it differ from 20CR and CMIP?** 20cr is filled with observations like HadISST which is derived from HadSST, which is also part of HadCRUT4. With CMIP historical experiment, HadCRUT4 has just the **climate trend** in common.
Reanalysis – 20th Century Reanalysis	**Reanalysis - What is this?** Data products that rely on both **observations and models** to estimate conditions using a single consistent assimilation scheme throughout	**20CR - What is this?** Reanalysis from NOAA **covering the 20th century** using data simulation and observation, it's a four-dimensional global atmospheric dataset of weather spanning 1836 to 2015 (using an ensemble filter)	**How does it differ from HadCRUT4 and CMIP?** Not purely based on observations (HadCRUT4), It's not model output (CMIP) either.
Climate Models – Historical CMIP5	**CMIP5 and ESMs - What is this?** Intercomparison Project between different climate models (Phase 5). ESMs = **Earth System Models** Include the atmosphere, ocean, land, ice and, particularly, the biosphere in an interactive way.	**Historical - What is this?** ESM simulations, which covers the time from 1850 to 2000 - 2015. Initialized with pre-industrial (1850) conditions. Should **simulate climate change** due to some CO2-input/parametrization (otherwise it's called pi-Control)	**How does it differ from HadCRUT4 and 20CR?** Hist vs. 20CR: historical runs are not pushed towards observations Hist vs. HadCRUT4: model data covers the whole world/grid for the whole time period & more variables then SST → no data gaps

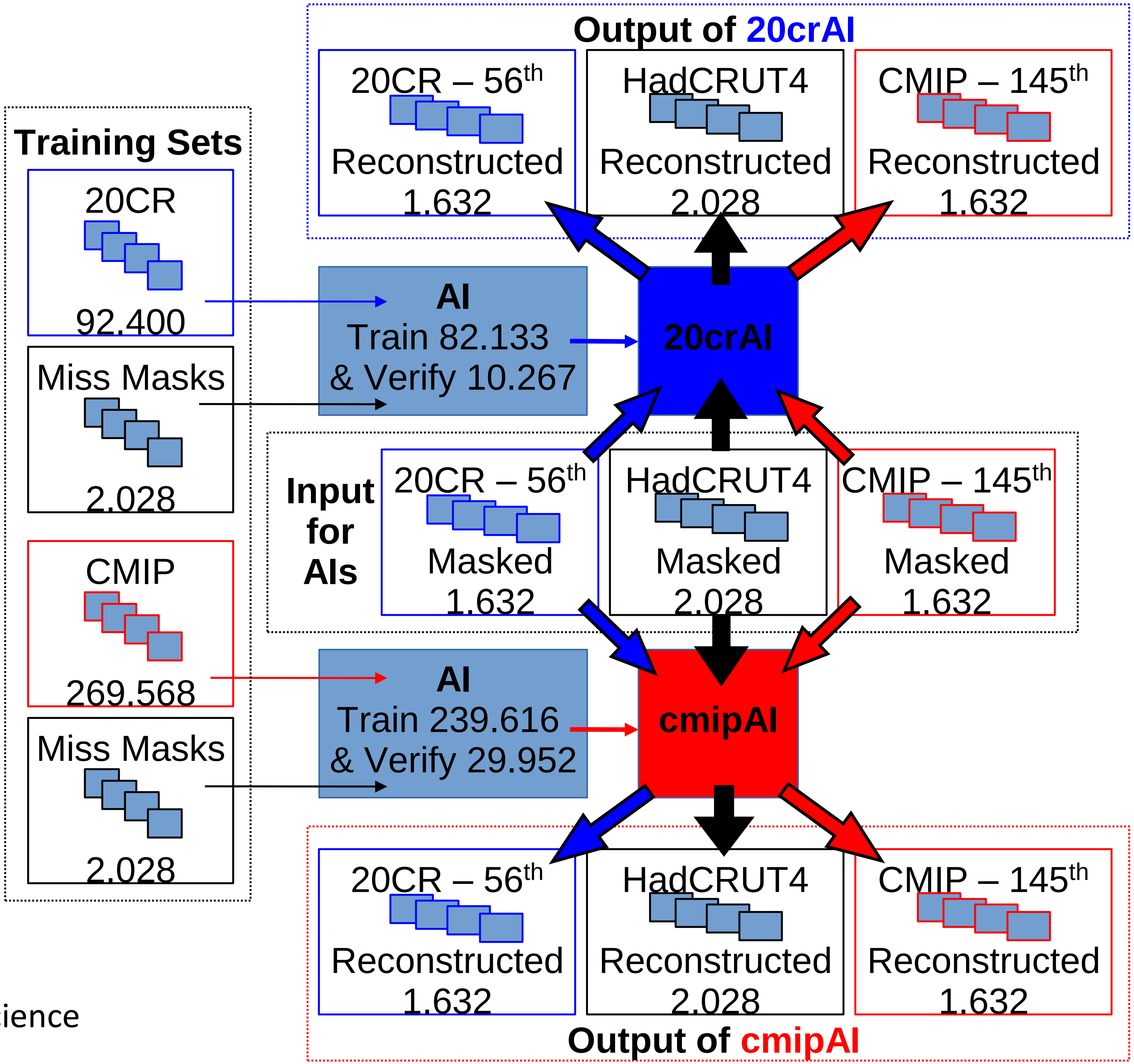
Transfer Learning

20CR Reanalysis
1870-2009
1 Model (Atmos.)
55+1 Ens. Member

HadCRUT4
1850-2018

CMIP5 Historical
1850-2005
35 Models (ESMs)
144+1 Ens. Member

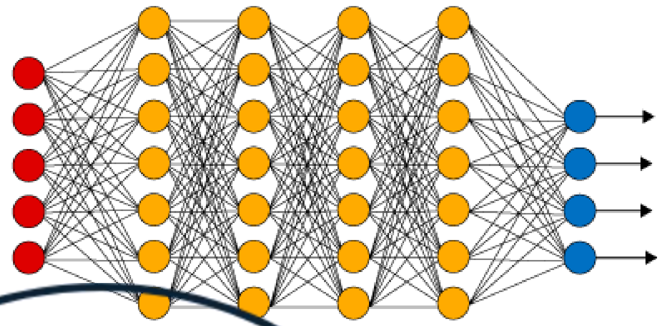
HadCRUT4
1850-2018



Setup of Machine Learning sets:
Training, Validation, Test
IMPORTANT: How to select each? Science and Evaluation! Don‘t cheat yourself.

ML Techniques

Deep Learning Neural Network



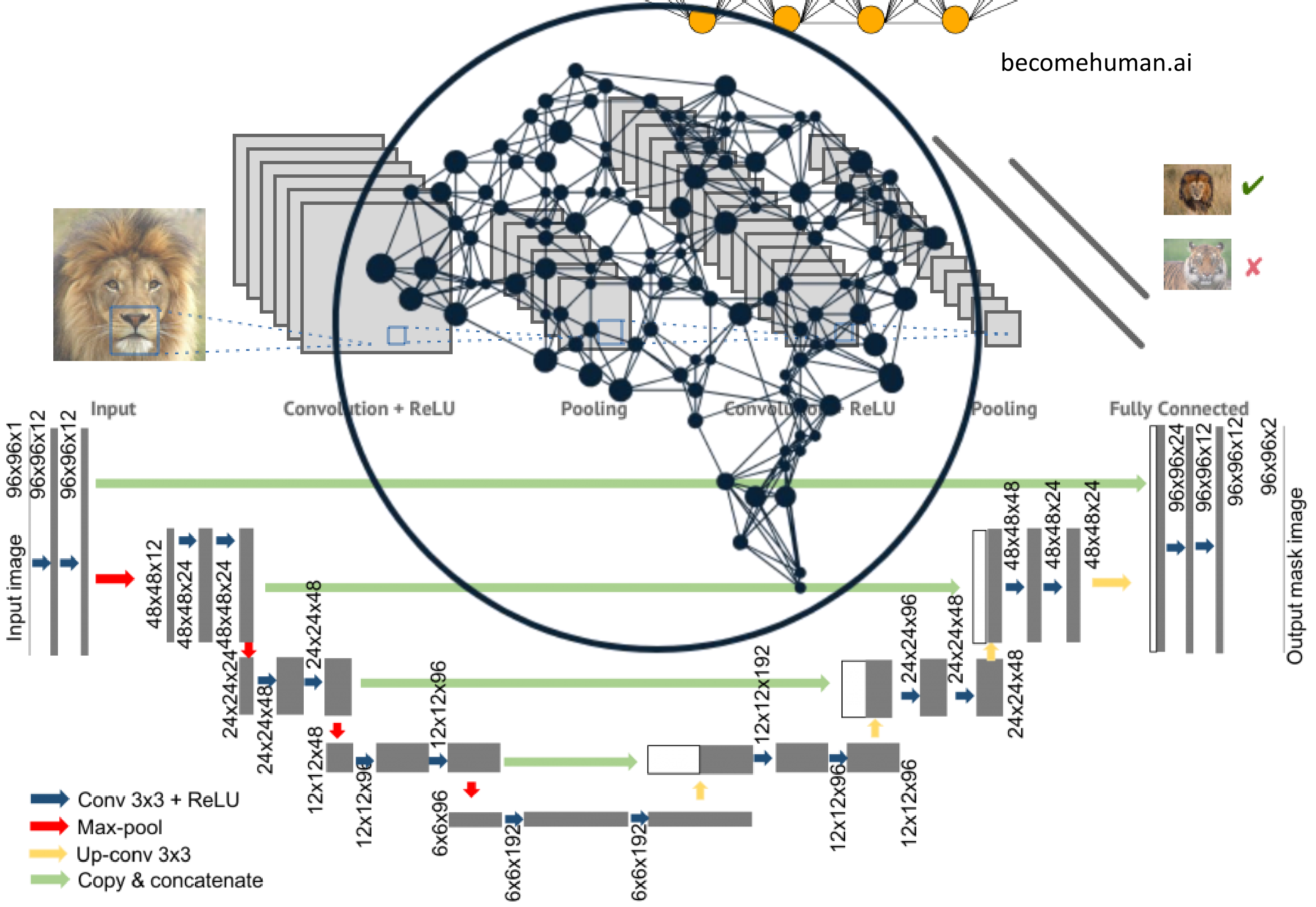
becomehuman.ai

Neural Networks

PyTorch,
Partial Convolution,
cuDNN CUDA-GPU accelerated

(Partial) Convolution

U-Net Architecture



Pre-Research

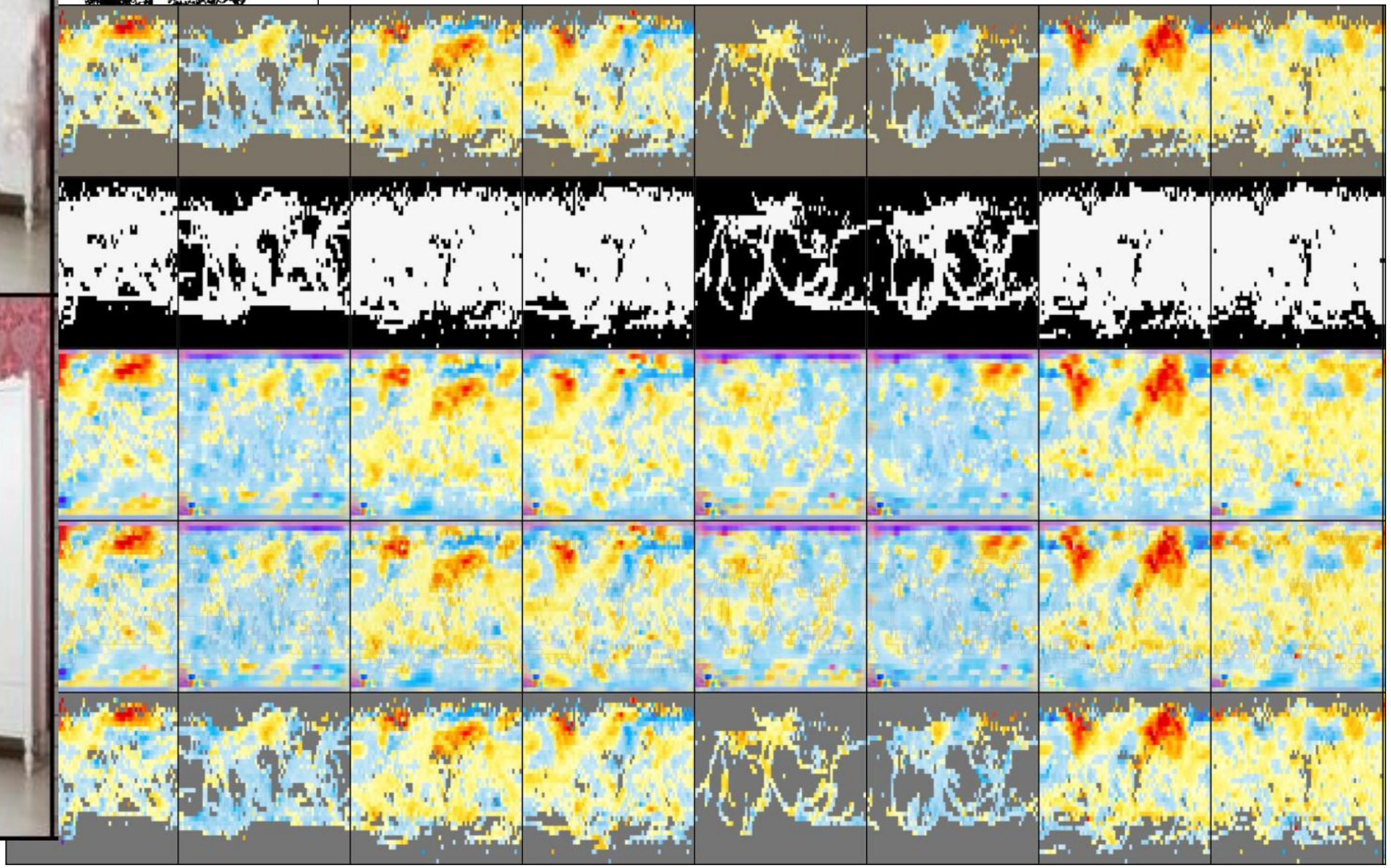
Find computer with GPUs
~1 day

Install Software, Download Pics
~3 days

Run Software in its original mode
~2 days

Produce JPGs from Climate Data and run
~3 days

Change AI software to read climate data (NetCDF) and run and tune
~12 days



Pre-Research

DKRZ

Shown is the learning process, each square and step shows **50 iterations** in the neural network, to create a related (climate) pattern.
From
Kadow et al 2020

Entdecken

Einstellungen

Nature Geoscience
8.353 Tweets

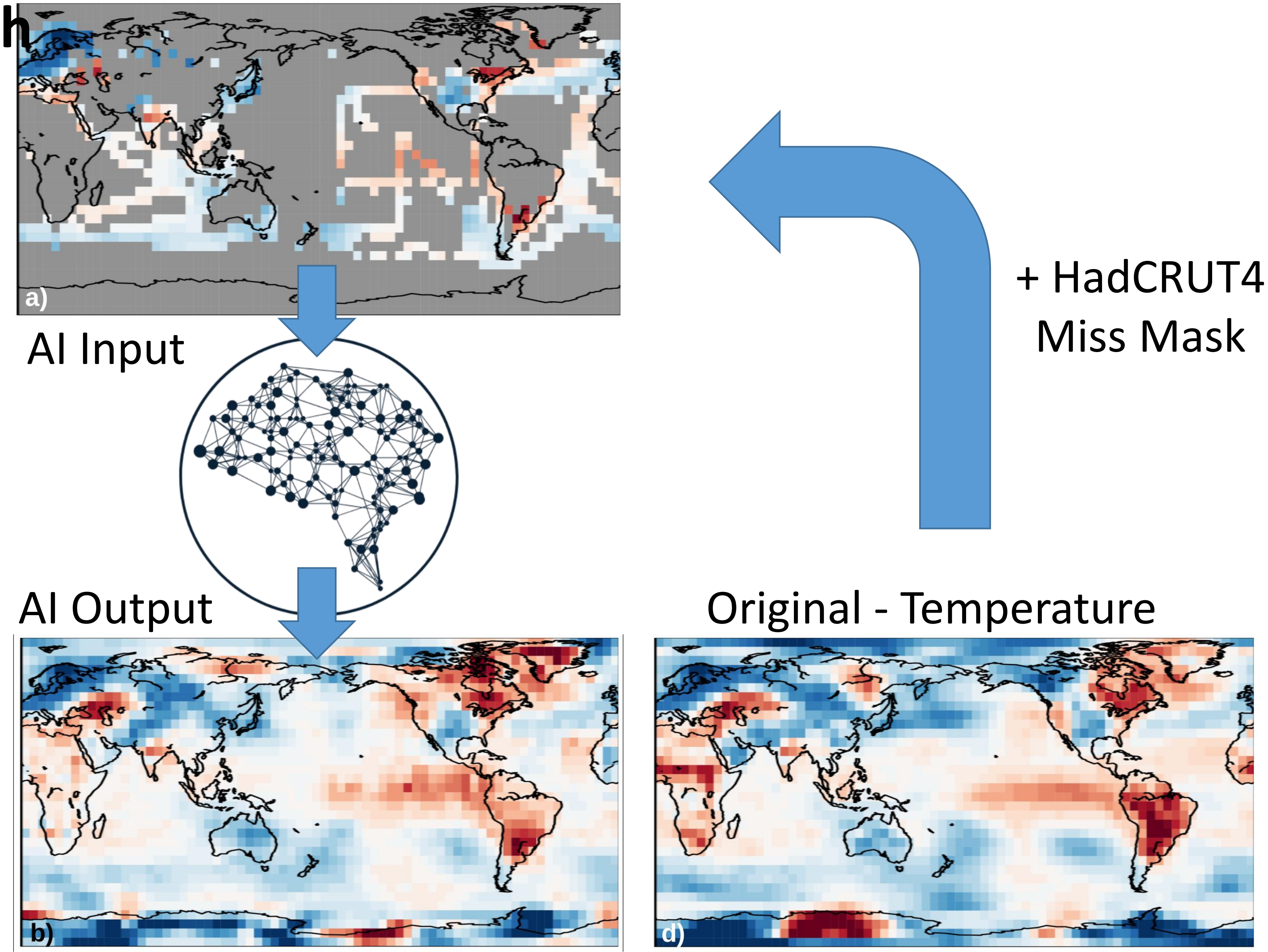
Folgen

Nature Geoscience @NatureGeosci · 1. Juni
NGeo: Artificial intelligence can reconstruct missing historical temperature data

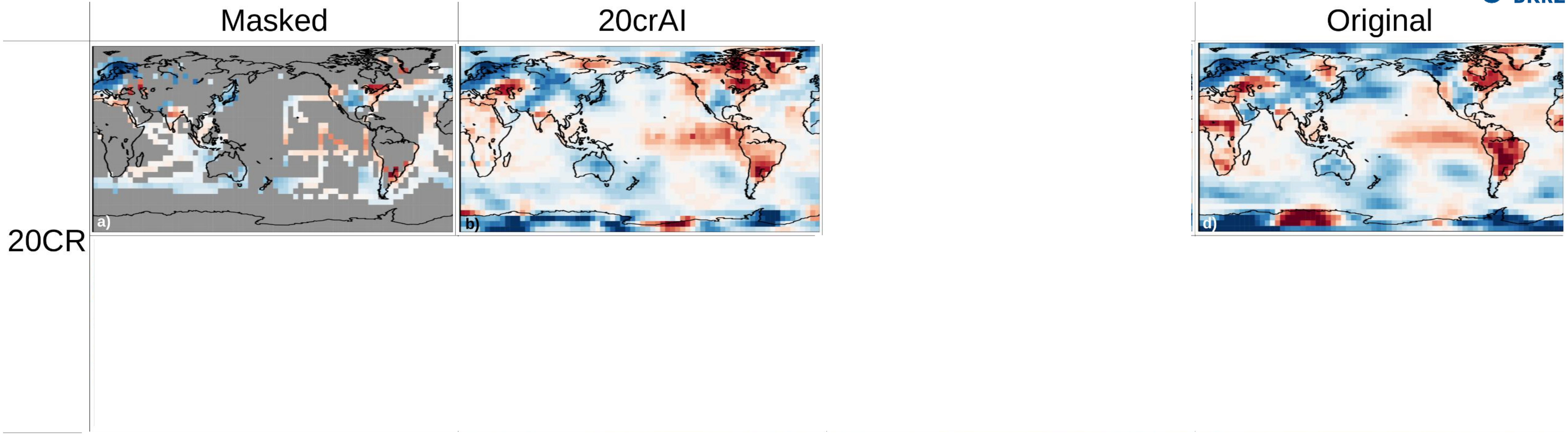
nature.com/articles/s4156...

2 3

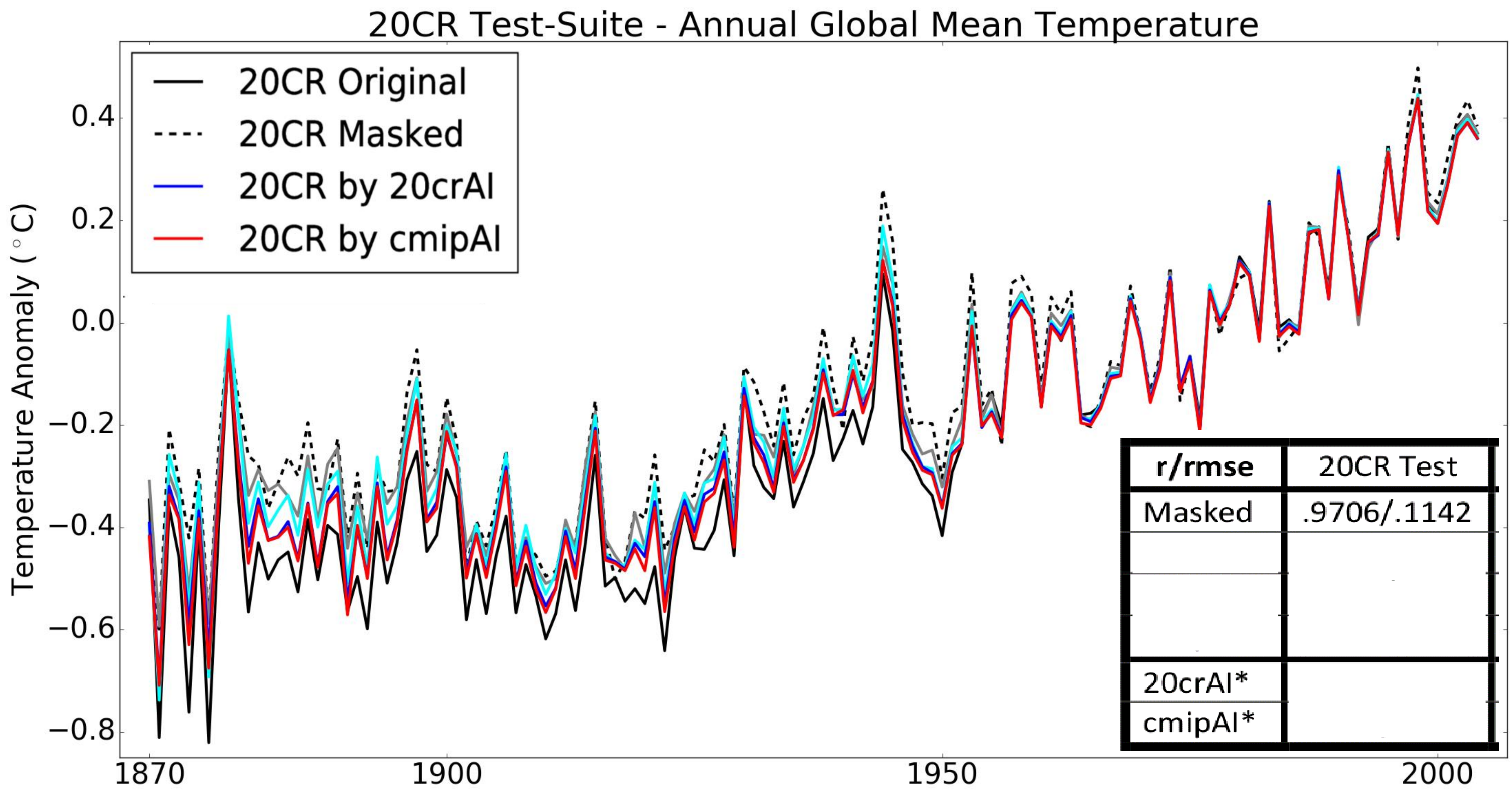
Research


a)
AI Input
+ HadCRUT4
Miss Mask
AI Output
Original - Temperature
b)
d)

Research



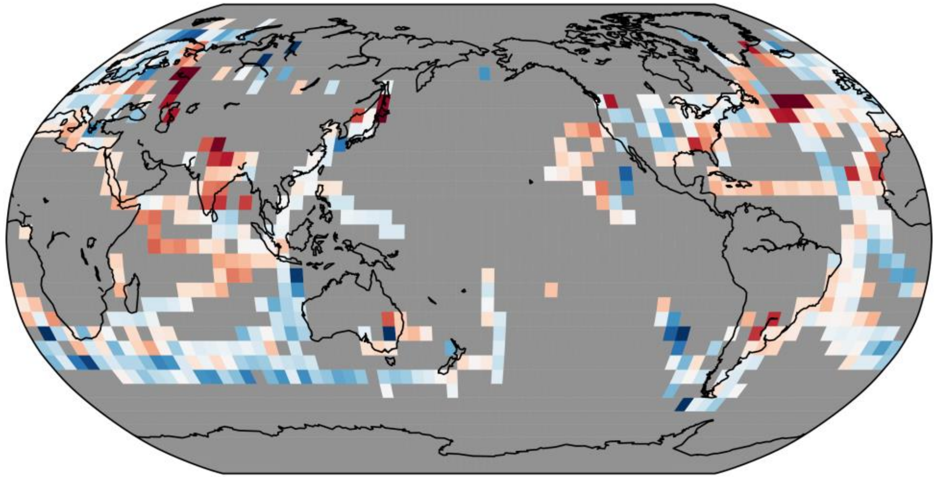
Research


20CR Test-Suite - Annual Global Mean Temperature
20CR Original
20CR Masked
20CR by 20crAI
20CR by cmipAI
Temperature Anomaly (°C)
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
1870
1900
1950
2000
r/rmse
20CR Test
Masked
.9706/.1142
20crAI*
cmipAI*

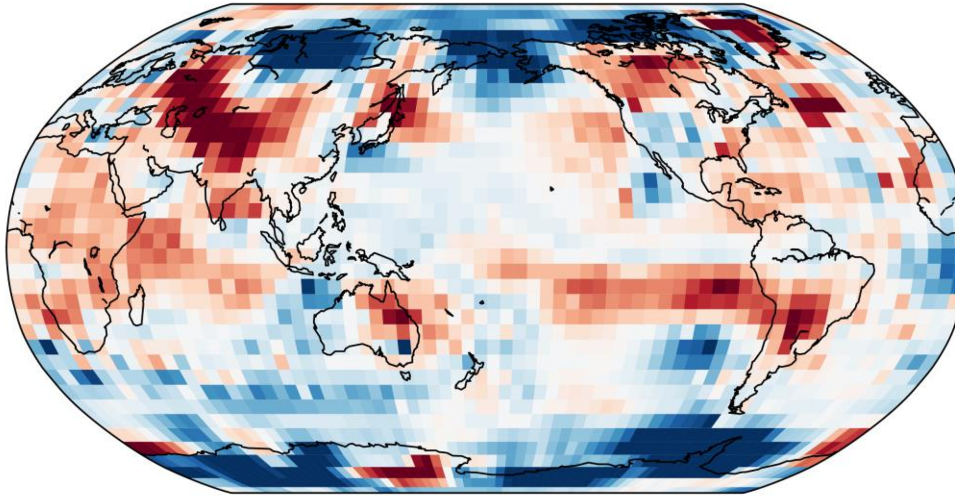
Temporal and Spatial – HadCRUT4

El Nino
July
1877

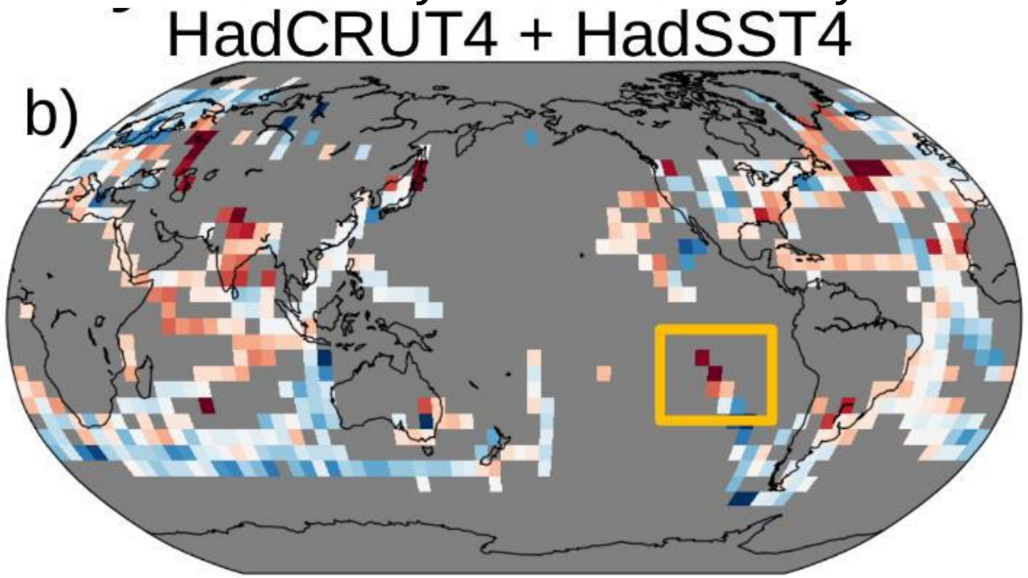
HadCRUT4 Original

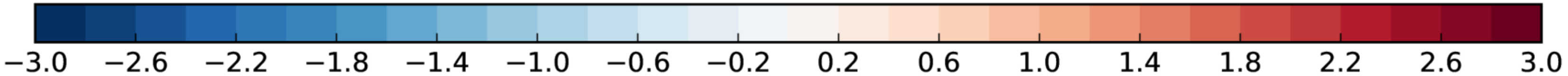


HadCRUT4 by 20crAI

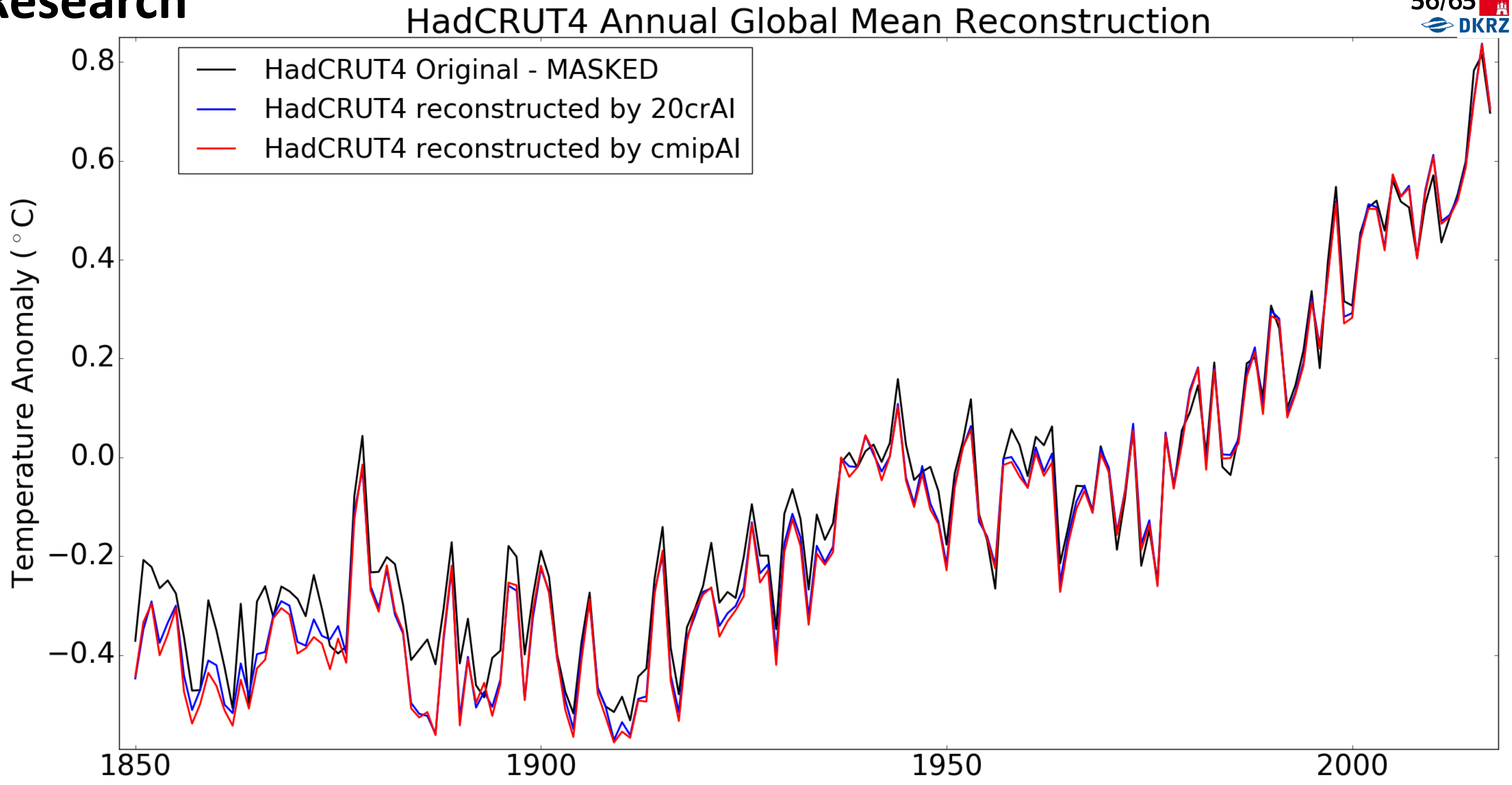


HadCRUT4 by Cowtan and Way

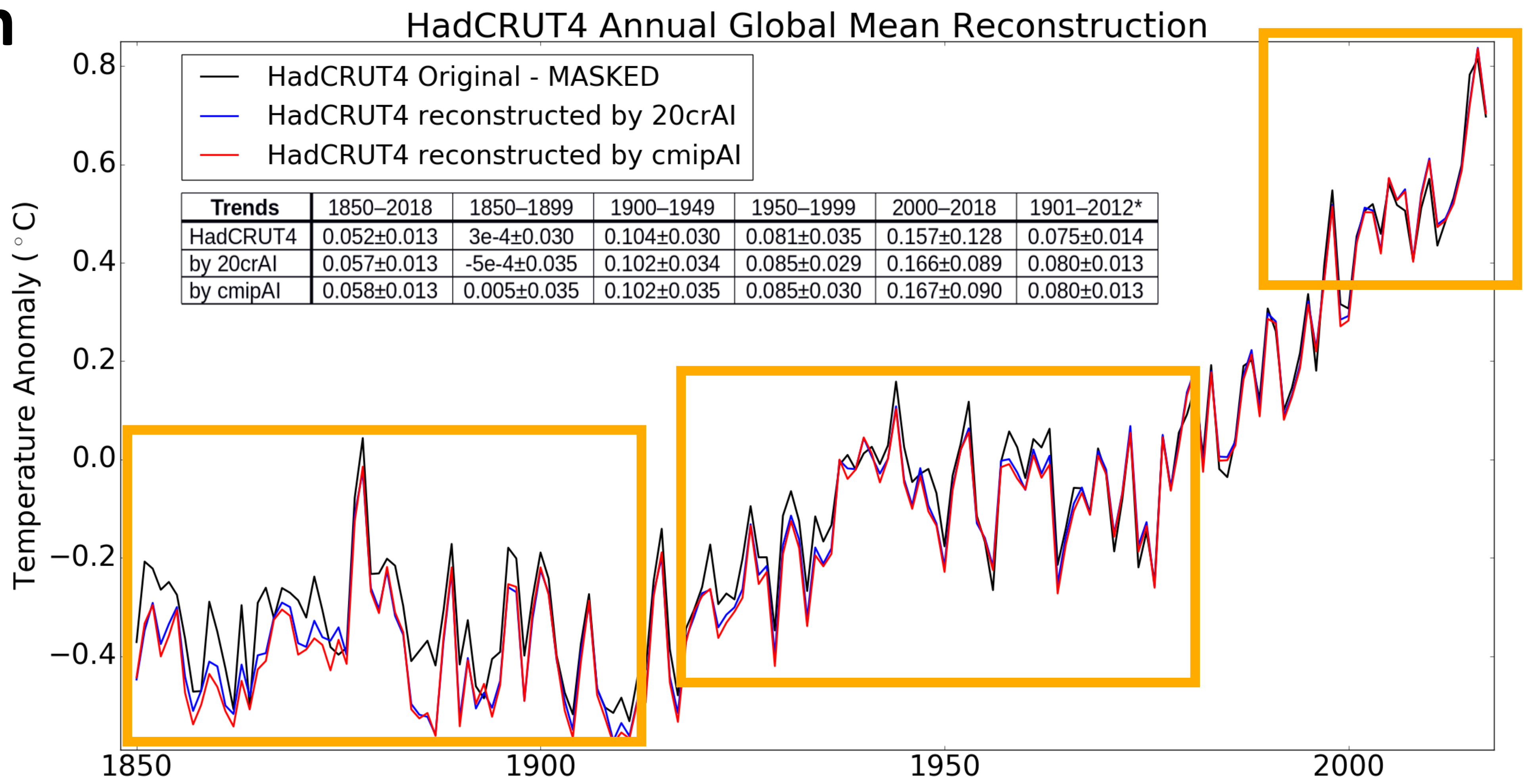




Research


HadCRUT4 Annual Global Mean Reconstruction
HadCRUT4 Original - MASKED
HadCRUT4 reconstructed by 20crAI
HadCRUT4 reconstructed by cmipAI
Temperature Anomaly (°C)
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
1850
1900
1950
2000

Research



Trends	1850–2018	1850–1899	1900–1949	1950–1999	2000–2018	1901–2012*
HadCRUT4	0.052±0.013	3e-4±0.030	0.104±0.030	0.081±0.035	0.157±0.128	0.075±0.014
by 20crAI	0.057±0.013	-5e-4±0.035	0.102±0.034	0.085±0.029	0.166±0.089	0.080±0.013
by cmipAI	0.058±0.013	0.005±0.035	0.102±0.035	0.085±0.030	0.167±0.090	0.080±0.013

- **cooler** global mean temperature from the **mid 19th to the early 20th century**
  - an **underestimation of** the global mean temperature **trend** between 1850 and 2018.
- The results of the AI reconstructions support other studies by also showing a **cooler** period in **the mid of the 20th century**.
- **Early 21th century**: Compared to HadCRUT4, both AIs agree on a **weaker hiatus** phase and a **stronger trend** including higher values for **2016**, the warmest year on record.

AI Reconstructs Photos with Realistic Results

ARTICLES
https://doi.org/10.1038/s41561-020-0582-5

nature geoscience

Check for updates

Artificial intelligence reconstructs missing climate information

Christopher Kadow[1,2], David Matthew Hall[3] and Uwe Ulbrich[2]

Historical temperature measurements are the basis of global climate datasets like HadCRUT4. This dataset contains many missing values, particularly for periods before the mid-twentieth century, although recent years are also incomplete. Here we demonstrate that artificial intelligence can skilfully fill these observational gaps when combined with numerical climate model data. We show that recently developed image inpainting techniques perform accurate monthly reconstructions via transfer learning using either 20CR (Twentieth-Century Reanalysis) or the CMIP5 (Coupled Model Intercomparison Project Phase 5) experiments. The resulting global annual mean temperature time series exhibit high Pearson correlation coefficients (≥0.9941) and low root mean squared errors (≤0.0547 °C) as compared with the original data. These techniques also provide advantages relative to state-of-the-art kriging interpolation and principal component analysis-based infilling. When applied to HadCRUT4, our method restores a missing spatial pattern of the documented El Niño from July 1877. With respect to the global mean tem-

DKRZ
DEUTSCHES KLIMARECHENZENTRUM

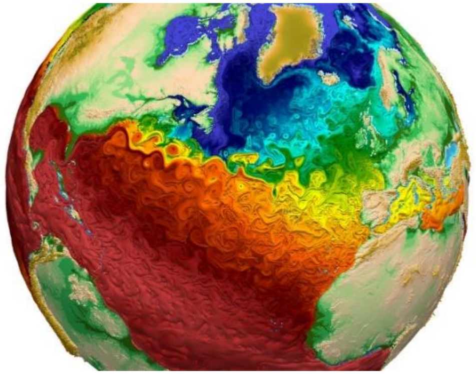
NVIDIA

Freie Universität Berlin

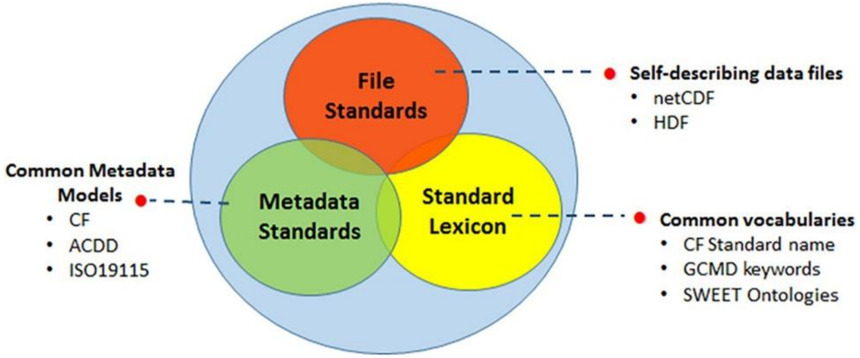
Research

AI reconstructs missing Climate Information

HPC Modeling



HPC Processing

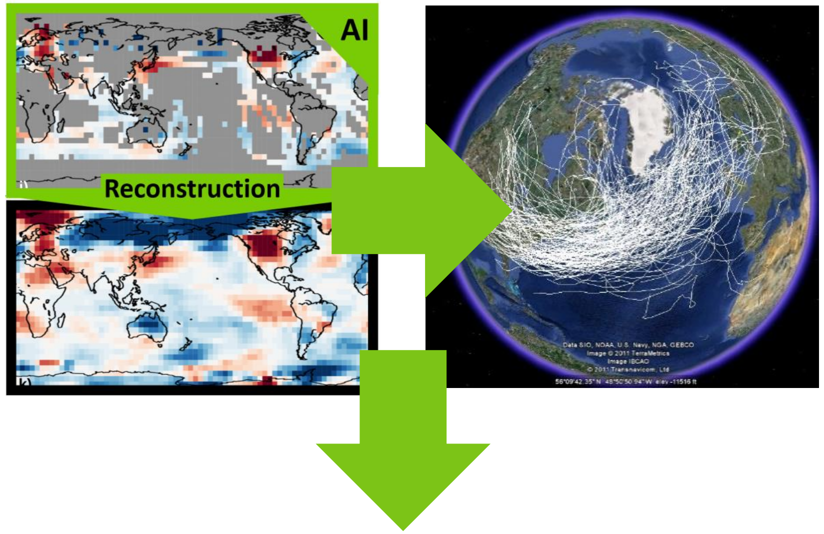


HPC AI/ML

- **Successful combination** of climate modeling, observation and artificial intelligence -> on and thanks to HPCs
- At the moment, many groups (try to) improve models with AI, here it is **vice versa**
- Missing values introduce structual biases, which can be **reduced by AI**
- Other studies are confirmed (trends, hiatus, etc.), but this study shows an **added value** in terms of temperol (global mean) and spatial structures (e.g. El Nino 1877).
- Data and technology will be continously **prepared for the community (on GitHub)**

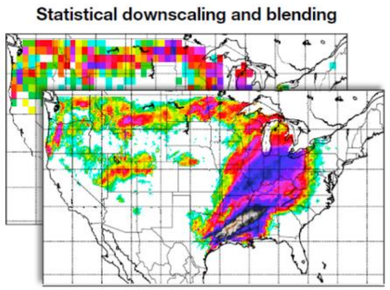
What is next?

Helmholtz-Zentrum Geesthacht
Zentrum für Material- und Küstenforschung

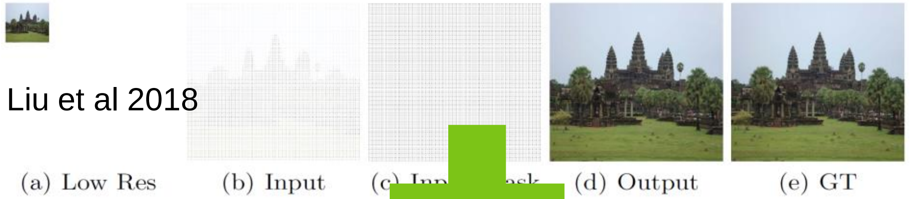


Other Variable
Other Frequency
?

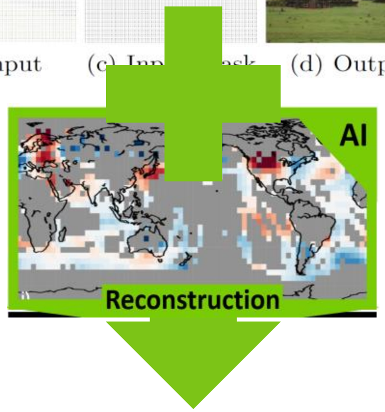
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG



Reichstein et al 2019

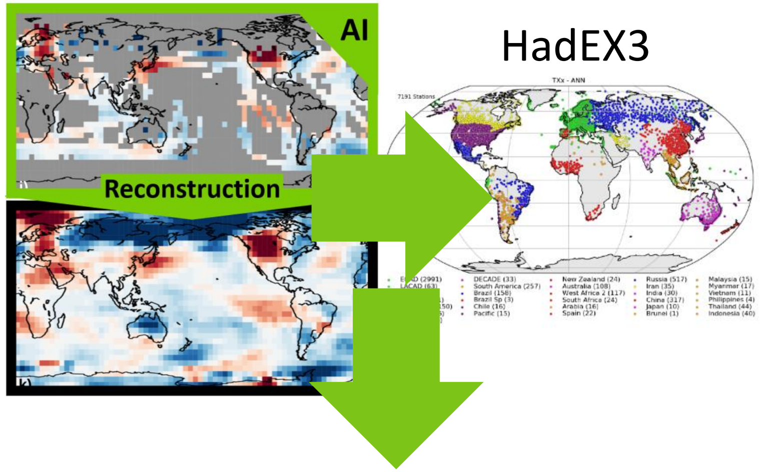


Liu et al 2018



Downscaling ?

BSC



Extremes? ?

What is next?

DeOldify

neural
.love

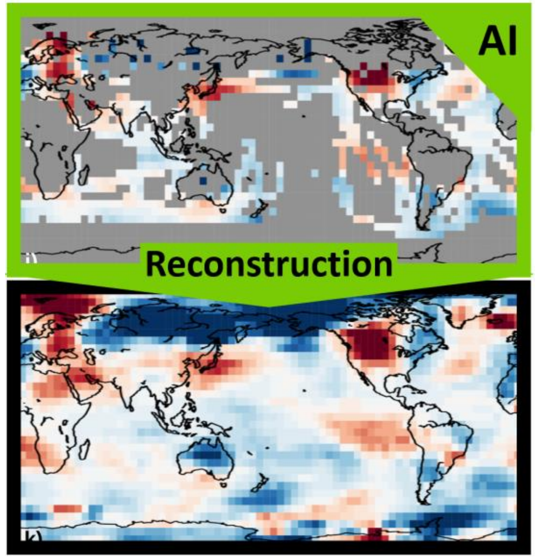
What is next?

DeOldify

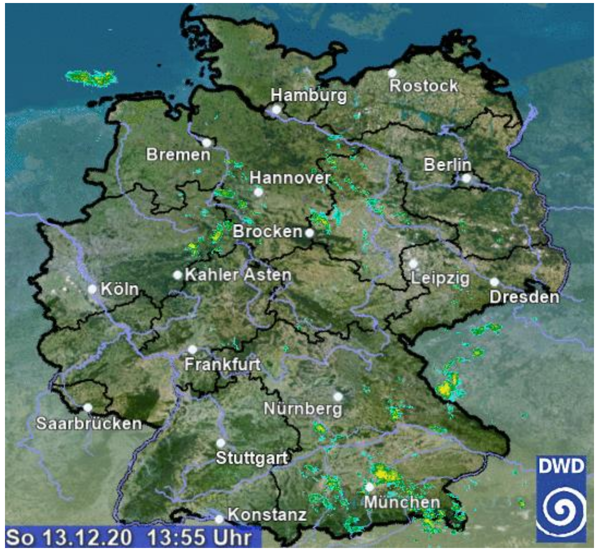
neural
.love

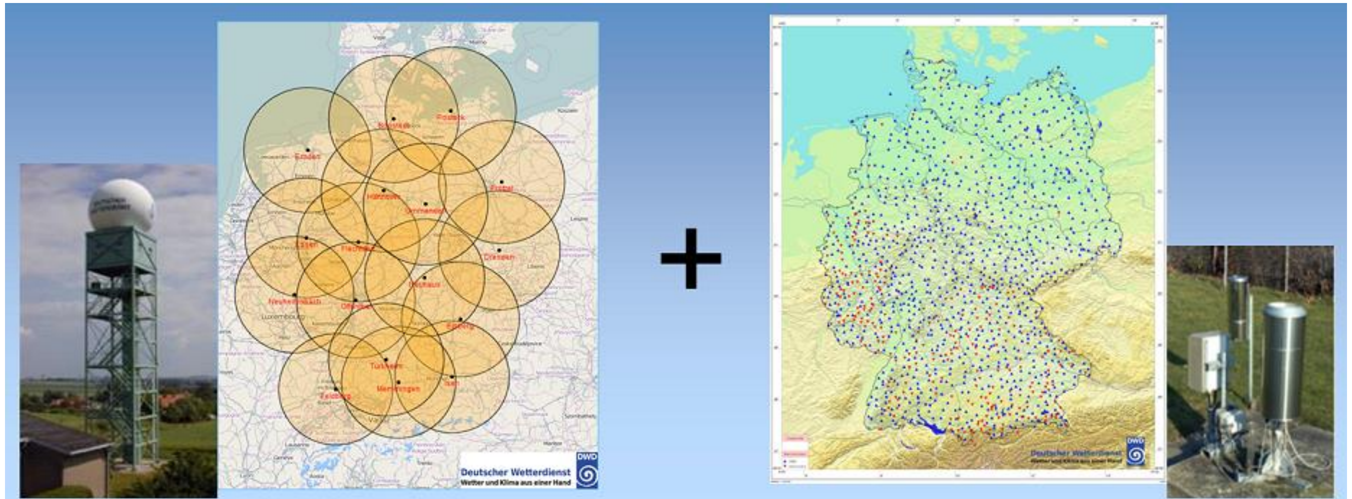
Master Thesis at DKRZ!?

Focus on Precipitation



- Is precipitation possible to reconstruct?
  - Same training method?
- From climate to weather and back:
  - Can we reconstruct radar data from station data?
  - Back in time where no radar existed?



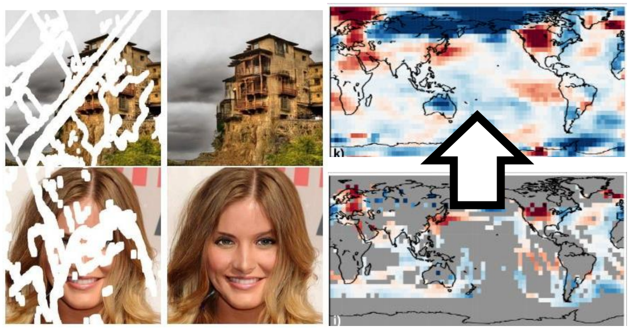


In cooperation with:

GERICS Climate Service Center Germany

Eine Einrichtung des Helmholtz-Zentrums Geesthacht

Conclusion



Past – Observation Reconstruction

Combination of climate modeling, observation and artificial intelligence. Successful re-fill of climate information. Interpolation plus pattern recognition is a strong tool for climate research.

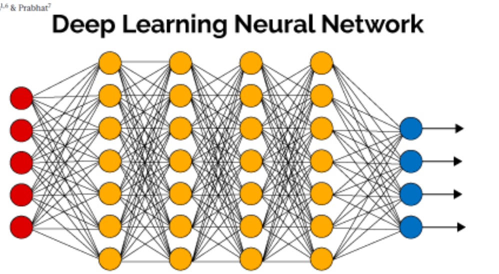
Just some thoughts

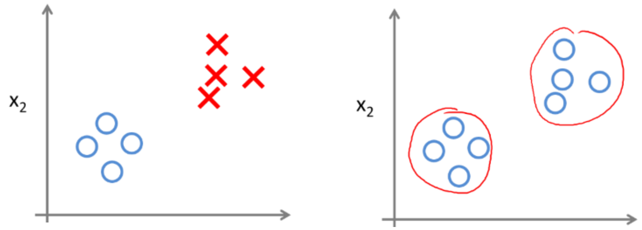
- Battle of the image inpainting community: do they care about a picture?
    - Can we put a climate benchmark set outthere?
- Speaking the same language: *difference of an analysis and a reanalysis?*
- The world on a square: pre-processing not optimal, convolutions on boundaries
- Not one code optimization: AI technology has a lot of potential left!? Hopefully this gets beaten soon.
- Transfer learning needs science: *e.g. you cannot train on missing values*
- What is happening next? Go on higher scales? Other variables?

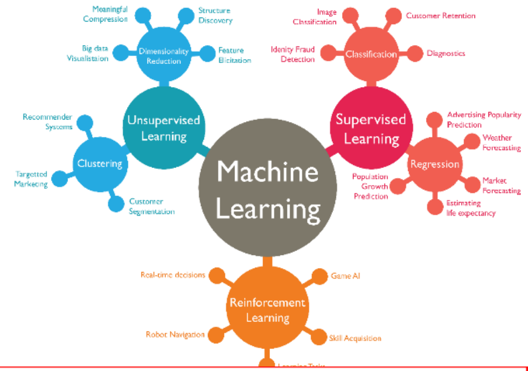
Summary

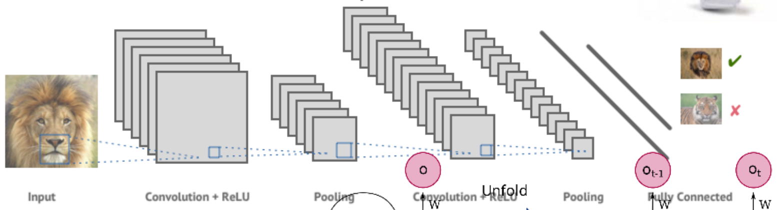
- General
  - DL, ML, AI? WTF? Literature
  - What is a Neural Network?
- Methods & Networks
  - Supervised Learning
  - Unsupervised Learning
  - Reinforcement Learning
  - Convolutional Neural Network
  - Recurrent Neural Network (!)
  - Generative Adversal Network (!)
- Hardware & Software
  - PCs, HPCs, Clouds
  - Tools, Frameworks, First Steps
- AI reconstructs missing Climate

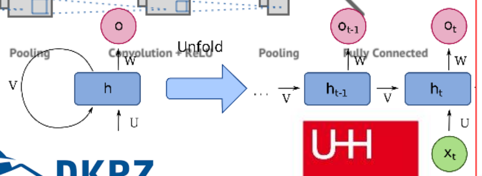
Master Thesis with/at DKRZ

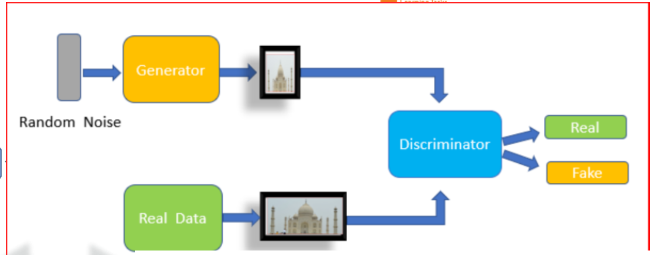


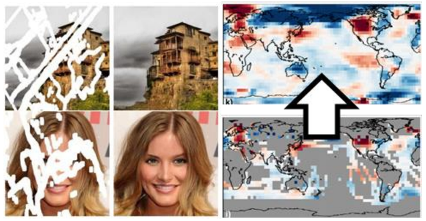


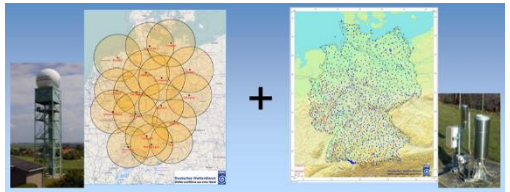












What is next?

Research

https://github.com/FREVA-CLINT/climatereconstructionAI

README.md

Climate Reconstru
Partial Convolutic

[Applied implementation] (https://

Official implementation is released

Note that this is an ongoing re-imp
input and output!

This is an unofficial pytorch impleme
Convolutions [Liu+, arXiv2018].

Requirements

- Python 3.6+
- Pytorch 0.4.1+

```
pip install -r requirements.txt
```

Search or jump to... | Pull requests | Issues | Marketplace | Explore

FREVA-CLINT / climatereconstructionAI

Watch 3 | Star 12 | Fork 7

Code | Issues | Pull requests | Actions | Projects | Wiki | Security 1 | Insights | Settings

We found potential security vulnerabilities in your dependencies.
You can see this message because you have been granted access to Dependabot alerts for this repository.

See Dependabot alerts

master | 1 branch | 2 tags | Go to file | Add file | Code

Christopher Kadow monthly ful grid reconstructions now 1850-2018 | 1ebef0c on 2 Jun | 49 commits

h5/script	First commit with all necessary software and data to reconstruct clim...	8 months ago
masks	First commit with all necessary software and data to reconstruct clim...	8 months ago
reconstructions	monthly ful grid reconstructions now 1850-2018	6 months ago
snapshots	First commit with all necessary software and data to reconstruct clim...	8 months ago

About

Software to train climate reconstruction technology (image inpaiting with partial convolutions) with numerical model output to re-fill missing values in observational datasets like HadCRUT4

Readme

View license

Releases 2

Technical Fact Sheet

- Research applied at Freie University Berlin HPC (ML) and DKRZ infrastructure (Data Handling)
- AI models were trained using 500.000 iterations with an additional 500.000 iterations for fine tuning.
- Applying a **batch size of 18** on a **NVIDIA Geforce 1080Ti** at approximately 17its / sec.
- On 1 Node -> 2 GPU cards with 3.584 cores per card