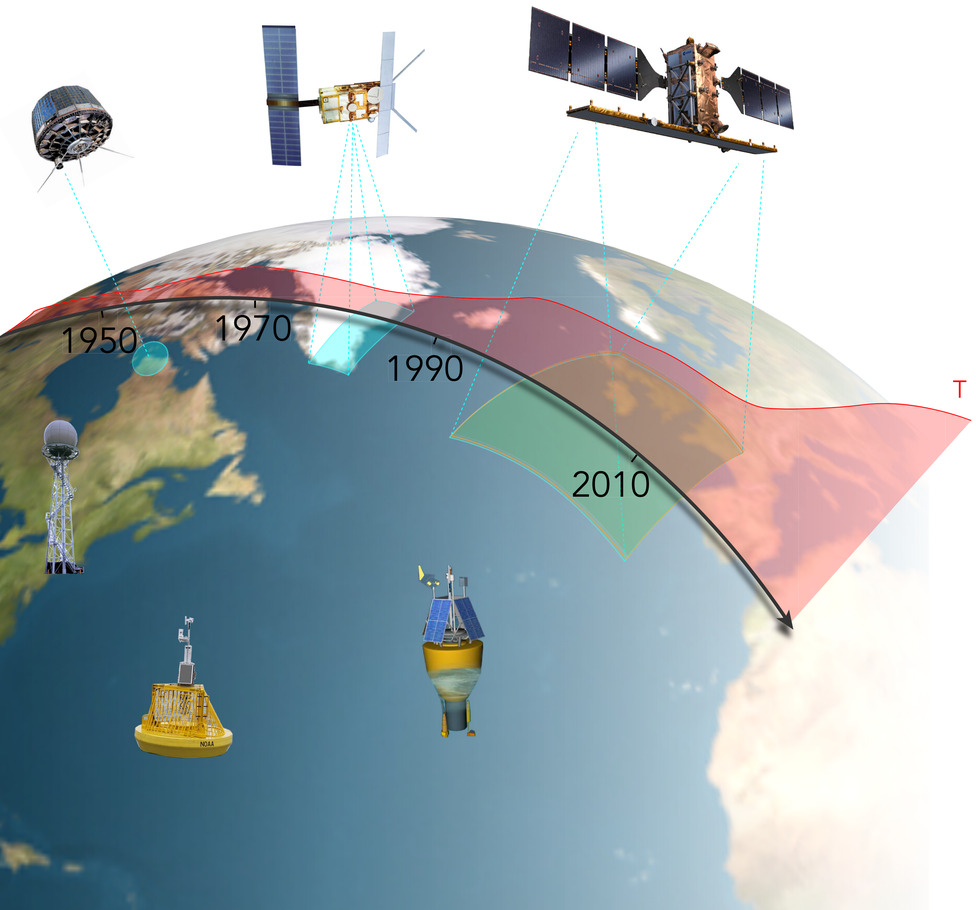
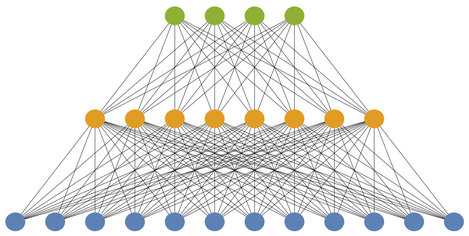
Representation Learning for Science

Christian Lessig, Otto-von-Guericke-Universität Magdeburg





large scale machine learning

scientific insight

Large scale representation learning

- Think of large language models: ChatGPT, GPT-X, PaLM, ...
  - Useful for scientists (reviews, grant applications, thesis appraisals, ...)

© Christian Lessig, 2023

Large scale representation learning

- Think of large language models: ChatGPT, GPT-X, PaLM, ...
  - Useful for scientists (reviews, grant applications, thesis appraisals, ...)
  - But not directly useful for science

© Christian Lessig, 2023

Large scale representation learning

- Learn a **domain-specific** but **task-independent** neural network that is useful for a range of applications

© Christian Lessig, 2023

Large scale representation learning

- Learn a **domain-specific** but **task-independent** neural network that is useful for a range of applications
  - Representation network provides transformation of network input to effective feature spaces
  - Self-supervised training on very large amounts of data with very large networks
  - Useful for downstream applications using tail network, fine-tuning, or in-context learning

© Christian Lessig, 2023

Large scale representation learning

- Learn a **domain-specific** but **task-independent** neural network that is useful for a range of applications
  - Representation network provides transformation of network input to effective feature spaces
  - **Self-supervised training** on **very large amounts of data** with **very large networks**
  - Useful for downstream applications using tail network, fine-tuning, or in-context learning

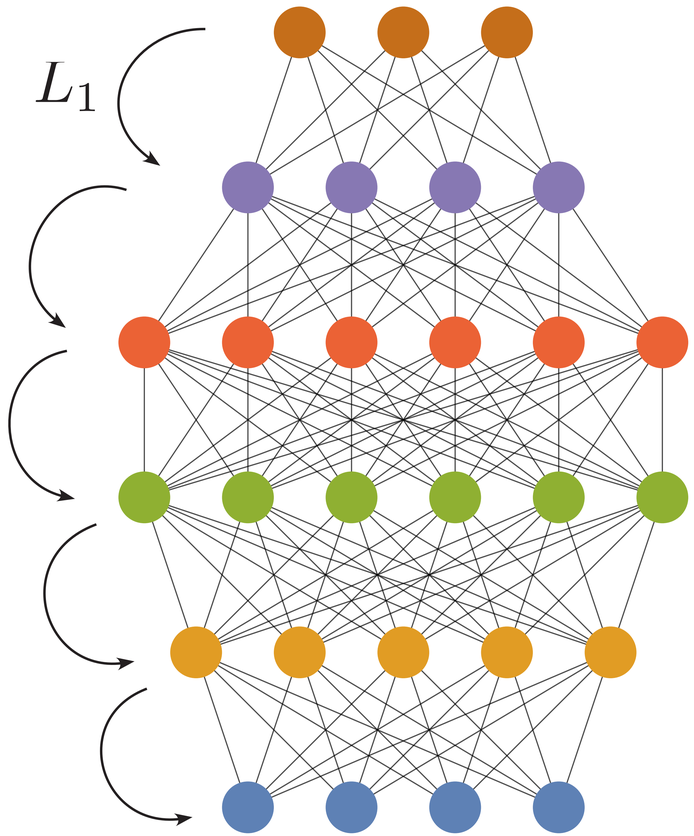
© Christian Lessig, 2023

Large scale representation learning

- Learn a **domain-specific** but **task-independent** neural network that is useful for a range of applications
  - Representation network provides transformation of network input to effective feature spaces
  - Self-supervised training on very large amounts of data with very large networks
  - Useful for downstream applications using tail network, fine-tuning, or **in-context learning**

© Christian Lessig, 2023

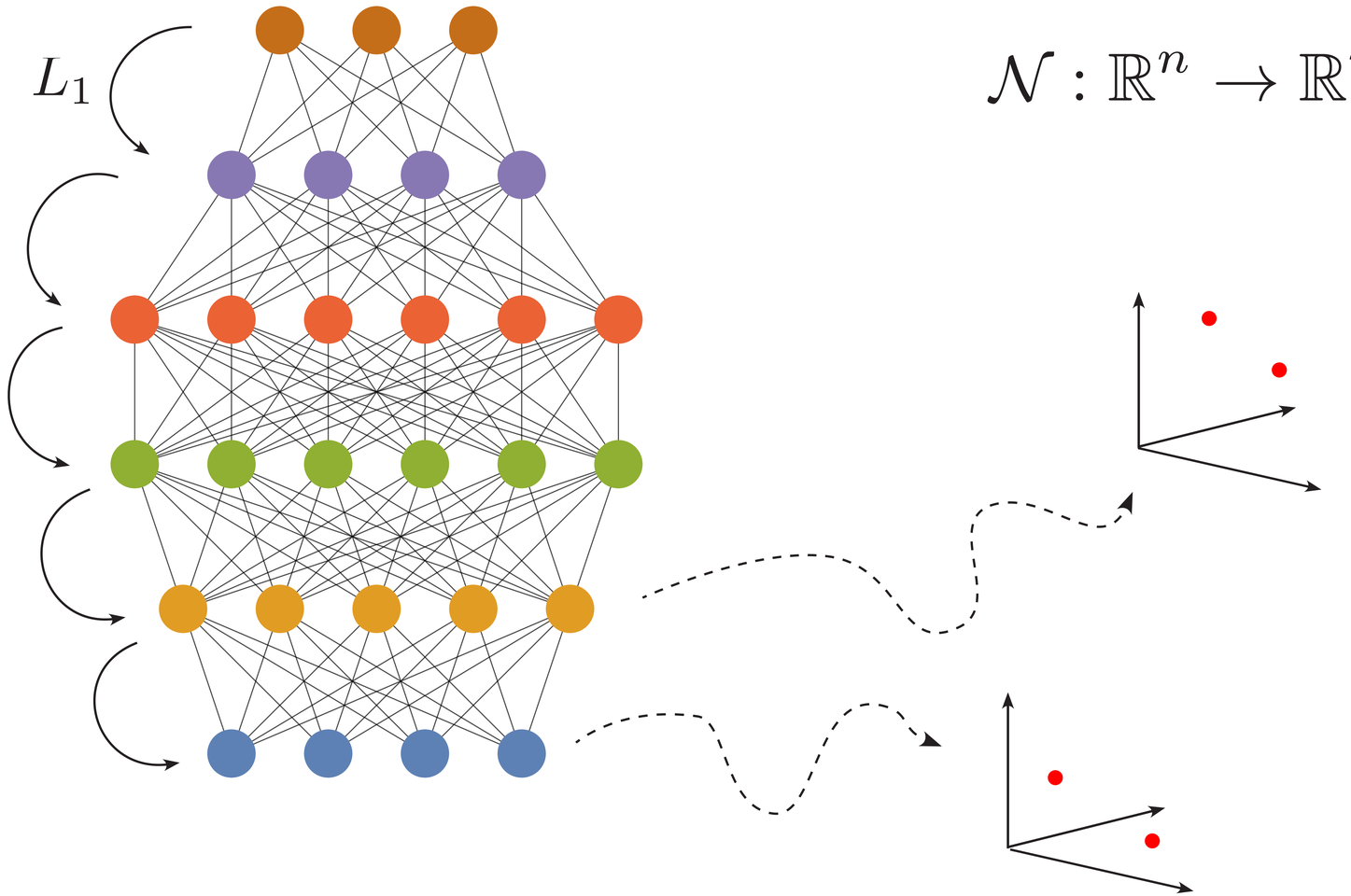
Large scale representation learning



$$\mathcal{N} : \mathbb{R}^n \to \mathbb{R}^m$$

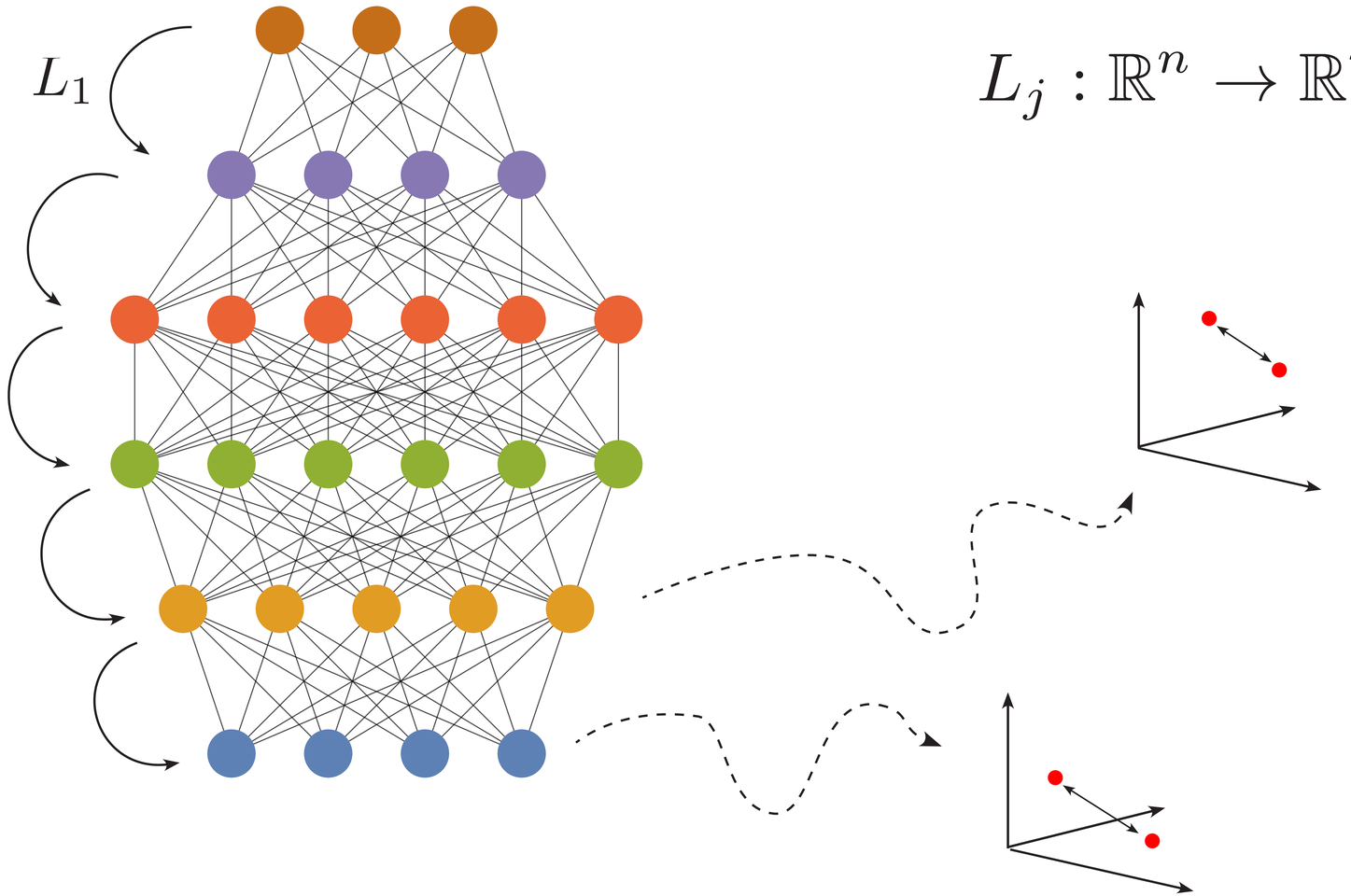
© Christian Lessig, 2023

Large scale representation learning



© Christian Lessig, 2023

Large scale representation learning



$$L_j : \mathbb{R}^n \to \mathbb{R}^m$$

Feature spaces deep in the network reveal intrinsic structure of the data

© Christian Lessig, 2023

Can we perform representation learning in and for science?

© Christian Lessig, 2023

Representation learning for science?

- Very large amounts of observational data
  - ERA5 reanalysis: 6+ PB
  - ESA MetOp-SG satellites: 8 x 864 GB/day
  - Data essentially completely unlabelled

© Christian Lessig, 2023

Representation learning for science?

- Very large amounts of observational data
  - ERA5 reanalysis: 6+ PB
  - ESA MetOp-SG satellites: 8 x 864 GB/day
  - Data essentially completely unlabelled

- Very similar situation in high energy physics, astronomy, DNA sequencing, ...

© Christian Lessig, 2023

Representation learning for science?

- Very large amounts of **observational** data
  - ERA5 reanalysis: 6+ PB
  - ESA MetOp-SG satellites: 8 x 864 GB/day
  - Data essentially completely unlabelled

- Very similar situation in high energy physics, astronomy, DNA sequencing, ...

© Christian Lessig, 2023

Representation learning for science?

- Very large amounts of **observational** data
  - ERA5 reanalysis: 6+ PB
  - ESA MetOp-SG satellites: 8 x 864 GB/day
  - Data essentially completely unlabelled

GPT-3: 10^{11} tokens
ERA5: 5^{14} tokens

- Very similar situation in high energy physics, astronomy, DNA sequencing, ...

© Christian Lessig, 2023

Representation learning for science?

- Very large amounts of **observational** data
- No complete classical model for system and dynamics
  - Central issue for forecasting and climate projections

© Christian Lessig, 2023

Representation learning for science?

- Very large amounts of **observational** data
- No complete classical model for system and dynamics
- Supervised training imposes scientific theory
  - Self-supervised training is more “open-minded“ about the relationships and structure in the data

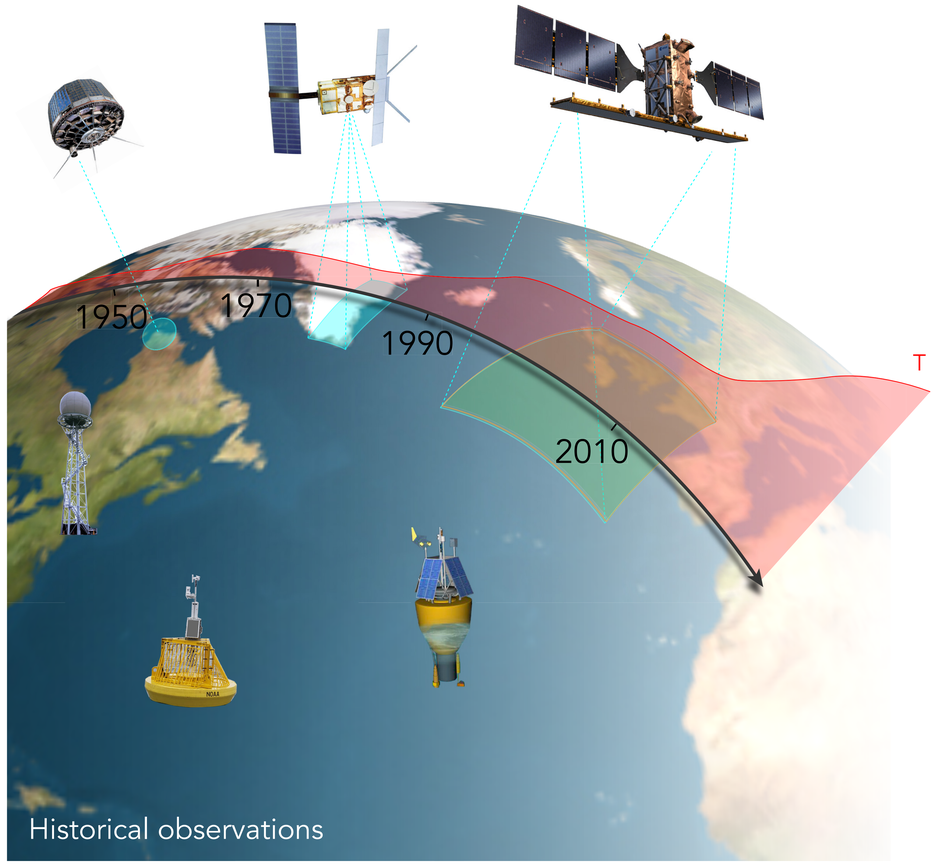
© Christian Lessig, 2023

AtmoRep

Large scale representation learning of atmospheric dynamics

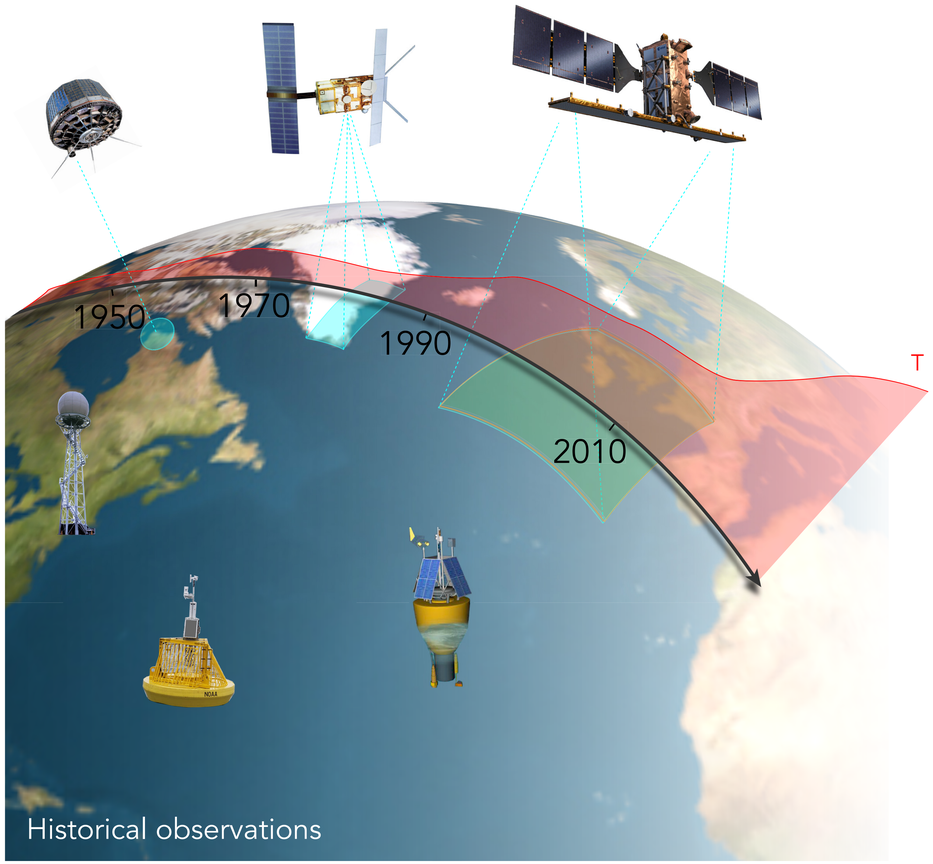
© Christian Lessig, 2023

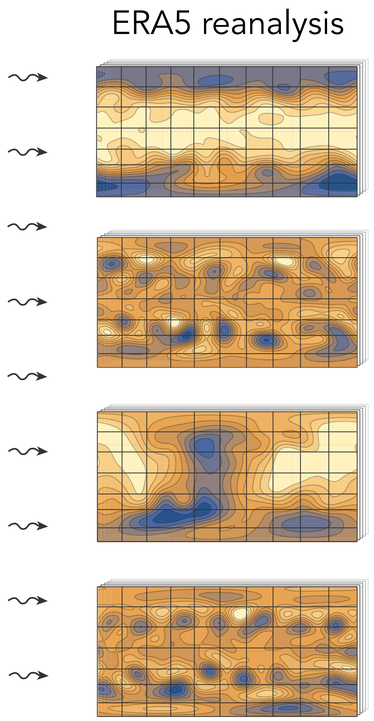
AtmoRep



© Christian Lessig, 2023

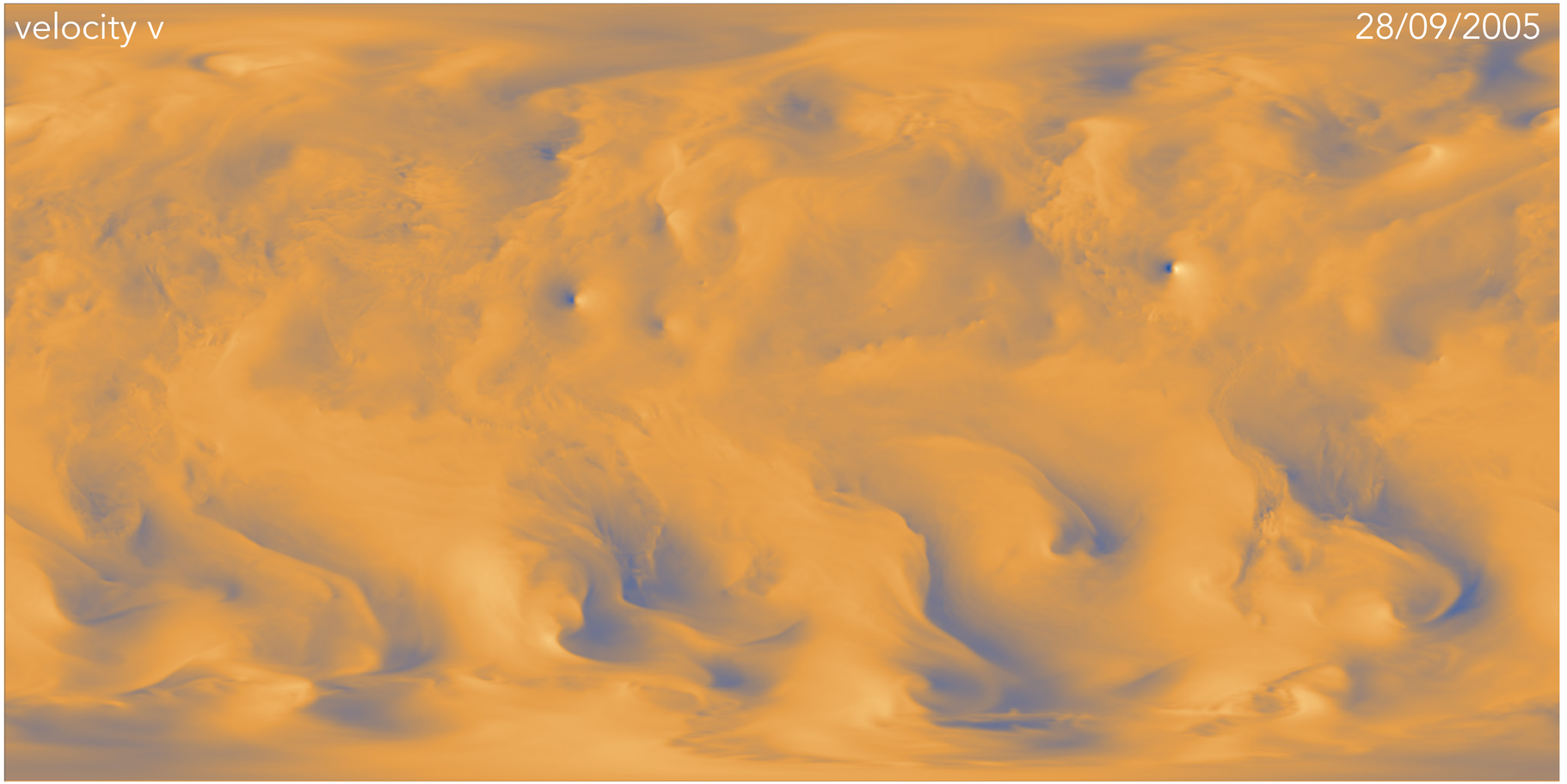
AtmoRep





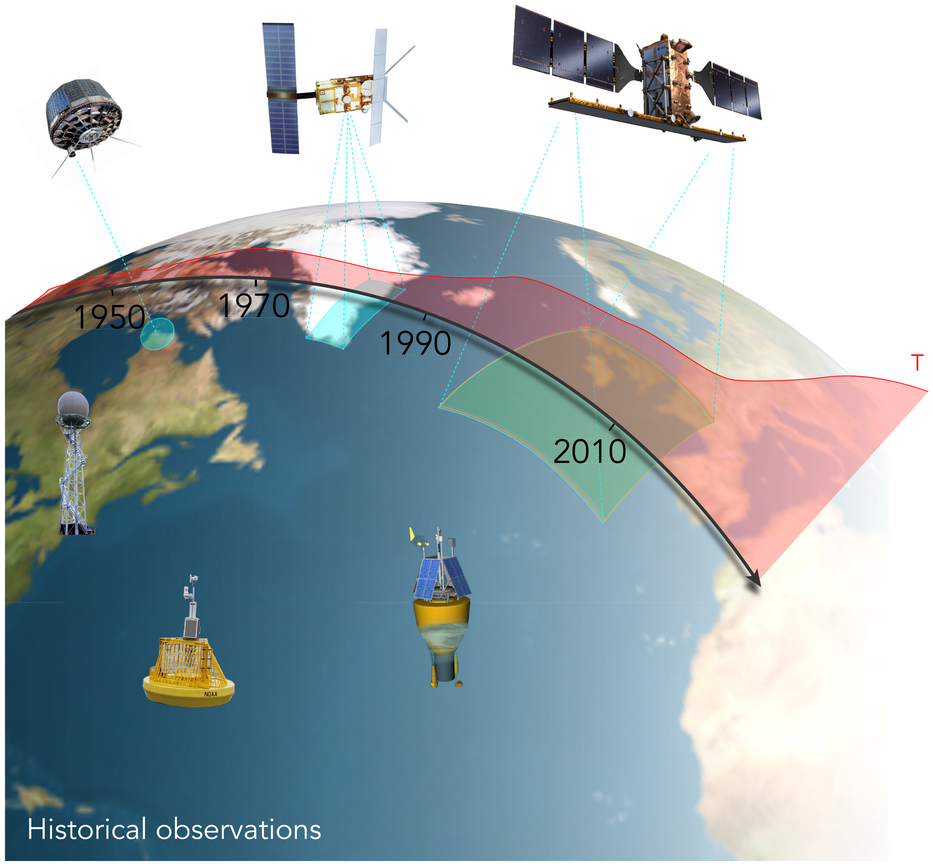
© Christian Lessig, 2023

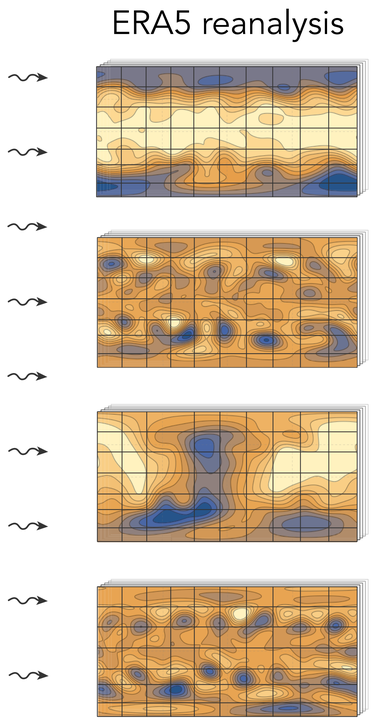
AtmoRep



© Christian Lessig, 2023

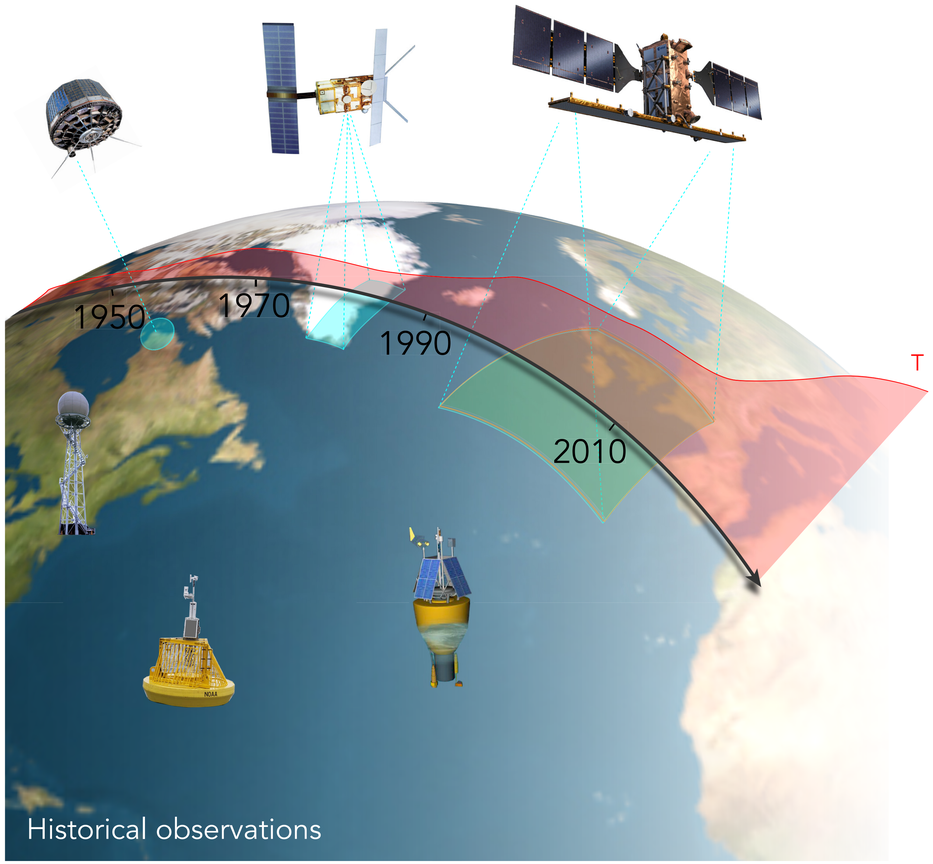
AtmoRep

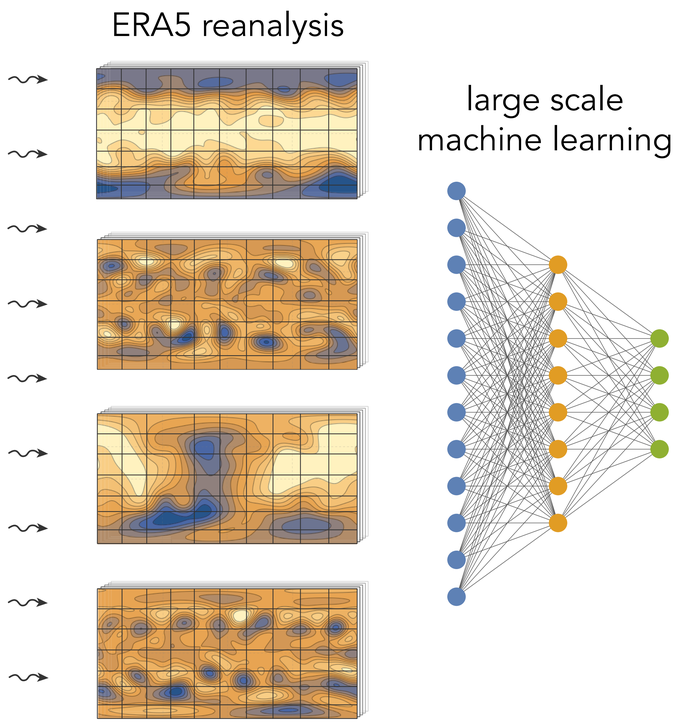




© Christian Lessig, 2023

AtmoRep

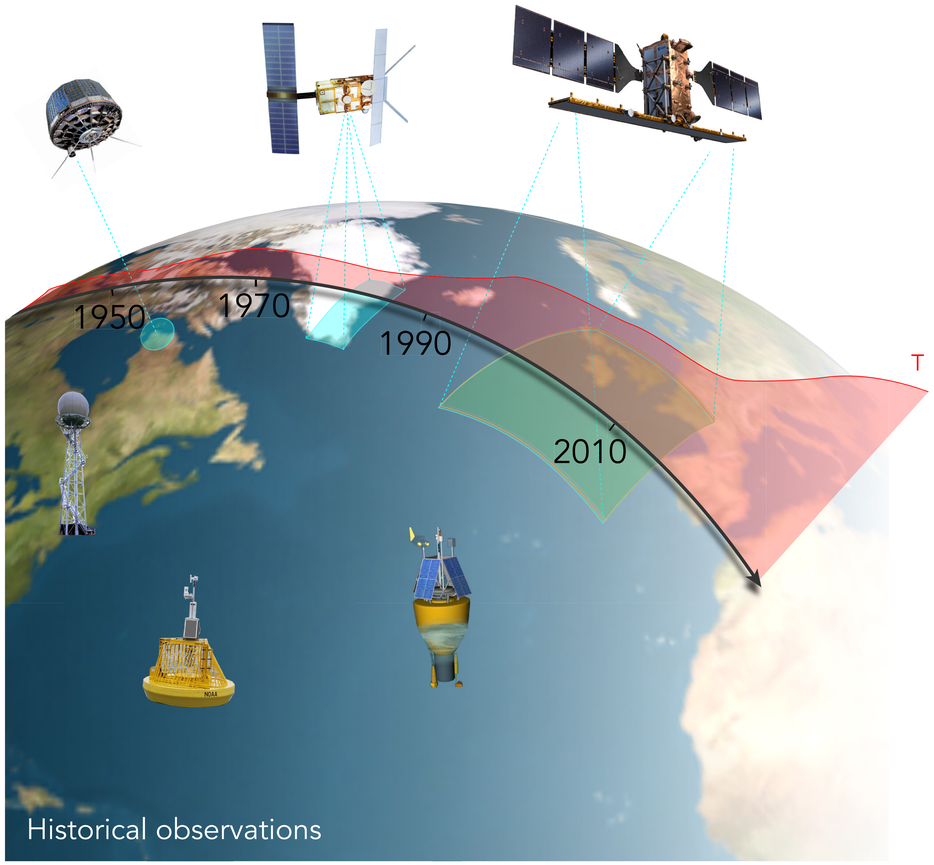


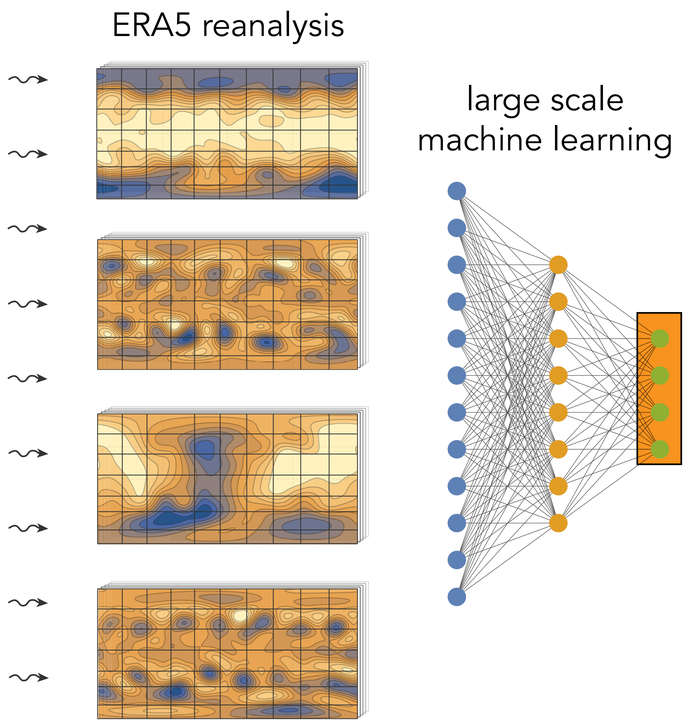


large scale
machine learning

© Christian Lessig, 2023

AtmoRep

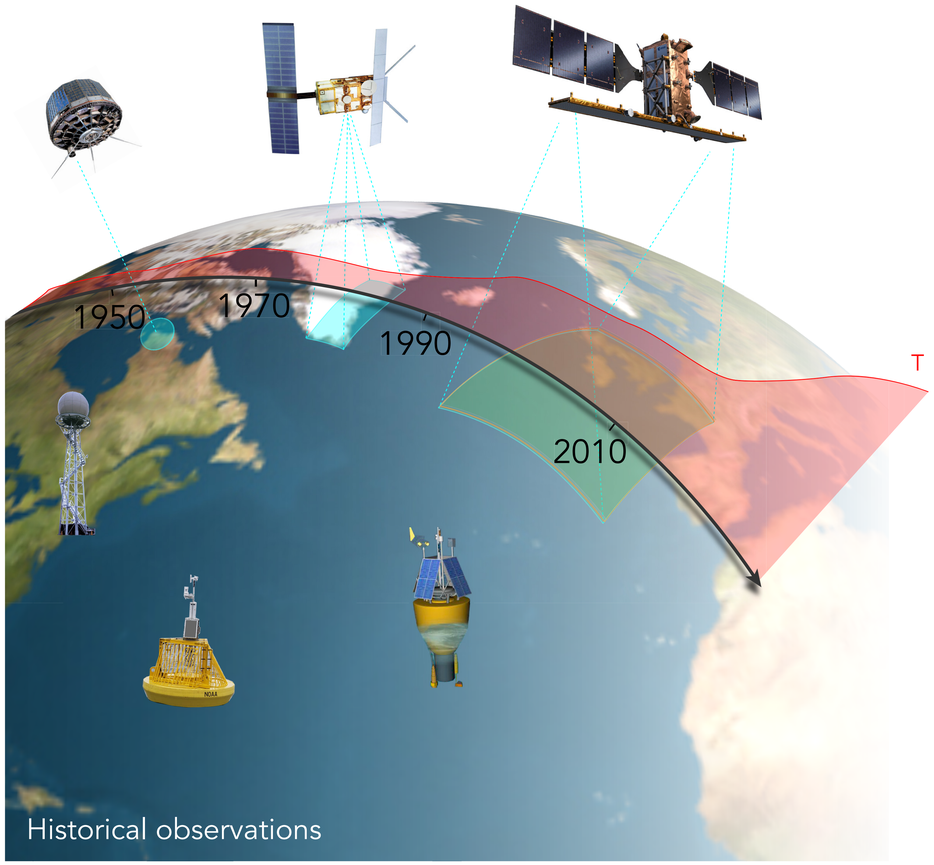


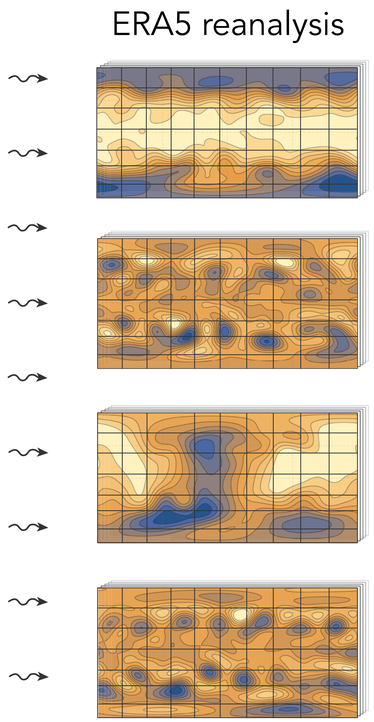


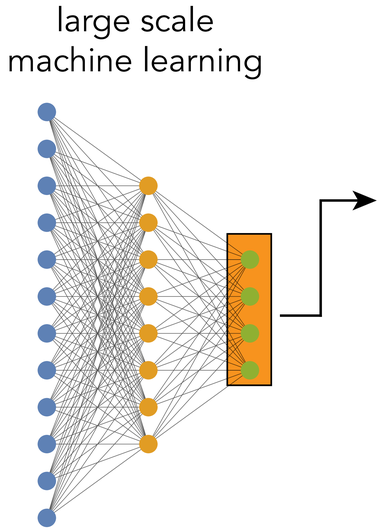
large scale
machine learning

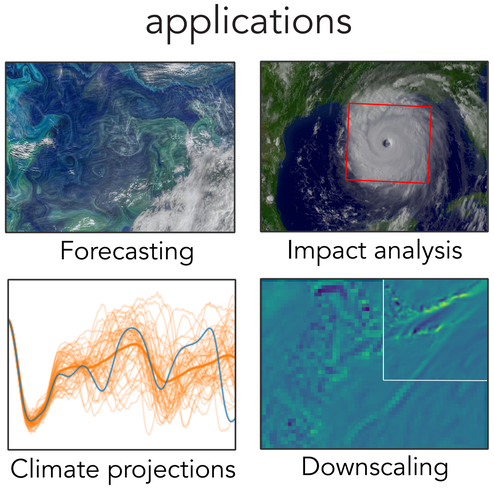
© Christian Lessig, 2023

AtmoRep



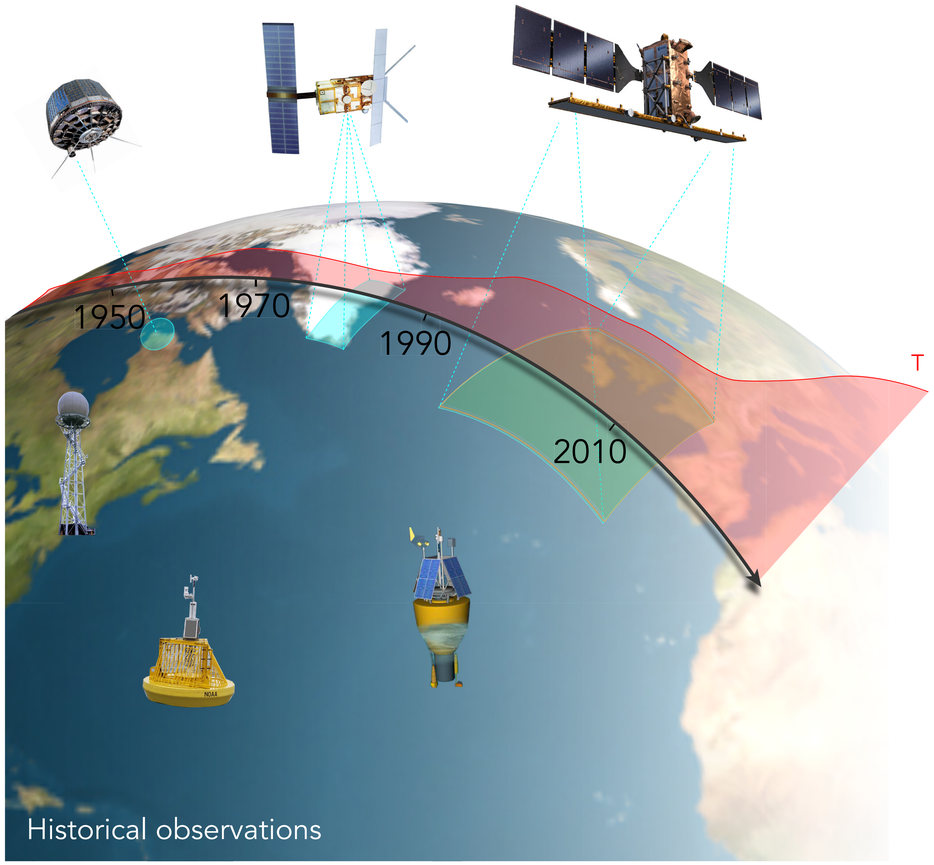




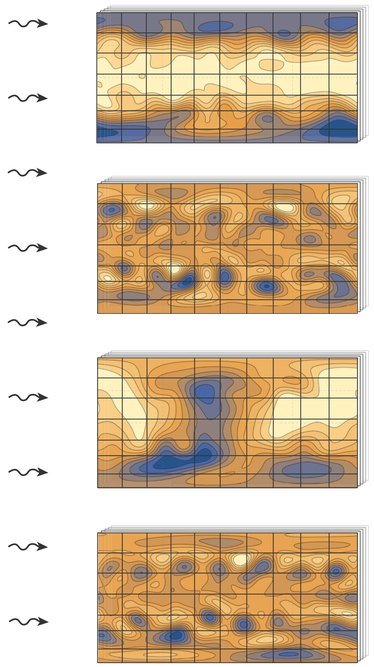


© Christian Lessig, 2023

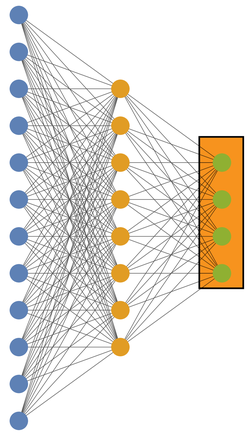
AtmoRep



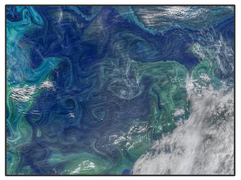
ERA5 reanalysis

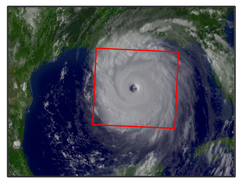


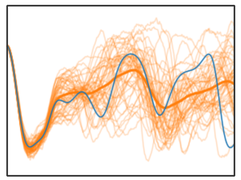
large scale machine learning

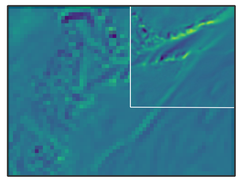


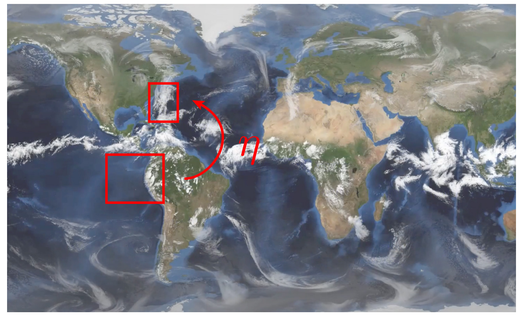
applications


Forecasting


Impact analysis


Climate projections

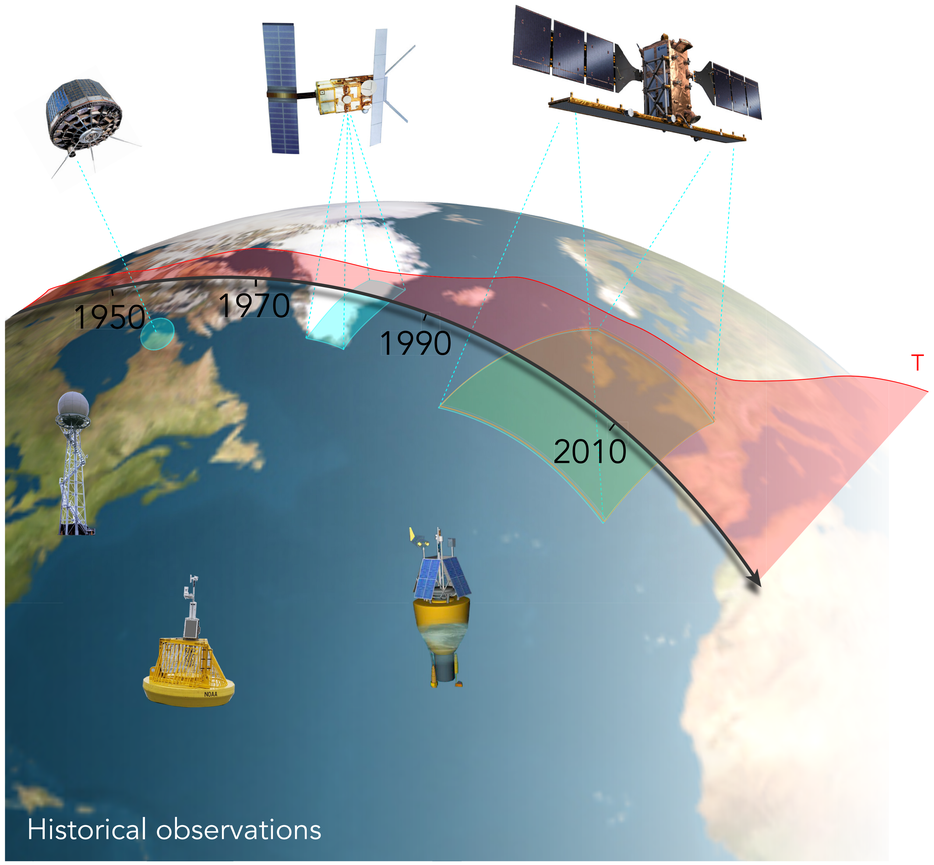

Downscaling

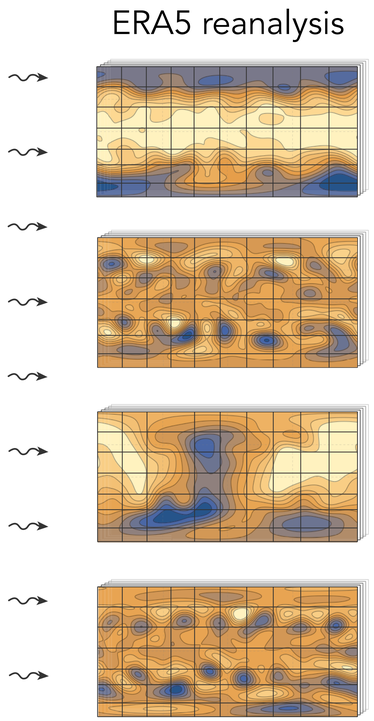


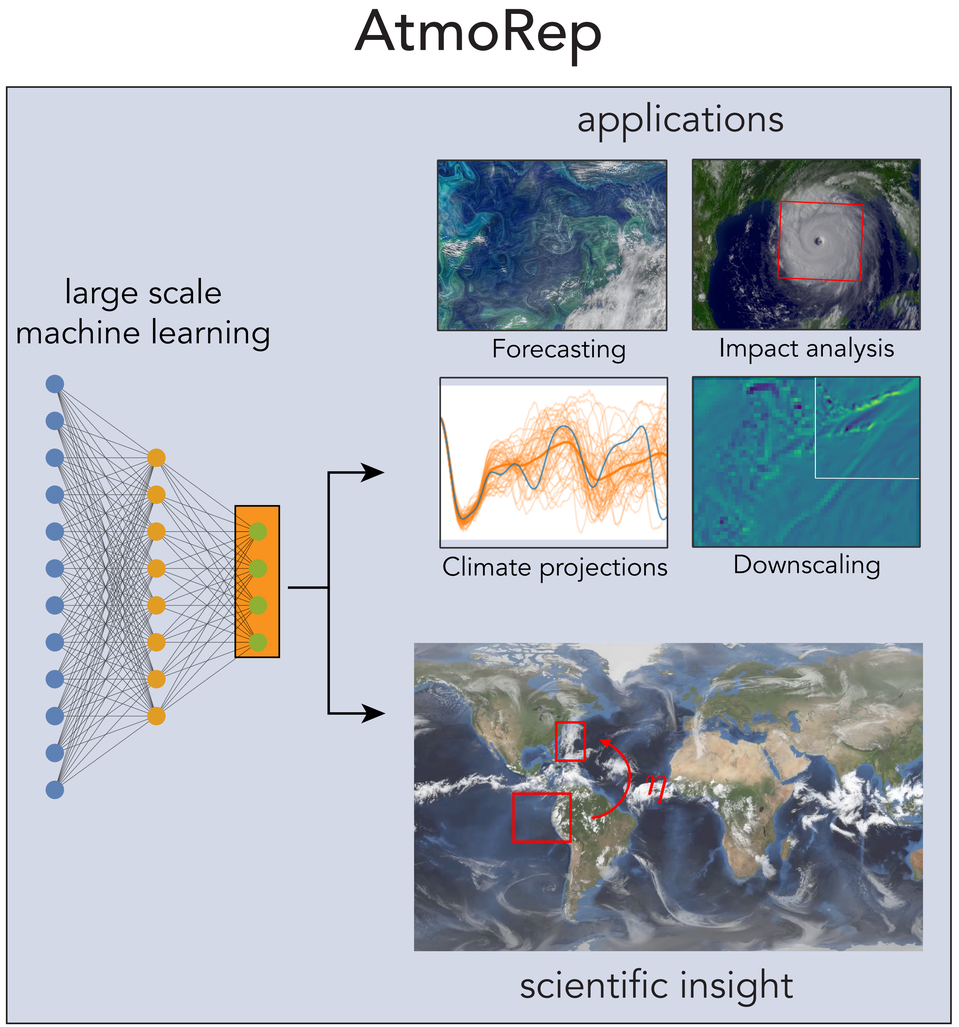
scientific insight

© Christian Lessig, 2023

AtmoRep







© Christian Lessig, 2023

A theoretical formulation

© Christian Lessig, 2023

A theoretical formulation

- Atmosphere as abstract stochastical dynamical system:

$$\bar{p}\left(\bar{y}|\bar{x}\right)$$

© Christian Lessig, 2023

A theoretical formulation

- Atmosphere as abstract stochastical dynamical system:

$$\bar{p}\left(\bar{y}|\bar{x}\right)$$

initial condition

$$\bar{x} = (\zeta, \mu, T, z, \cdots)$$

© Christian Lessig, 2023

A theoretical formulation

- Atmosphere as abstract stochastical dynamical system:

$$\bar{p}\left(\bar{y}|\bar{x}\right)$$

dependent state

initial condition

$$\bar{x} = (\zeta, \mu, T, z, \cdots)$$

© Christian Lessig, 2023

A theoretical formulation

- Atmosphere as abstract stochastical dynamical system:

$$\bar{p}\left(\bar{y}|\bar{x}\right)$$

tomorrow's weather — dependent state

today's weather — initial condition

$$\bar{x} = (\zeta, \mu, T, z, \cdots)$$

© Christian Lessig, 2023

A theoretical formulation

- Atmosphere as abstract stochastical dynamical system:

$$\bar{p}\left(\bar{y}|\bar{x}\right)$$

- Neural network model:

$$p_\theta\left(y|x,\alpha\right)$$

© Christian Lessig, 2023

A theoretical formulation

- Atmosphere as abstract stochastical dynamical system:

$$\bar{p}\left(\bar{y}|\bar{x}\right)$$

- Neural network model:

$$p_\theta\left(y|x,\alpha\right) \approx \bar{p}\left(\bar{y}|\bar{x}\right)$$

© Christian Lessig, 2023

A theoretical formulation

- Atmosphere as abstract stochastical dynamical system:

$$\bar{p}\big(\bar{y}|\bar{x}\big)$$

- Neural network model:

$$p_\theta\big(y|x,\alpha\big)$$

approx. initial condition
$x = (\zeta, \mu, T, z, \dots)$

auxiliary information
(e.g. global time)

© Christian Lessig, 2023

A theoretical formulation

- Atmosphere as abstract stochastical dynamical system:

$$\bar{p}\left(\bar{y}|\bar{x}\right)$$

- Neural network model:

$$\underbrace{p_\theta\left(y|x,\alpha\right)}$$

Standard formulation for generative models, e.g., large language models, Dall-E, diffusion models

© Christian Lessig, 2023

A theoretical formulation

- Atmosphere as abstract stochastical dynamical system:

$$\bar{p}\left(\bar{y}|\bar{x}\right)$$

- Neural network model:

$$\underbrace{p_\theta\left(y|x,\alpha\right)}$$

- forecasting
- downscaling
- model correction
- ...

Standard formulation for generative models, e.g., large language models, Dall-E, diffusion models

© Christian Lessig, 2023

A theoretical formulation

- Connects physics (expressed using mathematics) with machine learning model
  - General model allows to include all effects in the data

© Christian Lessig, 2023

A theoretical formulation

- Connects physics (expressed using mathematics) with machine learning model
  - › General model allows to include all effects in the data
- Intrinsically statistical/probabilistic formulation
  - › Fits the statistical/chaotic nature of the atmosphere

© Christian Lessig, 2023

A theoretical formulation

- Connects physics (expressed using mathematics) with machine learning model
  - › General model allows to include all effects in the data
- Intrinsically statistical/probabilistic formulation
  - › Fits the statistical/chaotic nature of the atmosphere
- Neural network model as factorization of $p_\theta\left(y|x,\alpha\right)$
  - › Loss (expectation maximization, ELBO, ...) has clear connection to physical model

© Christian Lessig, 2023

Multiformer

- Plug-and-play of fields
  - Fields can be added/removed with limited (or no) computational effort
- Cross-attention allows for explicit introspection of inter-action between fields
- Different physical fields with different properties have separate latent spaces (and transformations for these)

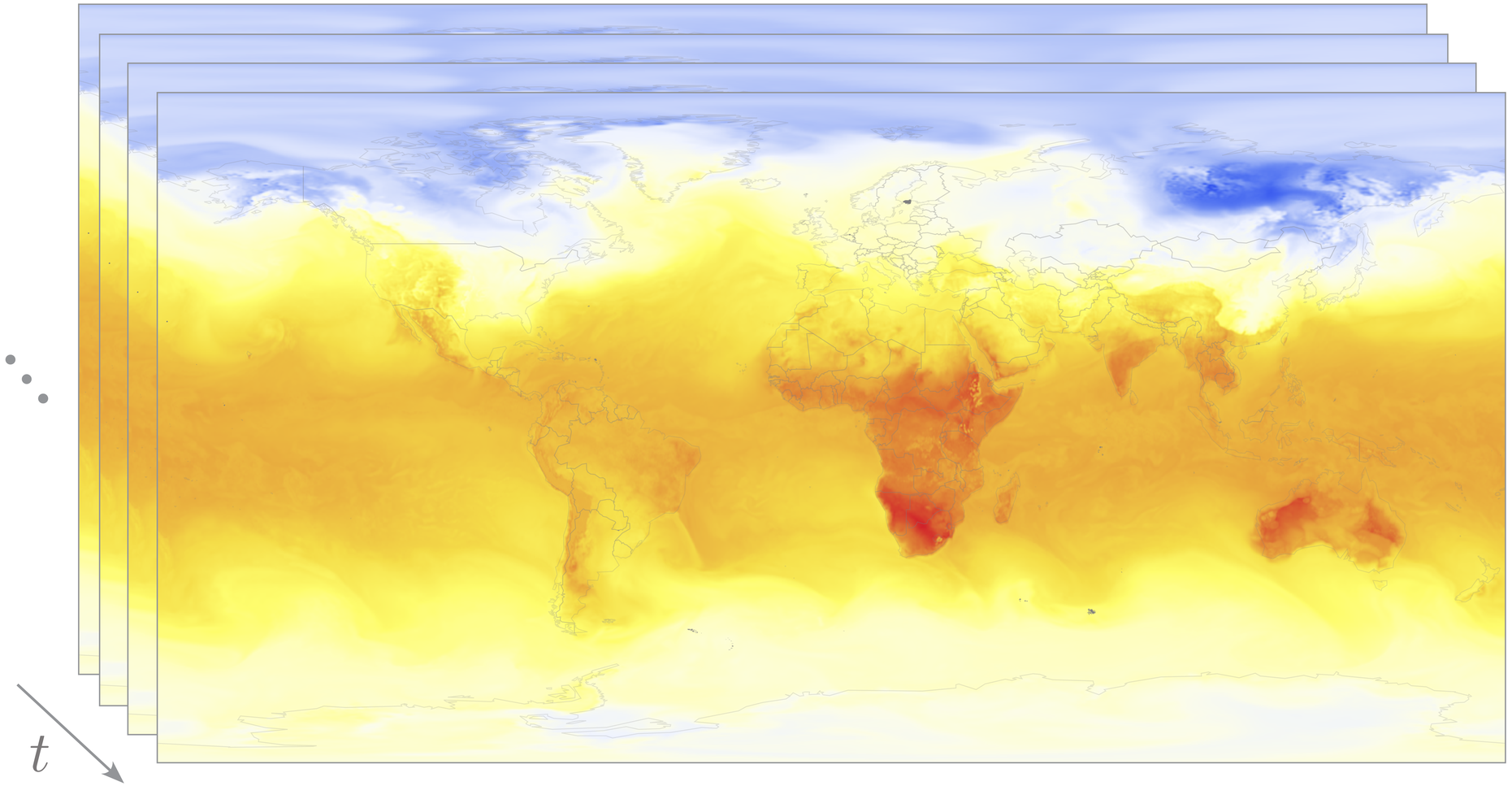
© Christian Lessig, 2023

AtmoRep network architecture

- Network is local in space-time

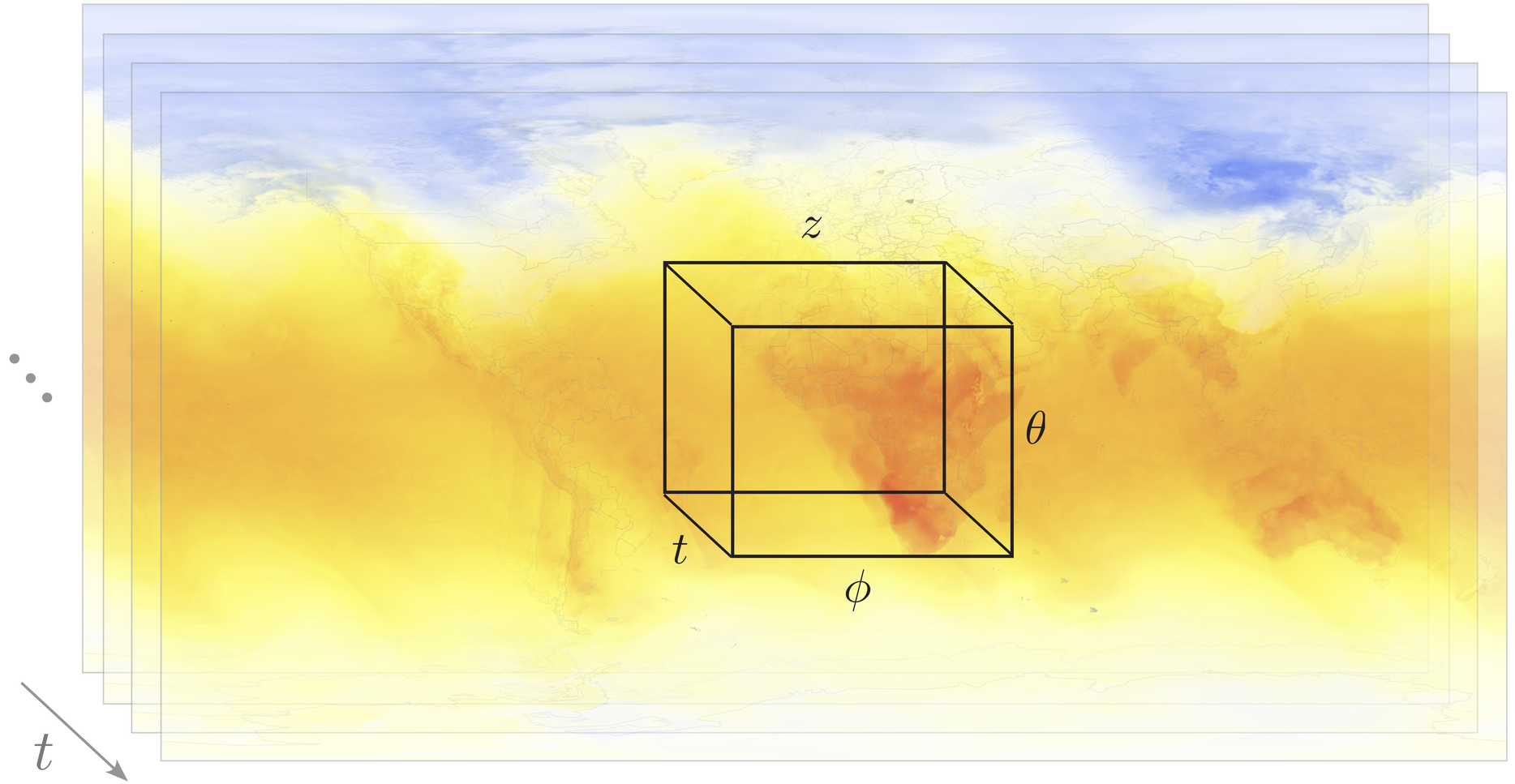
© Christian Lessig, 2023

AtmoRep network architecture



© Christian Lessig, 2023

AtmoRep network architecture



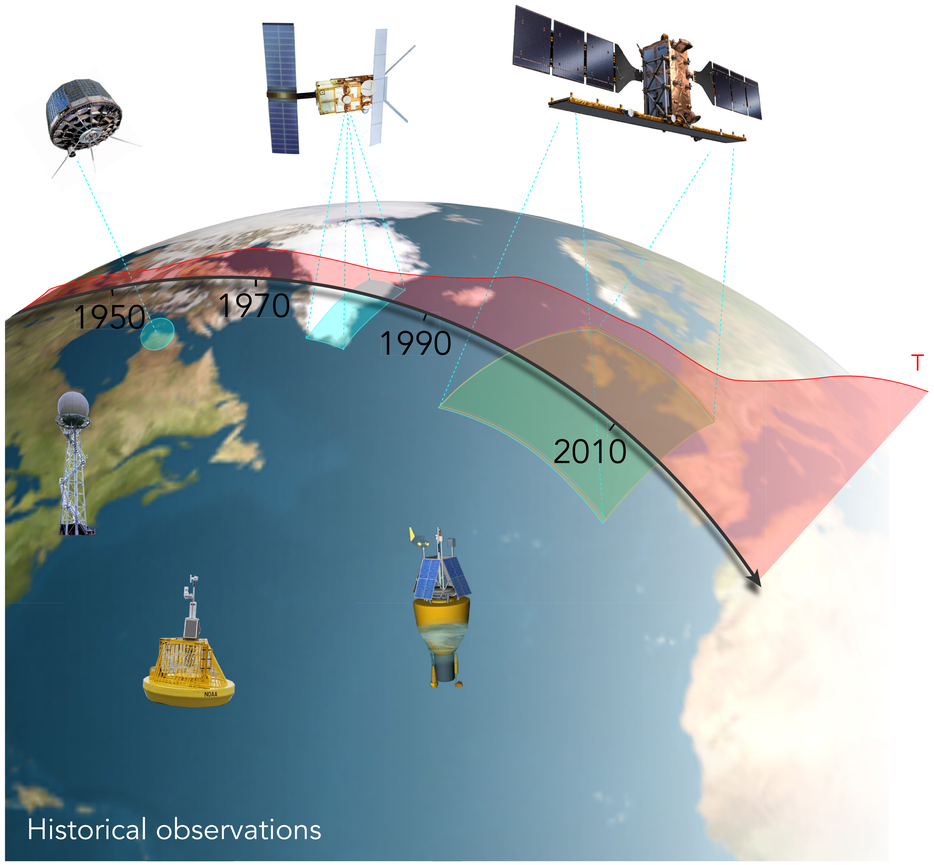
© Christian Lessig, 2023

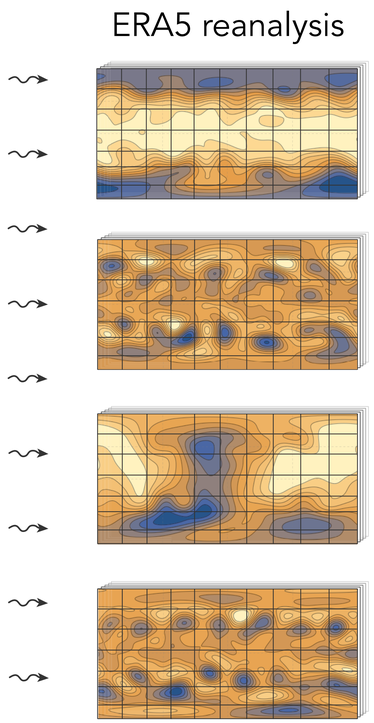
AtmoRep network architecture

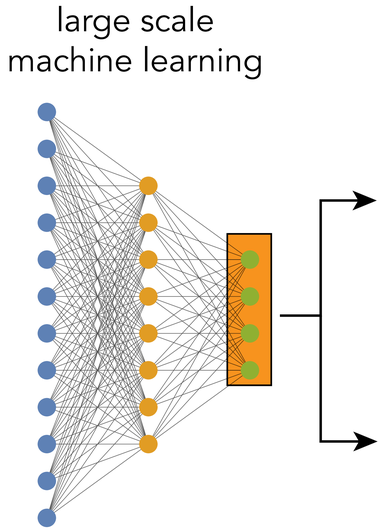
- Transformer as network architecture
  - Scales well to very large data-sets
  - Generative model (with decoder)
  - Attention maps provide (physical) interpretability

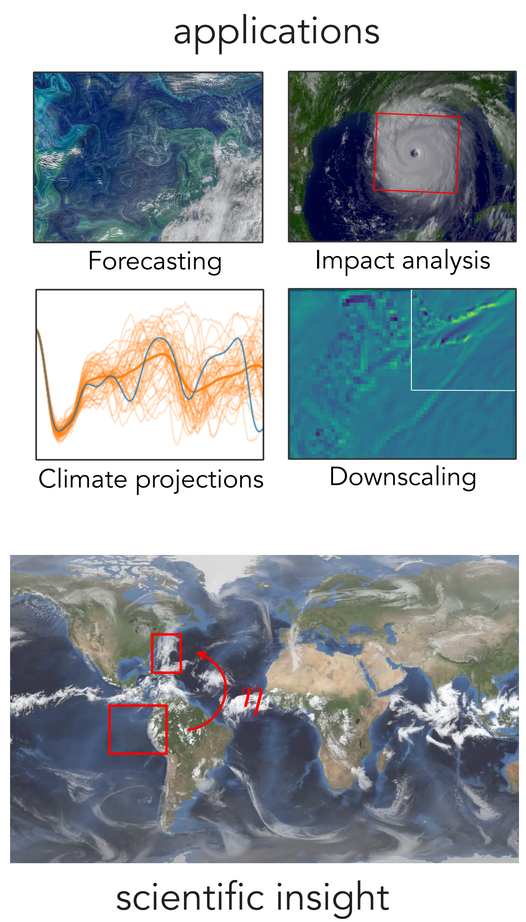
© Christian Lessig, 2023

AtmoRep



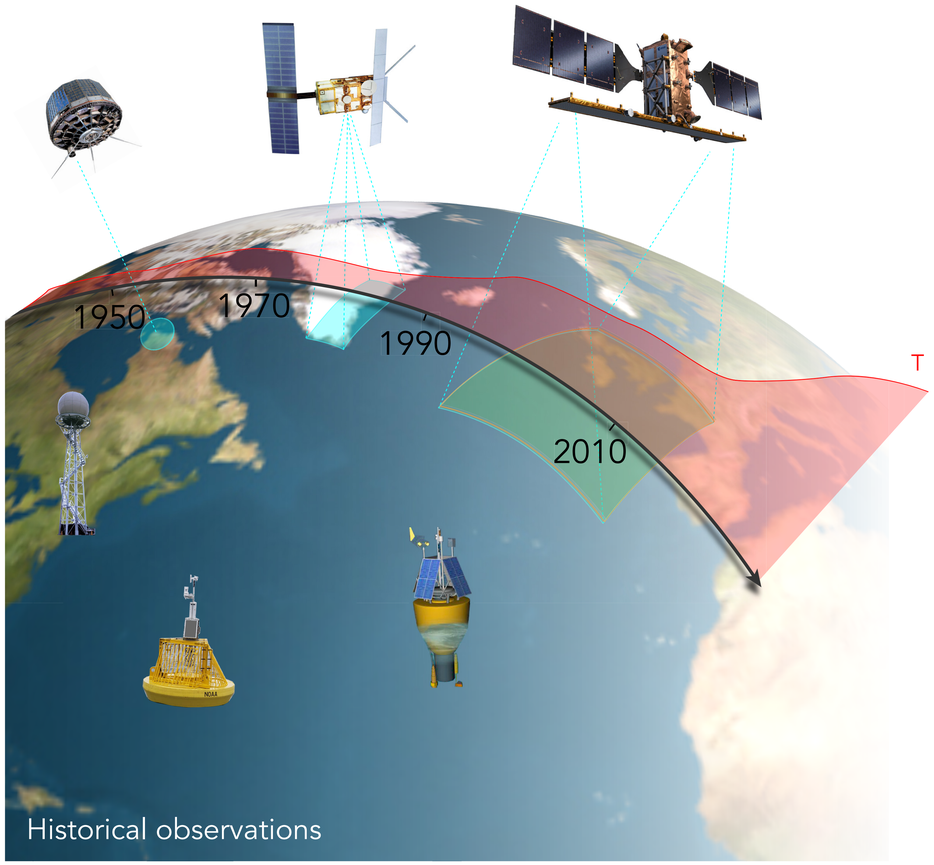


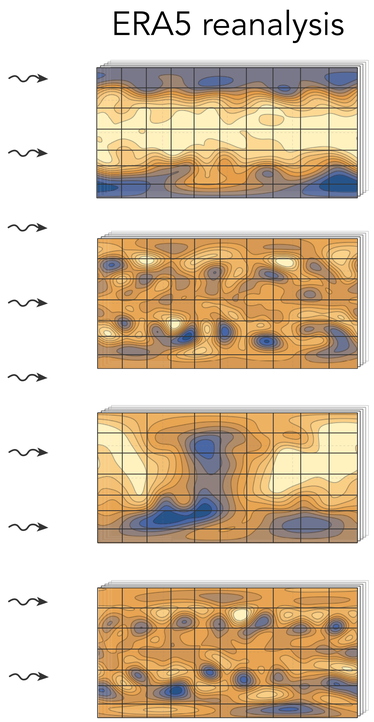




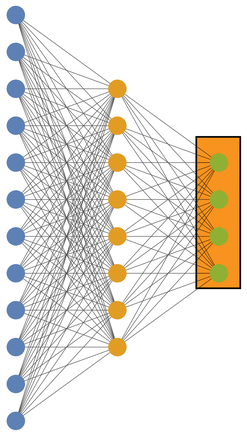
© Christian Lessig, 2023

AtmoRep

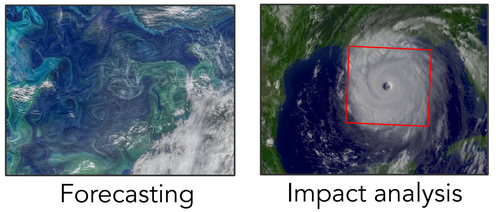


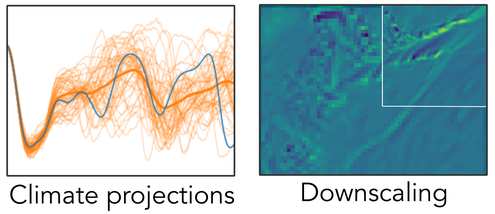


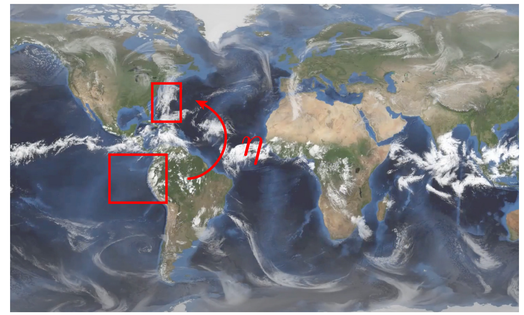
$p_\theta\left(y|x,\alpha\right)$



$p_\theta\left(y|x,\alpha,\tau\right)$







$p_\theta\left(y|x,\alpha\right) \approx \bar{p}\left(\bar{y}|\bar{x}\right)$

© Christian Lessig, 2023

Network architecture

What's the right network architecture?

© Christian Lessig, 2023

Network architecture

What's the right network architecture?

Doesn't matter!

© Christian Lessig, 2023

Network architecture

What's the right network architecture?

Doesn't matter!

© Christian Lessig, 2023

Network architecture

What's the right network architecture?

Doesn't matter!

(perhaps: use a transformer)

© Christian Lessig, 2023

Network architecture

"[...] architectural details such as network width or depth have minimal effects within a wide range."[1]

[1] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models, 2020.

© Christian Lessig, 2023

Network architecture

"The loss scales as a power-law with model size, dataset size, and the amount of compute used for training, with some trends spanning more than seven orders of magnitude. [...] architectural details such as network width or depth have minimal effects within a wide range."[1]

[1] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models, 2020.

© Christian Lessig, 2023

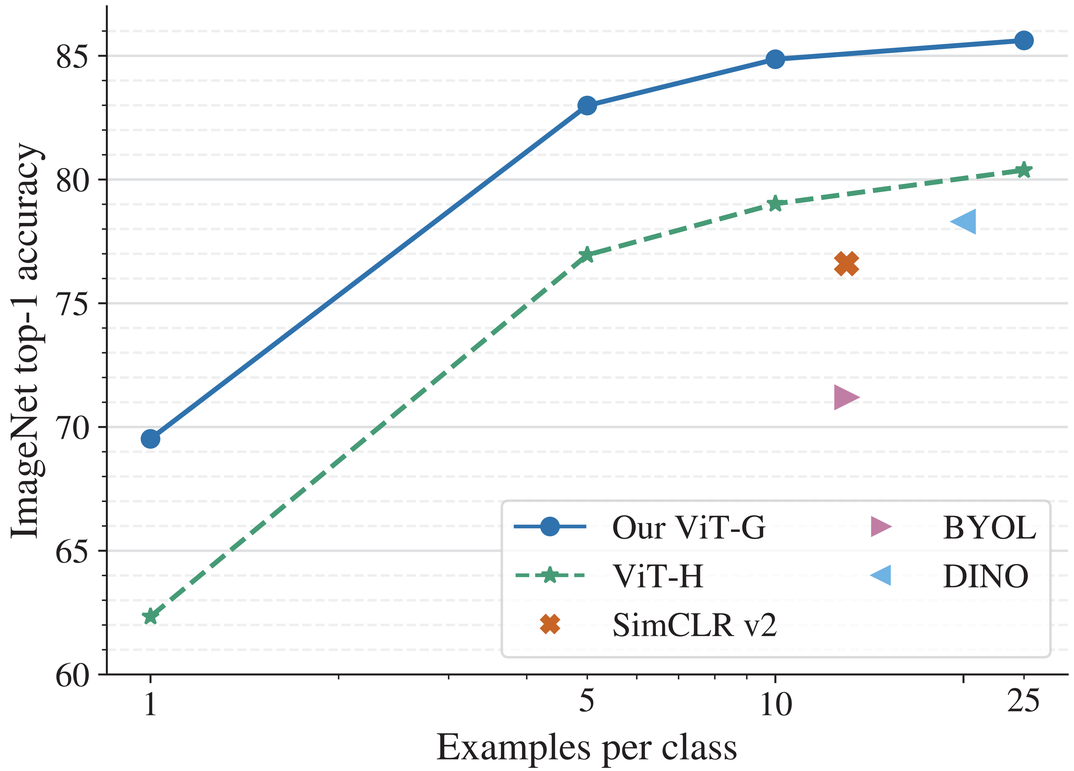
Network architecture

"The loss scales as a power-law with **model size**, **dataset size**, and the **amount of compute** used for training, with some trends spanning more than seven orders of magnitude. [...] architectural details such as network width or depth have minimal effects within a wide range."[1]

[1] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models, 2020.

© Christian Lessig, 2023

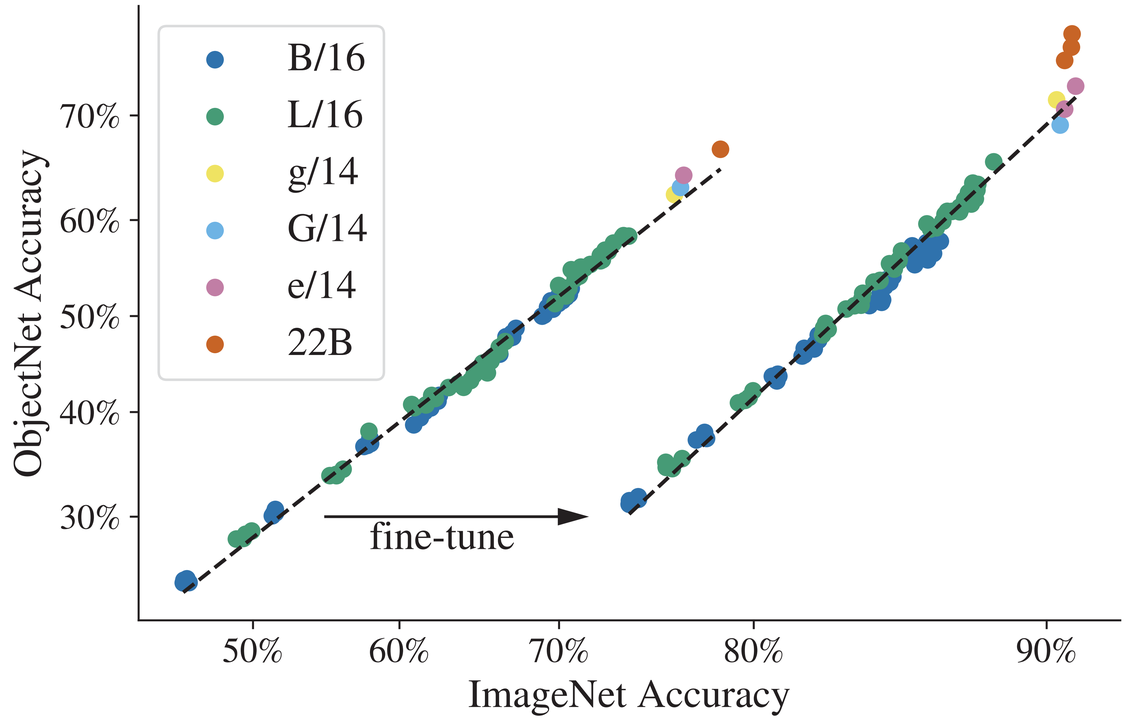
Network architecture



From: X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer. Scaling vision transformers, 2021.

© Christian Lessig, 2023

Network architecture

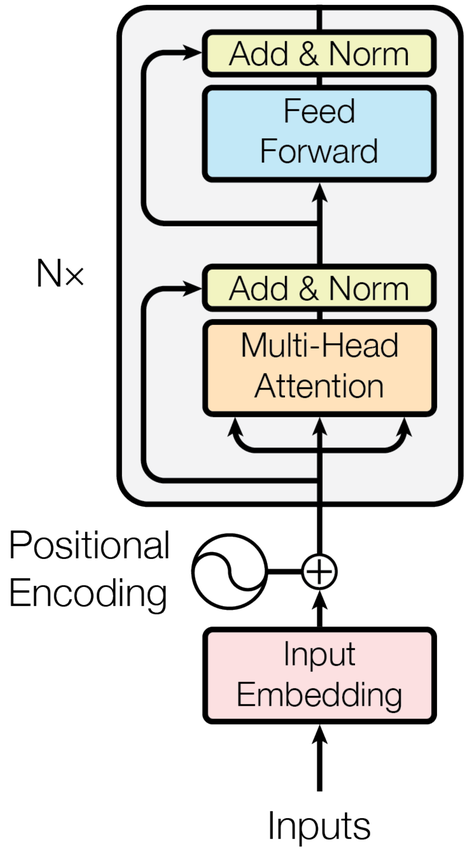


From: M. Dehghani, et al. Scaling vision transformers to 22 billion parameters, 2023.

© Christian Lessig, 2023

Network architecture

- Transformer encoder:

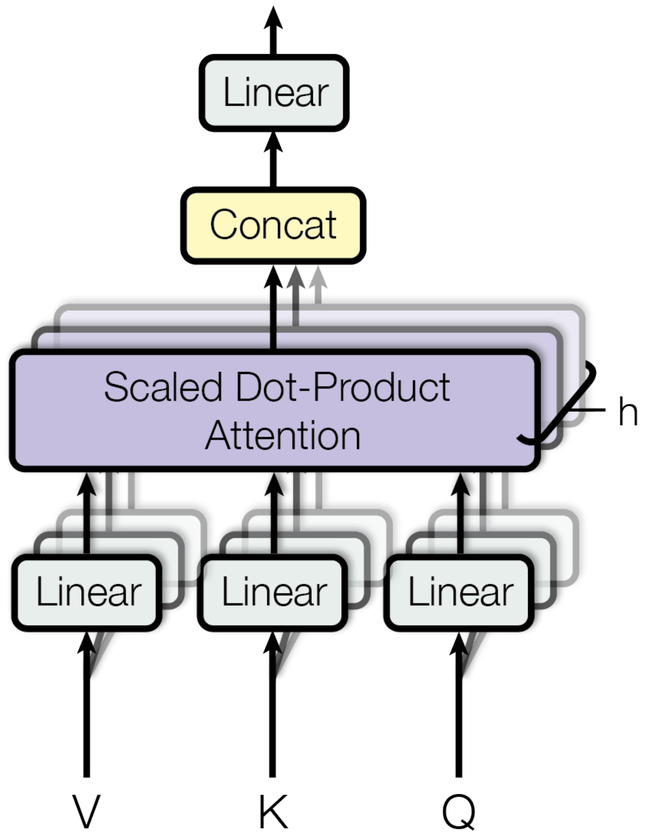


From: A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vish- wanathan, and R. Garnett, editors, NeurIPS, volume 30. Curran Associates, Inc., 2017.

© Christian Lessig, 2023

Network architecture

- Self attention block:

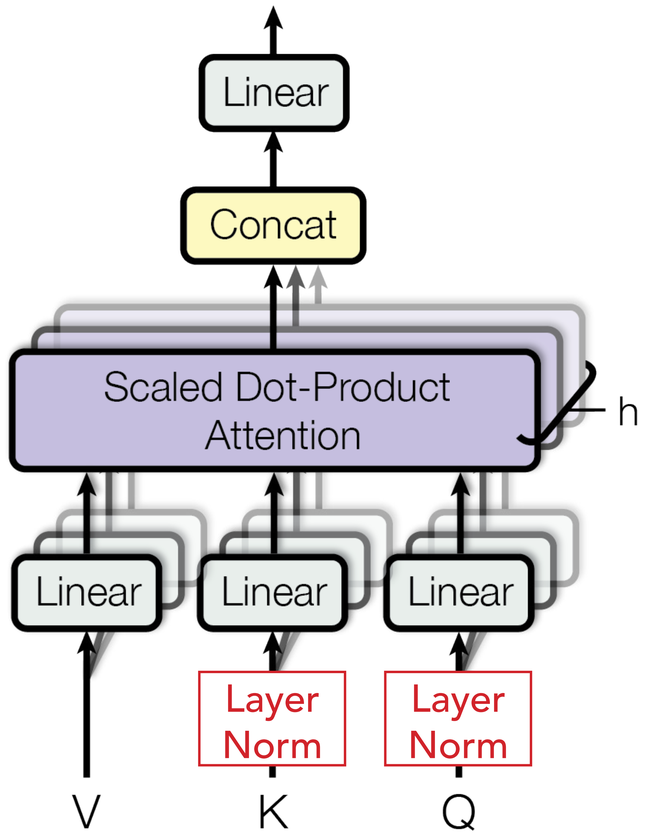


From: A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vish- wanathan, and R. Garnett, editors, NeurIPS, volume 30. Curran Associates, Inc., 2017.

© Christian Lessig, 2023

Network architecture

- Self attention block:

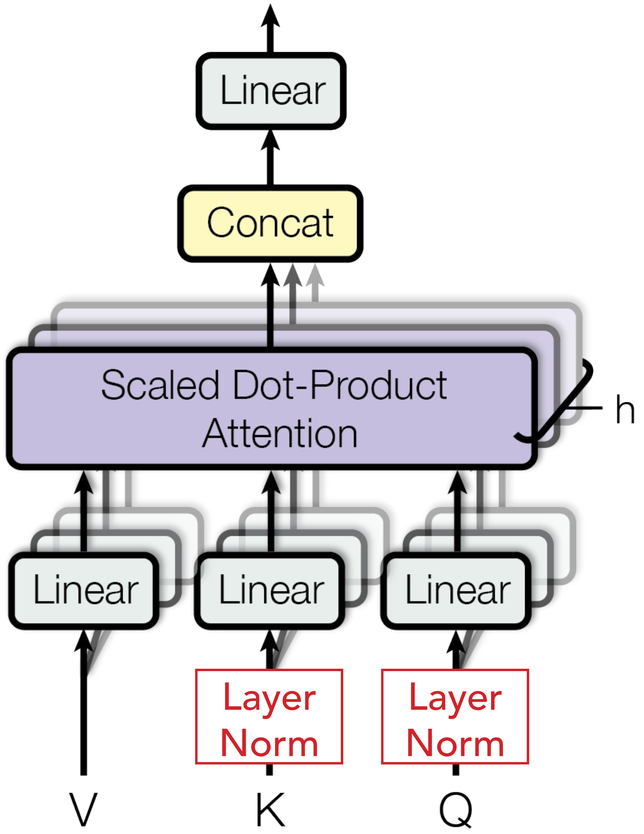


From: A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vish- wanathan, and R. Garnett, editors, NeurIPS, volume 30. Curran Associates, Inc., 2017.

© Christian Lessig, 2023

Network architecture

- Self attention block:



Key architectural change to go from 72% to 74.5% accuracy with 22 billion parameters

From: A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vish- wanathan, and R. Garnett, editors, NeurIPS, volume 30. Curran Associates, Inc., 2017.

© Christian Lessig, 2023

Network architecture: the bitter lesson

"We have to learn the bitter lesson that building in how we think we think does not work in the long run. The bitter lesson is based on the historical observations that 1) AI researchers have often tried to build knowledge into their agents, 2) this always helps in the short term, and is personally satisfying to the researcher, but 3) in the long run it plateaus and even inhibits further progress, and 4) breakthrough progress eventually arrives by an opposing approach based on scaling computation by search and learning."[1]

[1] Rich Sutton, 2019; http://www.incompleteideas.net/IncIdeas/BitterLesson.html

© Christian Lessig, 2023

Network architecture

1. Deep learning is a very large optimization problem
   - LayerNorms are key “trick“ to keep optimization well conditioned
   - If training is possible, this leads to effective and generalizing performance
   - Convergent training with trillions of parameters is, from a scientific point of view, a miracle

© Christian Lessig, 2023

Network architecture

2. Amount of training data usually more important than overall network architecture
   - Transformer: simple building blocks that run efficiently on state-of-the-art compute hardware
   - Pre-training to work on problems with little labeled data
   - Representation learning to obtain networks that are amenable to different problems in a domain
   - Scale network size with training data

© Christian Lessig, 2023

Network architecture

3. Loss and training protocol can play a critical role
   - Self-supervised training improves results often even when labels are available
   - Simple is usually better: how well can the loss be optimized?

© Christian Lessig, 2023

Network architecture: weather forecasting

- Meterologists spent a lot of time thinking about what small sub-problem is amenable to machine learning and what the optimal network architecture is

© Christian Lessig, 2023

Network architecture: weather forecasting

- Meterologists spent a lot of time thinking about what small sub-problem is amenable to machine learning and what the optimal network architecture is
- Training large networks to solve the entire problem by far outperforms everything else
  - Computer scientists at NVIDIA, Deepmind ...
  - Training on data curated by domain scientists

© Christian Lessig, 2023

Network architecture

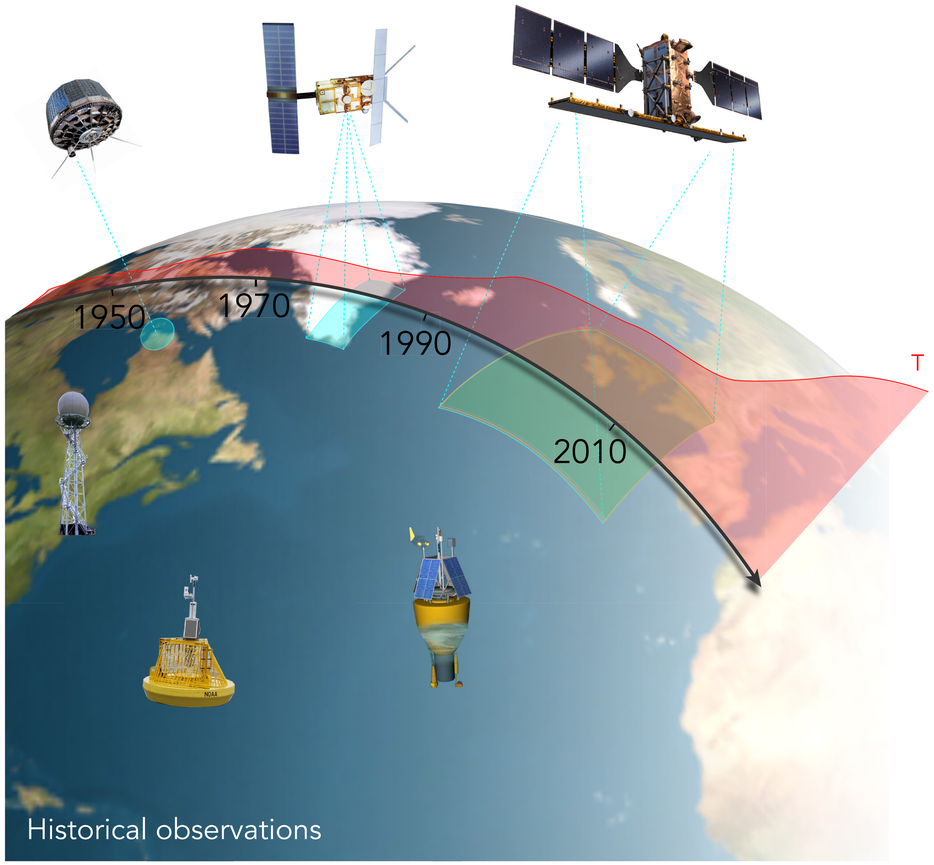
- What also matters ....

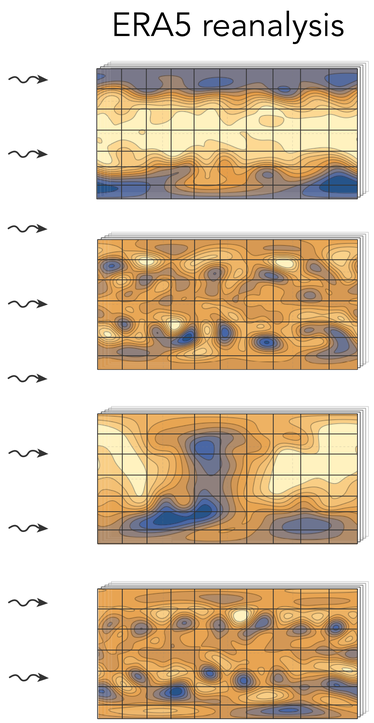
Scaling Vision Transformers to 22 Billion Parameters

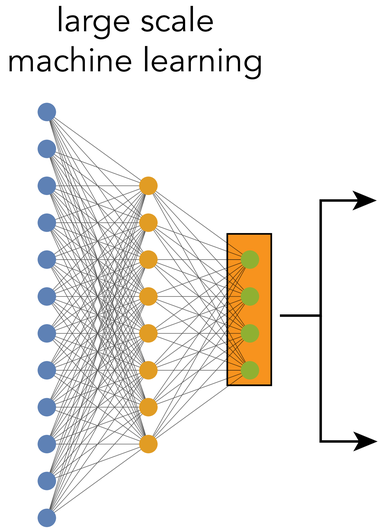
Mostafa Dehghani* **Josip Djolonga*** **Basil Mustafa*** **Piotr Padlewski*** **Jonathan Heek***
Justin Gilmer **Andreas Steiner** **Mathilde Caron** **Robert Geirhos** **Ibrahim Alabdulmohsin**
Rodolphe Jenatton **Lucas Beyer** **Michael Tschannen** **Anurag Arnab** **Xiao Wang**
Carlos Riquelme **Matthias Minderer** **Joan Puigcerver** **Utku Evci** **Manoj Kumar**
Sjoerd van Steenkiste **Gamaleldin F. Elsayed** **Aravindh Mahendran** **Fisher Yu**
Avital Oliver **Fantine Huot** **Jasmijn Bastings** **Mark Patrick Collier** **Alexey A. Gritsenko**
Vighnesh Birodkar **Cristina Vasconcelos** **Yi Tay** **Thomas Mensink** **Alexander Kolesnikov**
Filip Pavetić **Dustin Tran** **Thomas Kipf** **Mario Lučić** **Xiaohua Zhai** **Daniel Keysers**
Jeremiah Harmsen **Neil Houlsby***
Google Research

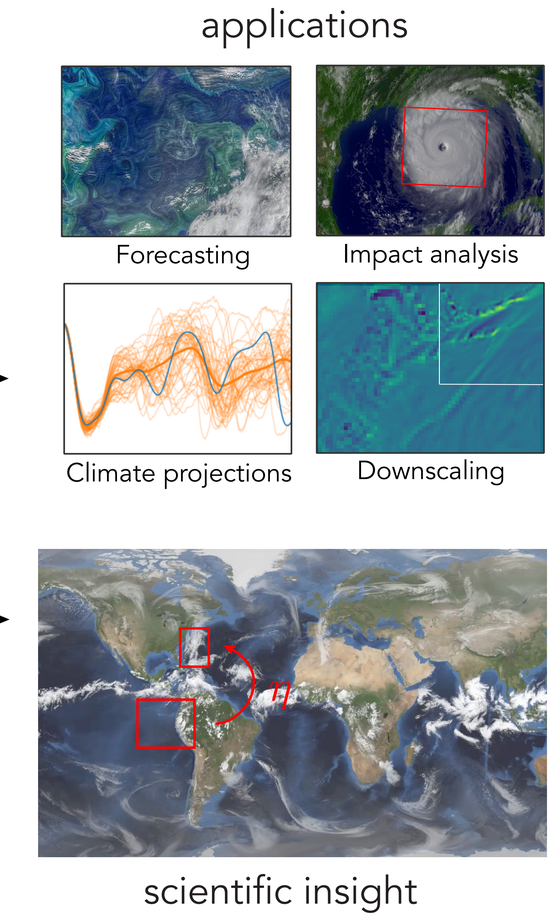
© Christian Lessig, 2023

AtmoRep









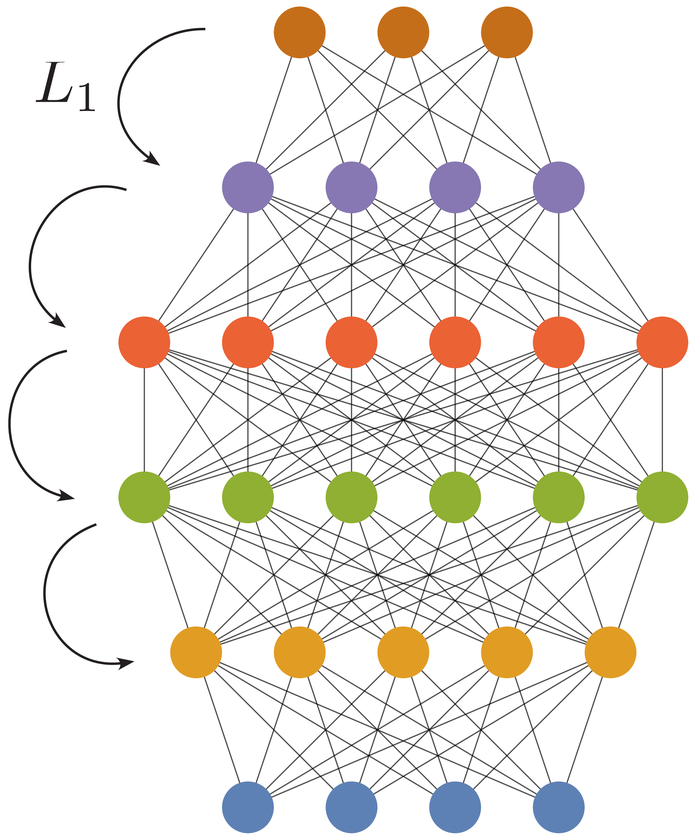
© Christian Lessig, 2023

What is a token?

- Key property of transformers: sequence of inputs is processed simultaneously
  - Language: input is sentence with words being tokens
  - Images: small image patches

© Christian Lessig, 2023

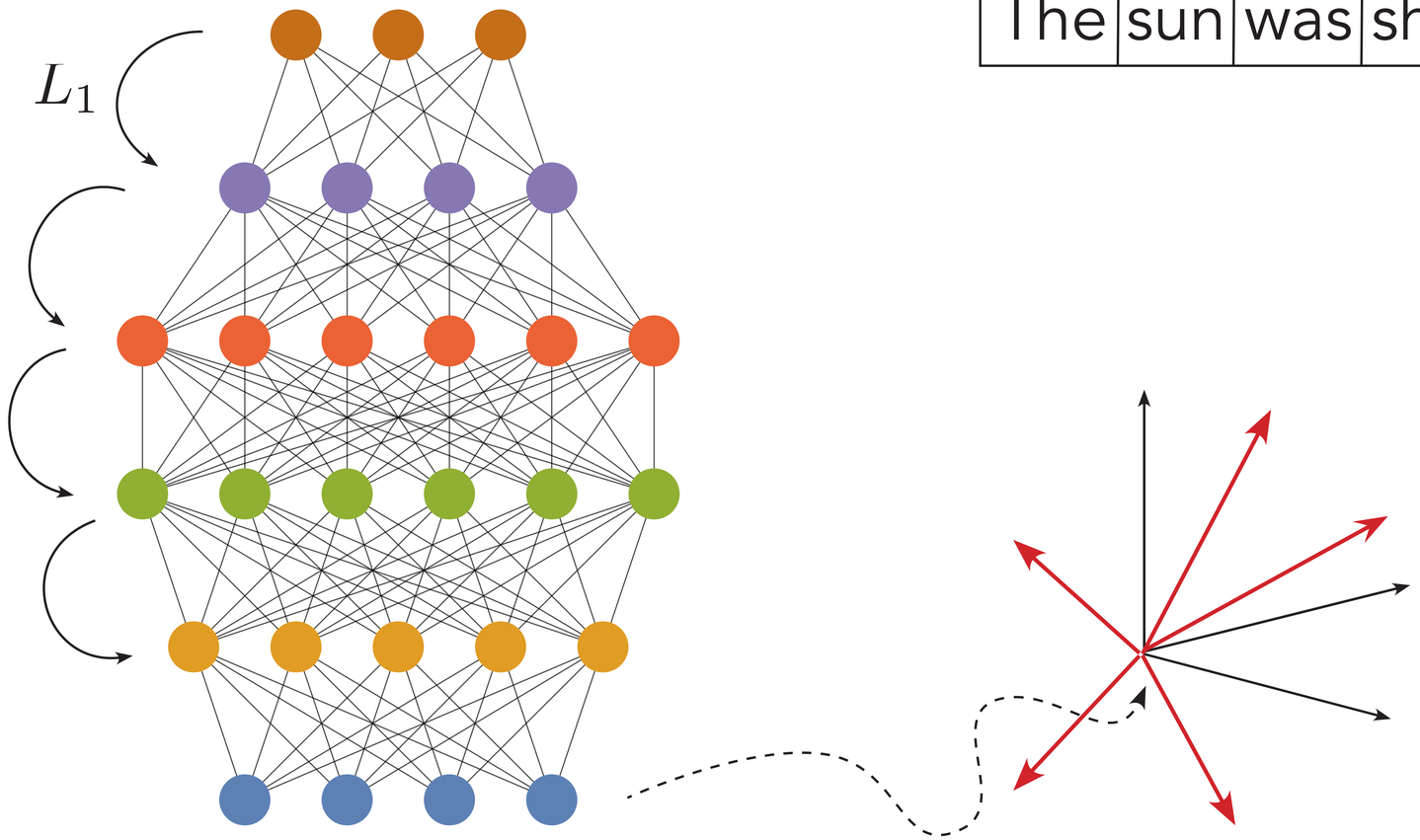
What is a token?



The	sun	was	shining	bright.

© Christian Lessig, 2023

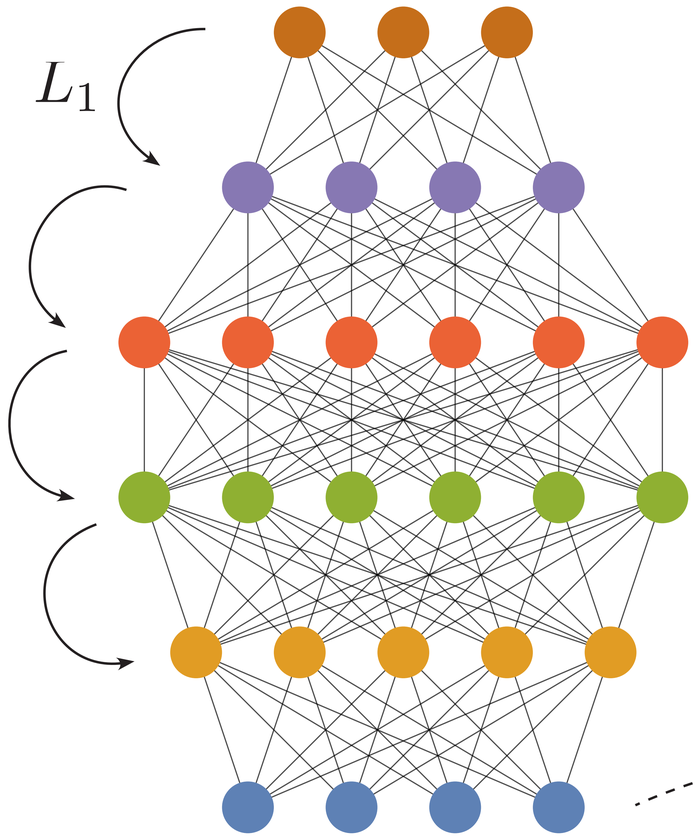
What is a token?



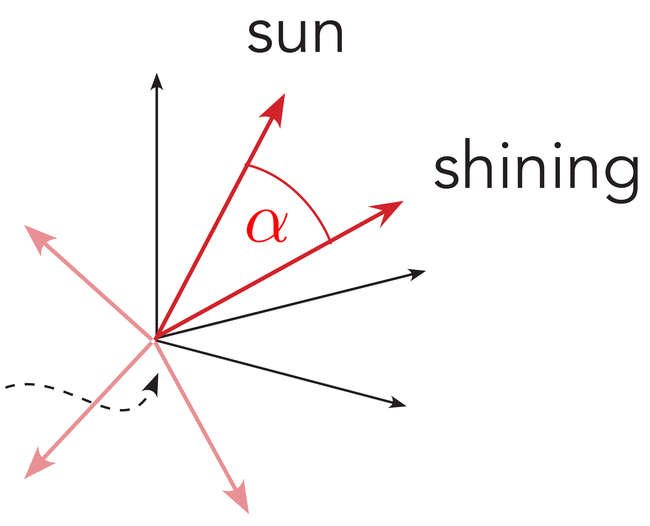
The	sun	was	shining	bright.

© Christian Lessig, 2023

What is a token?

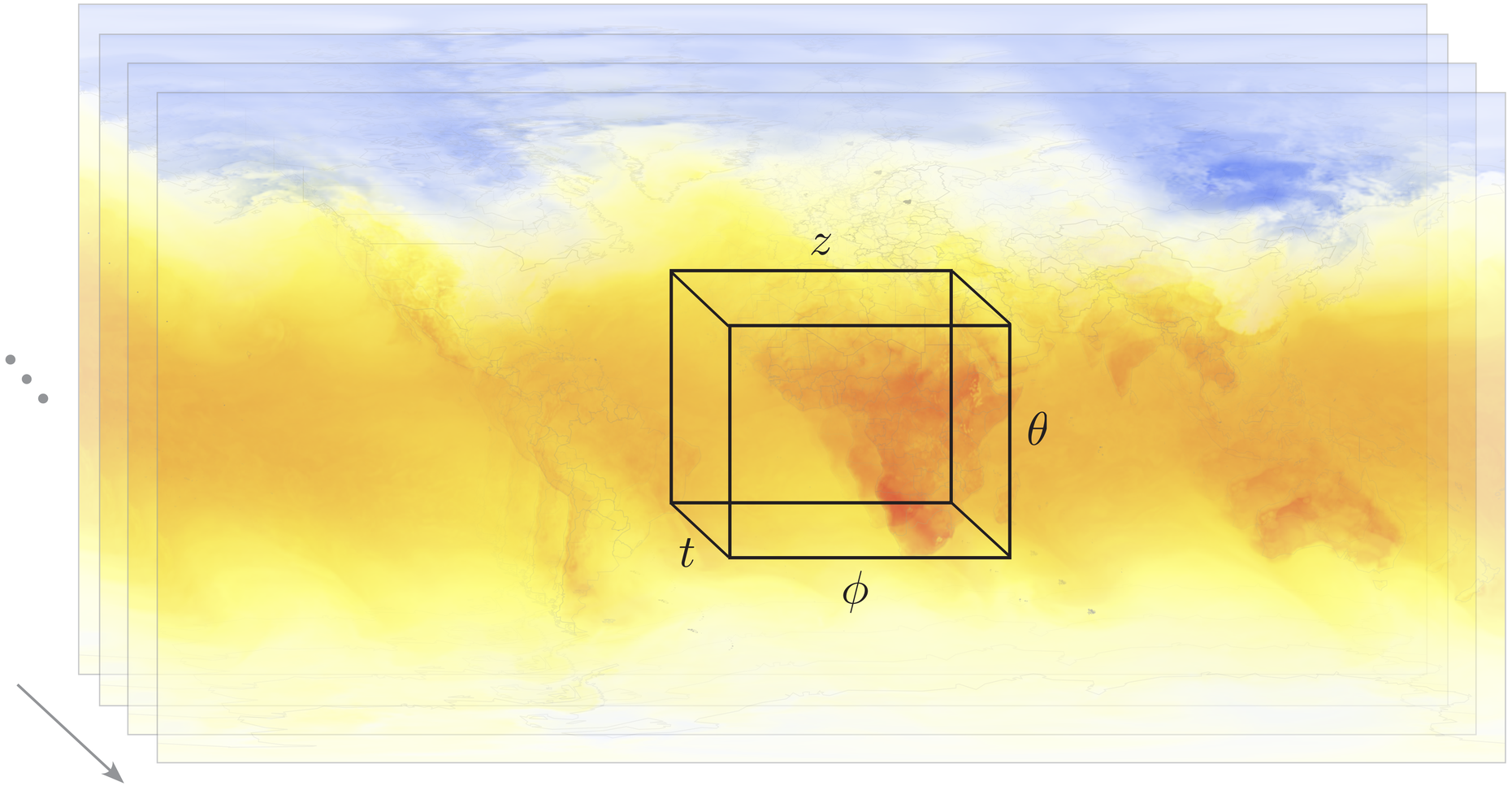


The	sun	was	shining	bright.



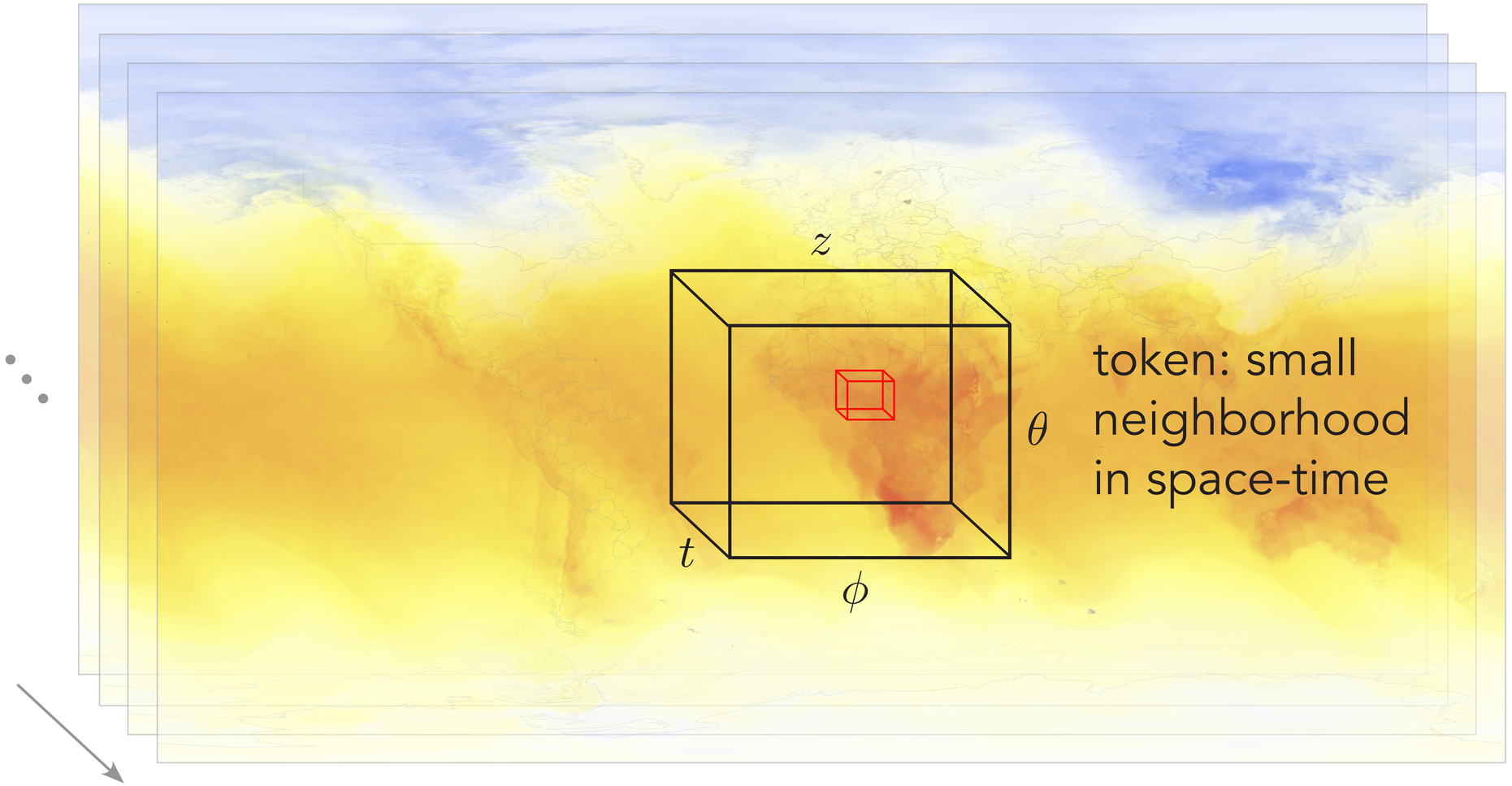
© Christian Lessig, 2023

What is a token?



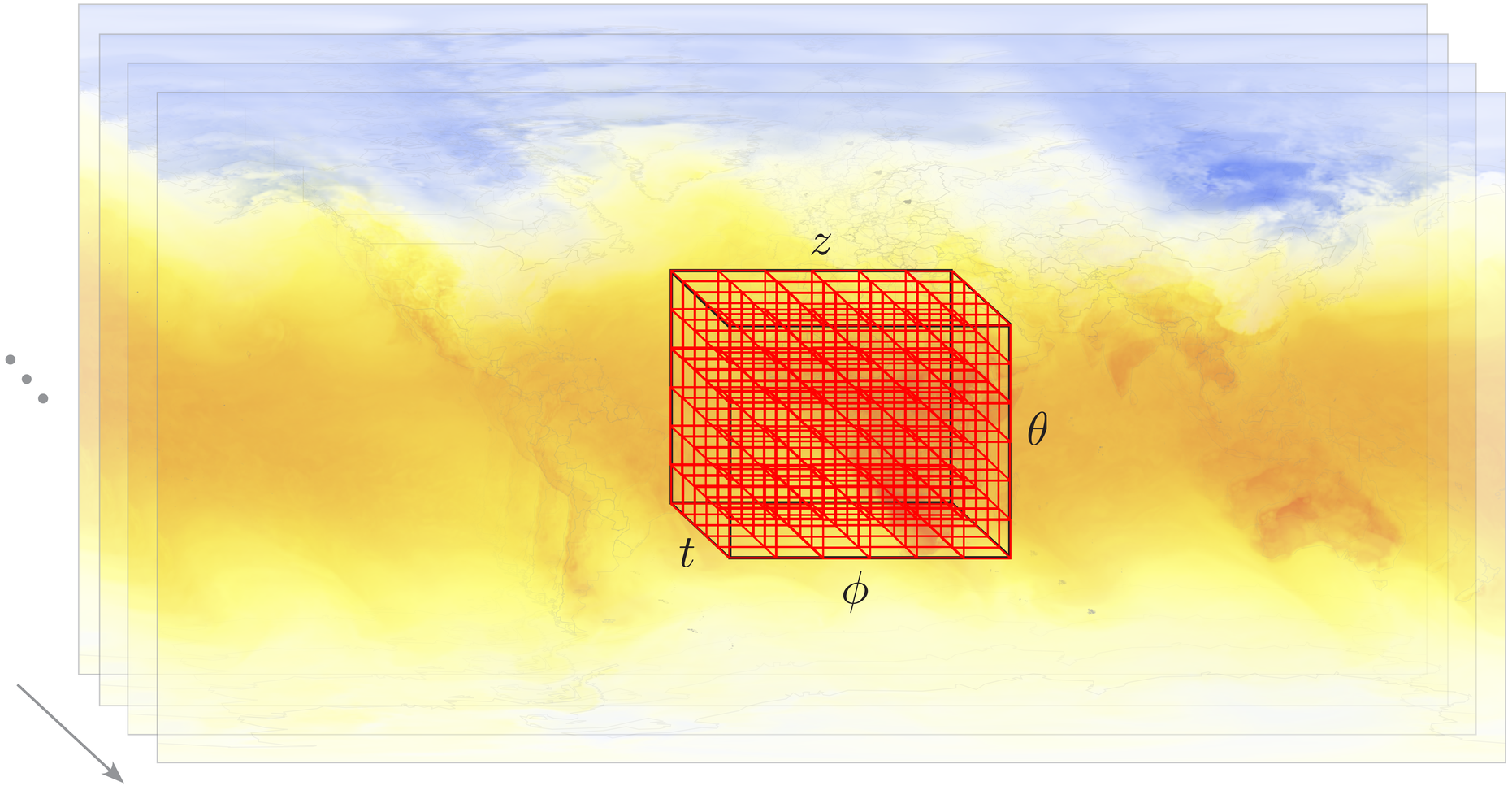
© Christian Lessig, 2023

What is a token?



© Christian Lessig, 2023

What is a token?



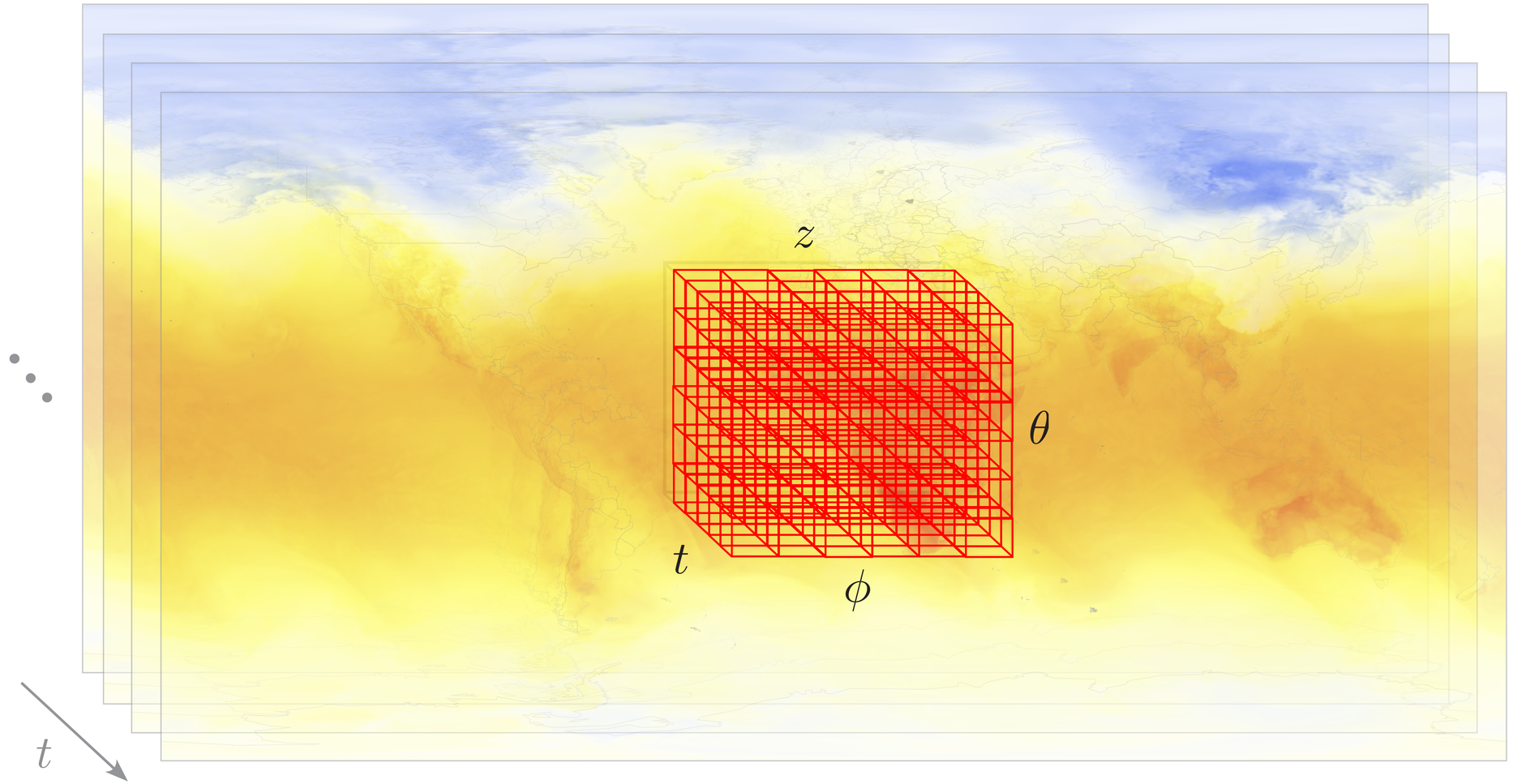
© Christian Lessig, 2023

Spatio-temporal BERT

- Self-supervised training with variation of BERT masked language (or token) model

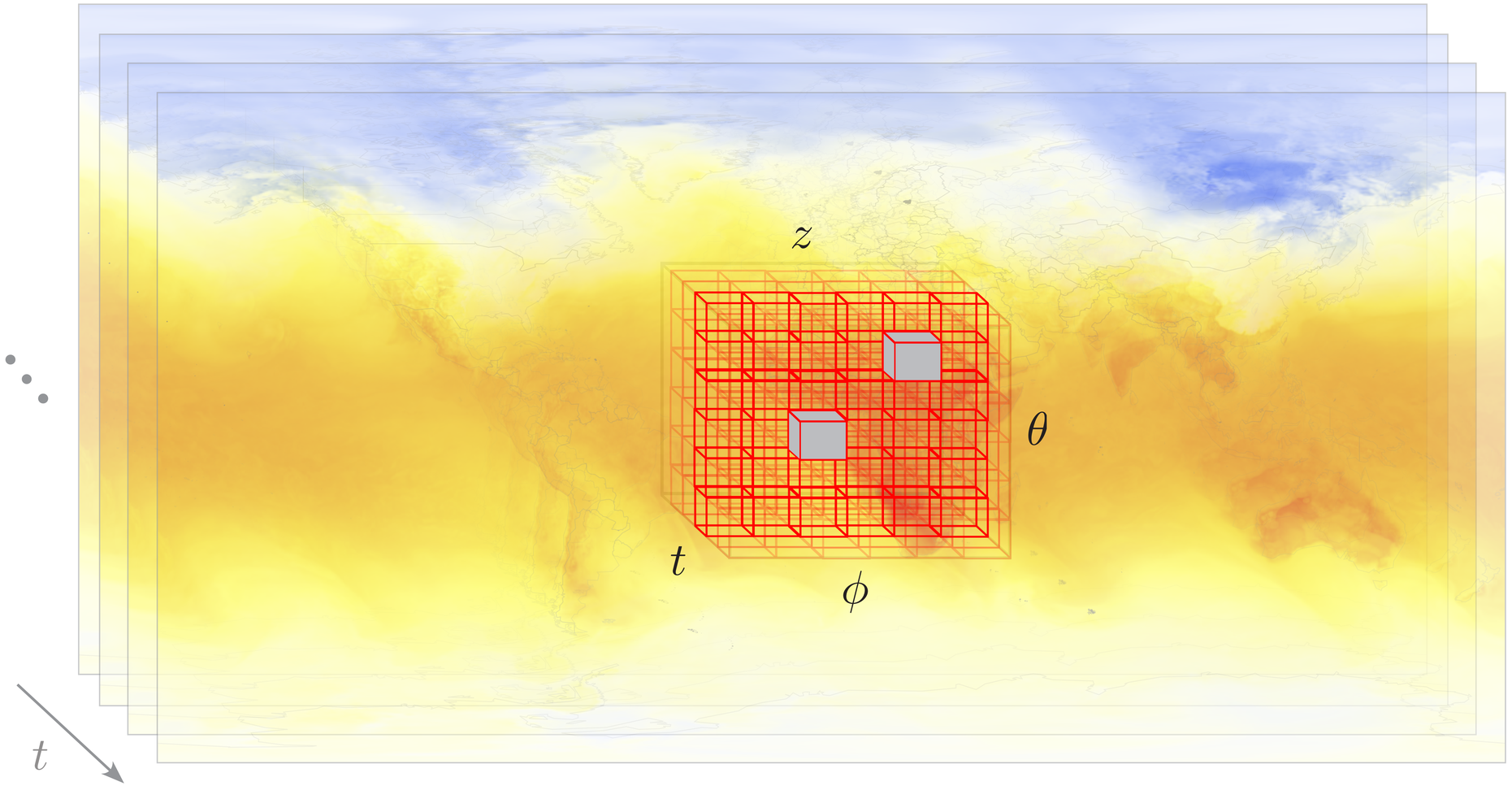
© Christian Lessig, 2023

Spatio-temporal BERT



© Christian Lessig, 2023

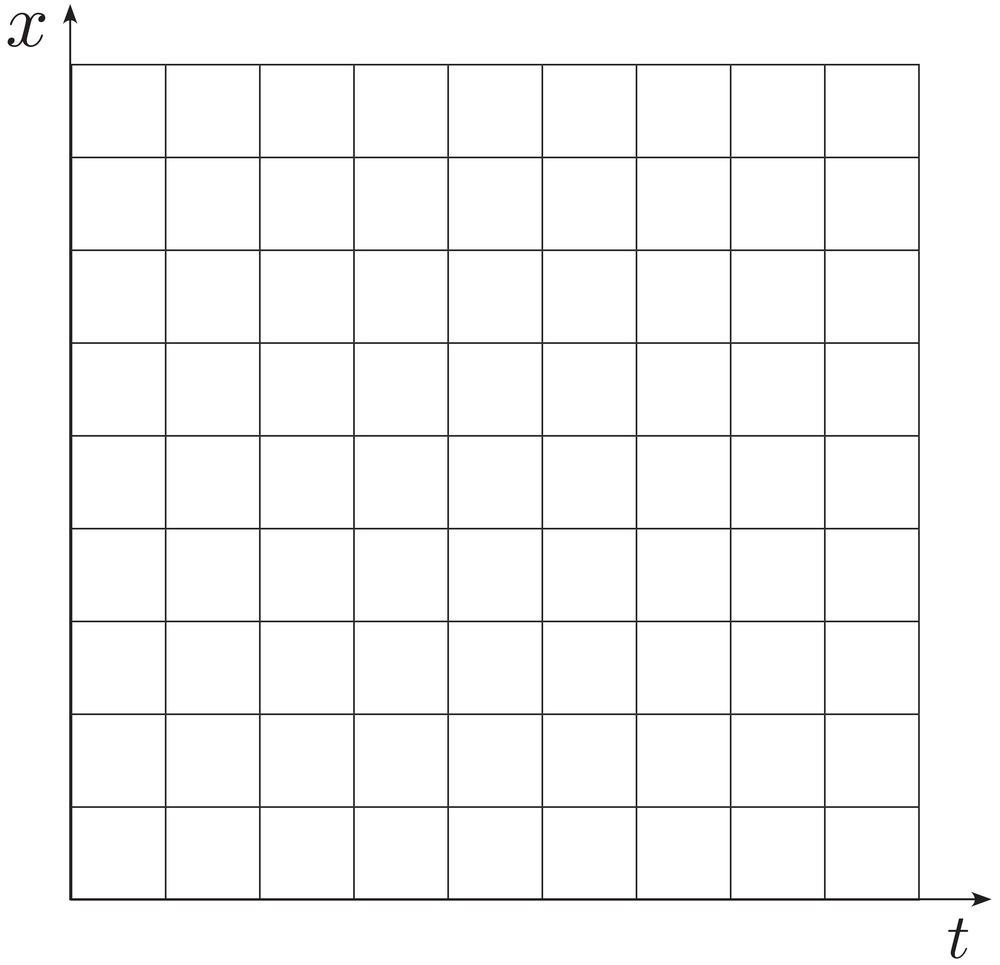
Spatio-temporal BERT



© Christian Lessig, 2023

Spatio-temporal BERT

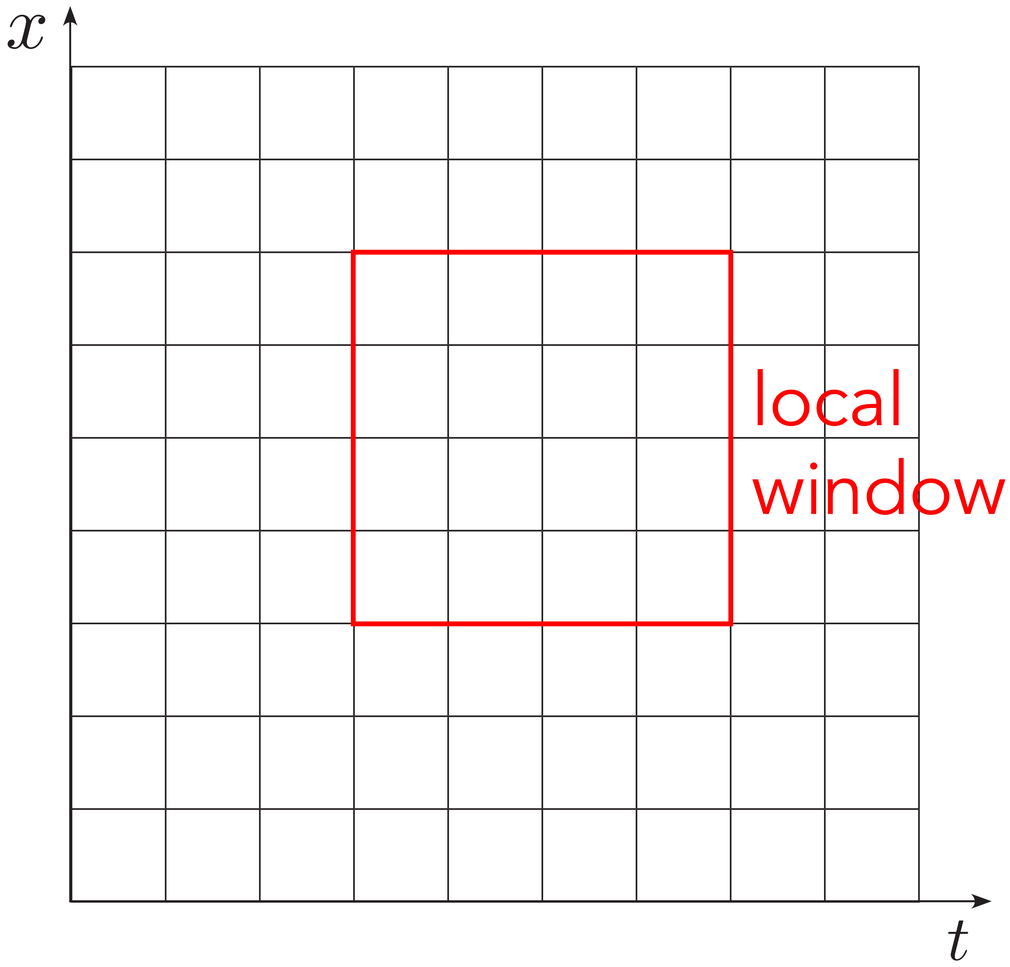
Flatland
view



© Christian Lessig, 2023

Spatio-temporal BERT

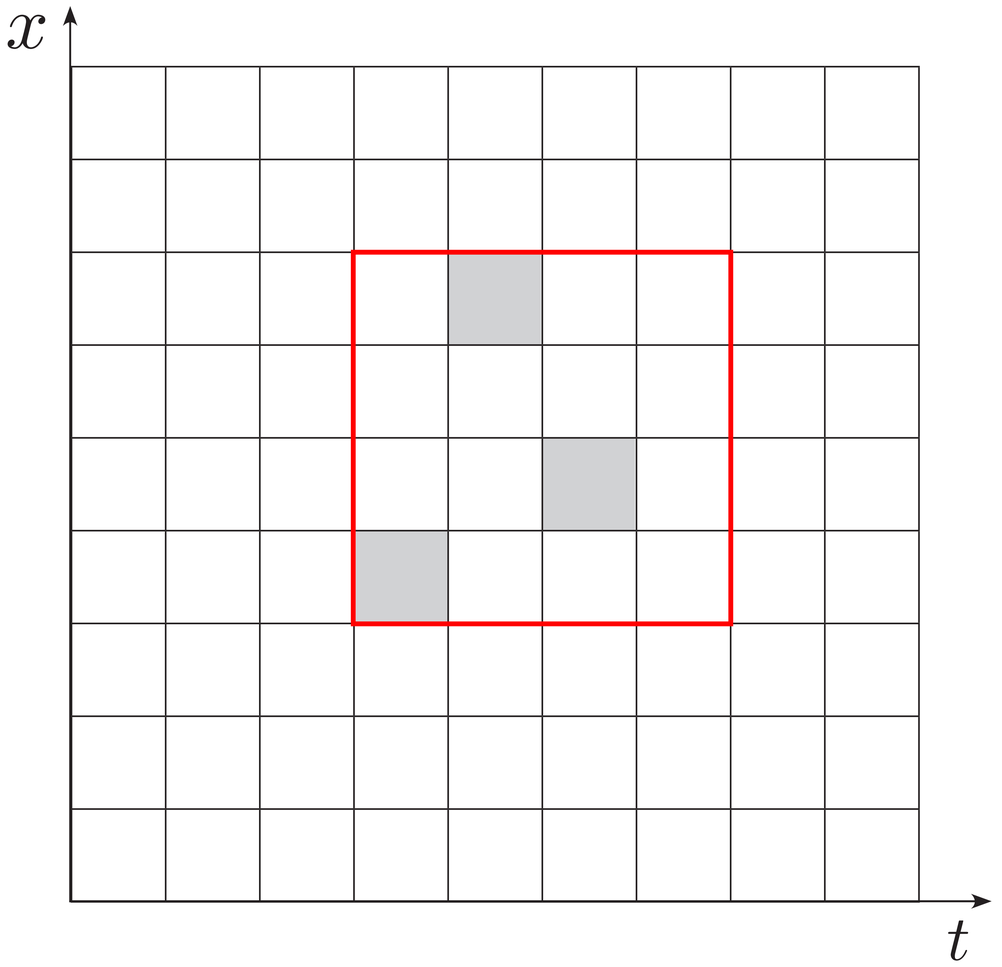
Flatland
view



© Christian Lessig, 2023

Spatio-temporal BERT

Flatland view of BERT



© Christian Lessig, 2023

Spatio-temporal BERT

- Self-supervised training with variation of BERT masked language language model
  - Natural interpretation as forecasting / hindcasting / interpolation

© Christian Lessig, 2023

Spatio-temporal BERT

- Self-supervised training with variation of BERT masked language language model
  - Natural interpretation as forecasting / hindcasting / interpolation
  - Random masking and distortions (noising, coarsening) ensures that a probabilistic model is learned

© Christian Lessig, 2023

Statistical loss

- Machine learning: Training on MSE loss is problematic in terms of training dynamics
  - One reason for overly smooth predicitons

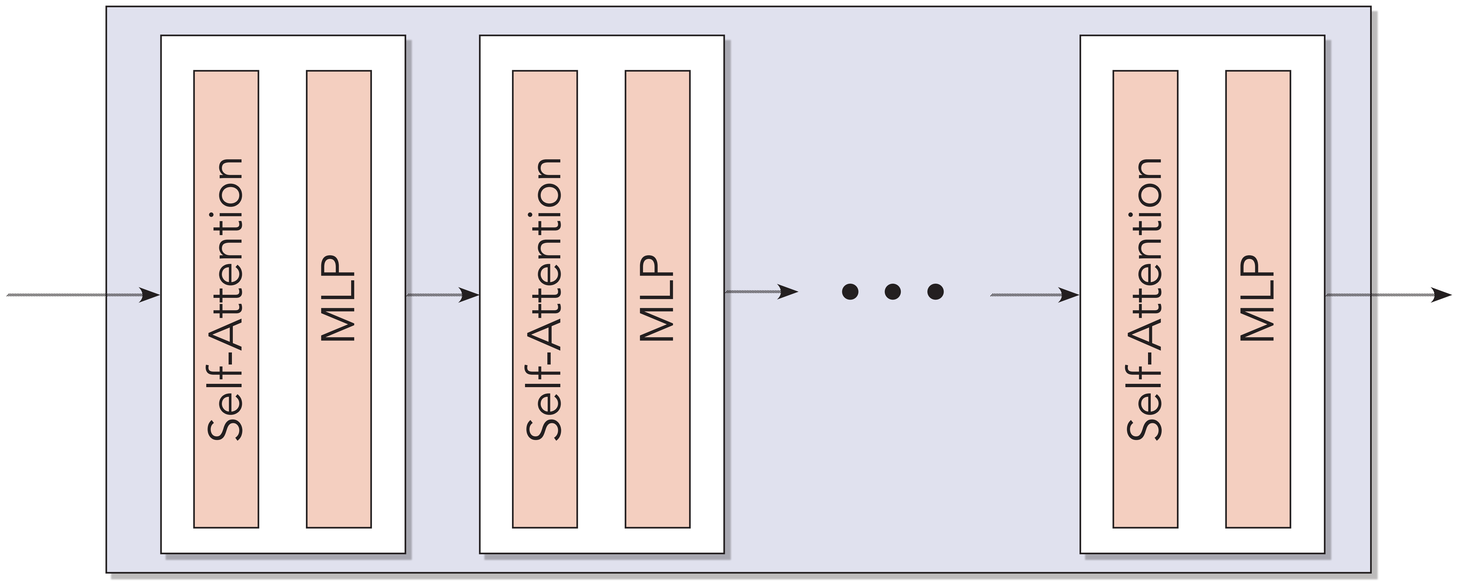
© Christian Lessig, 2023

Statistical loss

- Machine learning: Training on MSE loss is problematic in terms of training dynamics
- Training on just the mean is sub-optimal to learn a probabilitic/statistical representation of the dynamics and the system

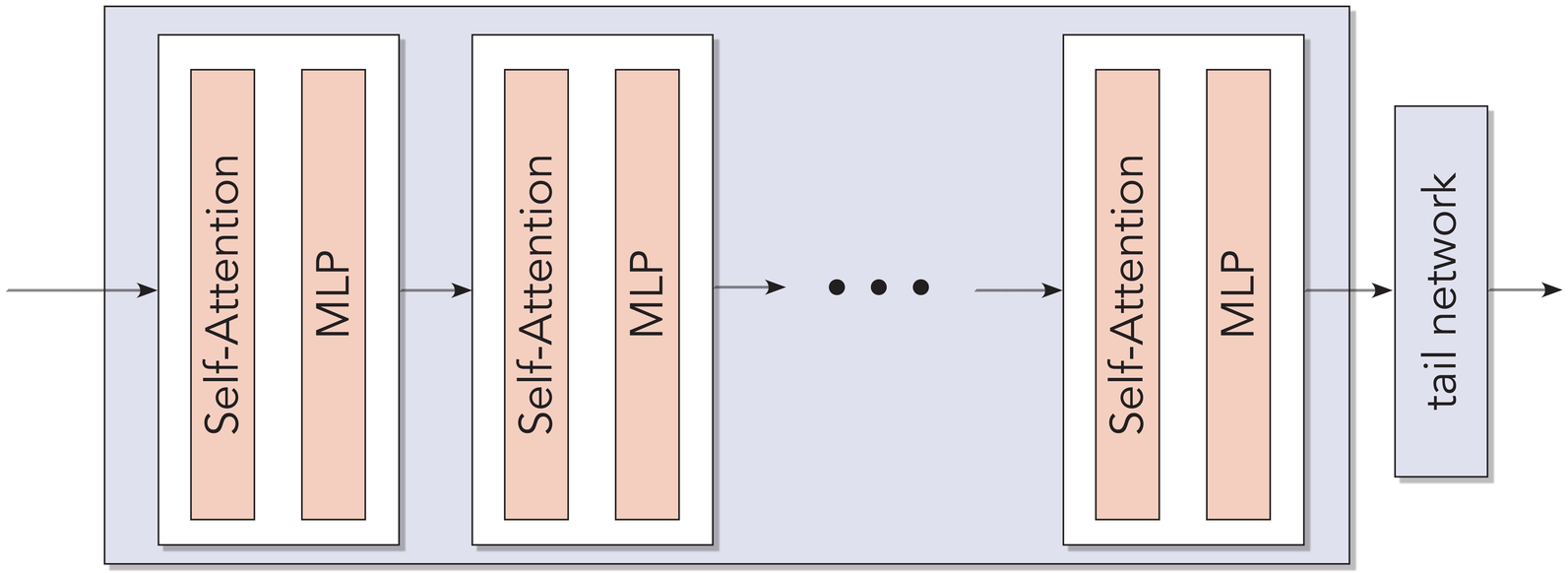
© Christian Lessig, 2023

Statistical loss



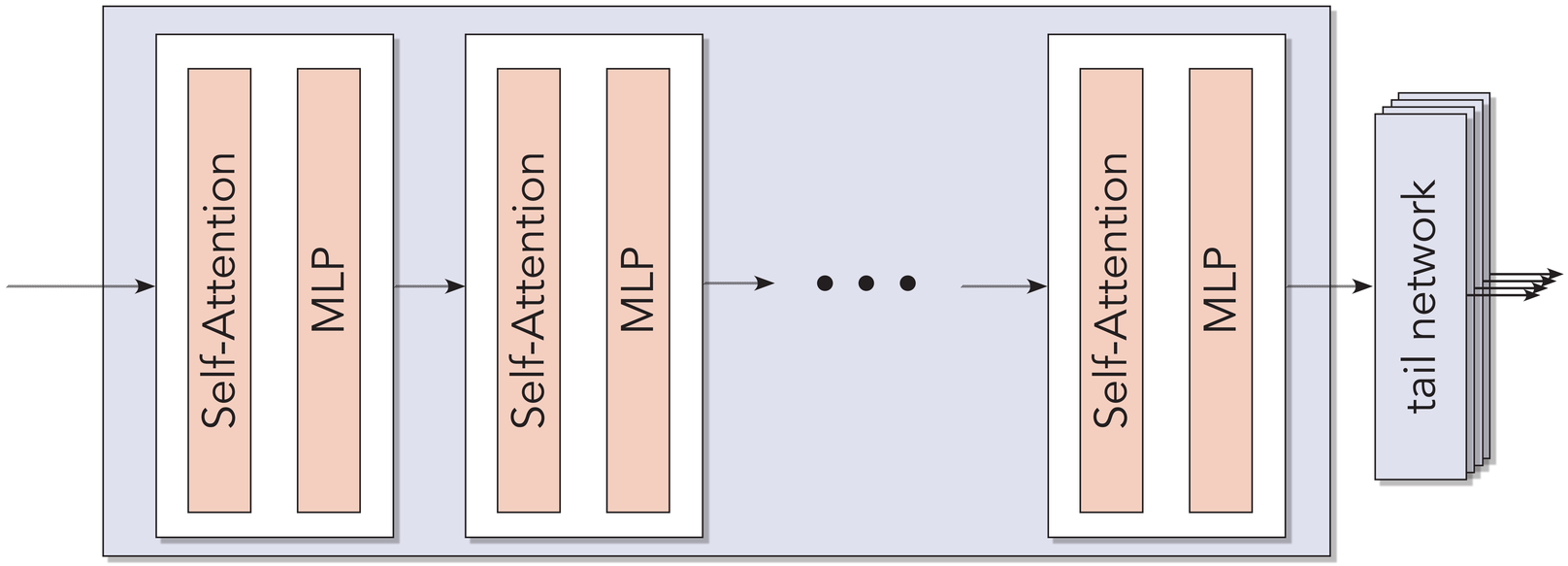
© Christian Lessig, 2023

Statistical loss



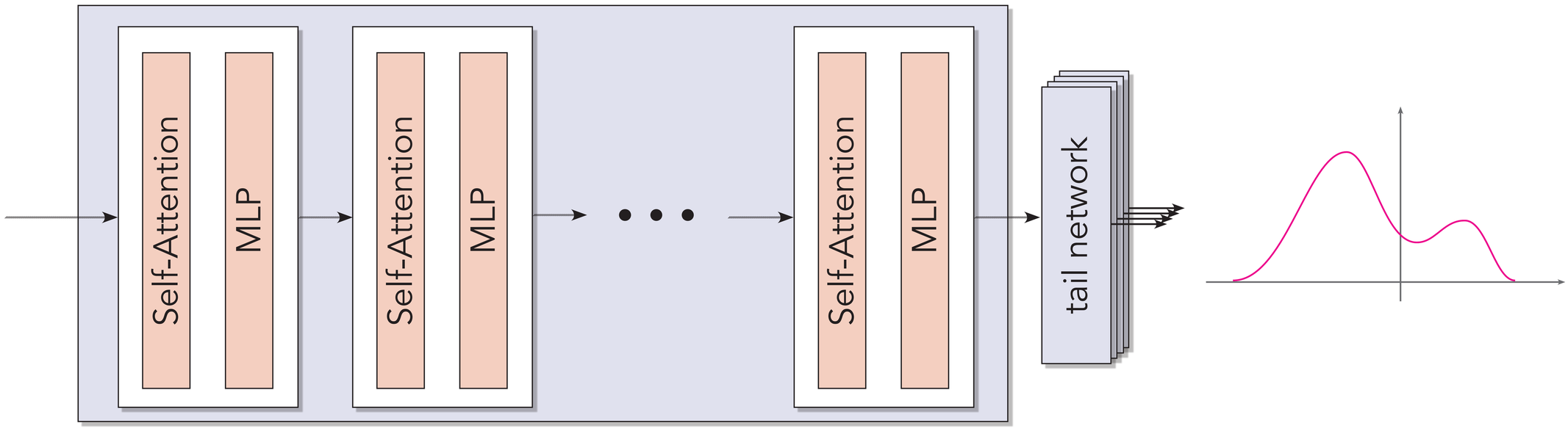
© Christian Lessig, 2023

Statistical loss



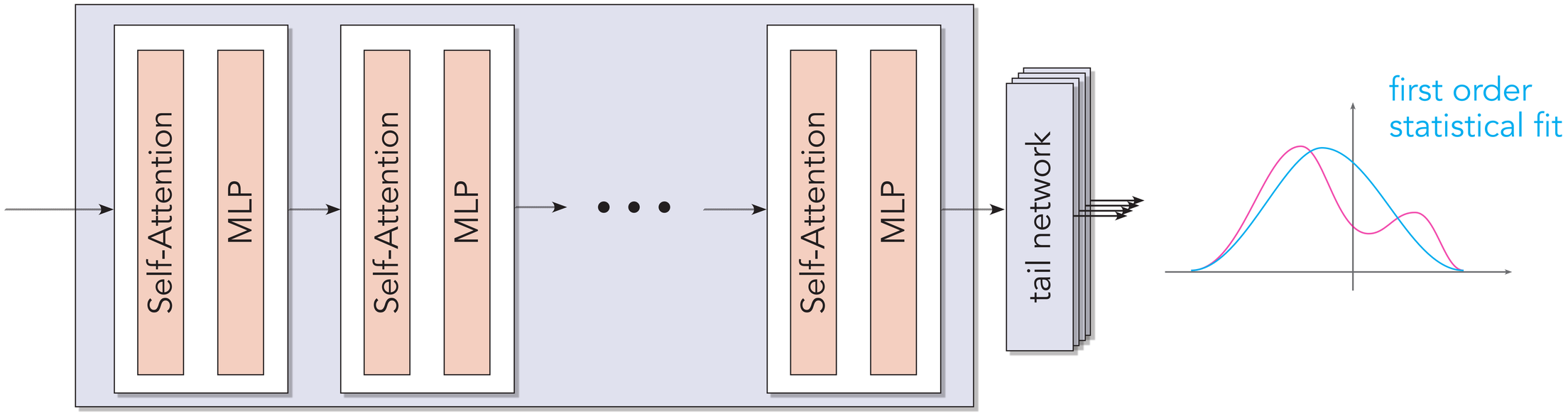
© Christian Lessig, 2023

Statistical loss



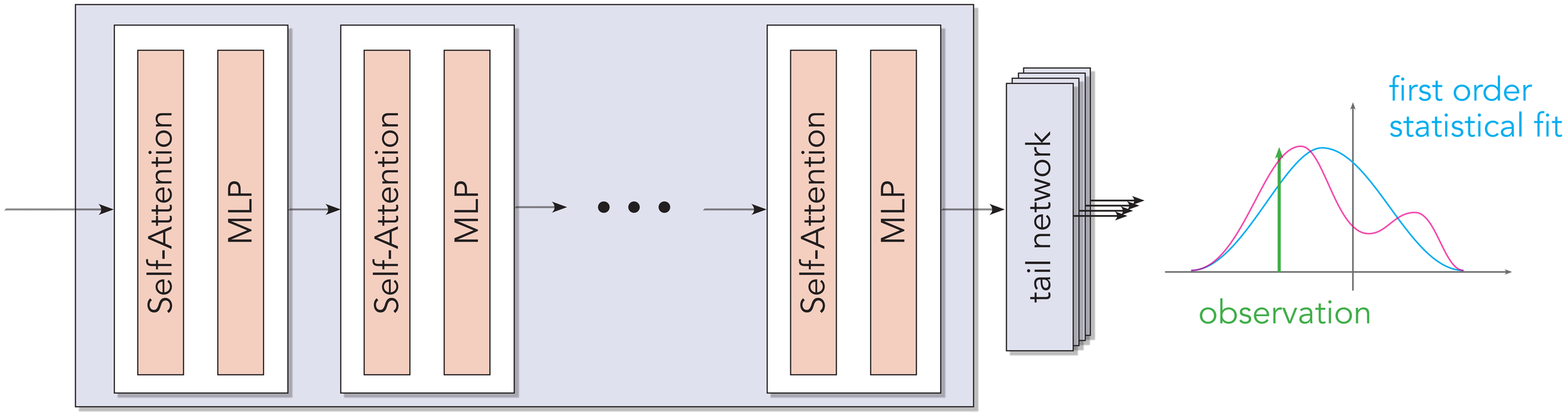
© Christian Lessig, 2023

Statistical loss



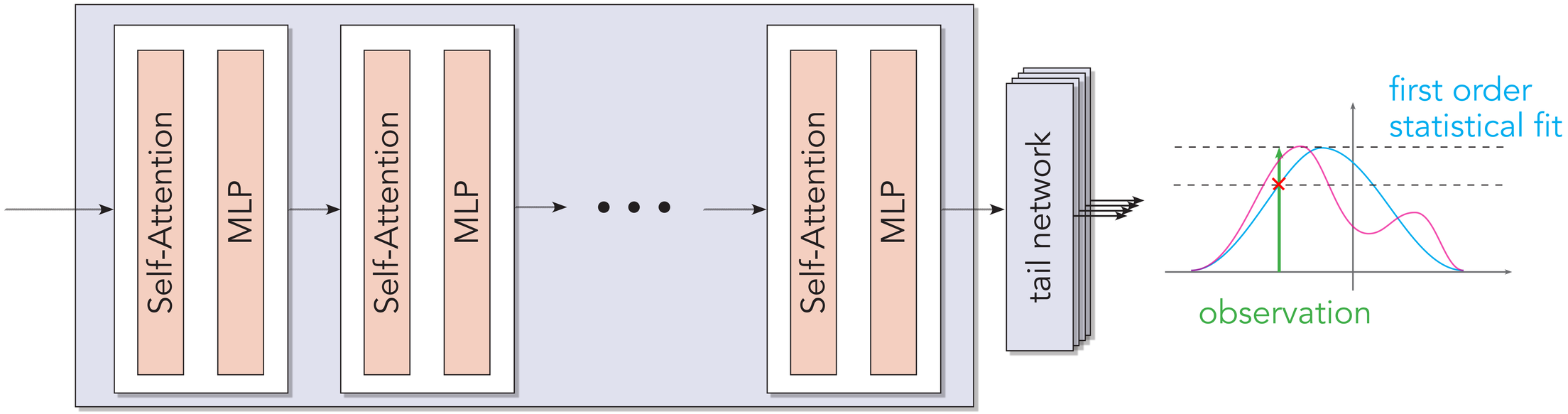
© Christian Lessig, 2023

Statistical loss



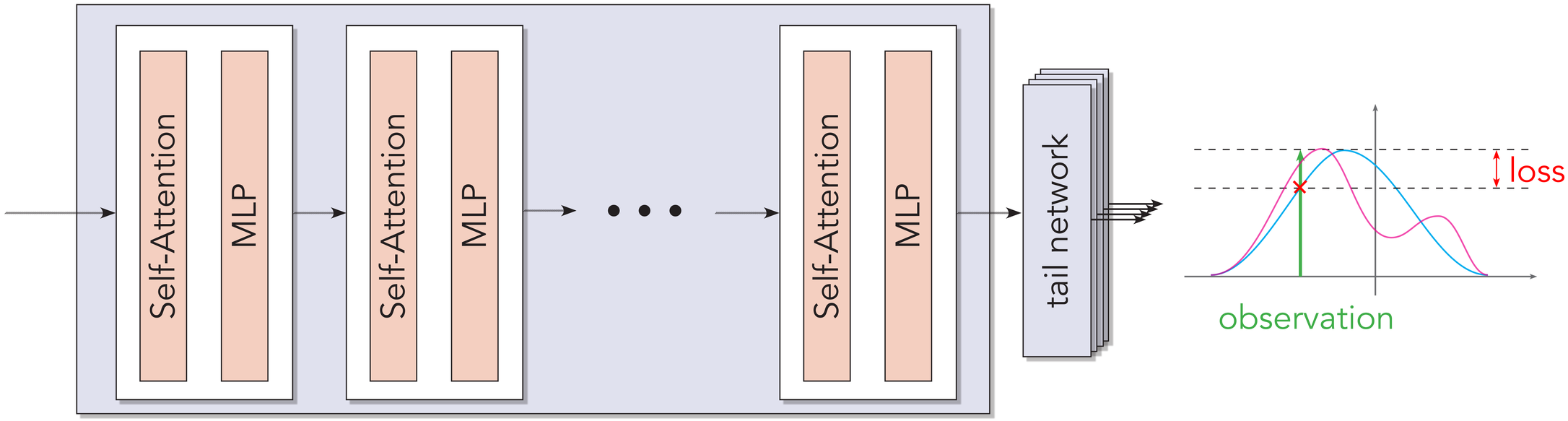
© Christian Lessig, 2023

Statistical loss



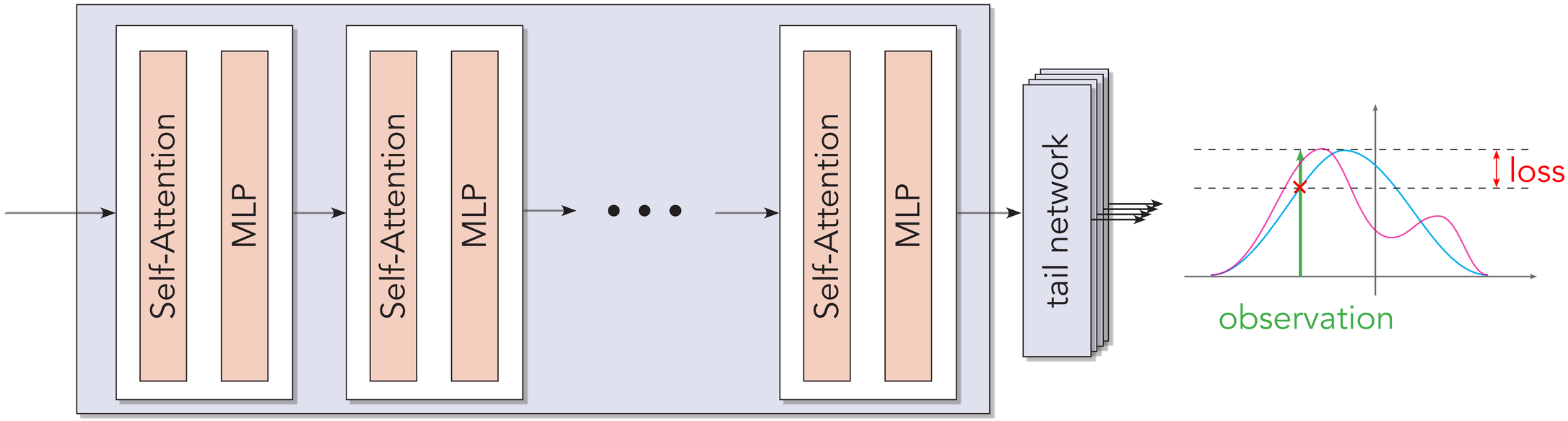
© Christian Lessig, 2023

Statistical loss



© Christian Lessig, 2023

Statistical loss



end-to-end training encourages statistical representation

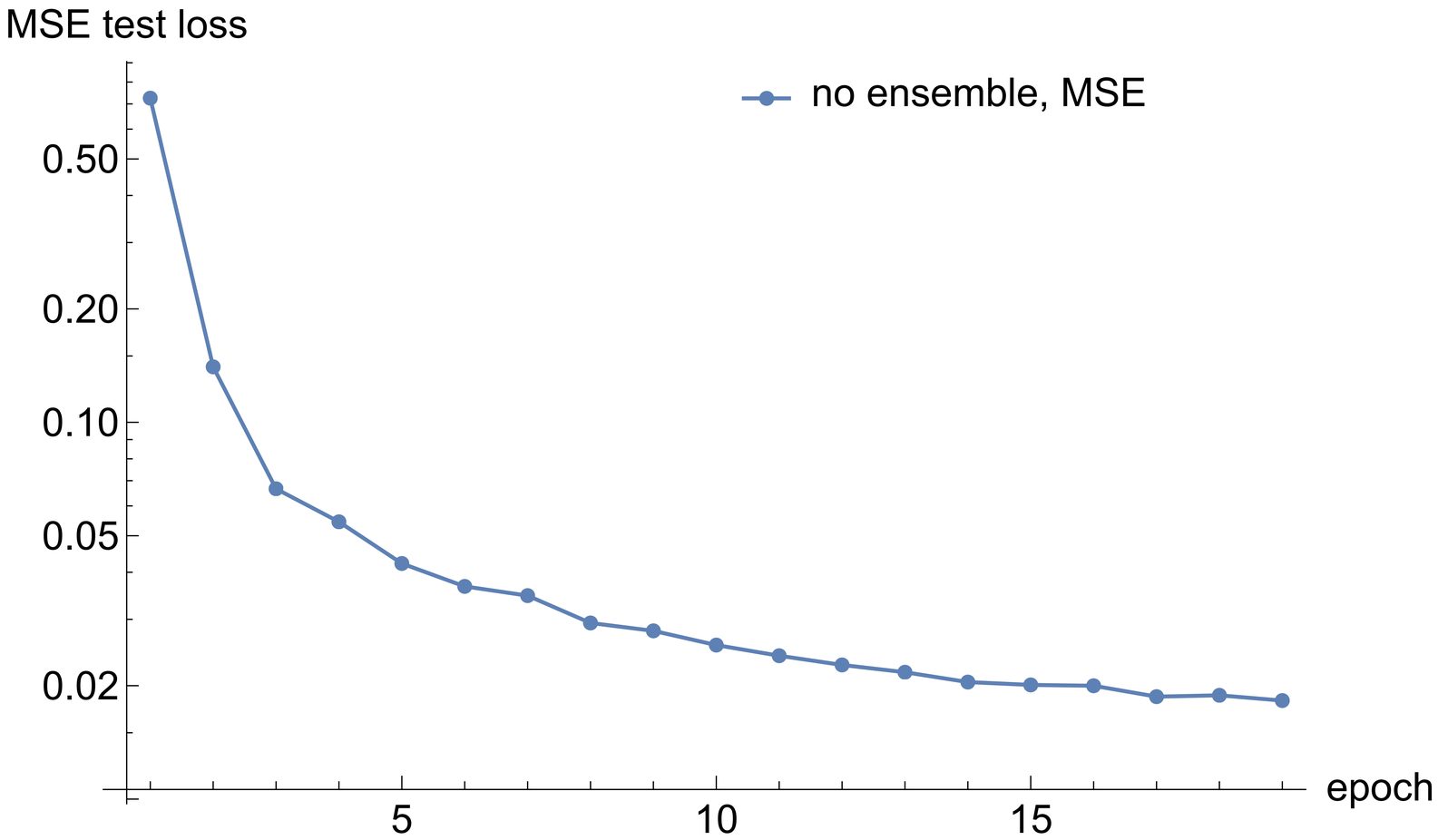
© Christian Lessig, 2023

Statistical loss: experiments

- BERT with conditional masking
- 975 hPa (high frequency) vorticity
- 40 years of training data

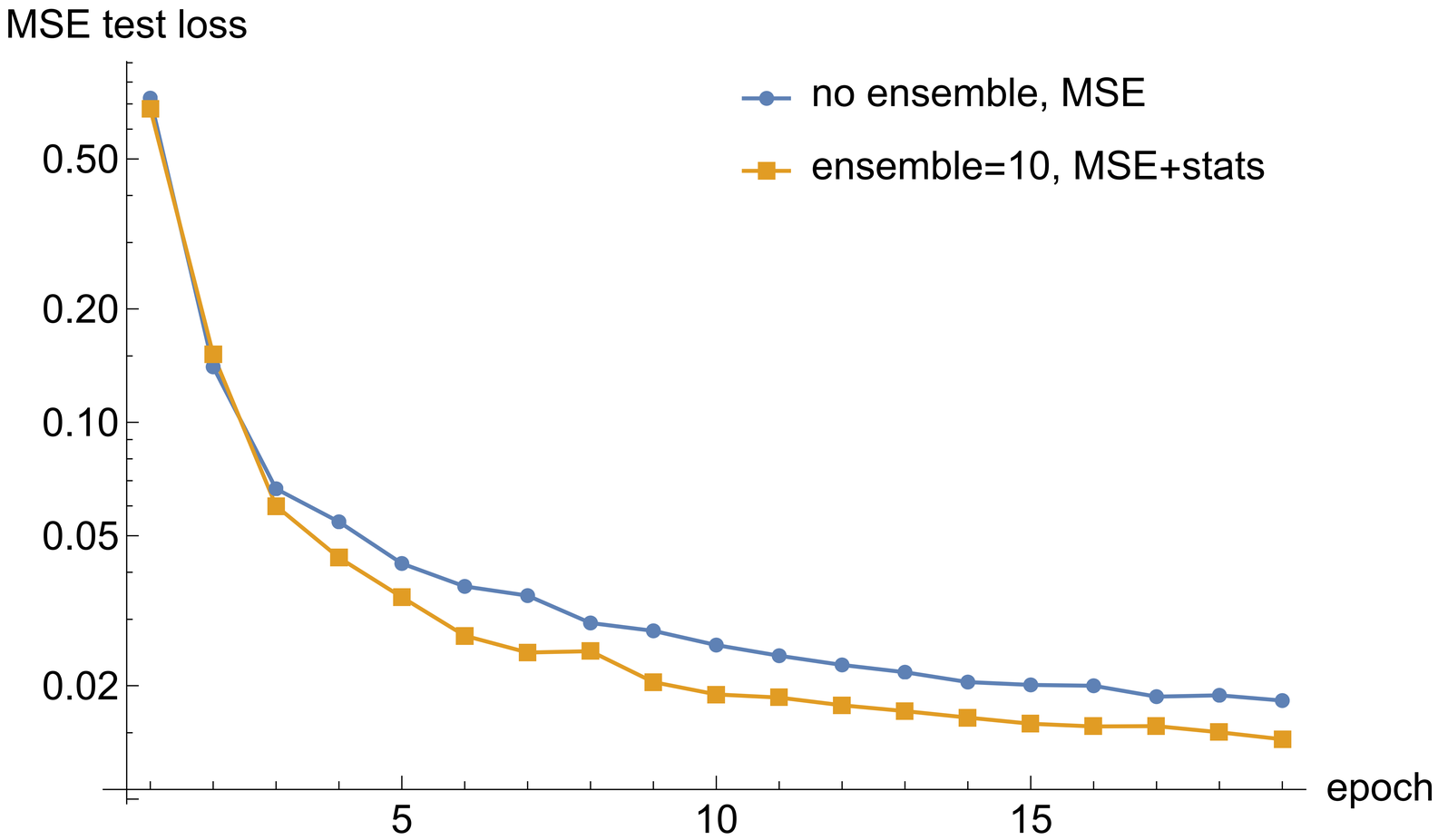
© Christian Lessig, 2023

Statistical loss



© Christian Lessig, 2023

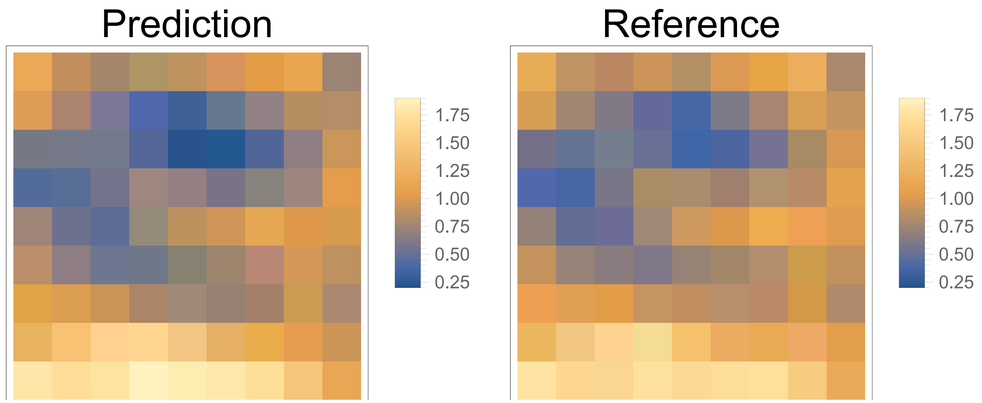
Statistical loss

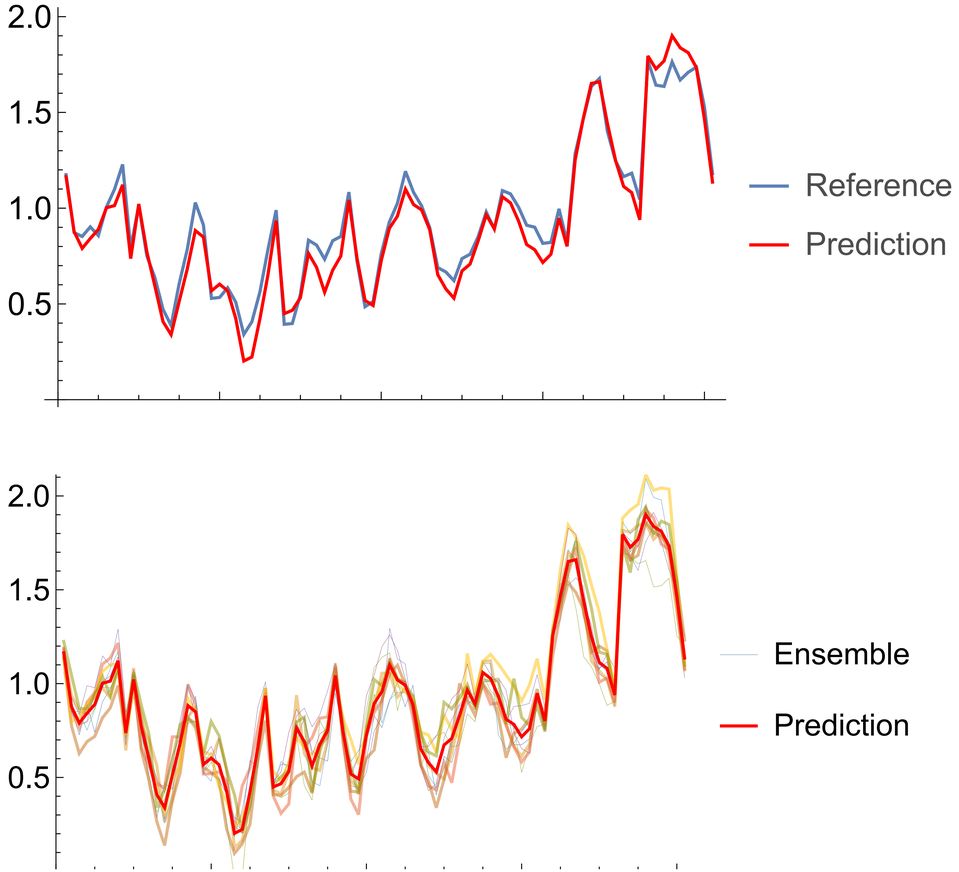


© Christian Lessig, 2023

Statistical loss

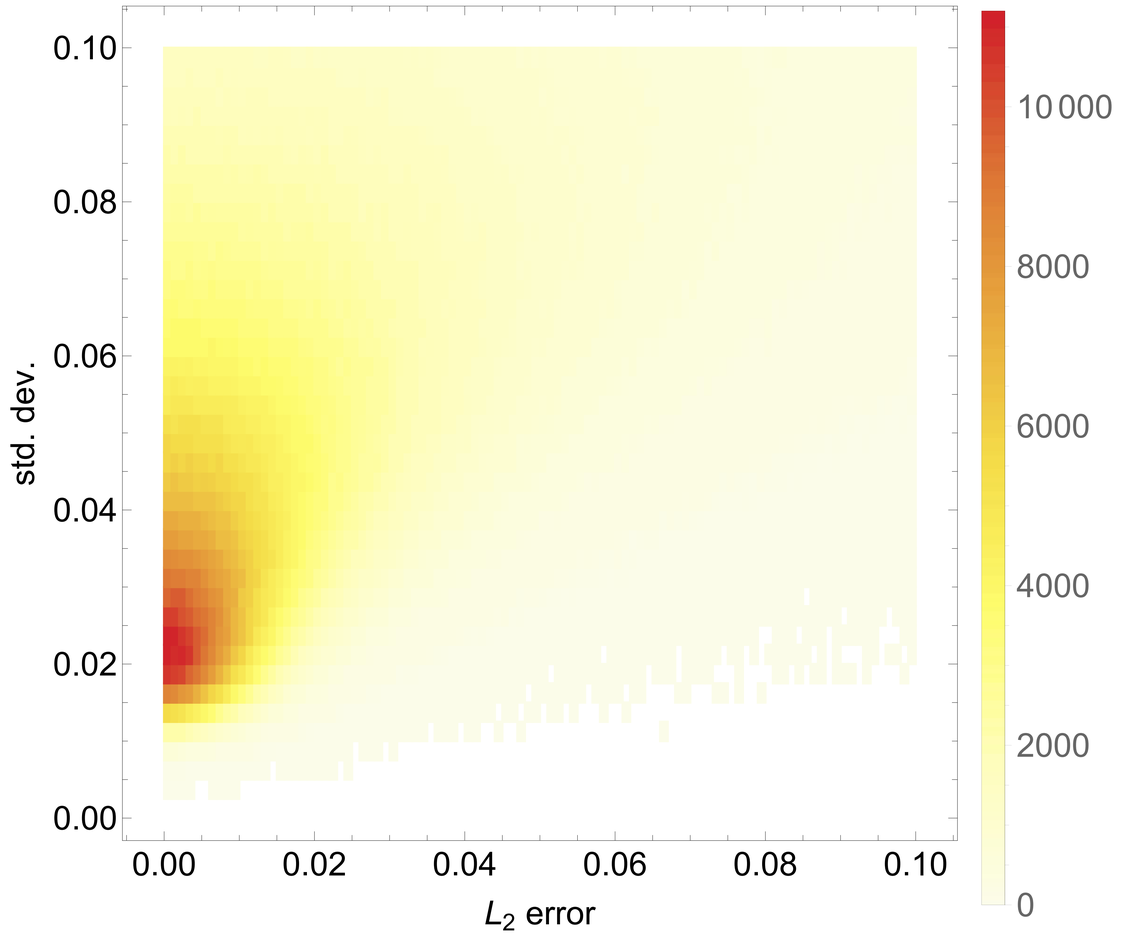
- Predictions:





© Christian Lessig, 2023

Statistical loss



2D Histogram of L_2 error vs. std. dev.

© Christian Lessig, 2023

AtmoRep: in-context learning

- In-context learning: ability to solve tasks without training with zero-/few-shot evaluation

© Christian Lessig, 2023

AtmoRep: in-context learning

- In-context learning: ability to solve tasks without training with zero-/few-shot evaluation
  - Language models: chat programs, translation, auto-correction, ... from training on next sentence prediction task
  - Natural language used to specify task

© Christian Lessig, 2023

AtmoRep: in-context learning

- In-context learning: ability to solve tasks without training with zero-/few-shot evaluation
  - Language models: chat programs, translation, auto-correction, ... from training on next sentence prediction task
  - Natural language used to specify task

What is in-context learning for AtmoRep $p_\theta\left(y|x,\alpha\right)$?

© Christian Lessig, 2023

AtmoRep: in-context learning

- The model $p_\theta\left(y|x,\alpha\right)$ implies that what we want to "control" is the output state y without re-learning

© Christian Lessig, 2023

AtmoRep: in-context learning

- The model $p_\theta\left(y|x,\alpha\right)$ implies that what we want to "control" is the output state y without re-learning

$$p_\theta\left(y|x,\alpha,\tau\right)$$

positional encoding
for output

© Christian Lessig, 2023

AtmoRep: in-context learning

- The model $p_\theta\left(y|x,\alpha\right)$ implies that what we want to "control" is the output state y without re-learning

$$p_\theta\left(y|x,\alpha,\tau\right)$$

  - τ : spatial and temporal location, resolution, quality
  - Few shot: "explain" τ to the network

© Christian Lessig, 2023

AtmoRep: in-context learning

- The model $p_\theta\left(y|x,\alpha\right)$ implies that what we want to "control" is the output state y without re-learning

$$p_\theta\left(y|x,\alpha,\tau\right)$$

  - τ : spatial and temporal location, resolution, quality
  - Few shot: "explain" τ to the network

- Does this make the network a scientific model?

© Christian Lessig, 2023

“I confess even to this day that I still don’t understand quantum mechanics, and I’m not even sure I really know how to use it all that well. And a lot of this has to do with the fact that I still don’t understand it.”

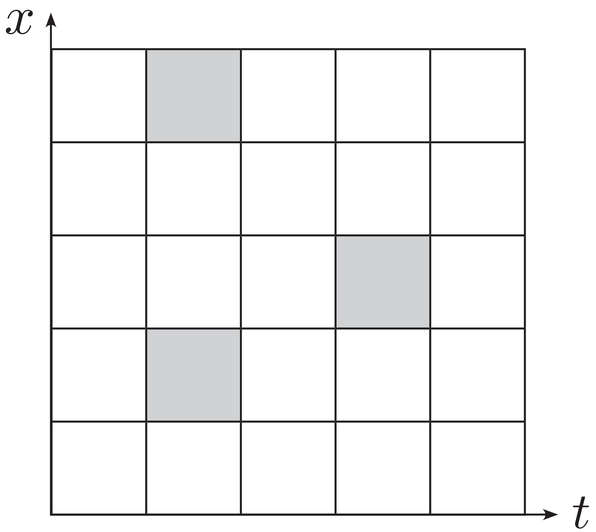
John Clauser, 2002

Quoted from https://www.nytimes.com/2022/10/04/science/nobel-prize-physics-winner.html

© Christian Lessig, 2023

Zero shot forecasting

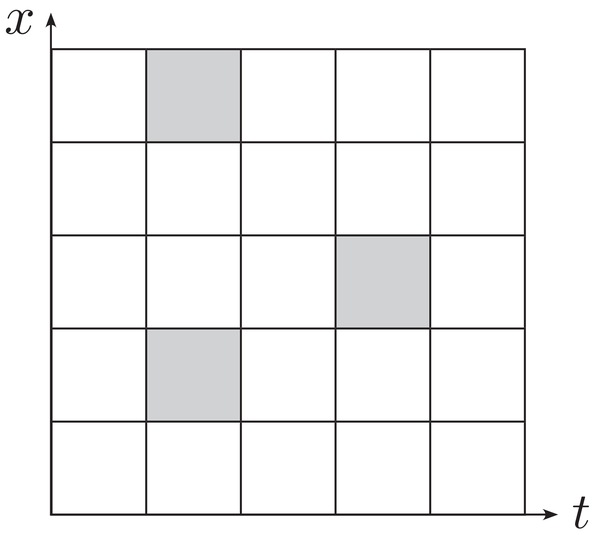
BERT



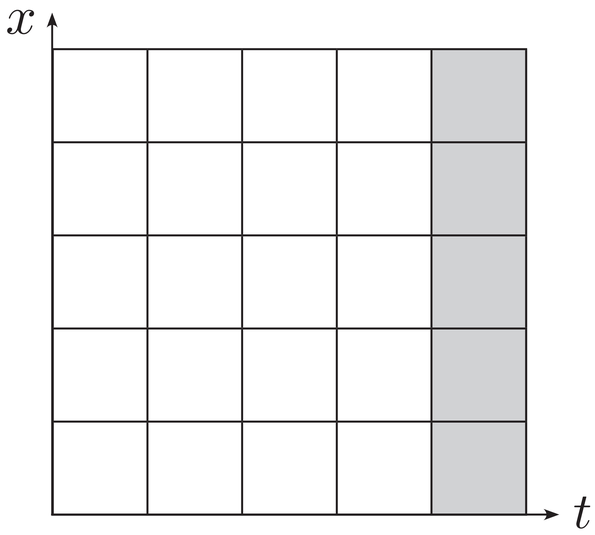
© Christian Lessig, 2023

Zero shot forecasting

BERT



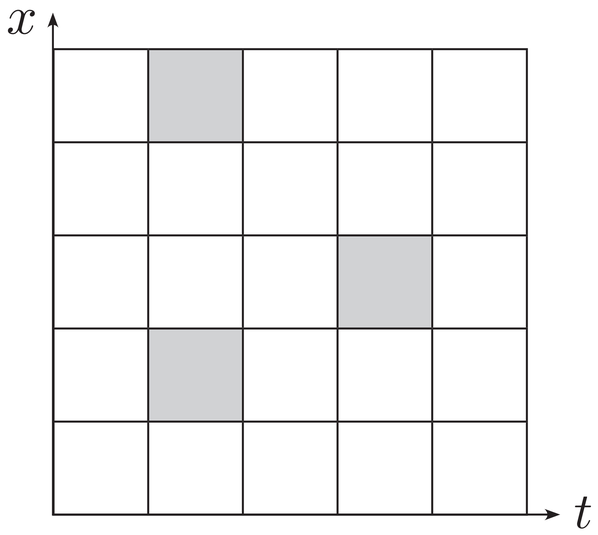
BERT-Forecast



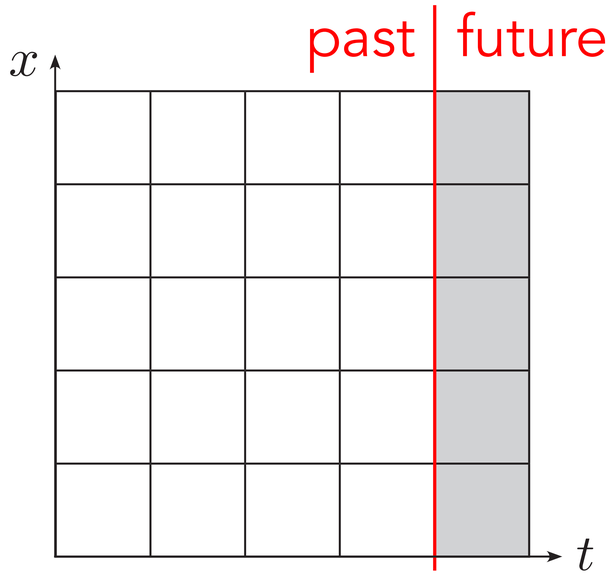
© Christian Lessig, 2023

Zero shot forecasting

BERT

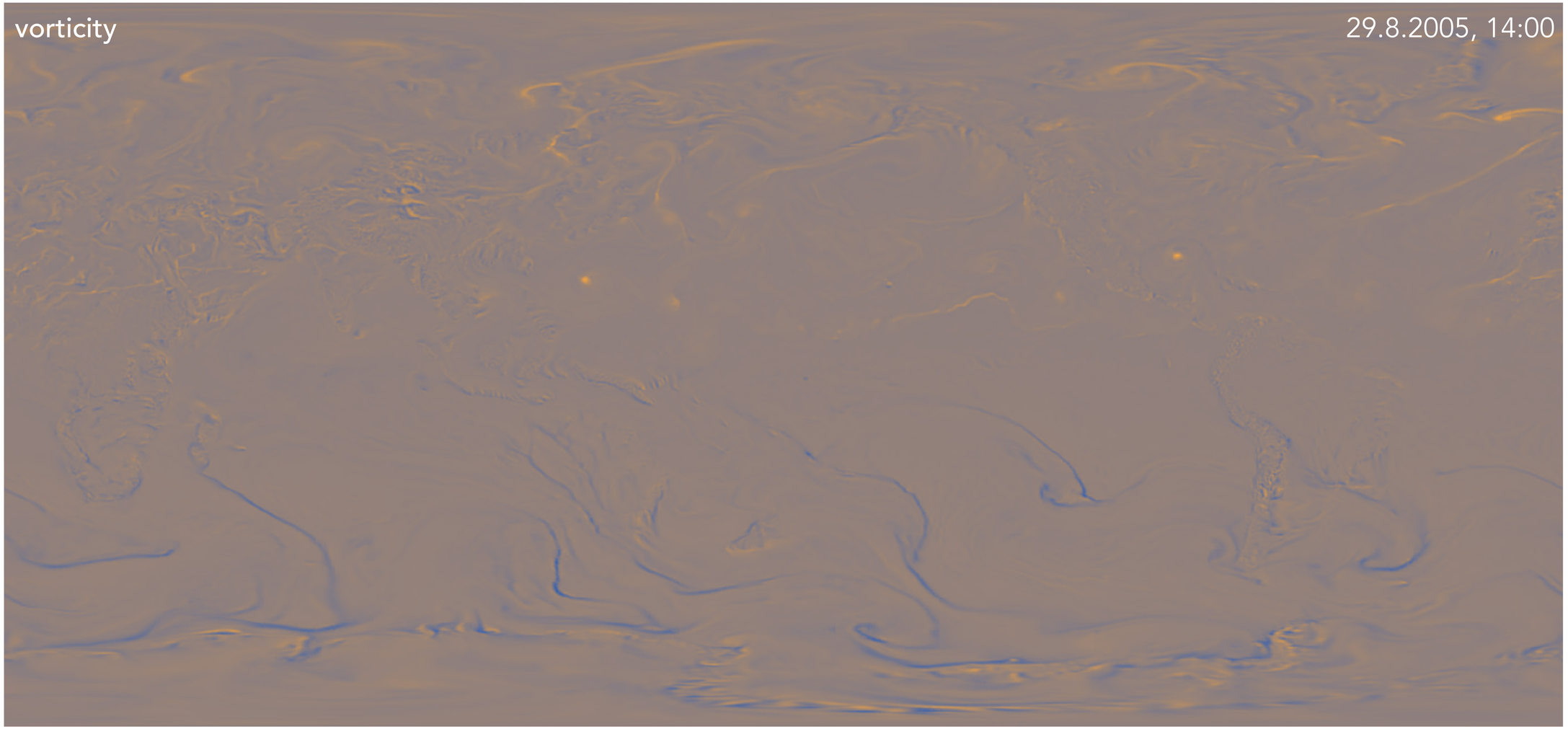


BERT-Forecast



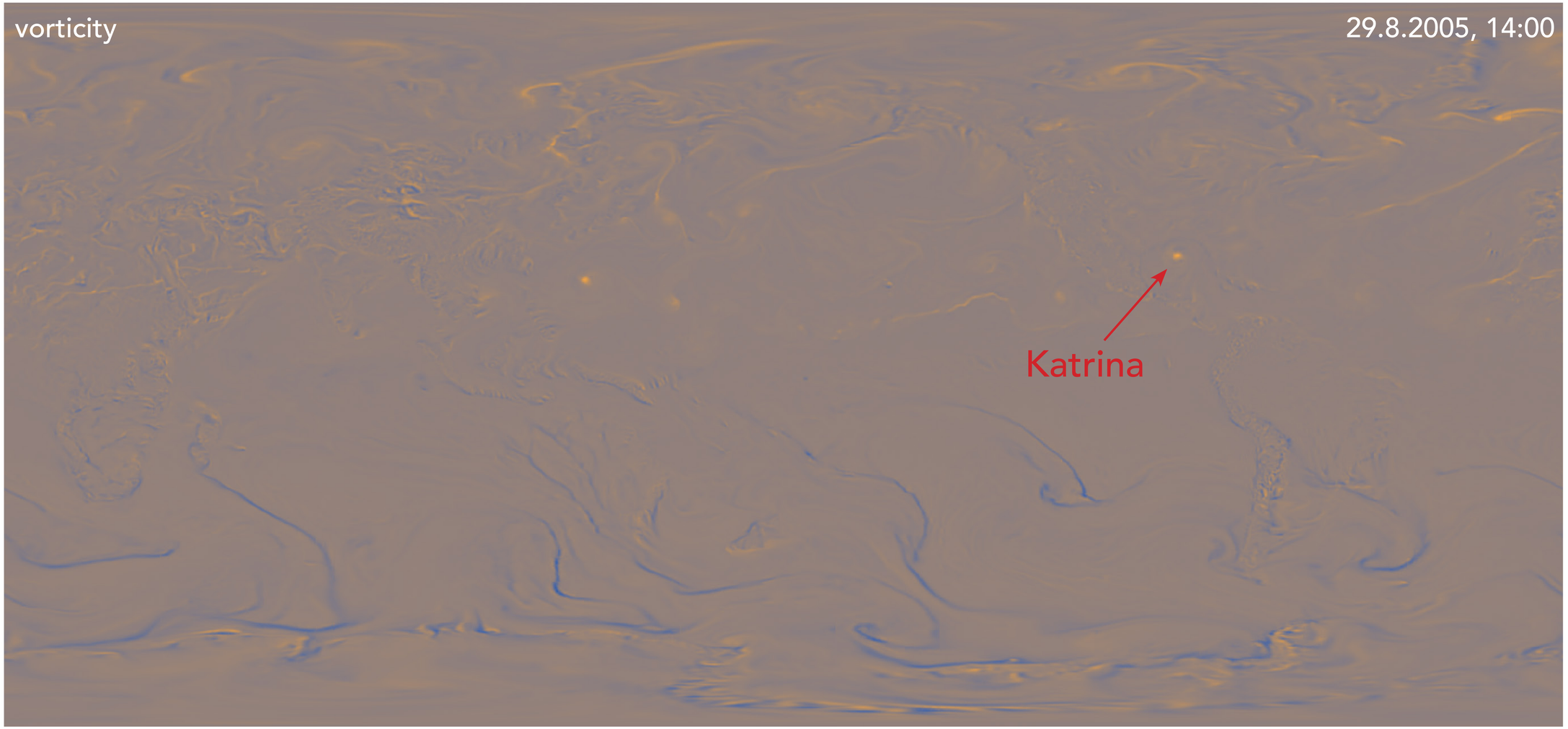
© Christian Lessig, 2023

Zero shot forecasting



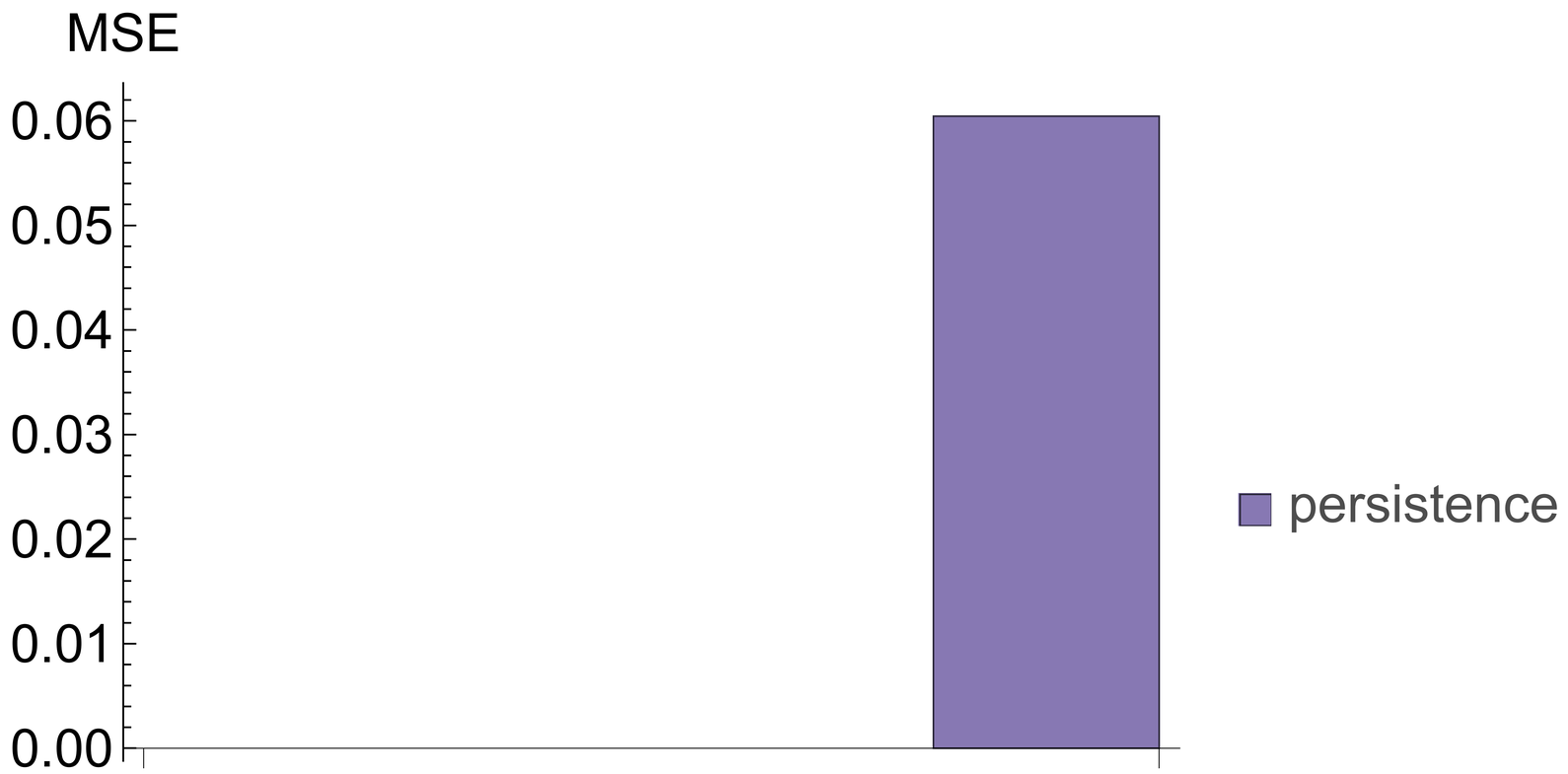
© Christian Lessig, 2023

Zero shot forecasting



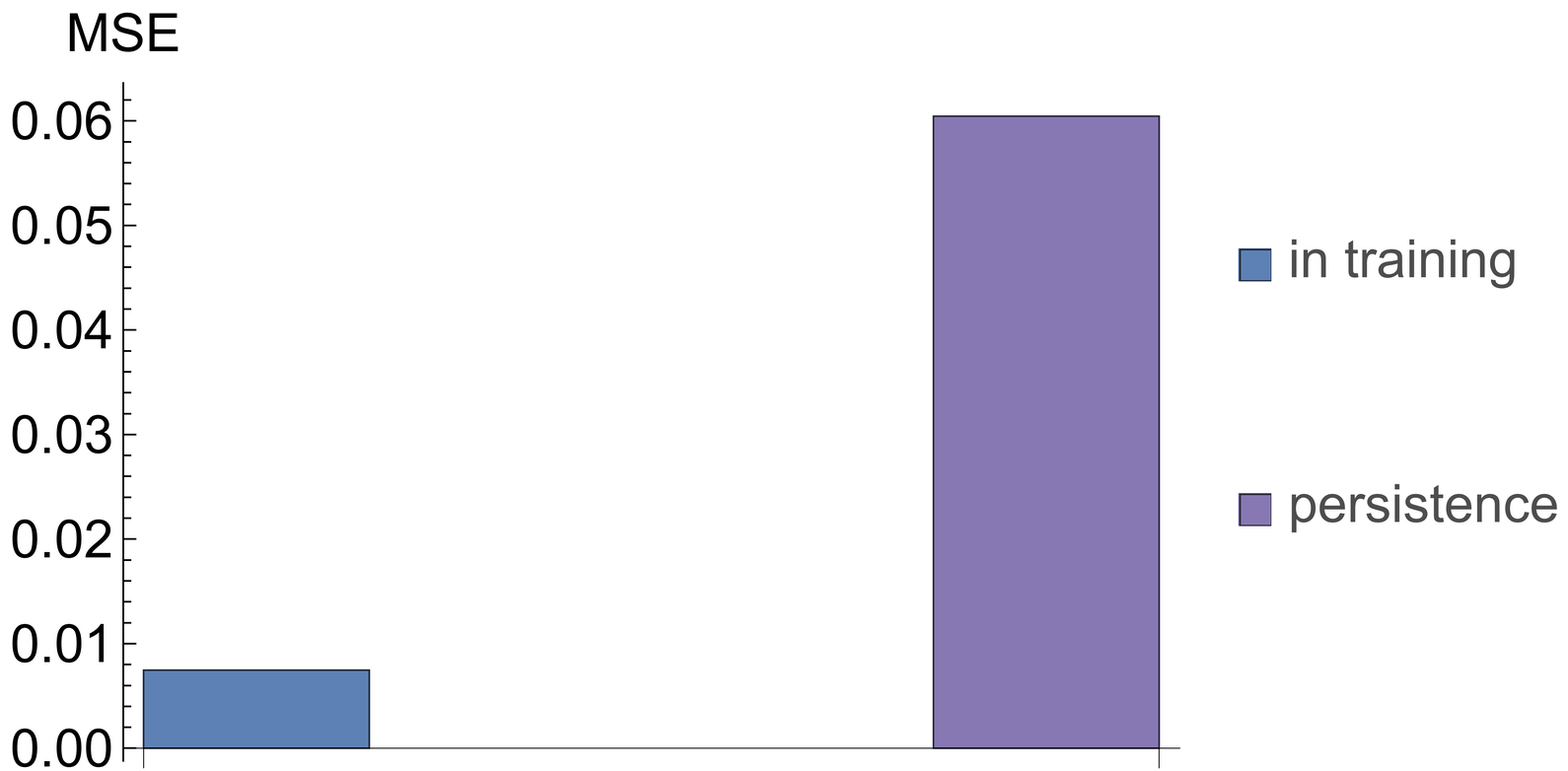
© Christian Lessig, 2023

Zero shot forecasting



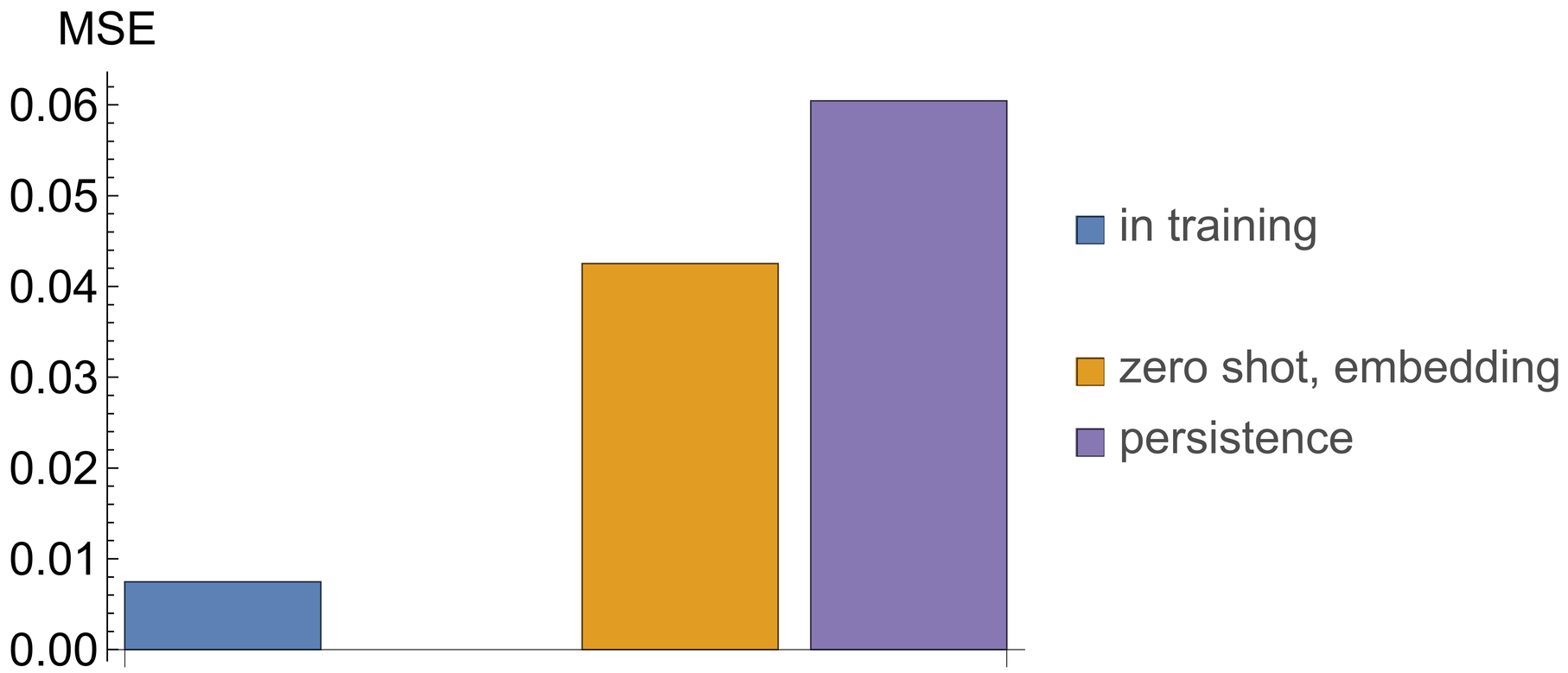
© Christian Lessig, 2023

Zero shot forecasting



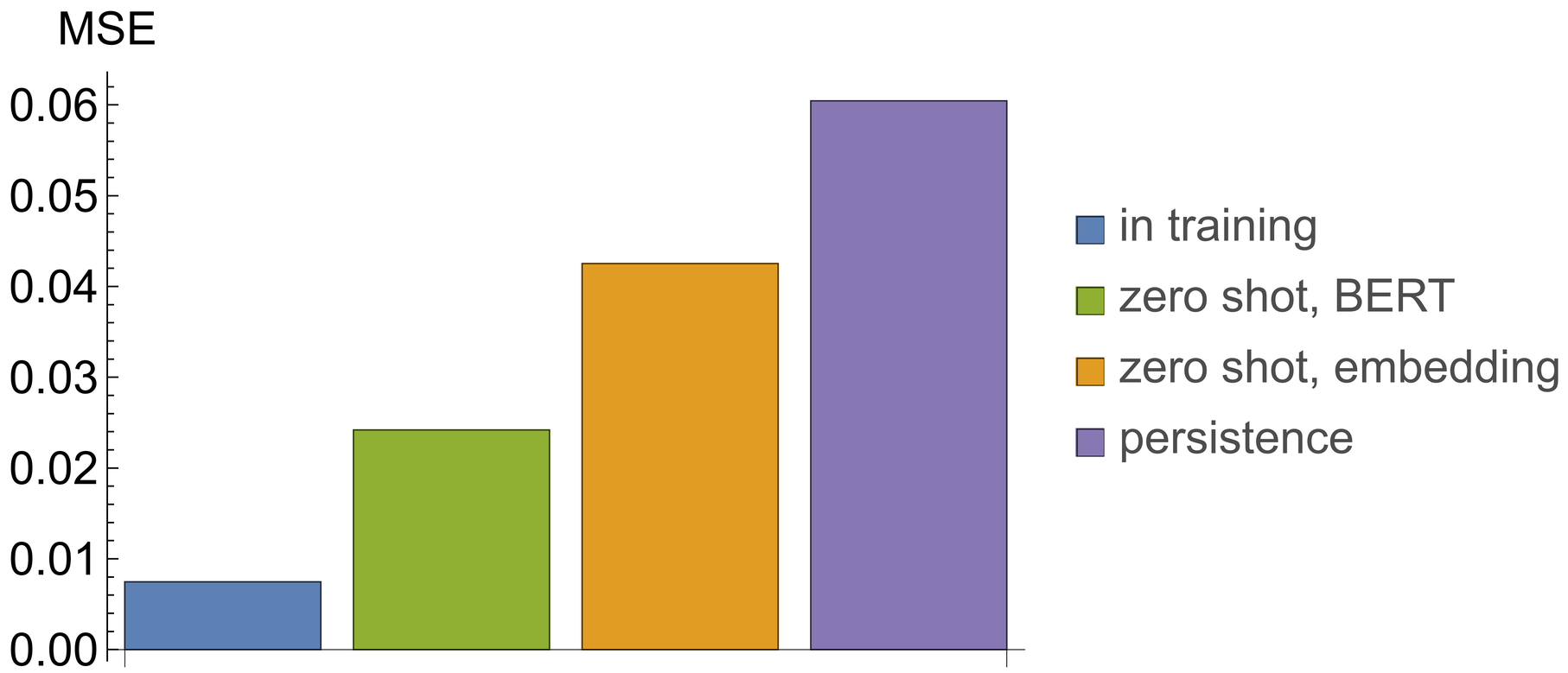
© Christian Lessig, 2023

Zero shot forecasting



© Christian Lessig, 2023

Zero shot forecasting



© Christian Lessig, 2023

Representation learning for science

- Learn a **domain-specific** but **task-independent** neural network that is useful for a range of applications
  - Self-supervised training on very large amounts of data with very large networks
- Many scientific disciplines have large amounts of unlabelled observational data
  - BERT-like masked auto-encoder task is very general and meaningful in many domains

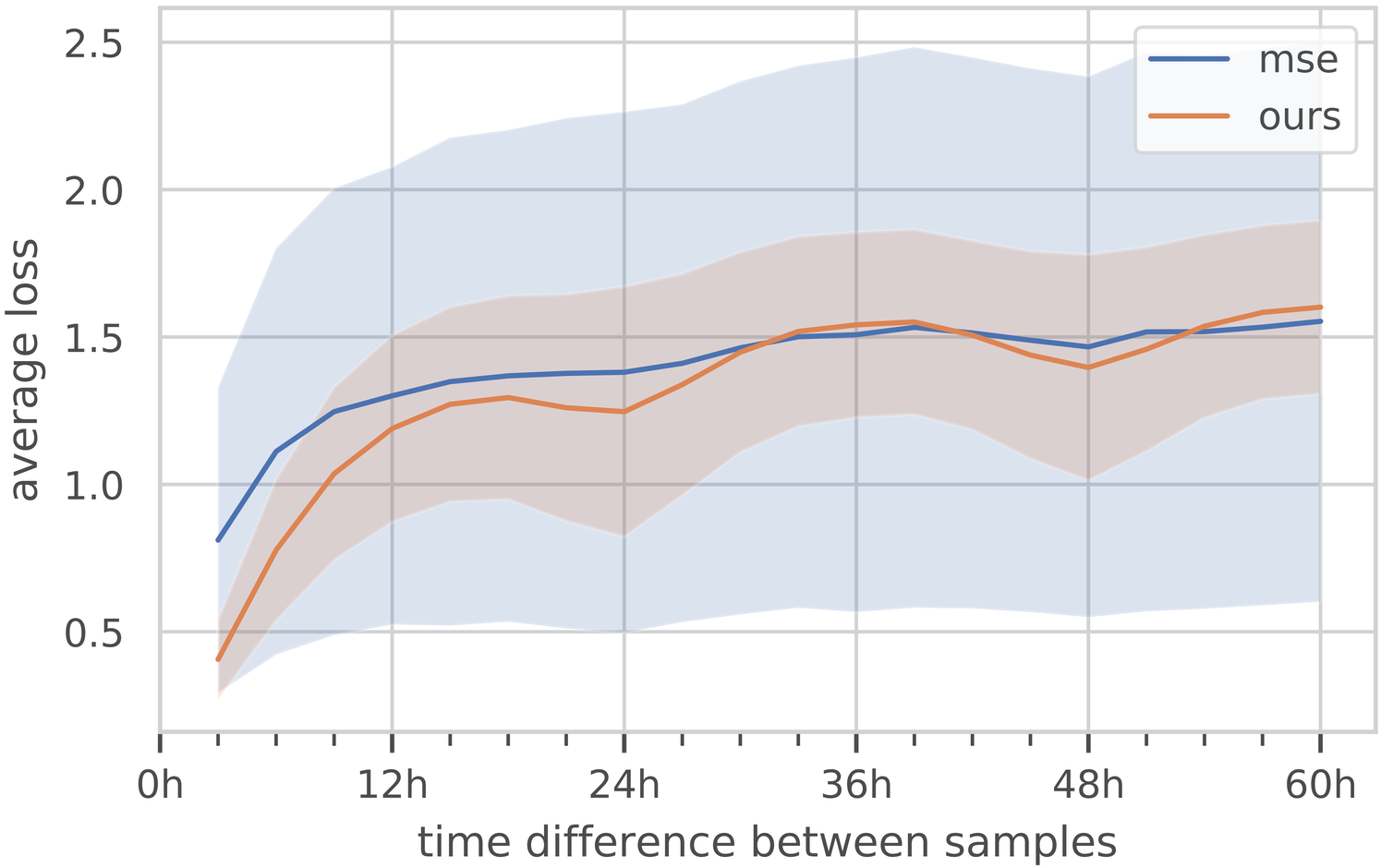
© Christian Lessig, 2023

Representation learning for science

- Can large representation networks trained on observational data serve as scientific models?
  - Self-supervised training to obtain “untainted“ model
  - Discover unknowns relationships in data through in-context learning?
  - Data curation as central task for scientists?

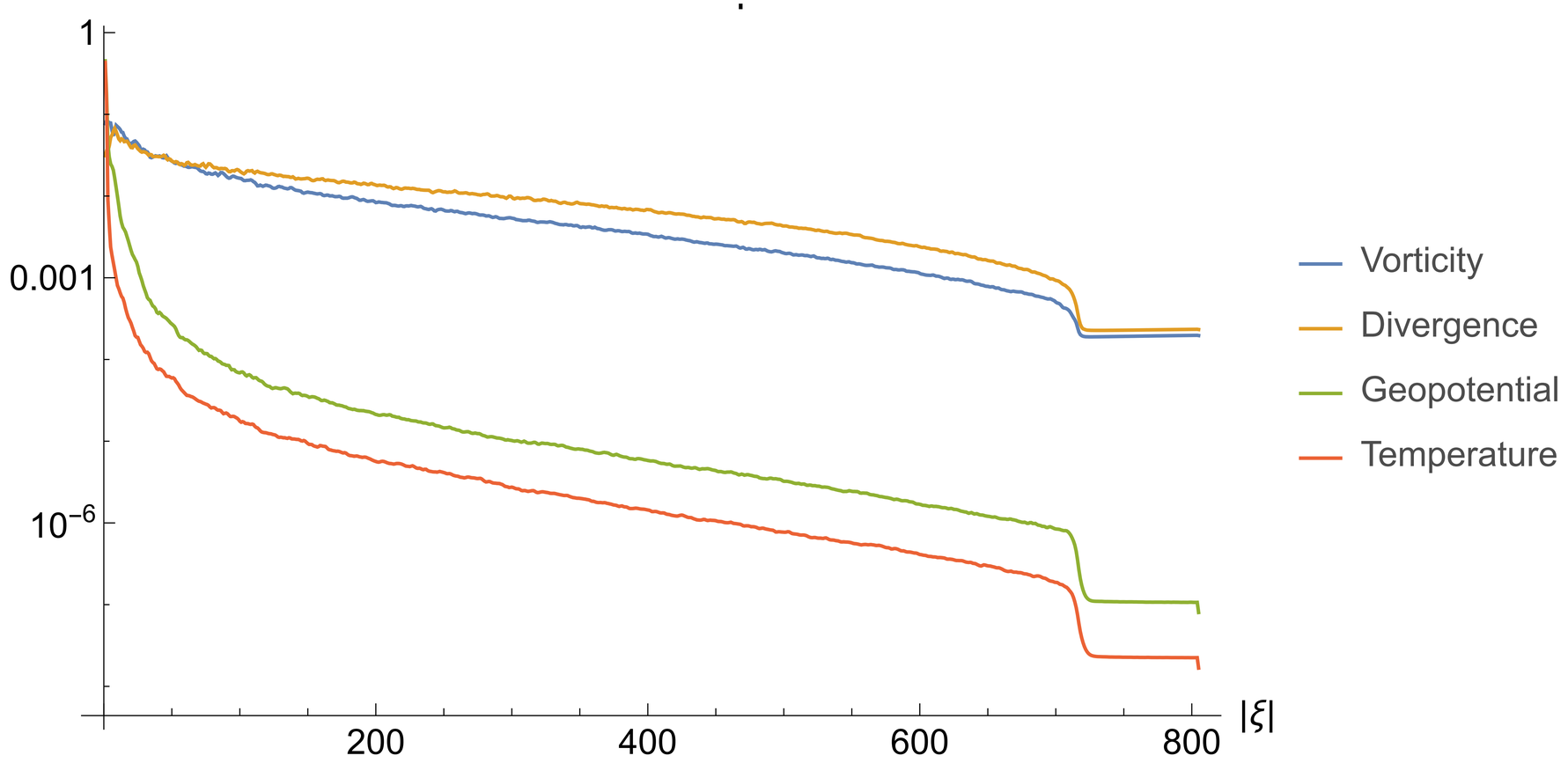
© Christian Lessig, 2023

AtmoDist: evaluation



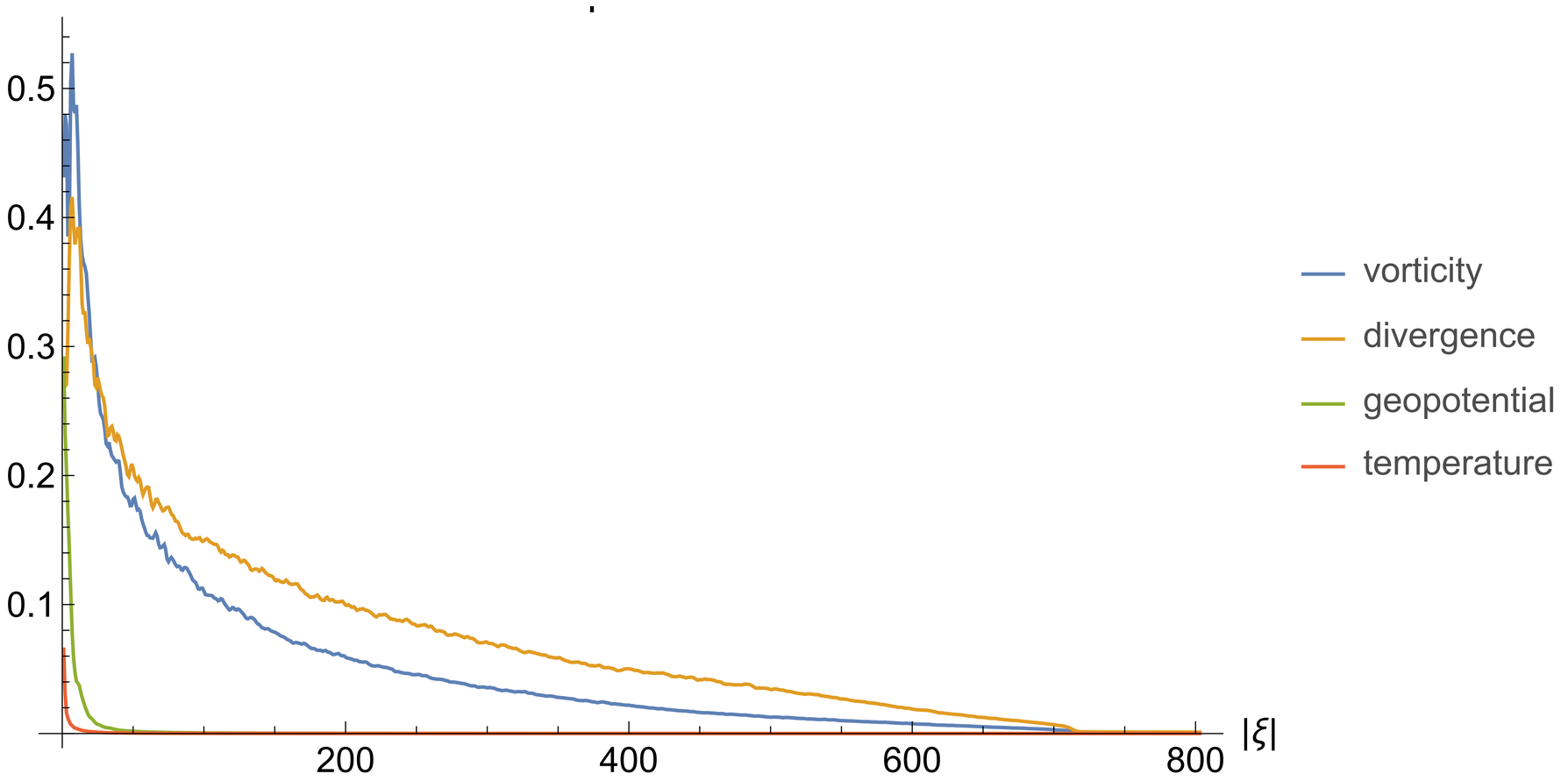
© Christian Lessig, 2023

AtmoRep data



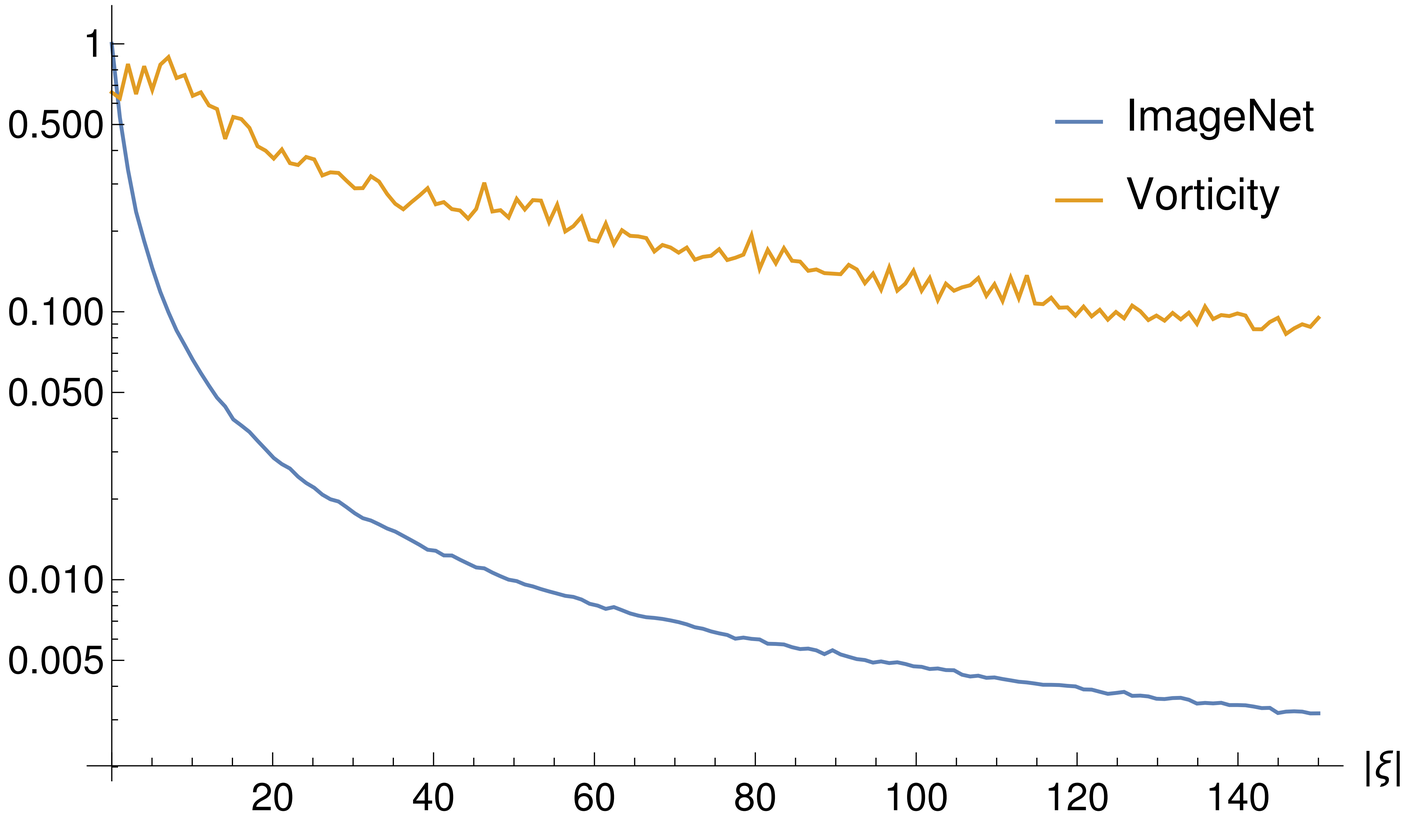
© Christian Lessig, 2023

AtmoRep data



© Christian Lessig, 2023

ERA5 versus ImageNet



© Christian Lessig, 2023

ERA5 versus ImageNet

stream function
velocity potential



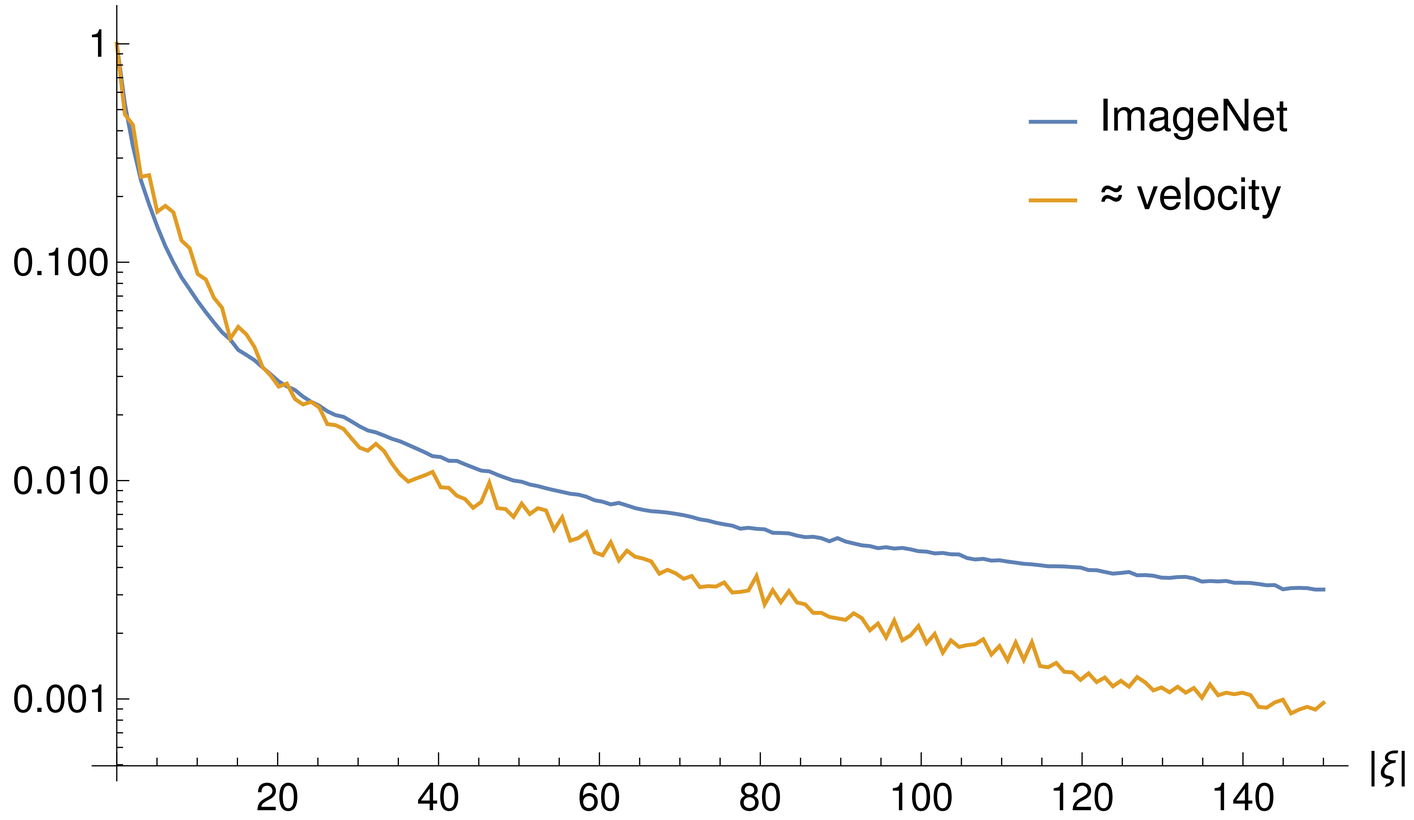
velocity
vector field



vorticity
divergence

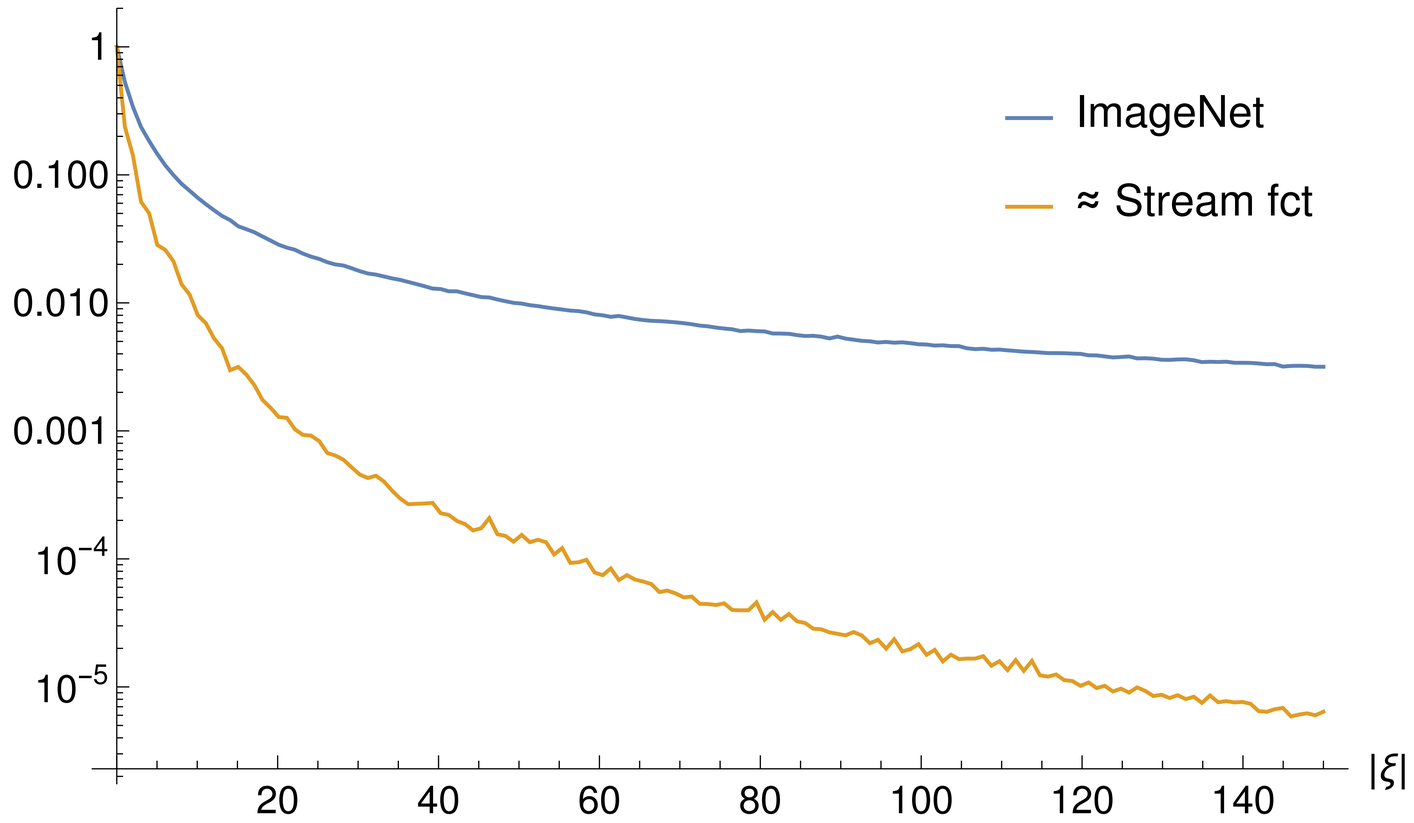
© Christian Lessig, 2023

ERA5 versus ImageNet



© Christian Lessig, 2023

ERA5 versus ImageNet



© Christian Lessig, 2023

ERA5 versus ImageNet

stream function
velocity potential



velocity
vector field



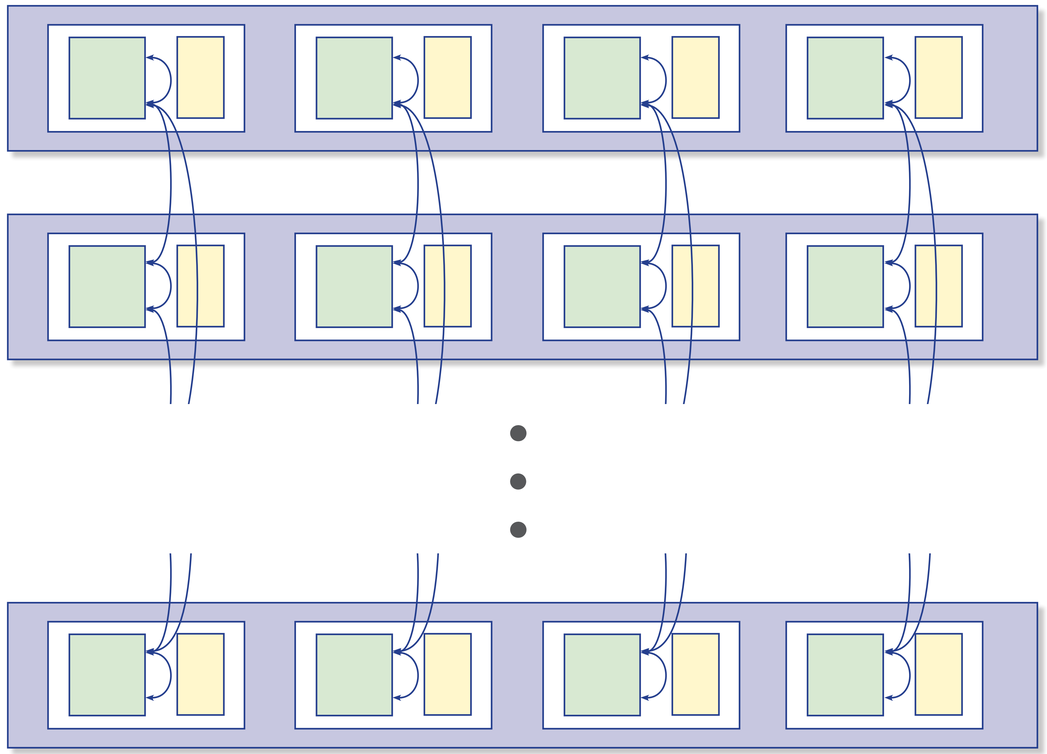
vorticity
divergence

Work with velocity with norm that emphasizes small scale detail

© Christian Lessig, 2023

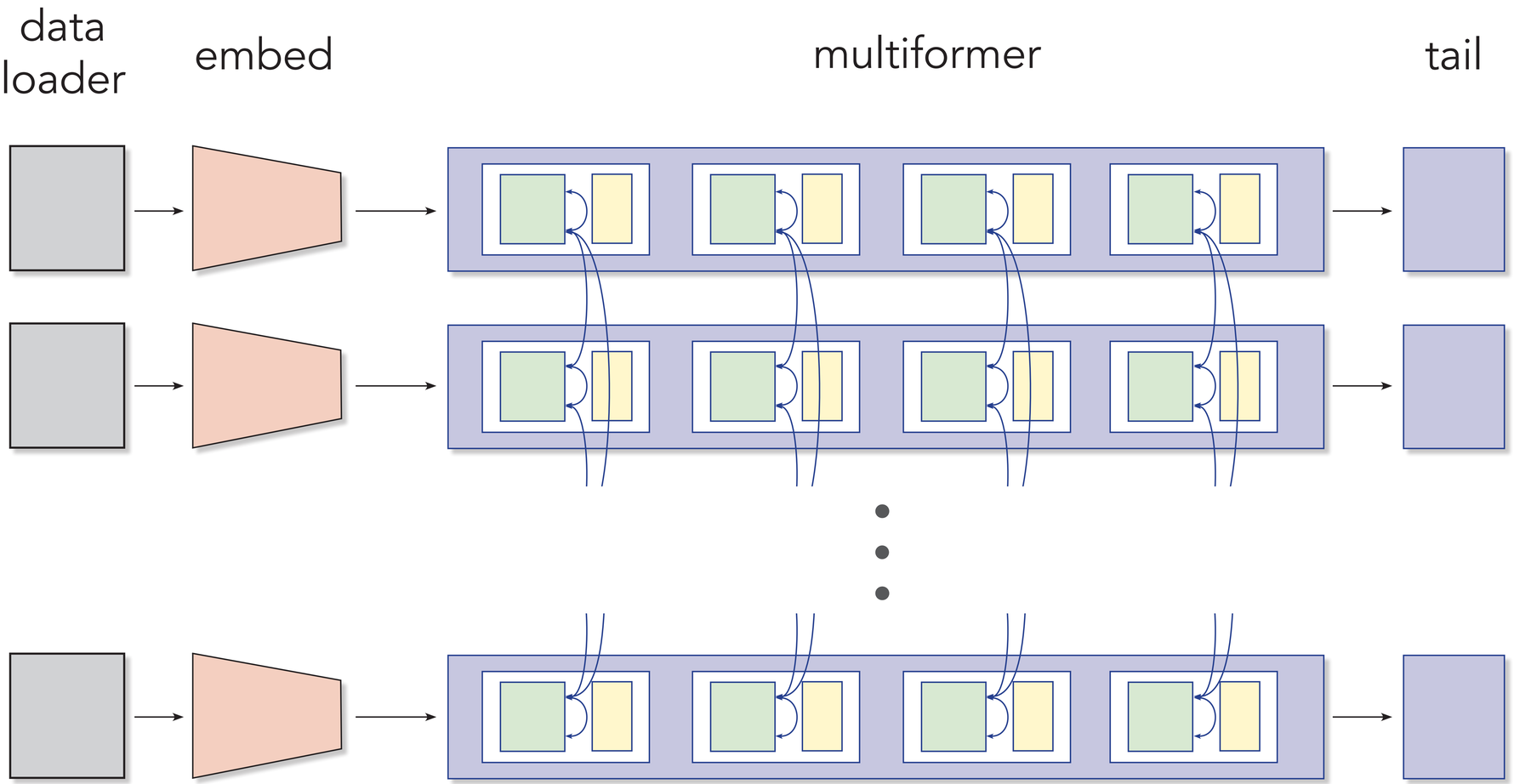
Embedding of tokens

multiformer



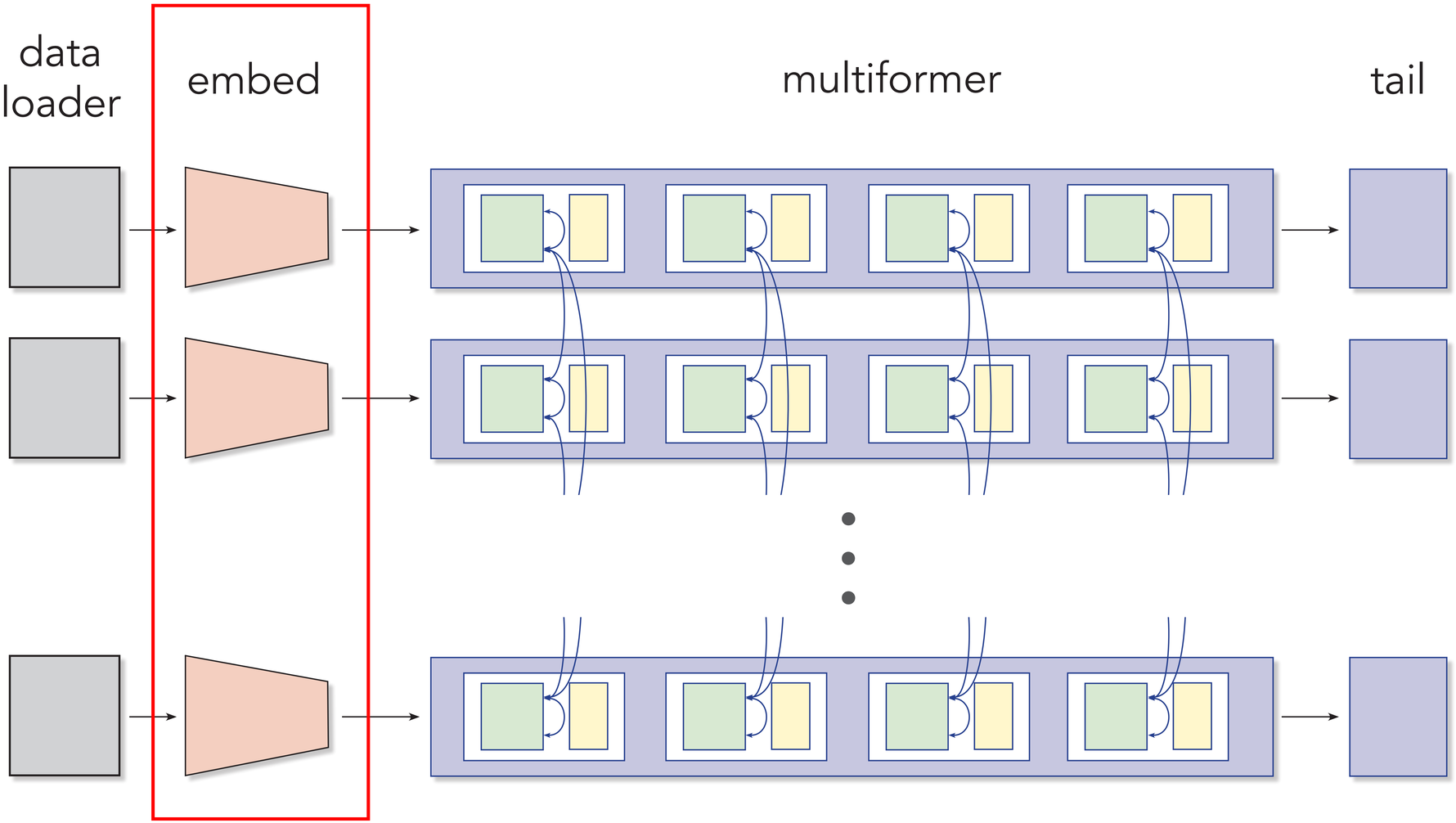
© Christian Lessig, 2023

Embedding of tokens



© Christian Lessig, 2023

Embedding of tokens



© Christian Lessig, 2023

Embedding network

- Multiformer models longer range effects and field interactions in a rich latent space

© Christian Lessig, 2023

Embedding network

- Multiformer models longer range effects and field interactions in a rich latent space
  - Embedding network provides rich encoding of input field

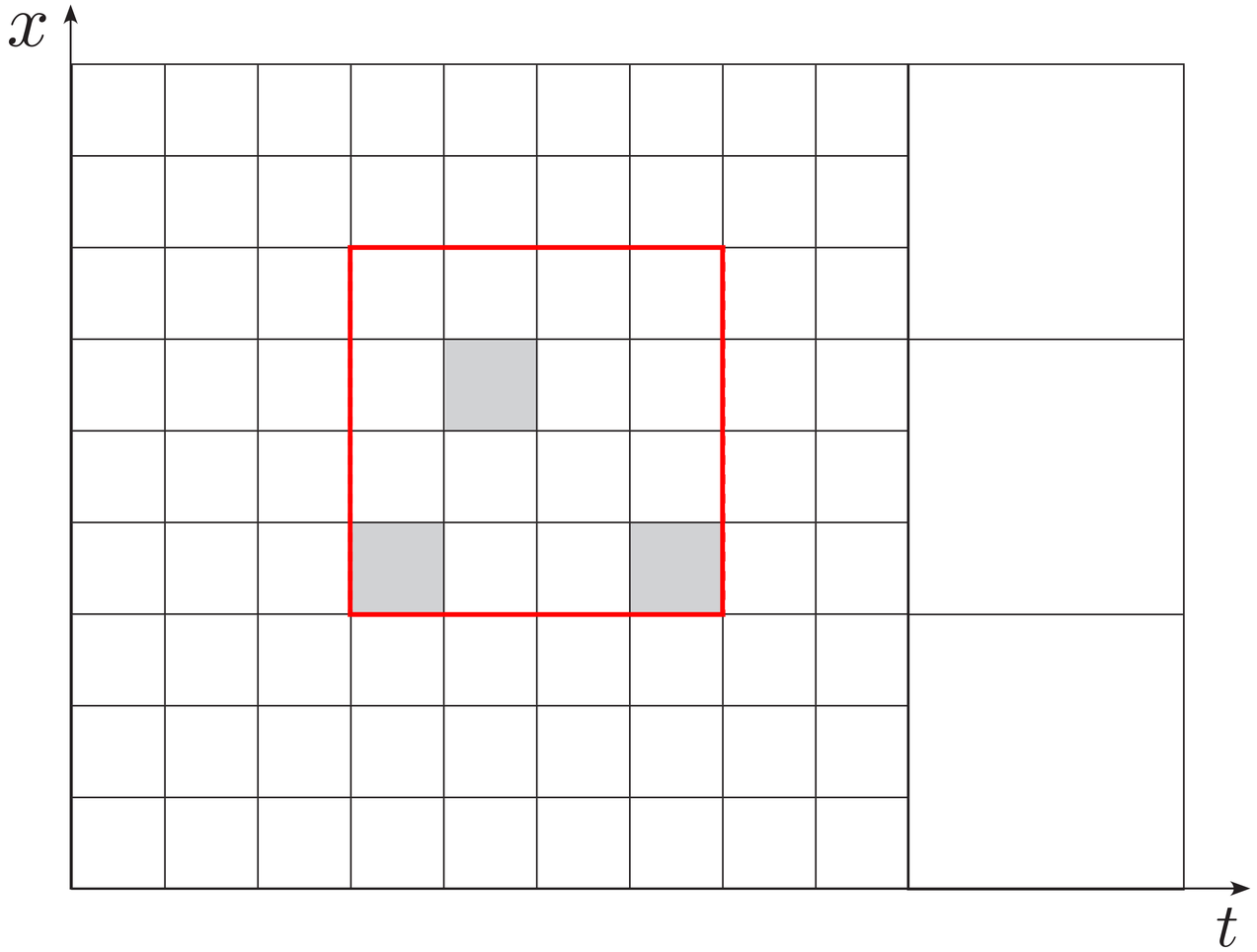
© Christian Lessig, 2023

Embedding network

- Multiformer models longer range effects and field interactions in a rich latent space
  - Embedding network provides rich encoding of input field
  - Embedding network allows for multi-resolution representation per field, i.e. different token sizes

© Christian Lessig, 2023

Embedding of tokens



© Christian Lessig, 2023

Embedding of tokens

- Multiformer models longer range effects and field interactions in a rich latent space
  - Embedding network provides rich encoding of input field
  - Embedding network allows for multi-resolution representation per field, i.e. different token sizes

© Christian Lessig, 2023

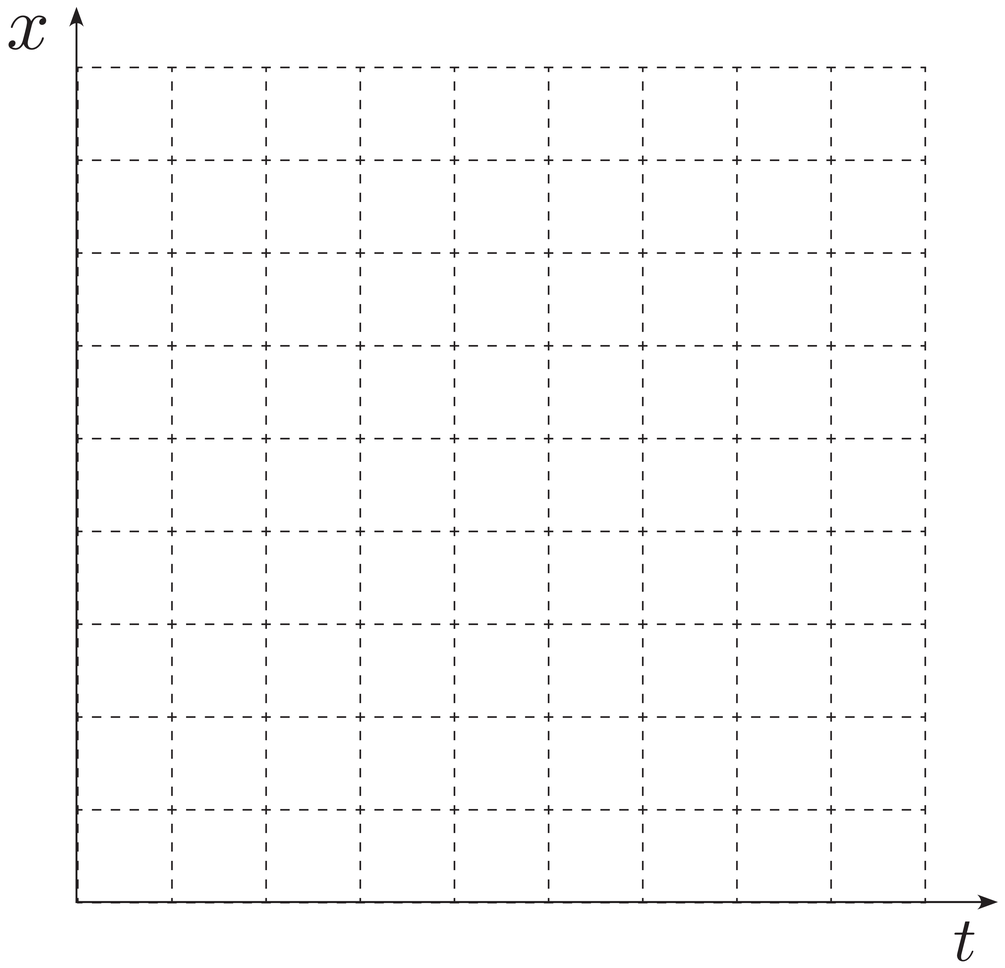
Embedding of tokens

- Multiformer models longer range effects and field interactions in a rich latent space
  - Embedding network provides rich encoding of input field
  - Embedding network allows for multi-resolution representation per field, i.e. different token sizes

⇒ Use transformer as embedding network

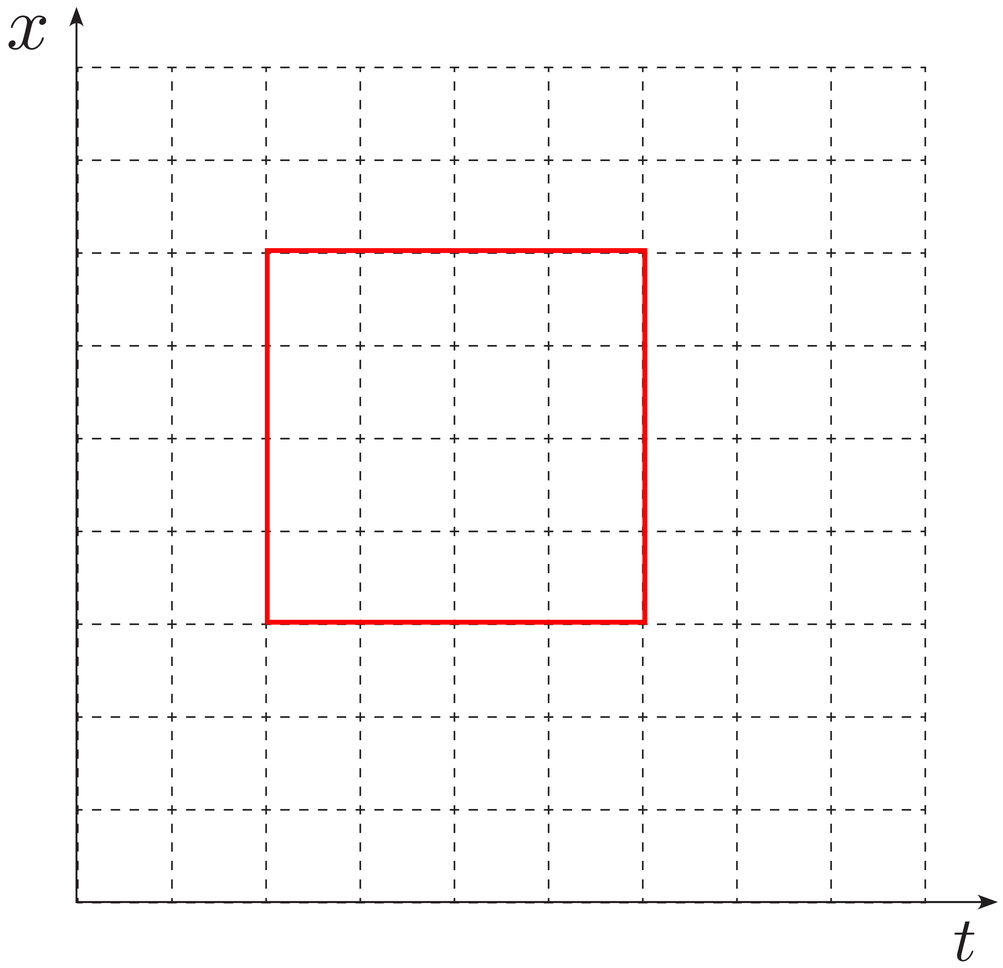
© Christian Lessig, 2023

Embedding of tokens



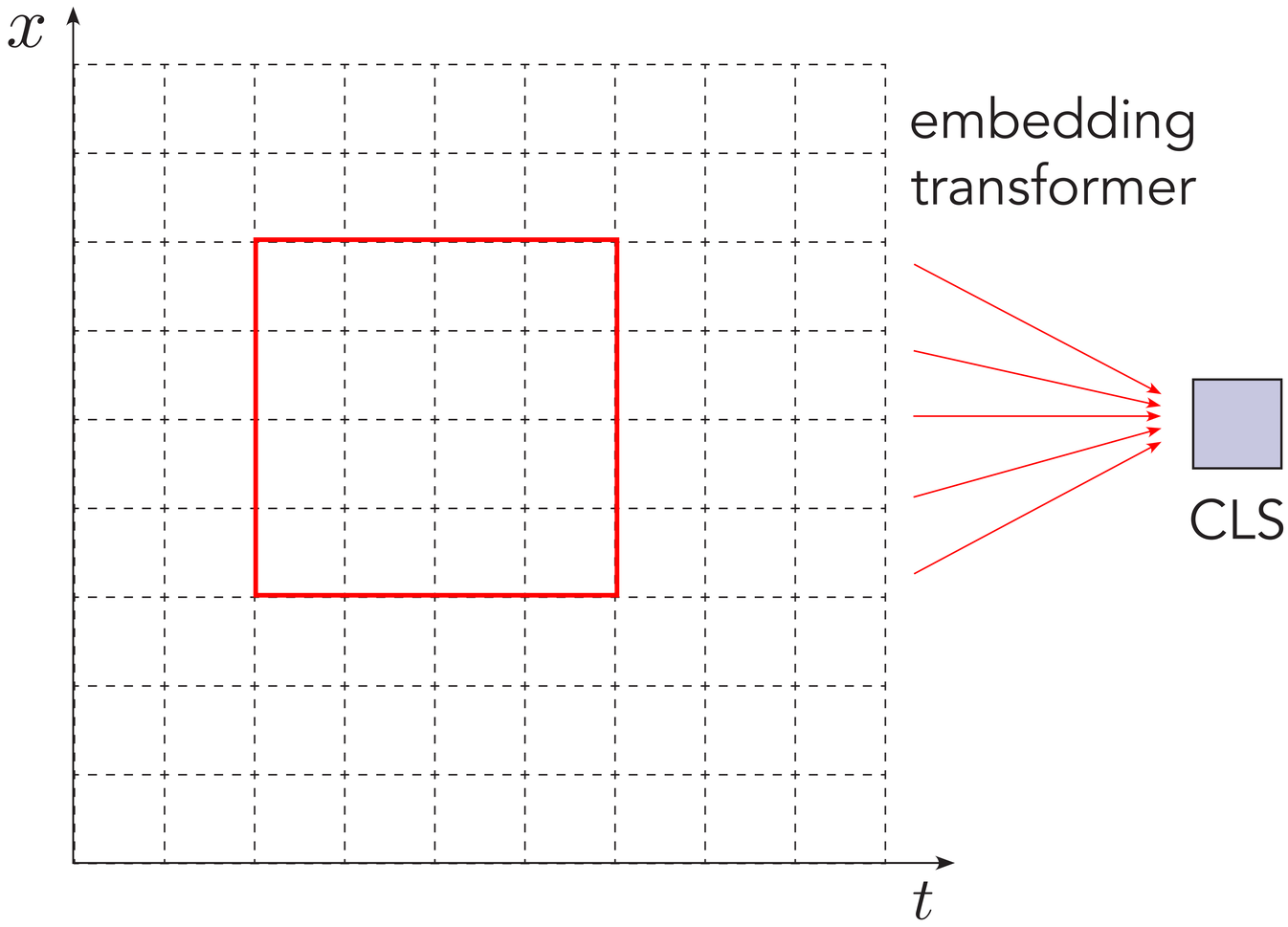
© Christian Lessig, 2023

Embedding of tokens



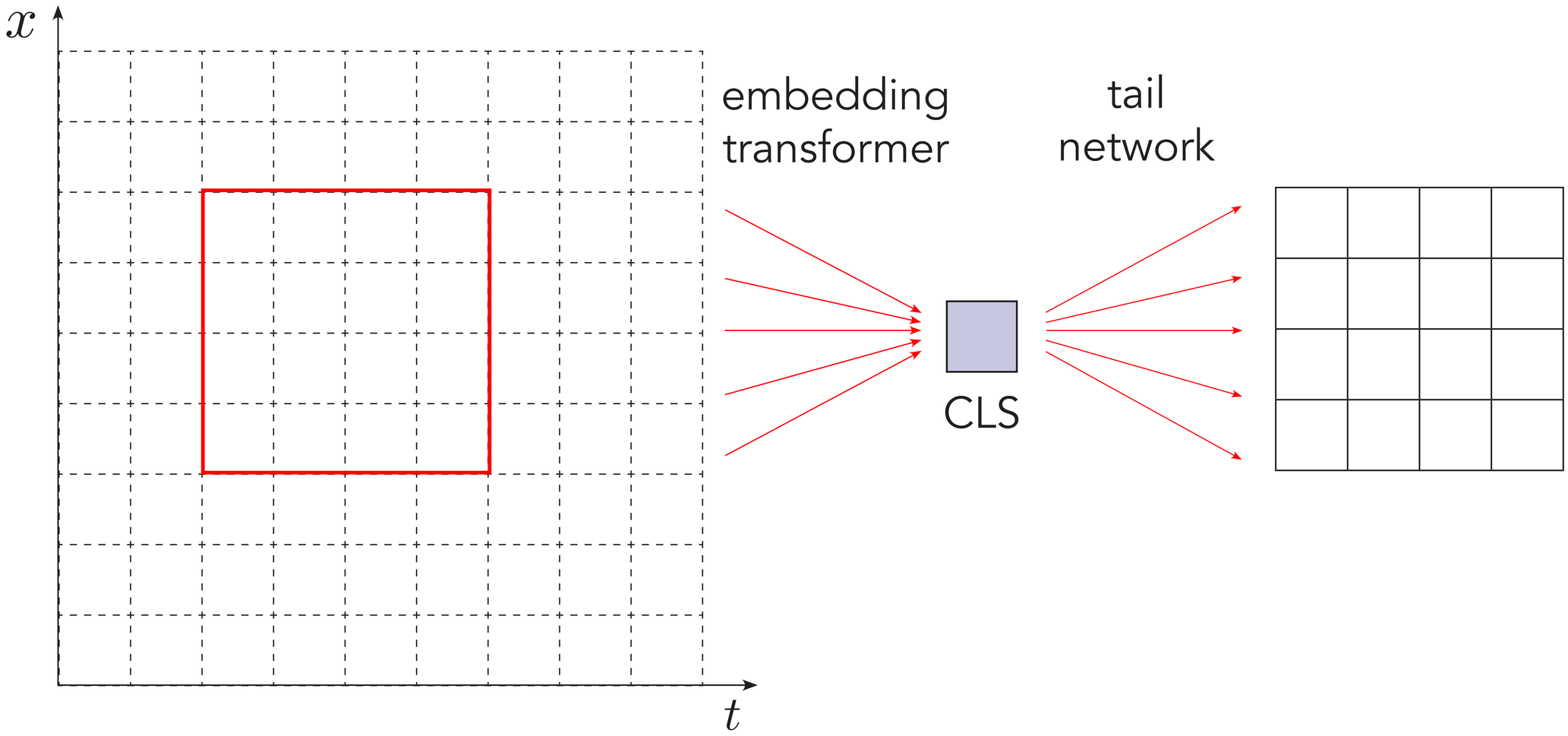
© Christian Lessig, 2023

Embedding of tokens



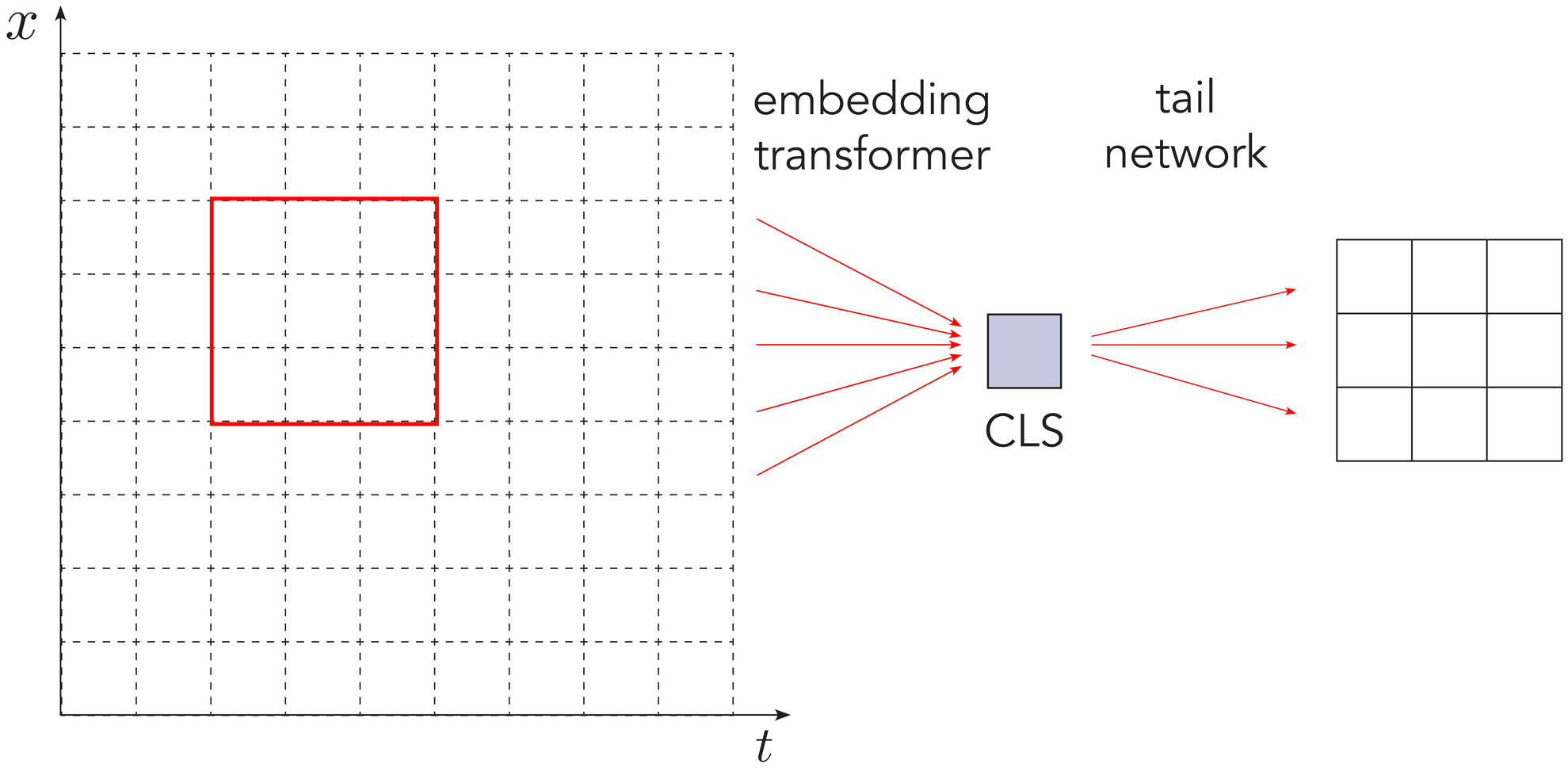
© Christian Lessig, 2023

Embedding of tokens



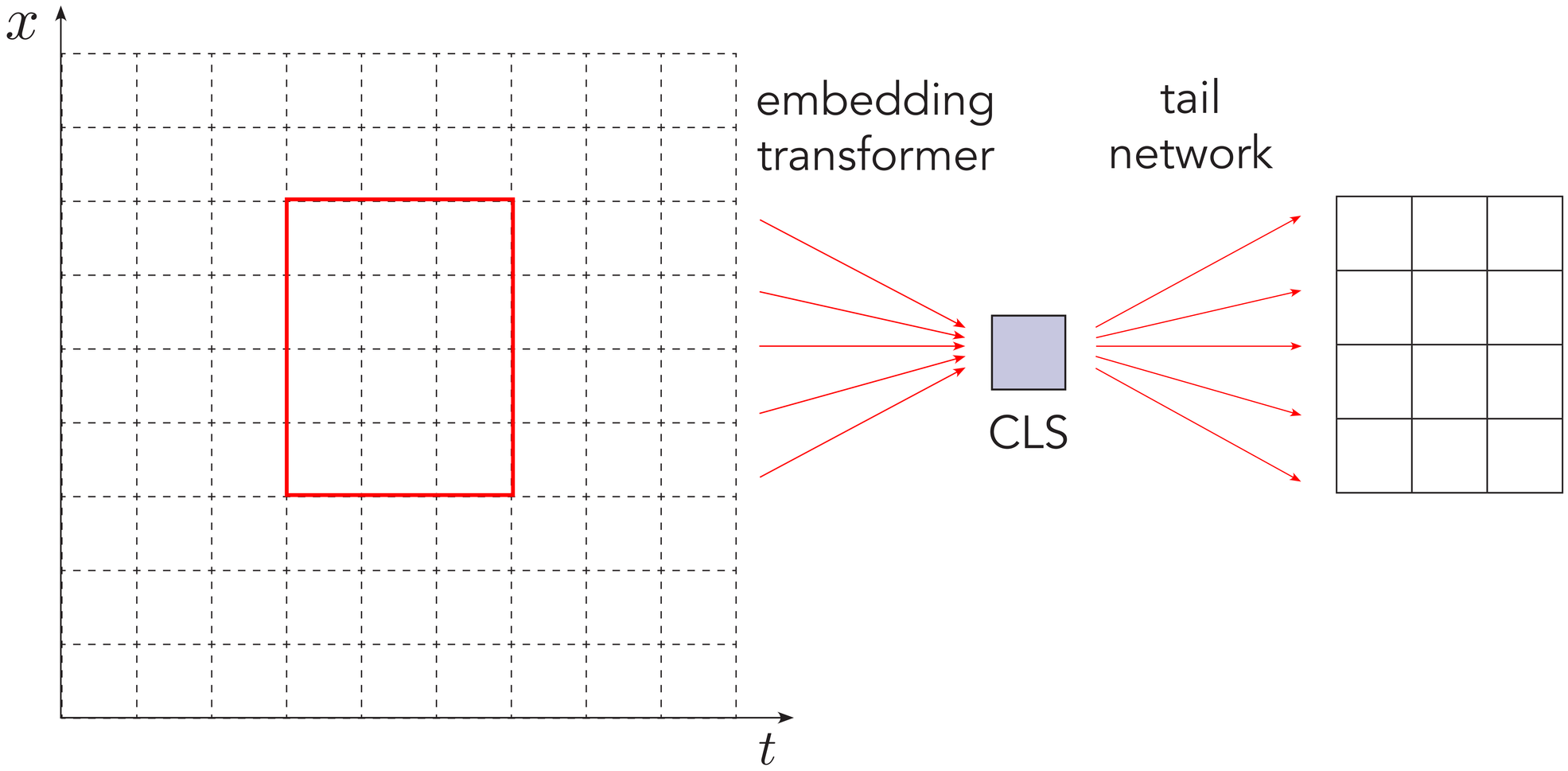
© Christian Lessig, 2023

Embedding of tokens



© Christian Lessig, 2023

Embedding of tokens



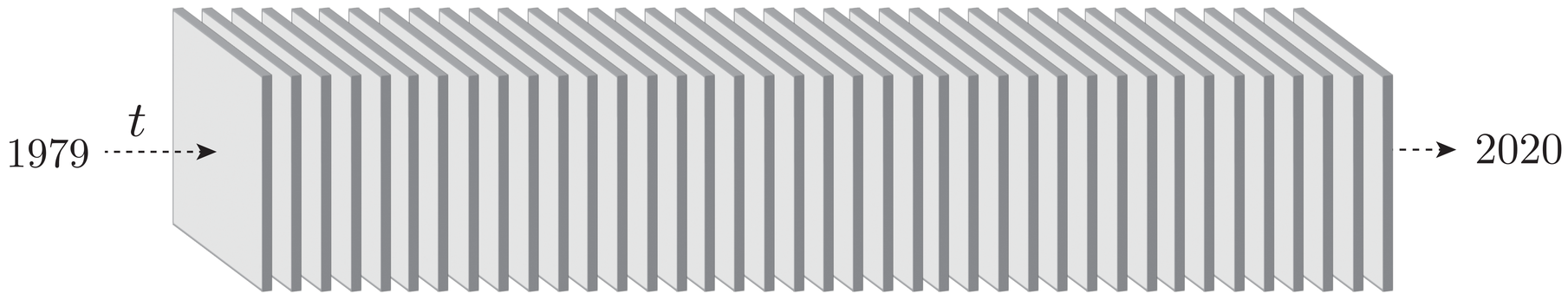
© Christian Lessig, 2023

Training

- Unbiased hierarchical Monte Carlo sampling of all possible ERA5 space-time cubes

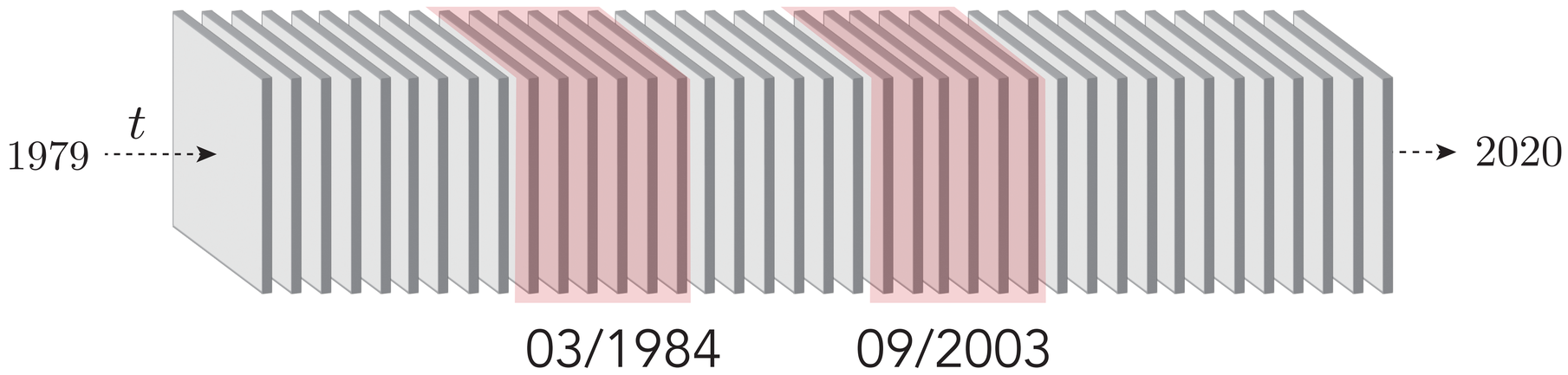
© Christian Lessig, 2023

Training



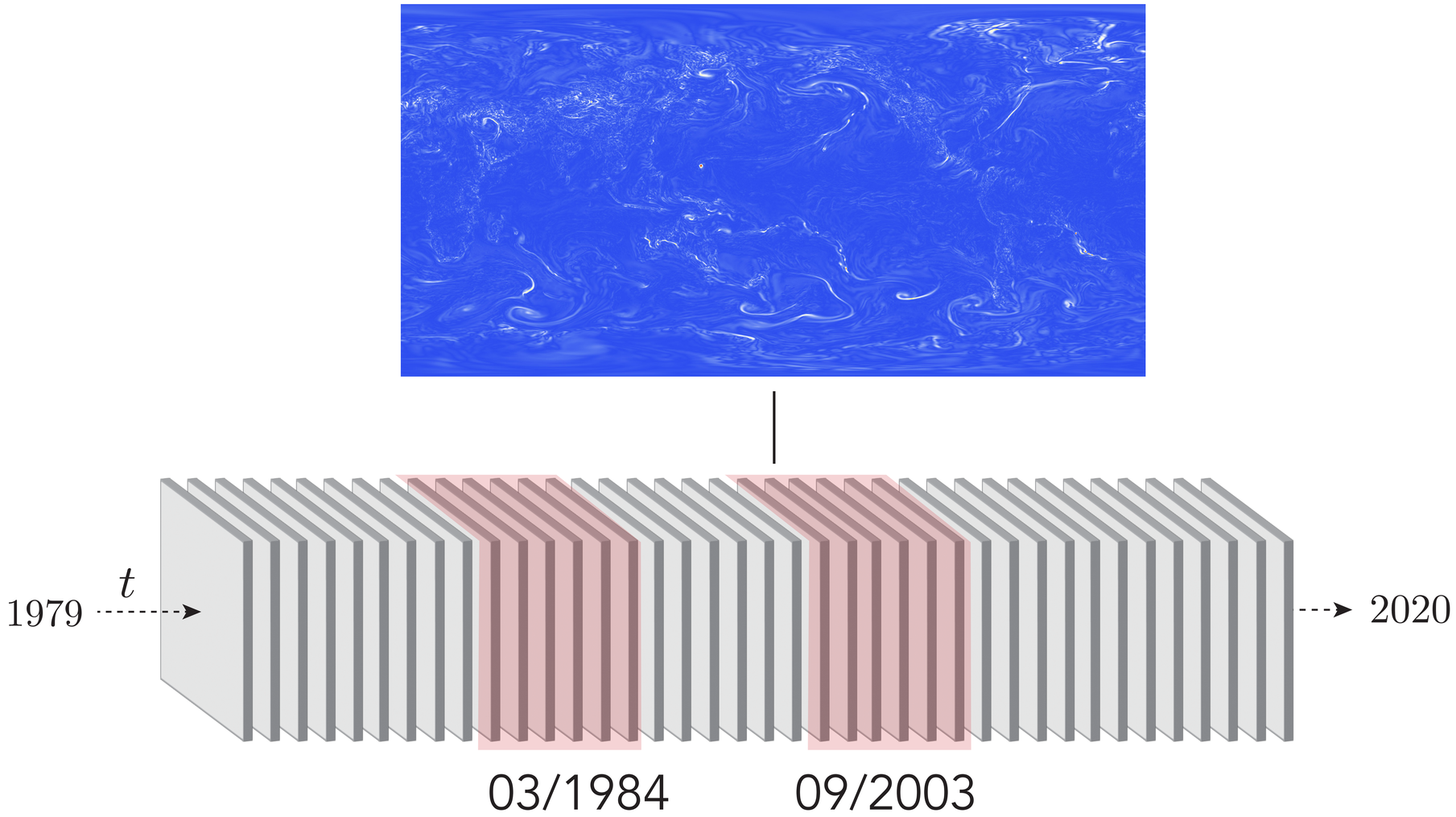
© Christian Lessig, 2023

Training



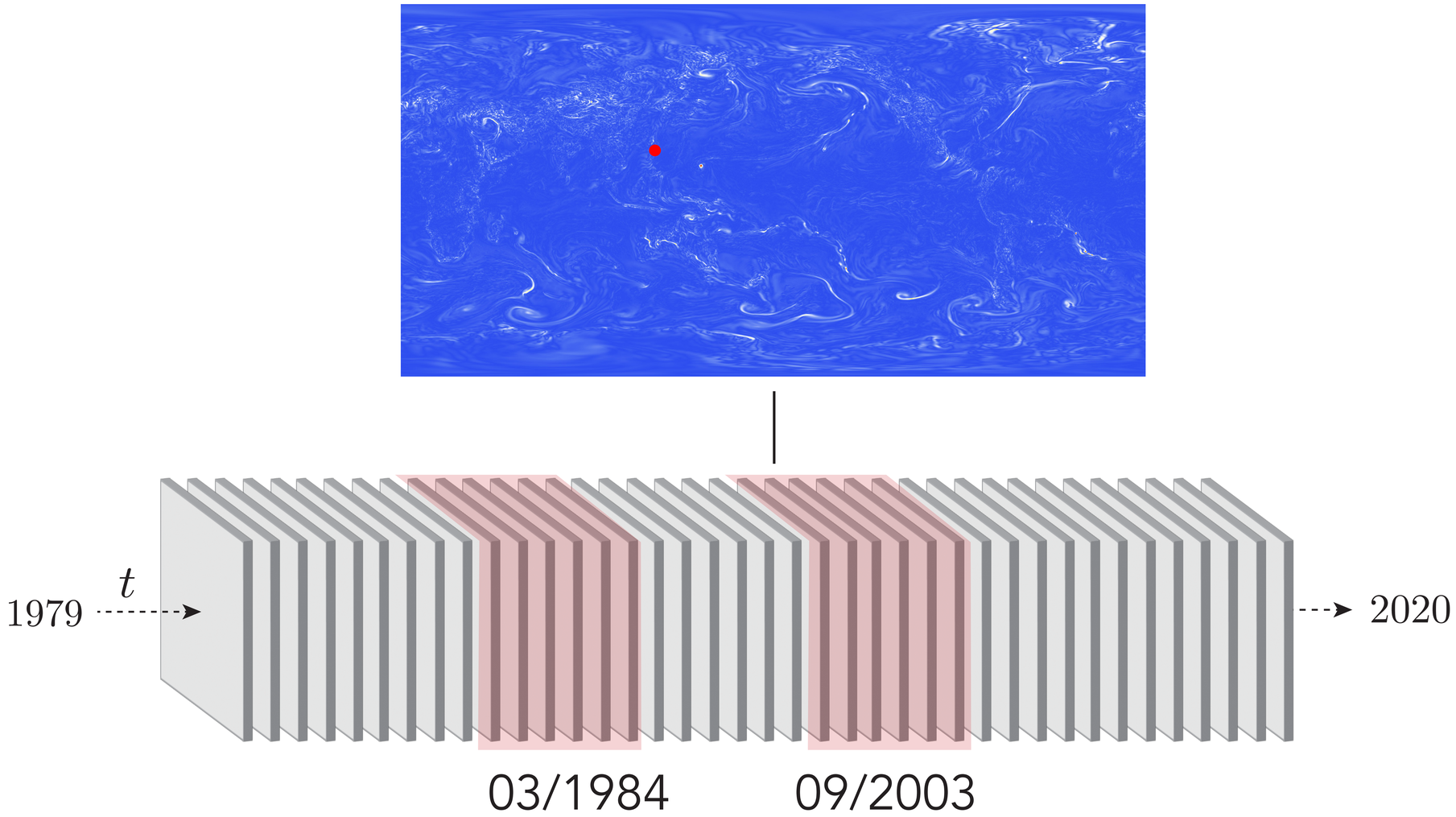
© Christian Lessig, 2023

Training



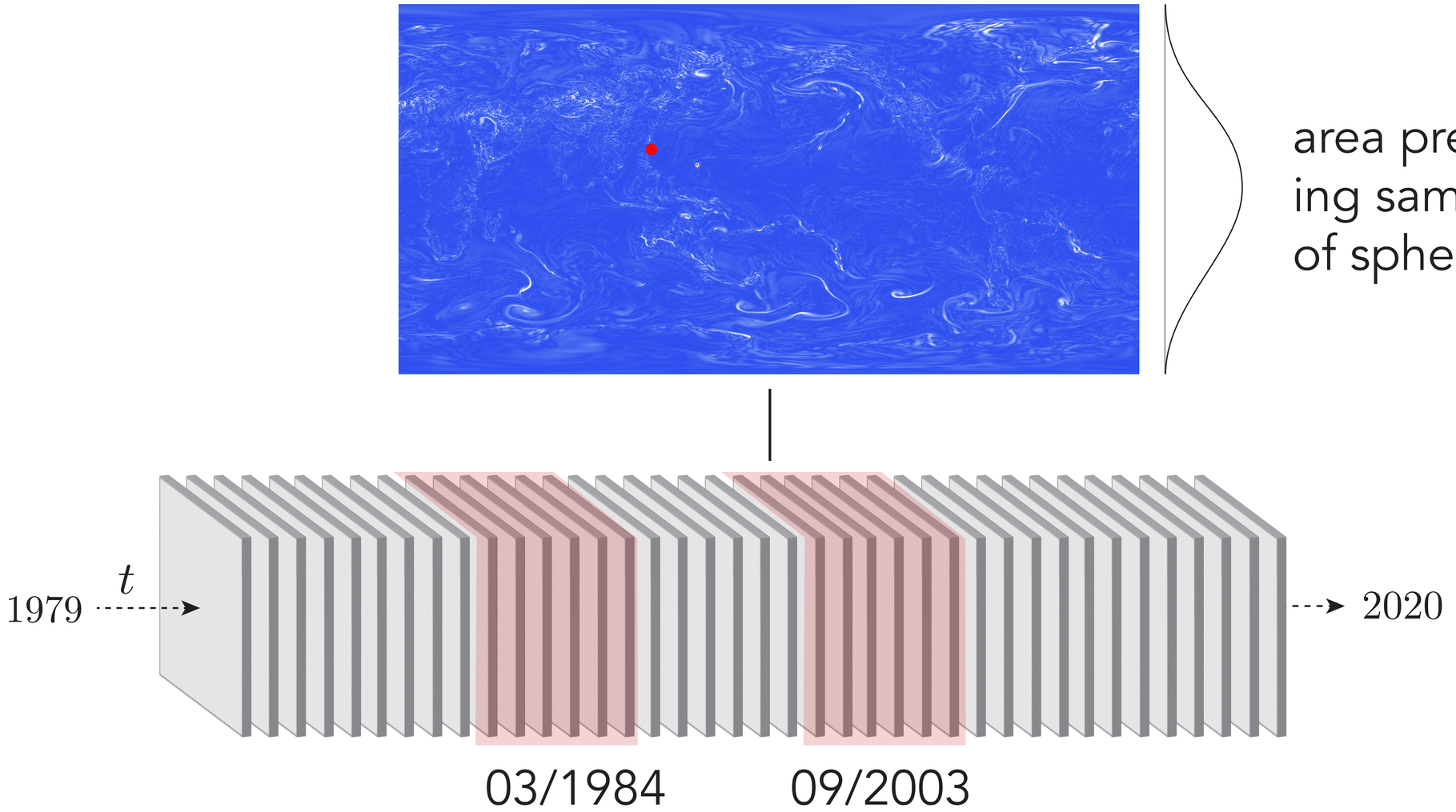
© Christian Lessig, 2023

Training



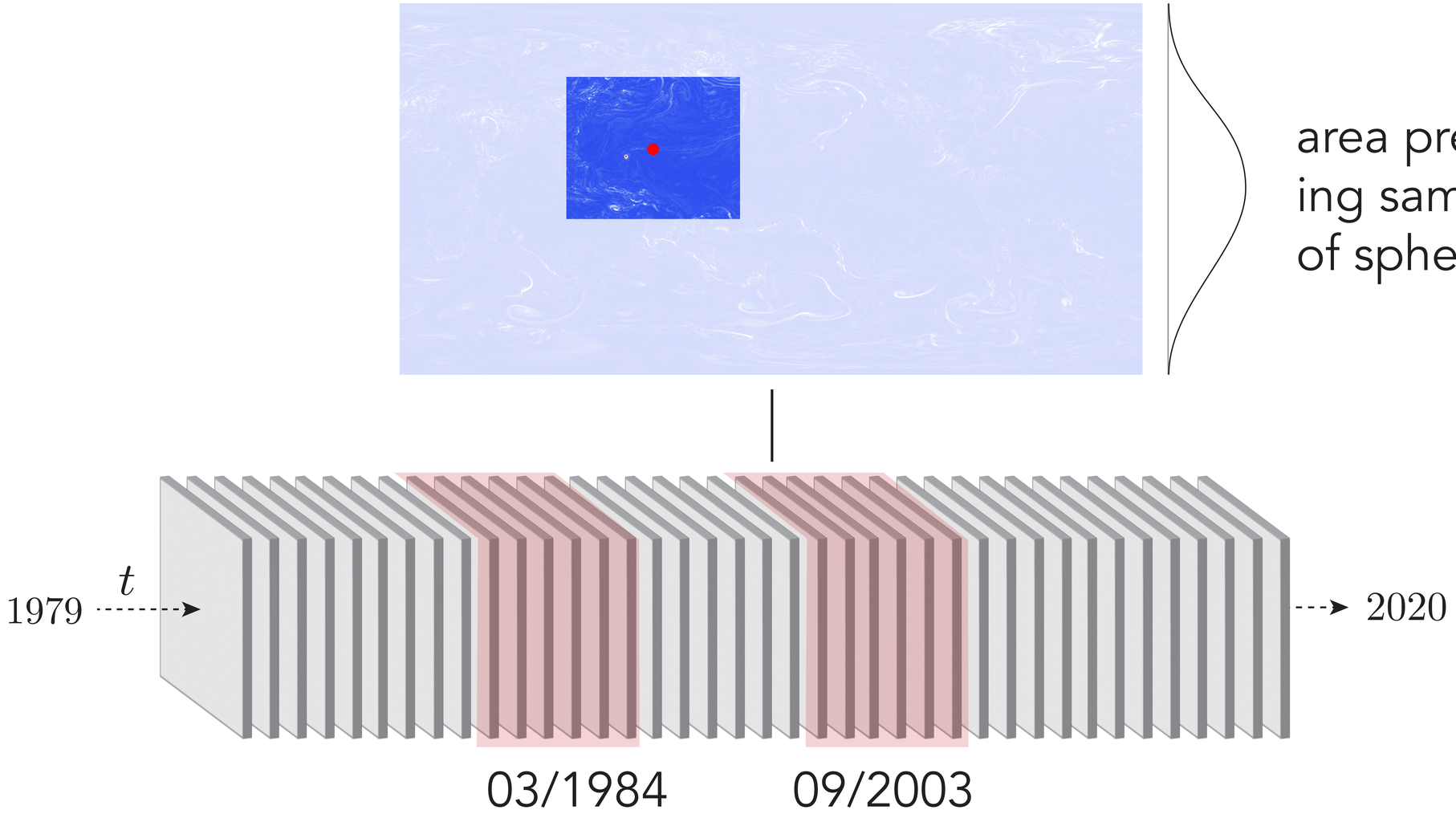
© Christian Lessig, 2023

Training



© Christian Lessig, 2023

Training



© Christian Lessig, 2023

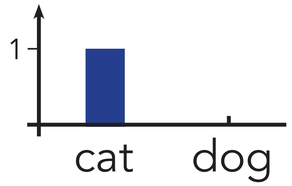
Statistical loss



© Christian Lessig, 2023

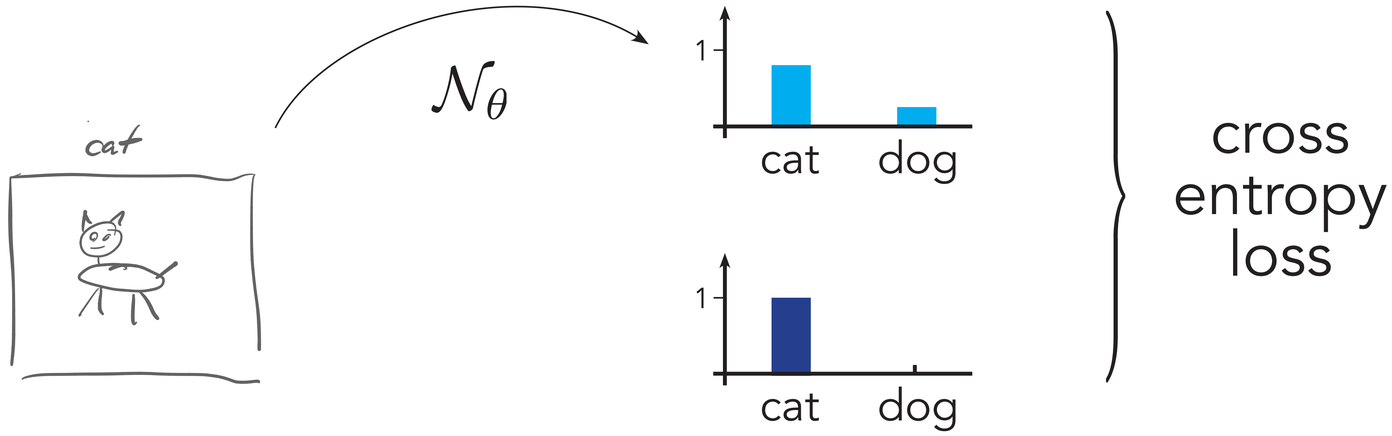
Statistical loss





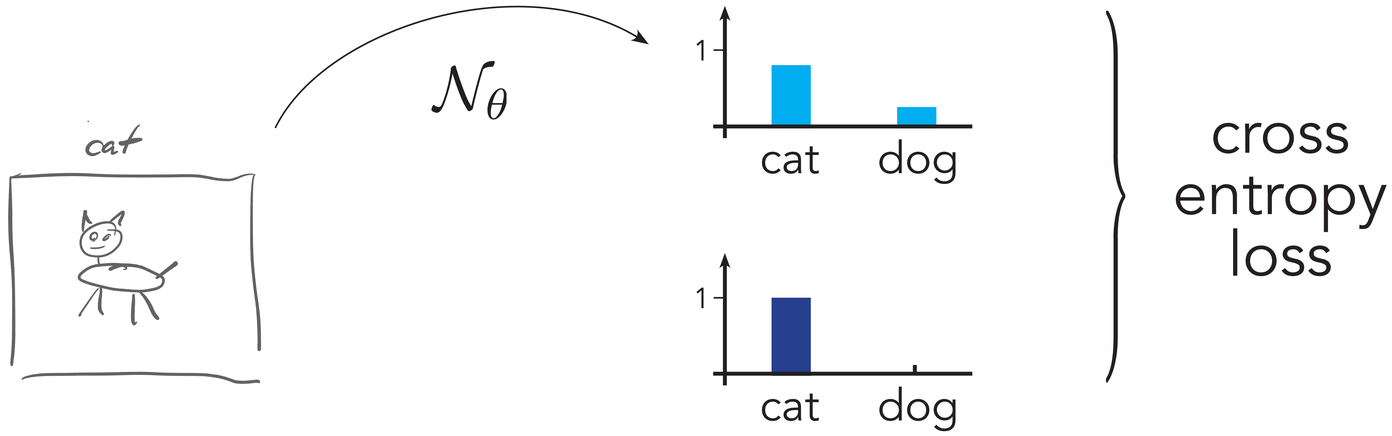
© Christian Lessig, 2023

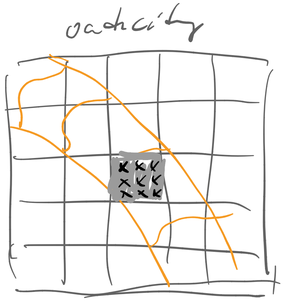
Statistical loss



© Christian Lessig, 2023

Statistical loss



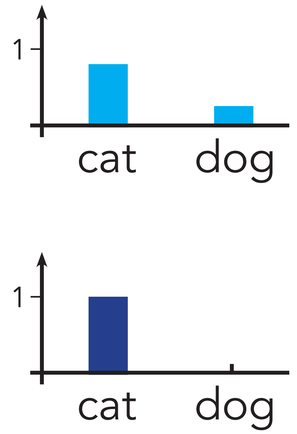


© Christian Lessig, 2023

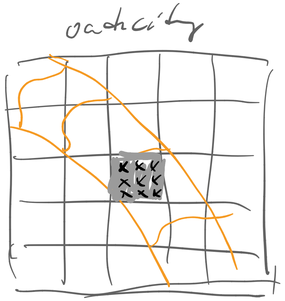
Statistical loss



$\mathcal{N}_\theta$



cross entropy loss



?

© Christian Lessig, 2023

Statistical loss

- Statistical loss:

$$\mathcal{L}_{\text{stats}} = \left| 1 - \int_{\mathrm{R}} \delta_y(x)\, G_{\tilde{\mu},\tilde{\sigma}}(x)\, \mathrm{d}x \right|^2$$

© Christian Lessig, 2023

Statistical loss

- Statistical loss:

$$\mathcal{L}_{\text{stats}} = \left| 1 - \int_{\mathrm{R}} \delta_y(x)\, G_{\tilde{\mu},\tilde{\sigma}}(x)\, \mathrm{d}x \right|^2$$

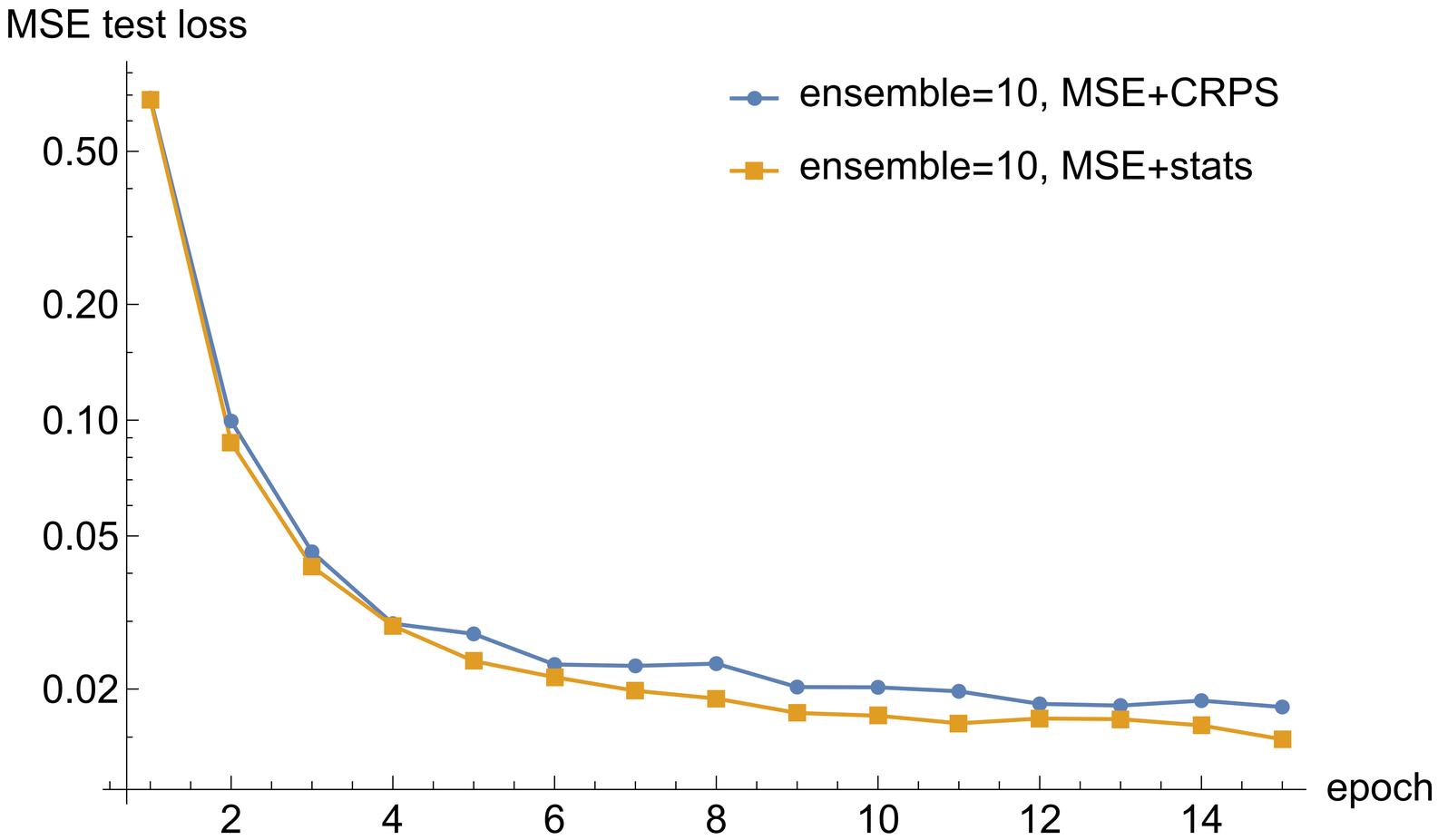
- CRPS:[1]

$$\mathcal{L}_{\text{CRPS}} = \int_{\mathrm{R}} \left| H_y(x)\, \mathrm{erf}_{\tilde{\mu},\tilde{\sigma}}(x) \right|^2 \mathrm{d}x$$

[1] S. Rasp and S. Lerch. Neural networks for postprocessing ensemble weather forecasts. Monthly Weather Review, 146(11):3885 – 3900, 2018.

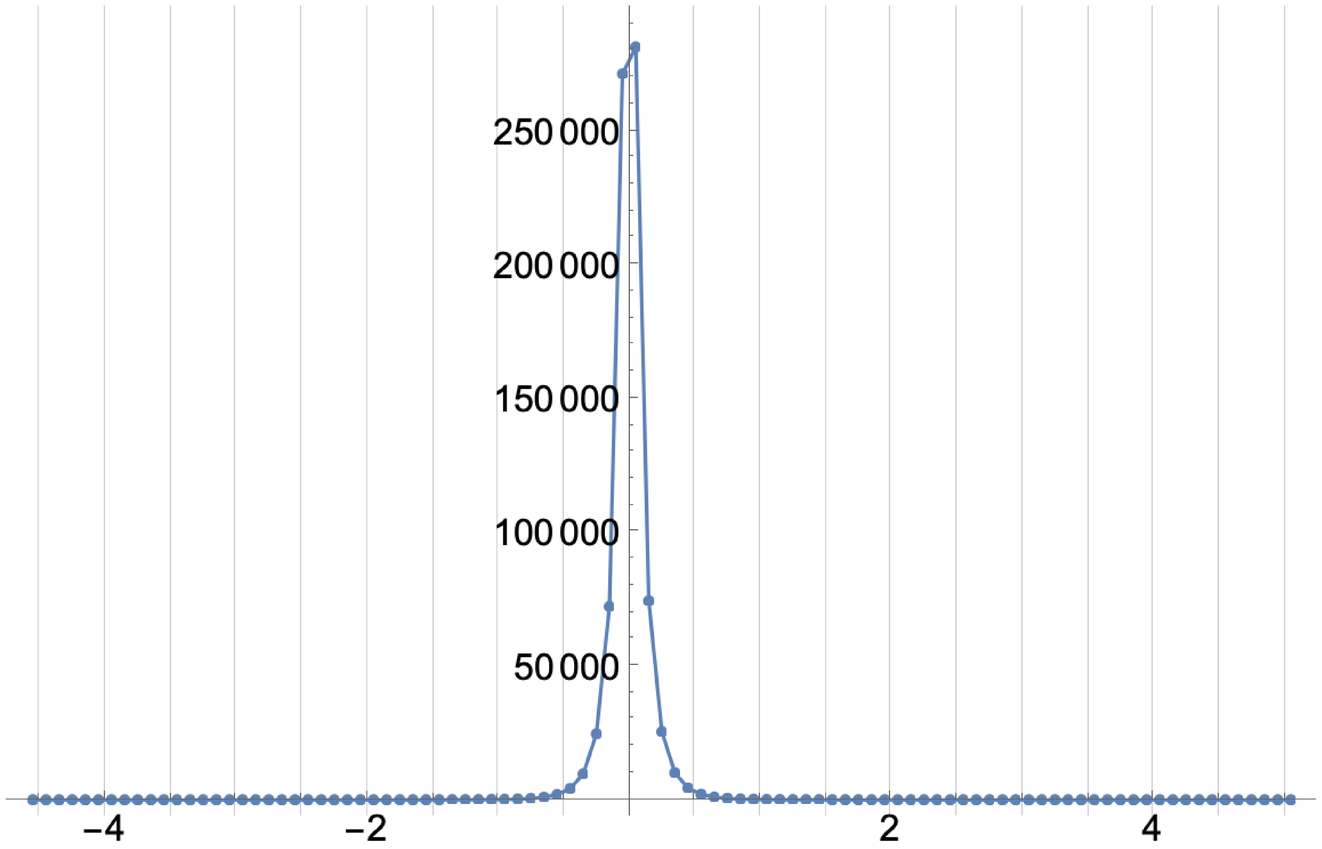
© Christian Lessig, 2023

Statistical loss



© Christian Lessig, 2023

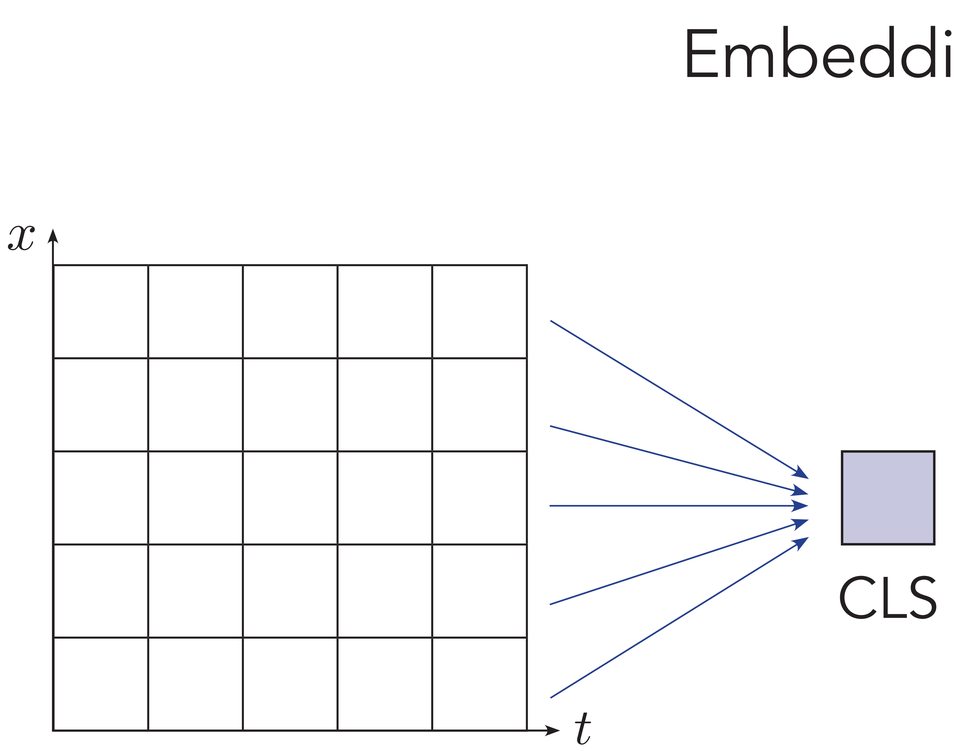
Statistical loss



Histogram of ensemble errors

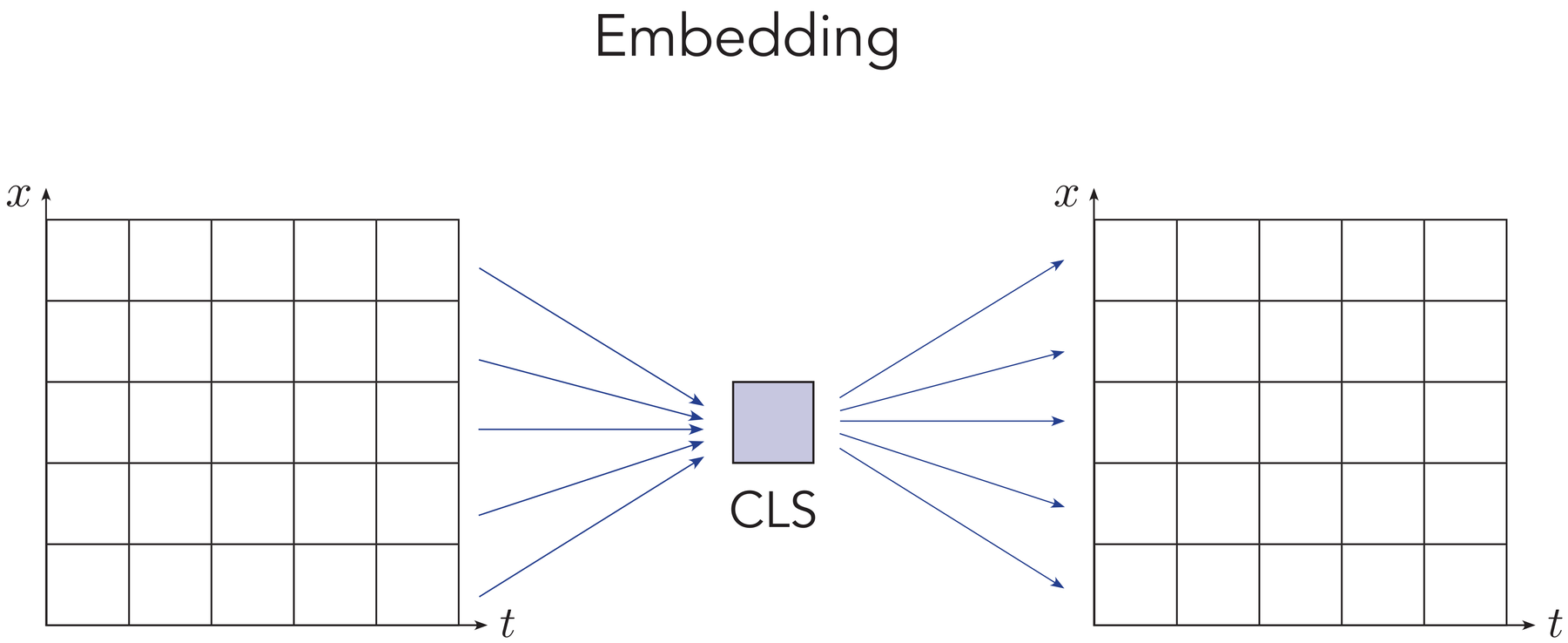
© Christian Lessig, 2023

Zero shot evaluation



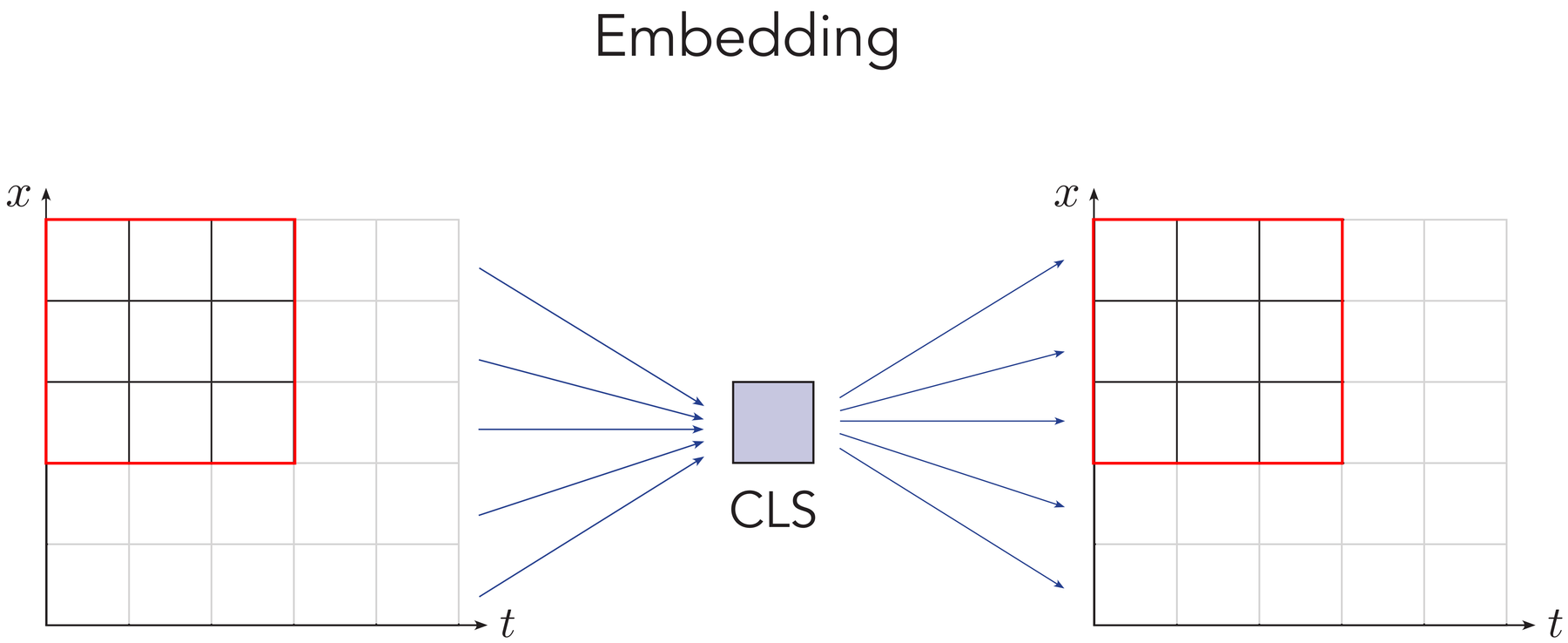
© Christian Lessig, 2023

Zero shot evaluation



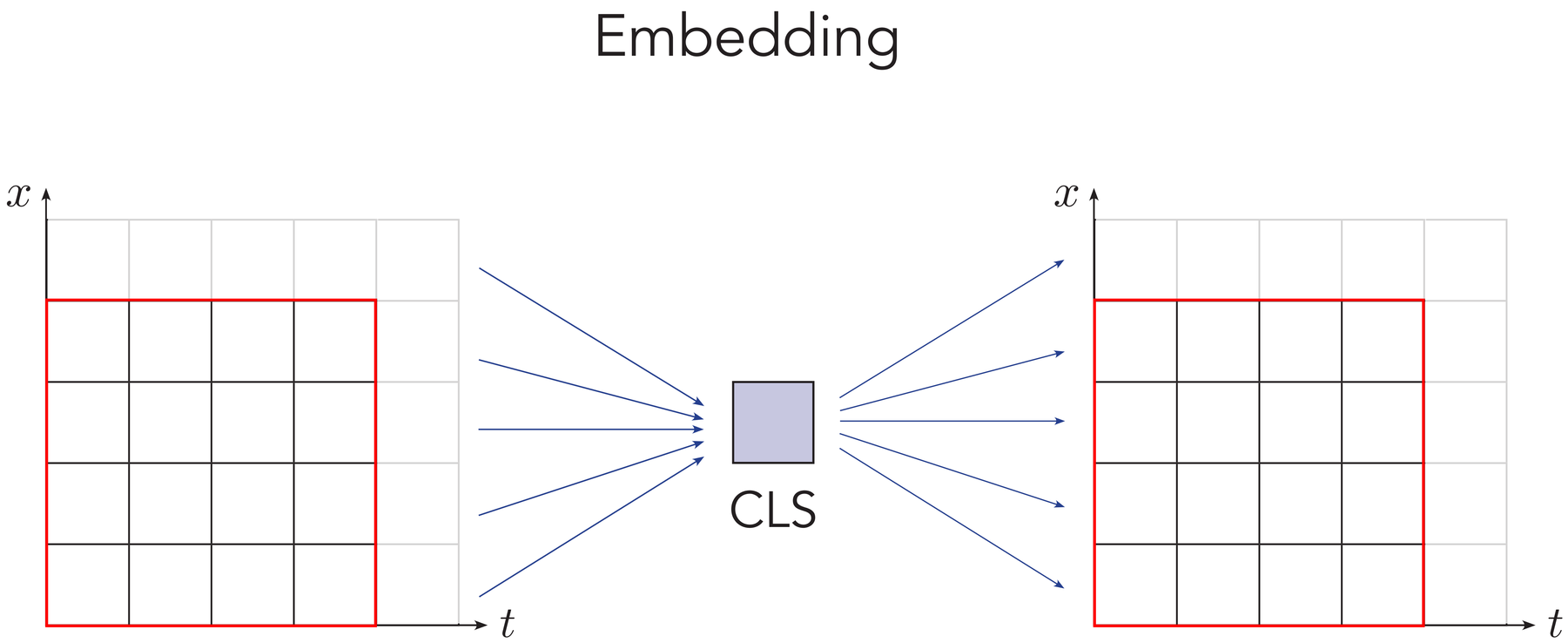
© Christian Lessig, 2023

Zero shot evaluation



© Christian Lessig, 2023

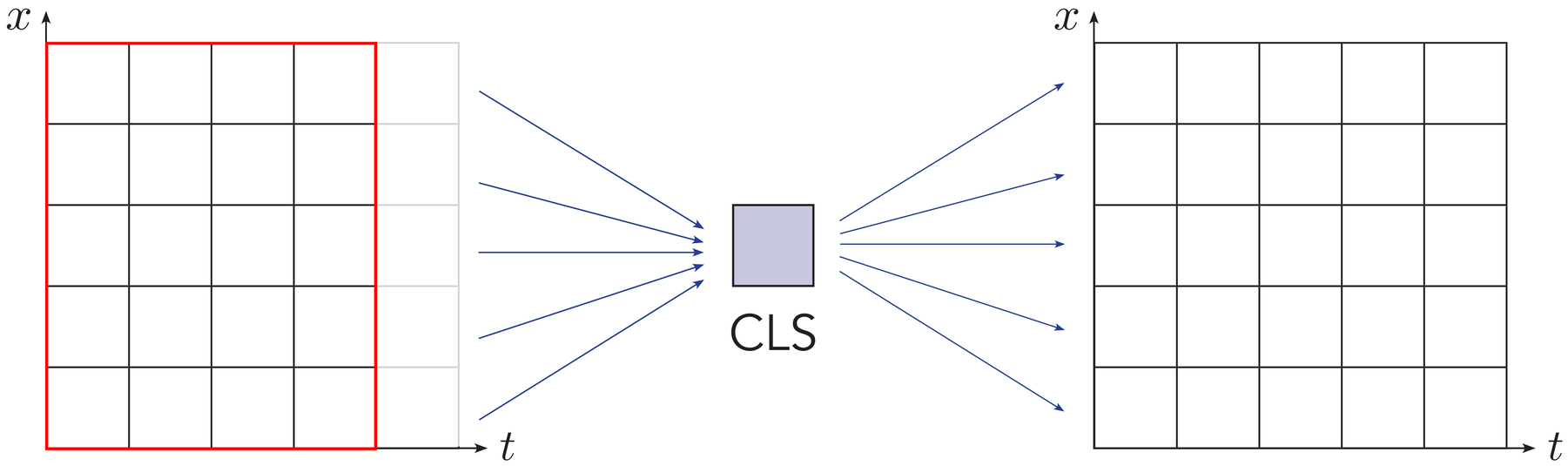
Zero shot evaluation



© Christian Lessig, 2023

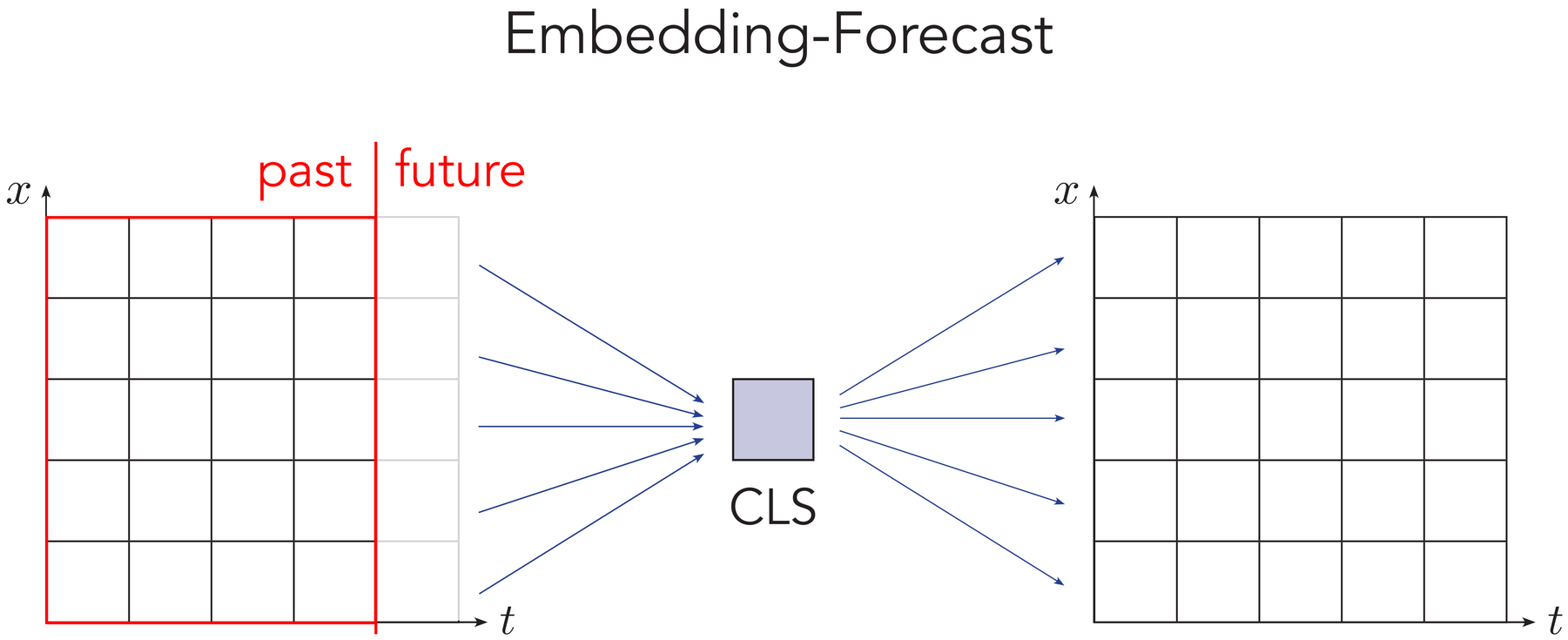
Zero shot evaluation

Embedding-Forecast



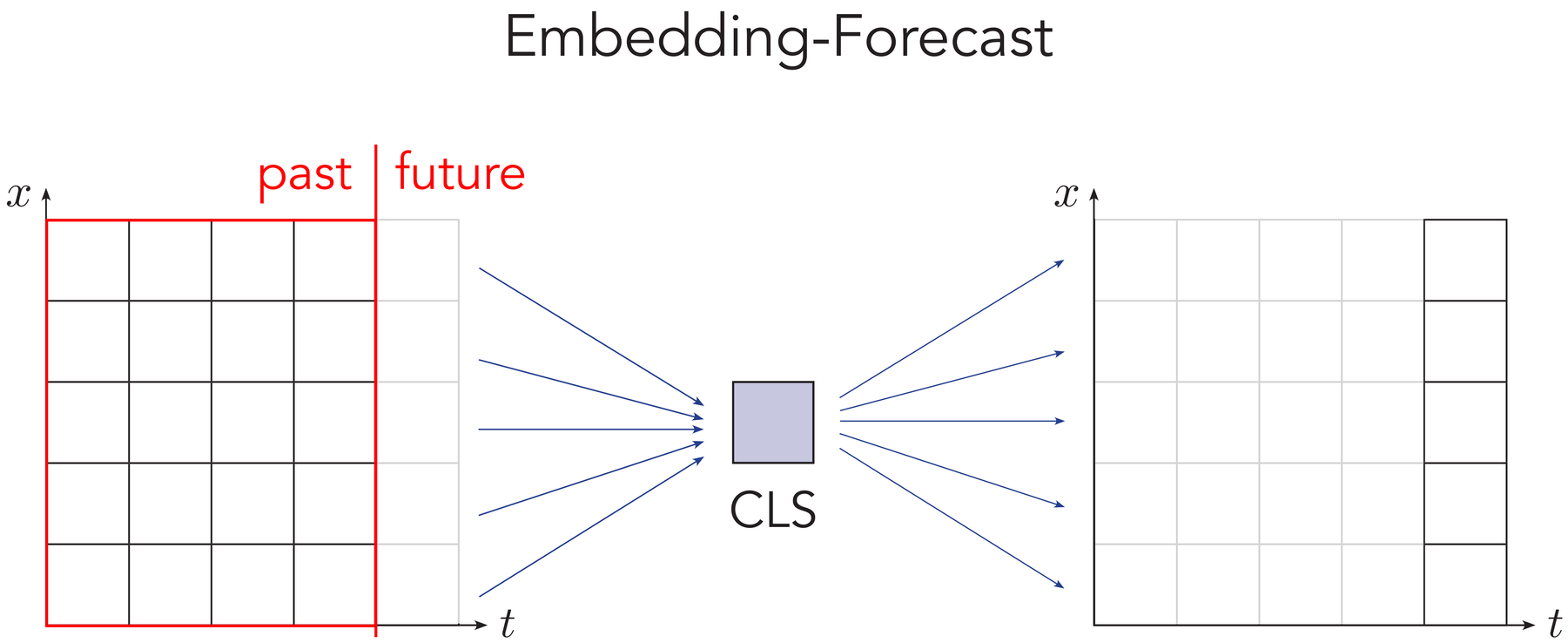
© Christian Lessig, 2023

Zero shot evaluation



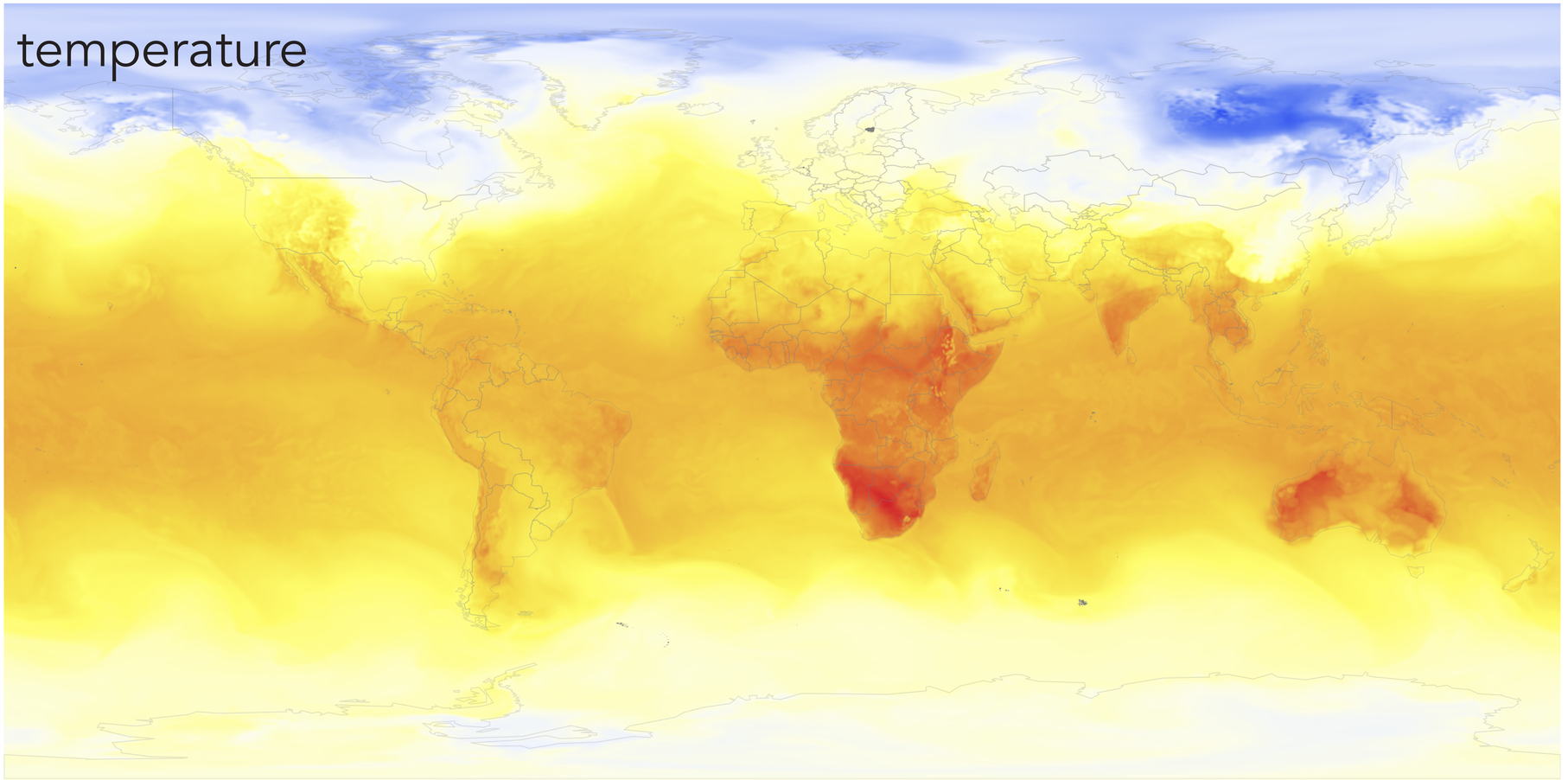
© Christian Lessig, 2023

Zero shot evaluation



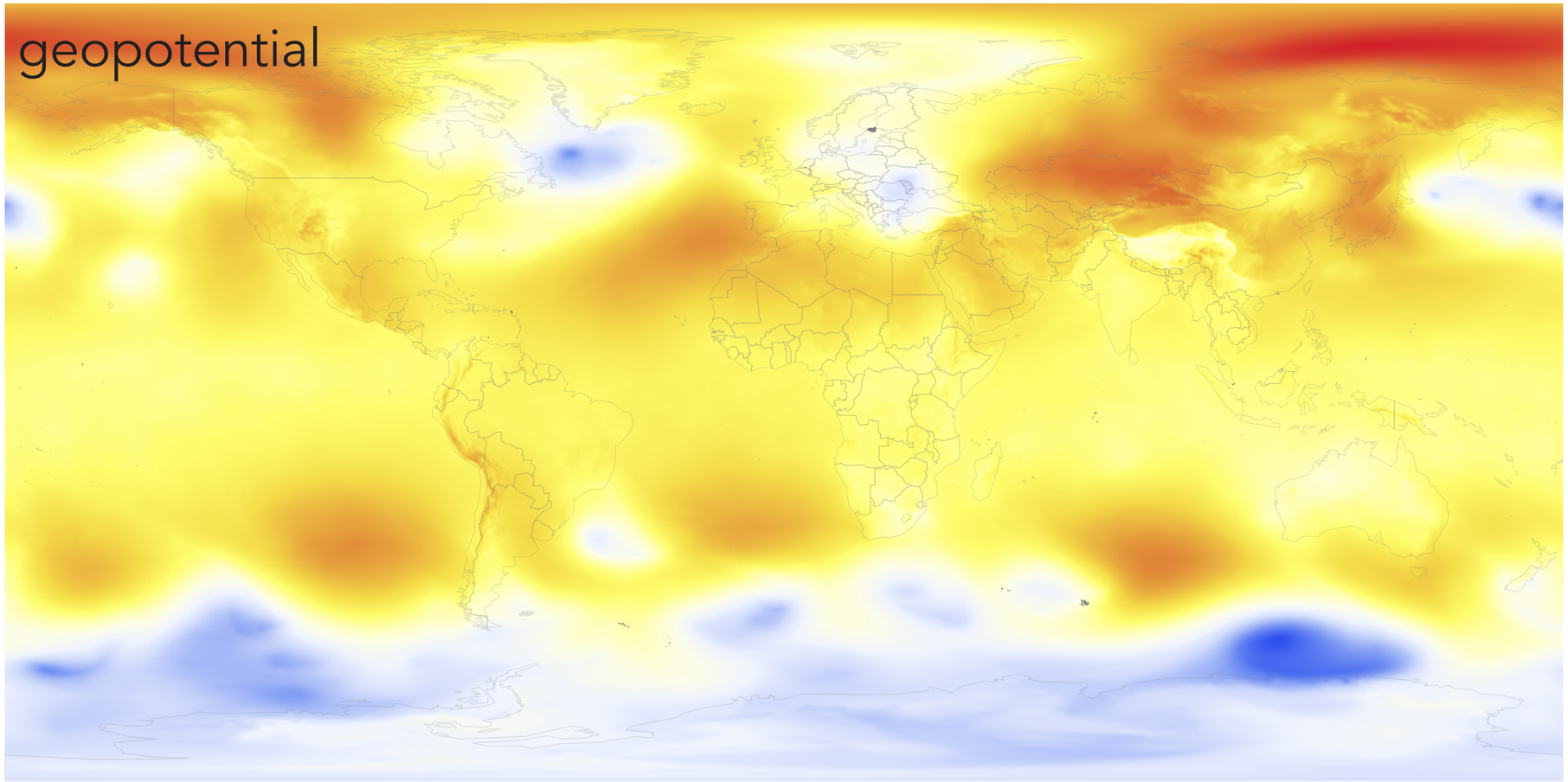
© Christian Lessig, 2023

AtmoRep data



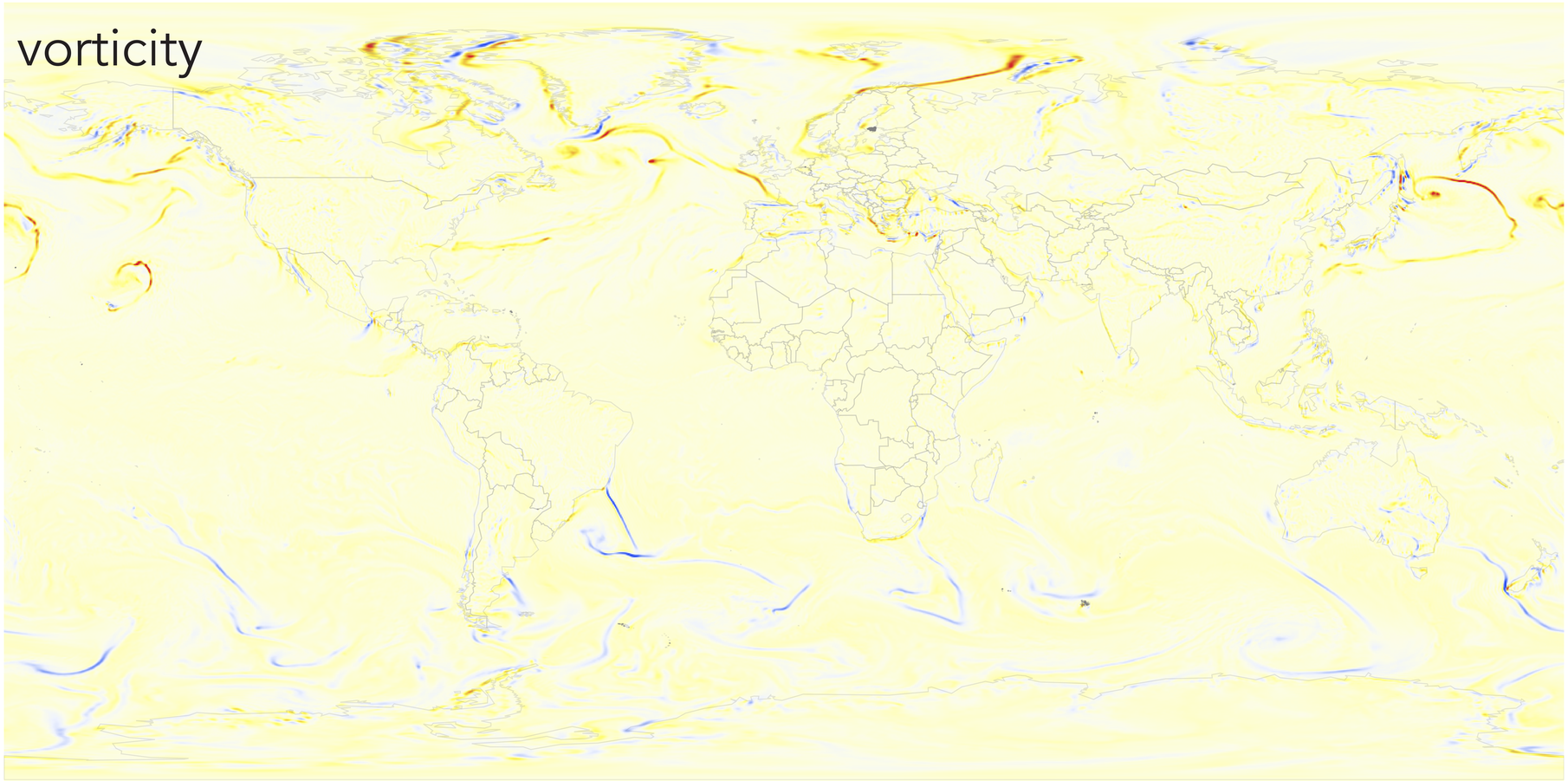
© Christian Lessig, 2023

AtmoRep data



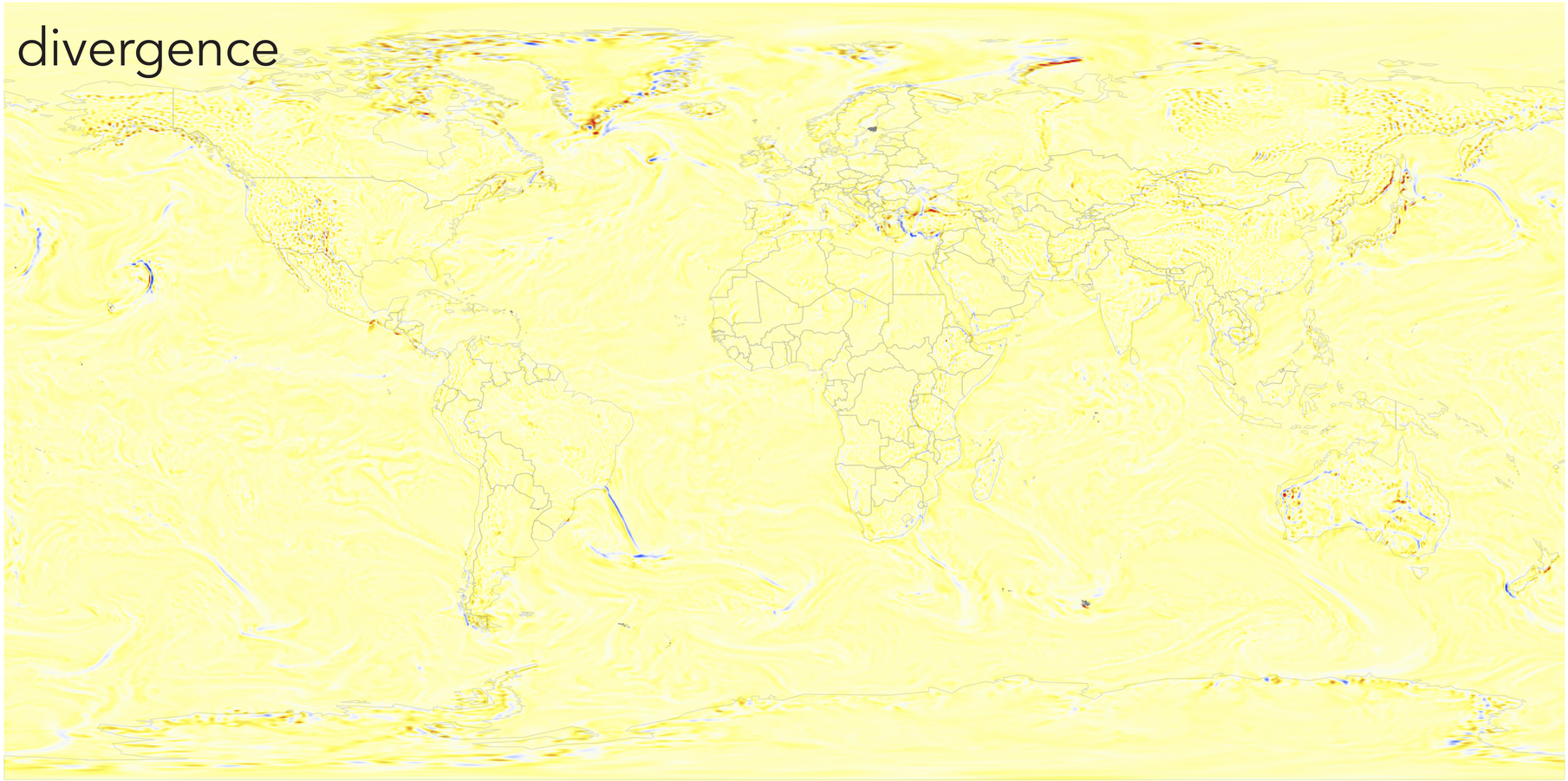
© Christian Lessig, 2023

AtmoRep data



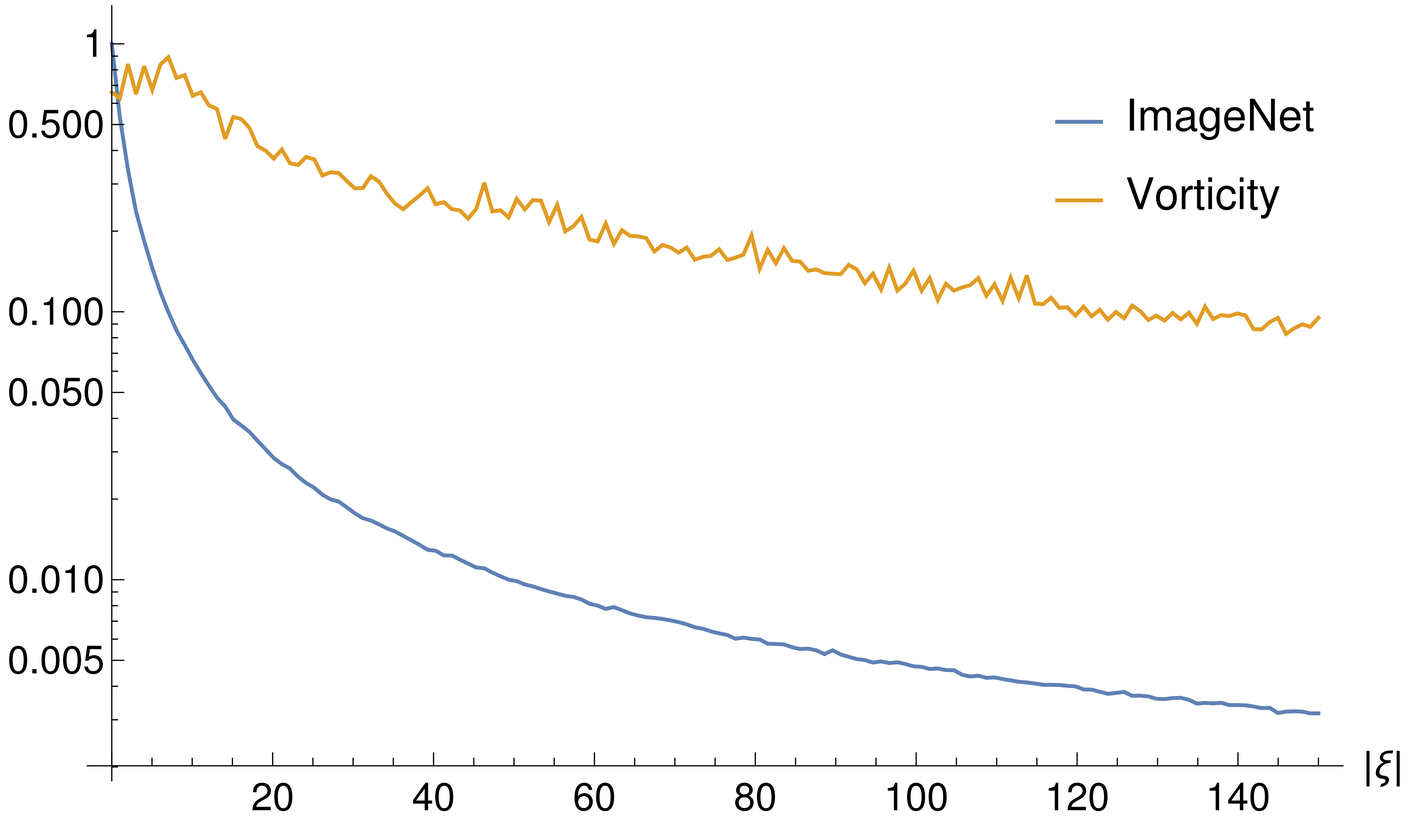
© Christian Lessig, 2023

AtmoRep data



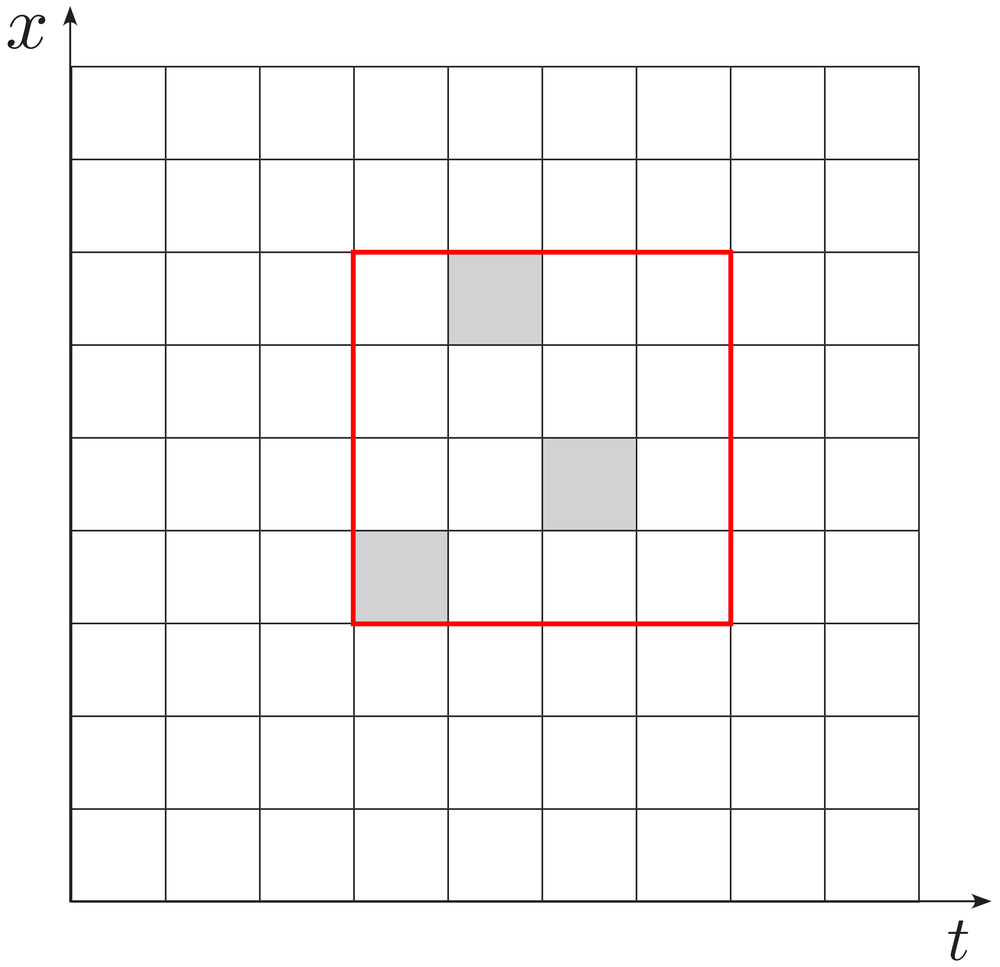
© Christian Lessig, 2023

AtmoRep data



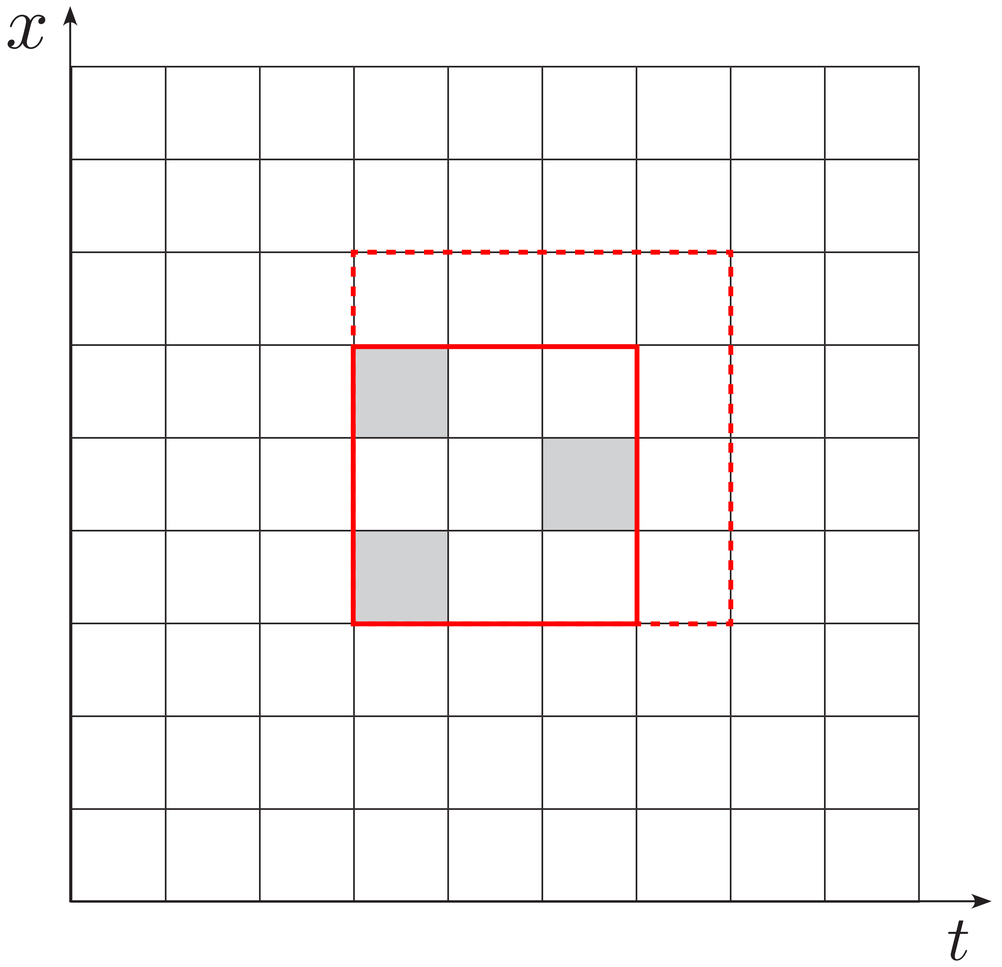
© Christian Lessig, 2023

Spatio-temporal BERT



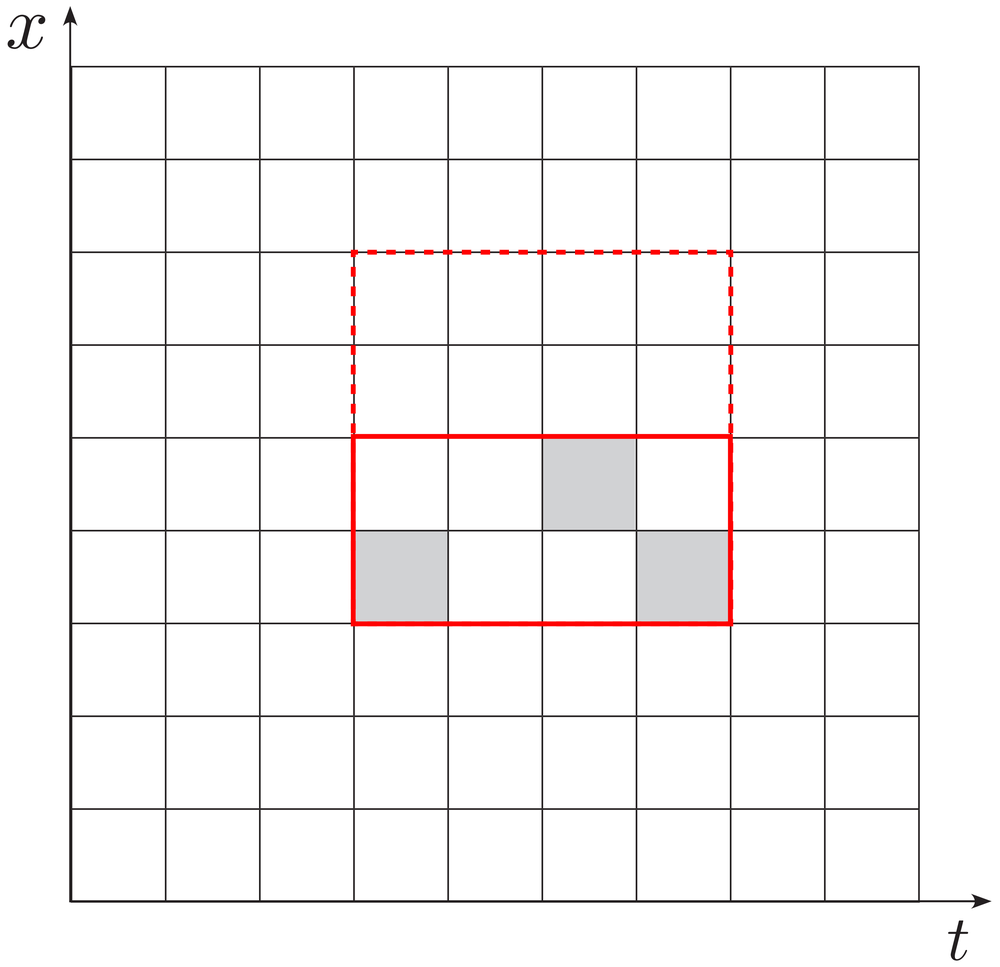
© Christian Lessig, 2023

Spatio-temporal BERT



© Christian Lessig, 2023

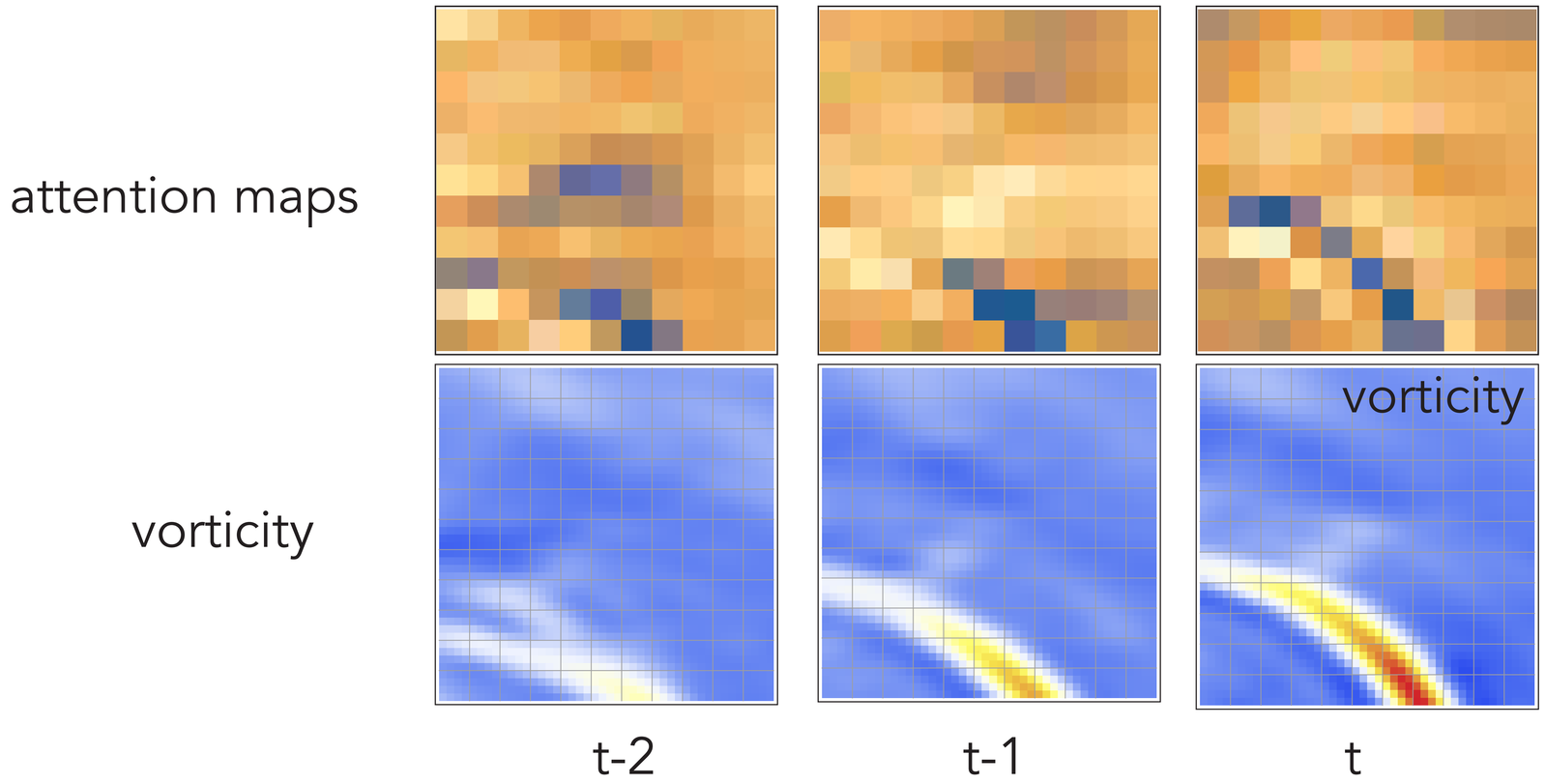
Spatio-temporal BERT



© Christian Lessig, 2023

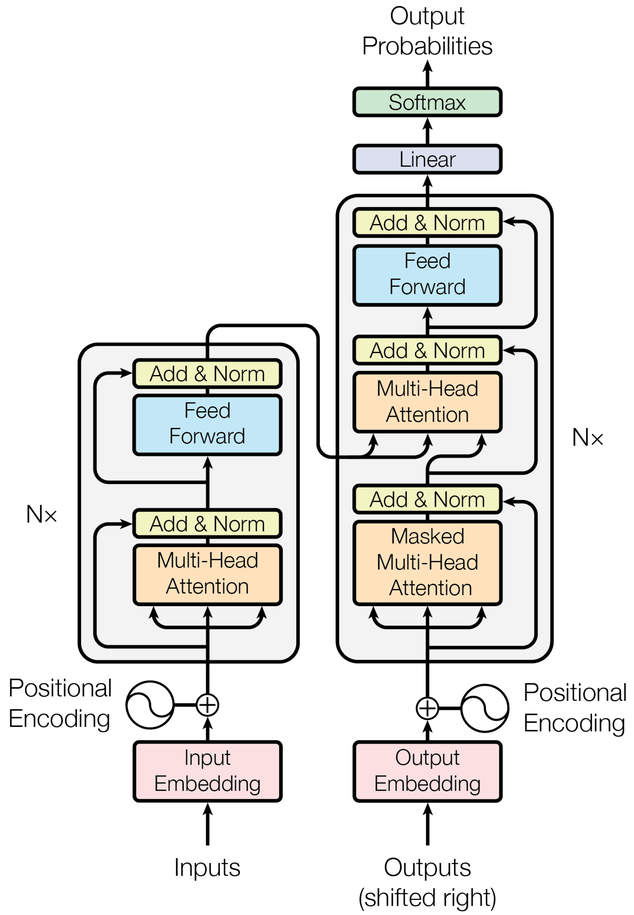
Statistical loss

- Attention maps:



© Christian Lessig, 2023

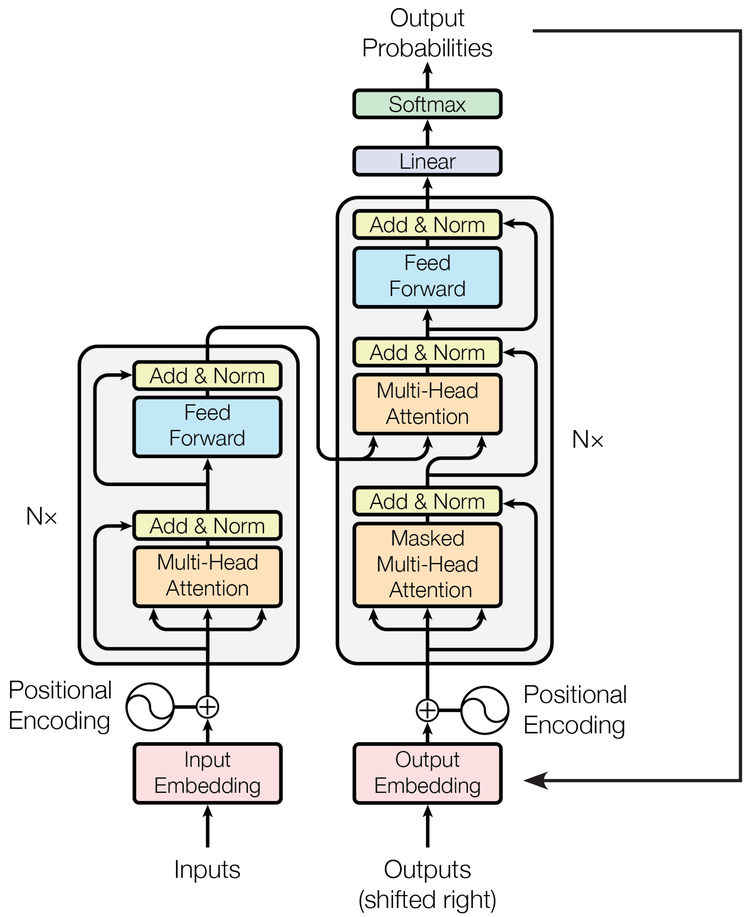
Forecasting / projections



A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.

© Christian Lessig, 2023

Forecasting / projections

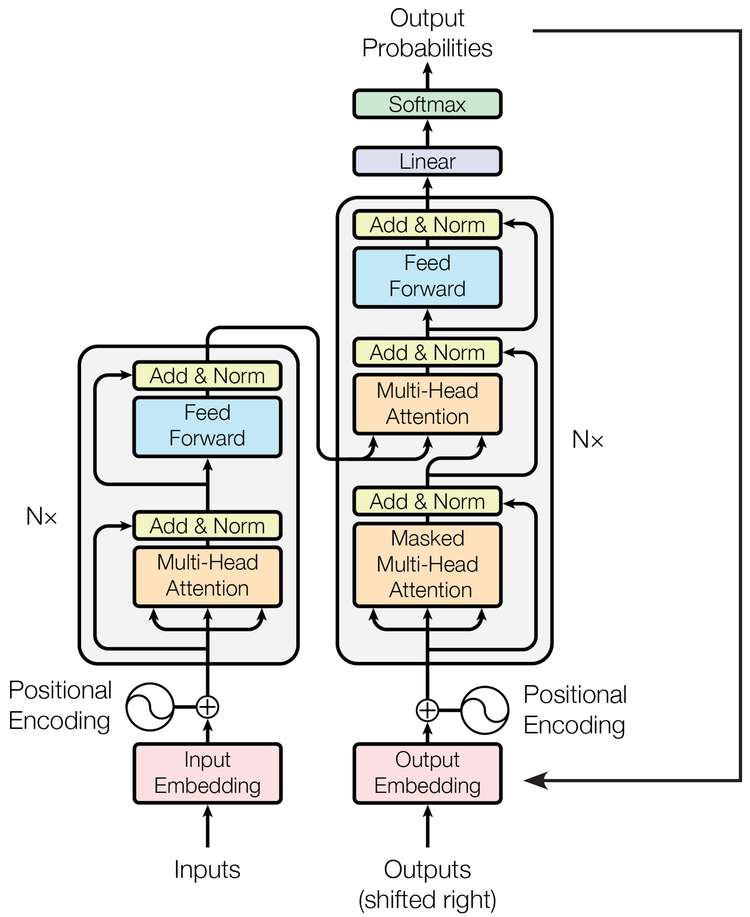


autoregressive, generative modeling

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.

© Christian Lessig, 2023

Forecasting / projections



autoregressive, generative modeling

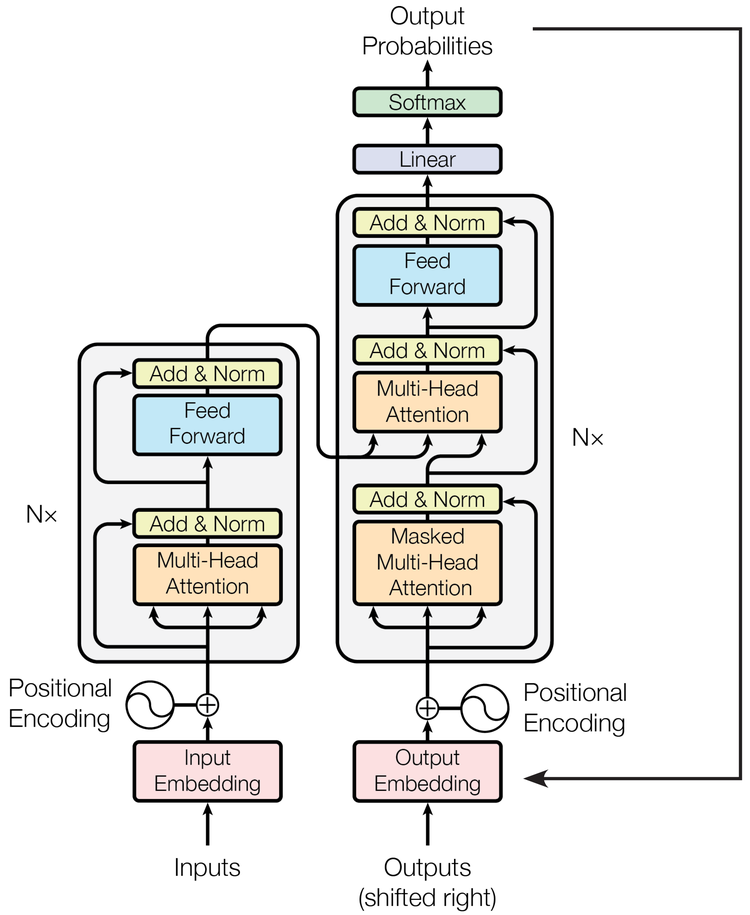
akin to time stepping loop (roll out) for forecasting/projections

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.

© Christian Lessig, 2023

Forecasting / projections

coarse scale/simple classical model



autoregressive, generative modeling

akin to time stepping loop (roll out) for forecasting/projections

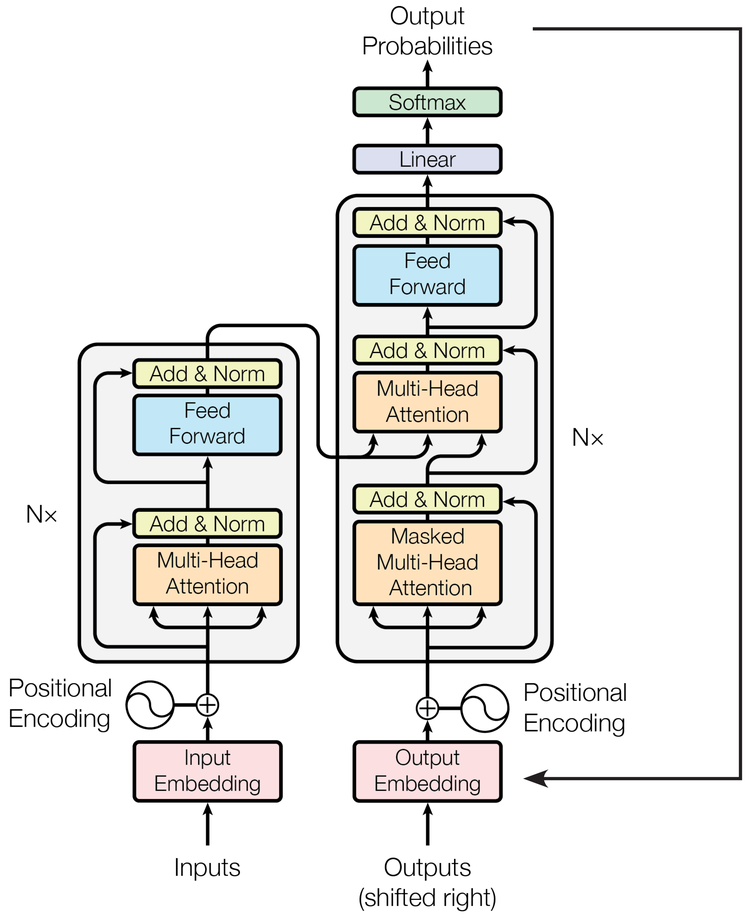
A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.

© Christian Lessig, 2023

Forecasting / projections

coarse scale/simple classical model

slow climate variables



autoregressive, generative modeling

akin to time stepping loop (roll out) for fore-casting/projections

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.

© Christian Lessig, 2023

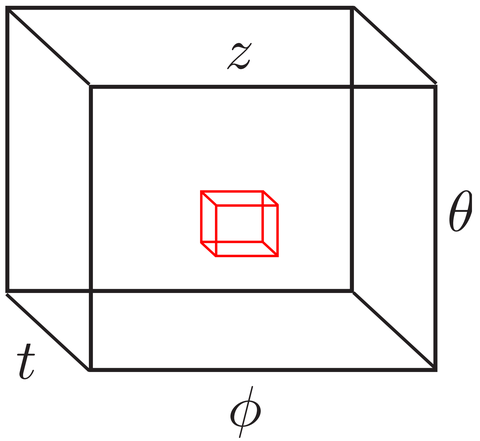
Training

- Unbiased hierarchical Monte Carlo sampling of all possible ERA5 space-time cubes
  - Random sampling of (year,month) tuples corresponding to individual files
  - Random sampling of space-time cubes in tuples
  - Trivially parallelizable with one data loader per field
- Area preserving sampling for sphere/Earth to compensate for distortion of regular grid

© Christian Lessig, 2023

What is a token?

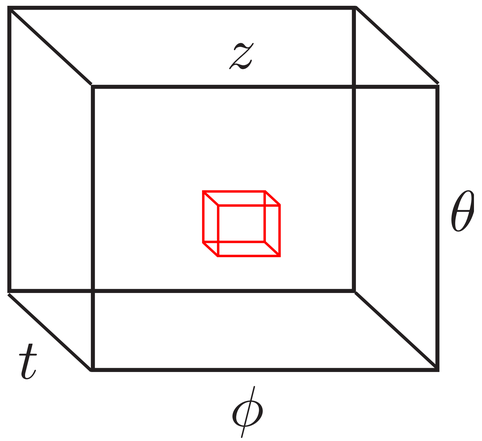
- Token is small neighborhood in space-time
  - Small for token attention / inter-action to be informative
  - Big enough so token has rich internal strurcture



© Christian Lessig, 2023

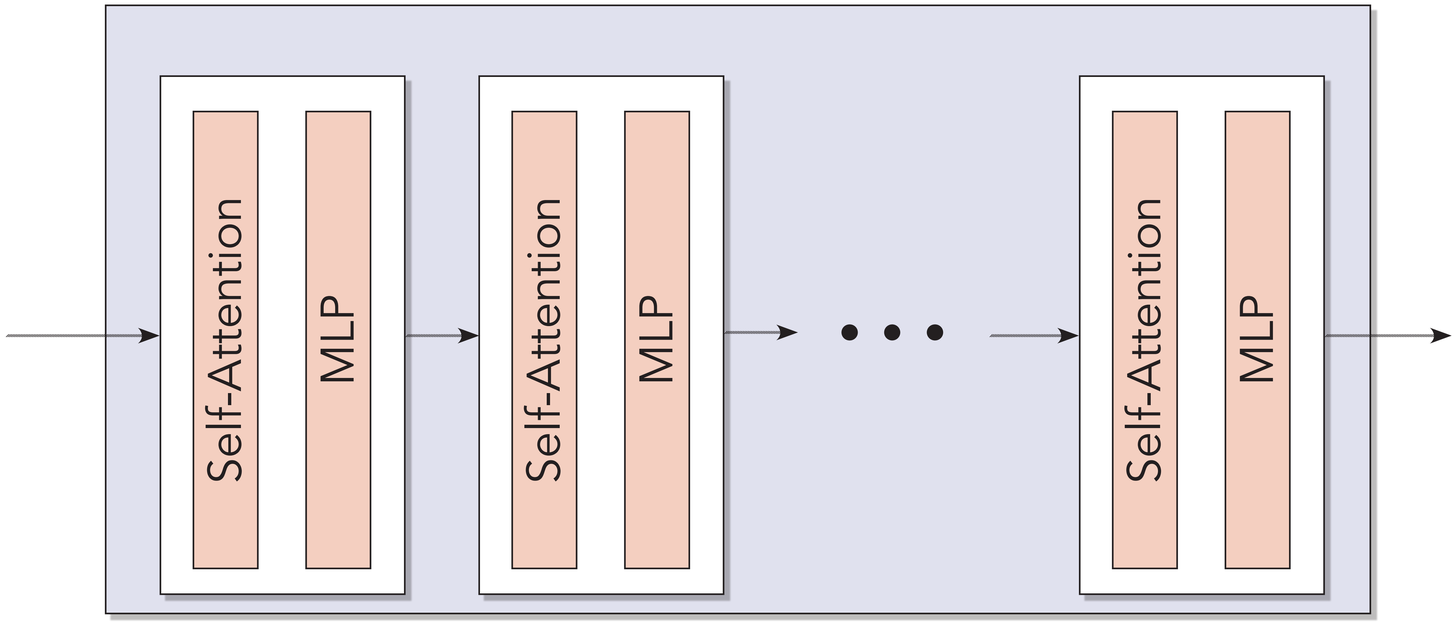
What is a token?

- Token is small neighborhood in space-time
  - Small for token attention / inter-action to be informative
  - Big enough so token has rich internal strurcture
- Token size is field-dependent



© Christian Lessig, 2023

Multiformer

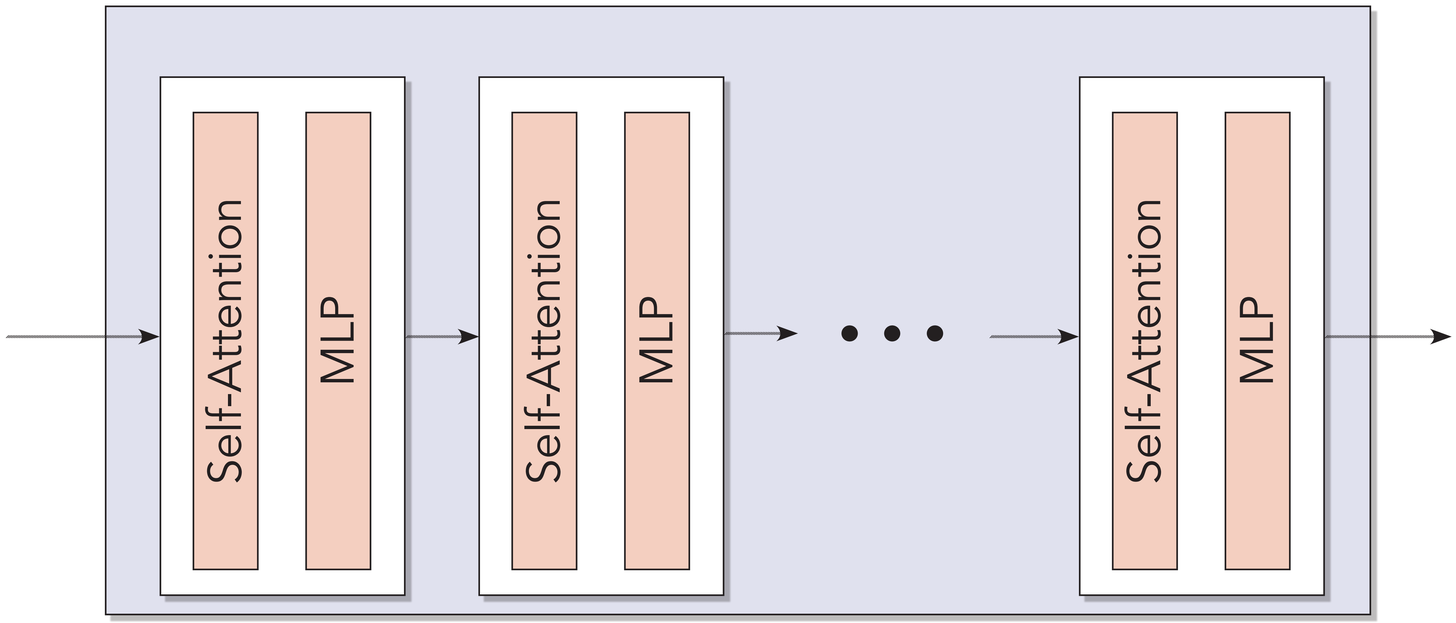


© Christian Lessig, 2023

Multiformer

Self
attention

$$\sigma\left(Q\,K^T\right)V$$

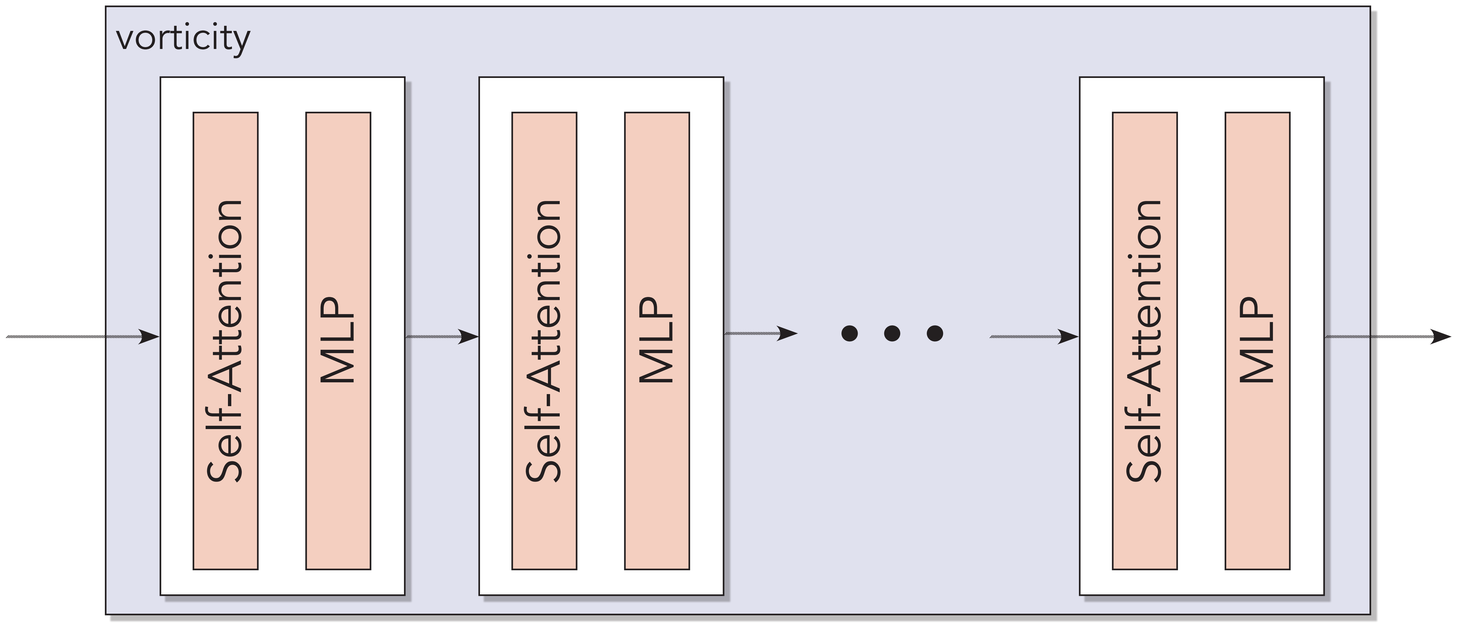


© Christian Lessig, 2023

Multiformer

Self
attention

$$\sigma\left(Q\,K^T\right)V$$

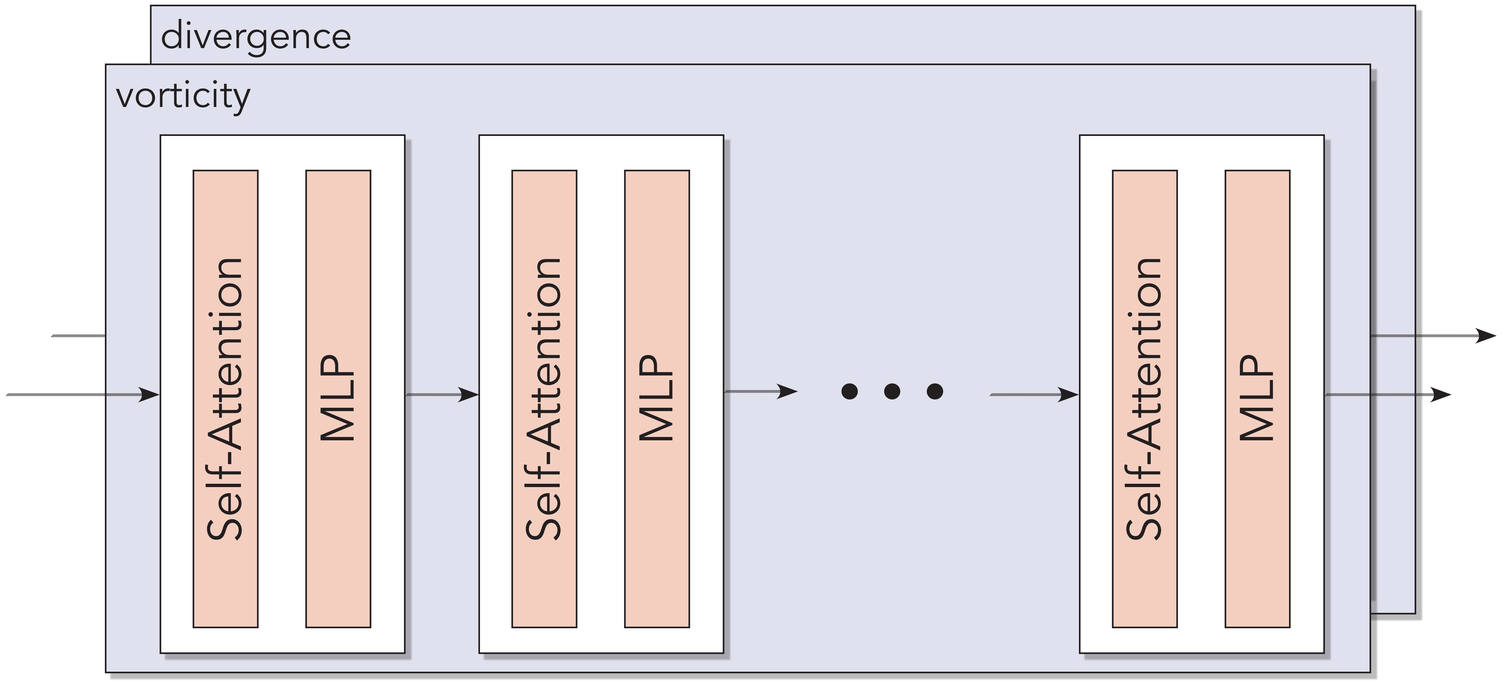


© Christian Lessig, 2023

Multiformer

Self
attention

$$\sigma\left(Q\,K^T\right)V$$

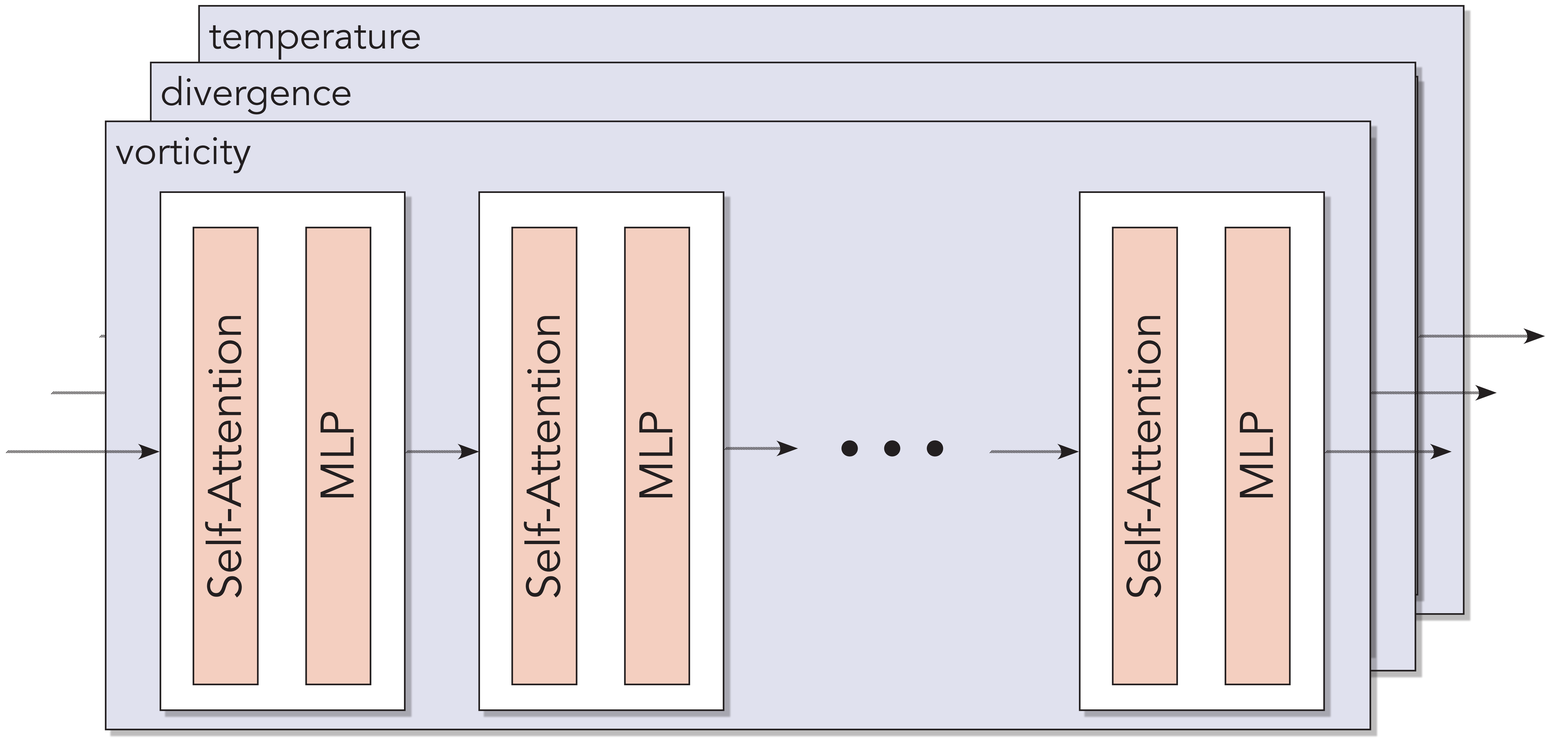


© Christian Lessig, 2023

Multiformer

Self
attention

$$\sigma\left(Q\,K^T\right)V$$

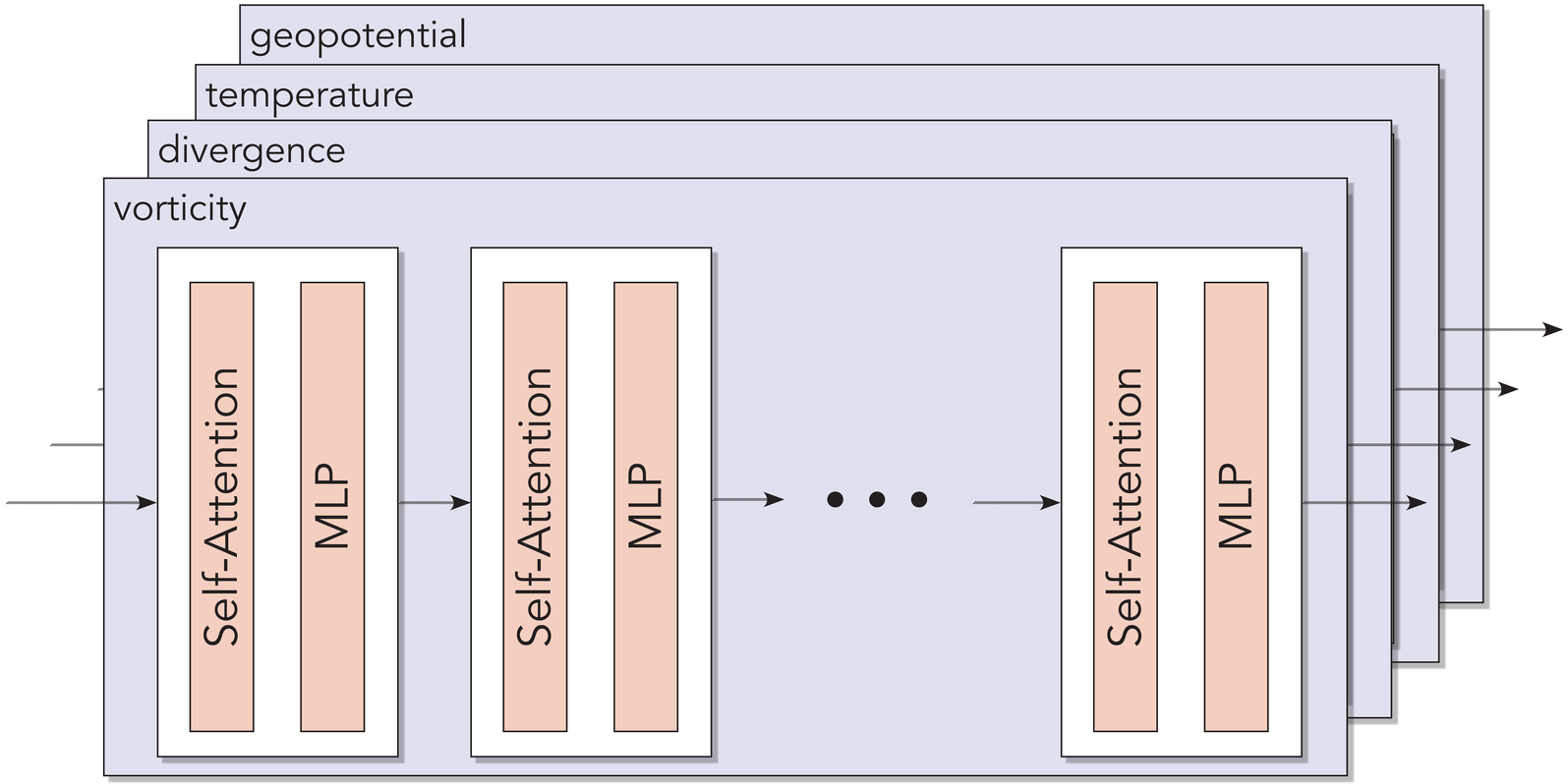


© Christian Lessig, 2023

Multiformer

Self
attention

$$\sigma\left(Q\,K^T\right)V$$

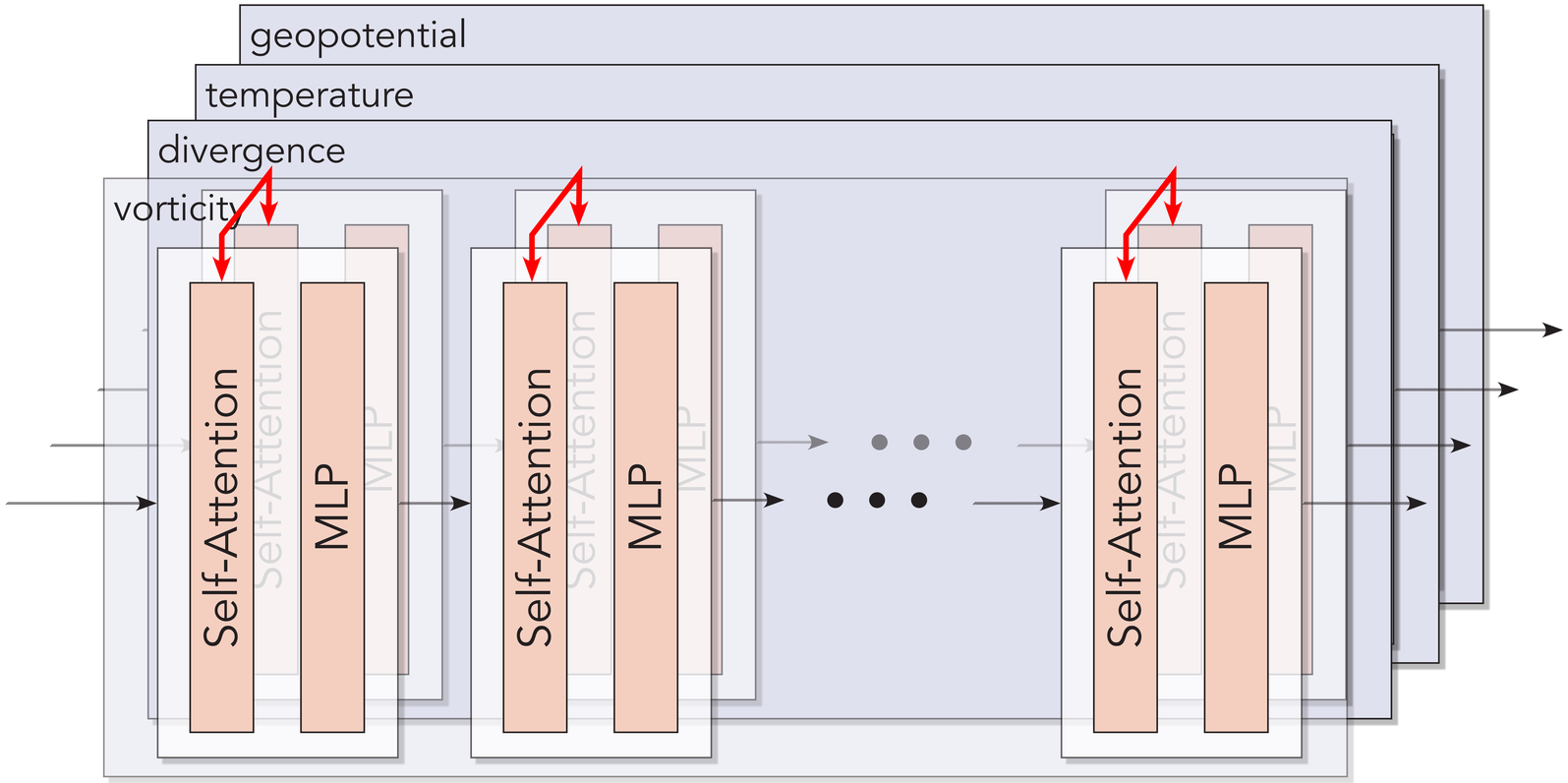


© Christian Lessig, 2023

Multiformer

Self
attention

$$\sigma\left(Q\,K^T\right)V$$



© Christian Lessig, 2023

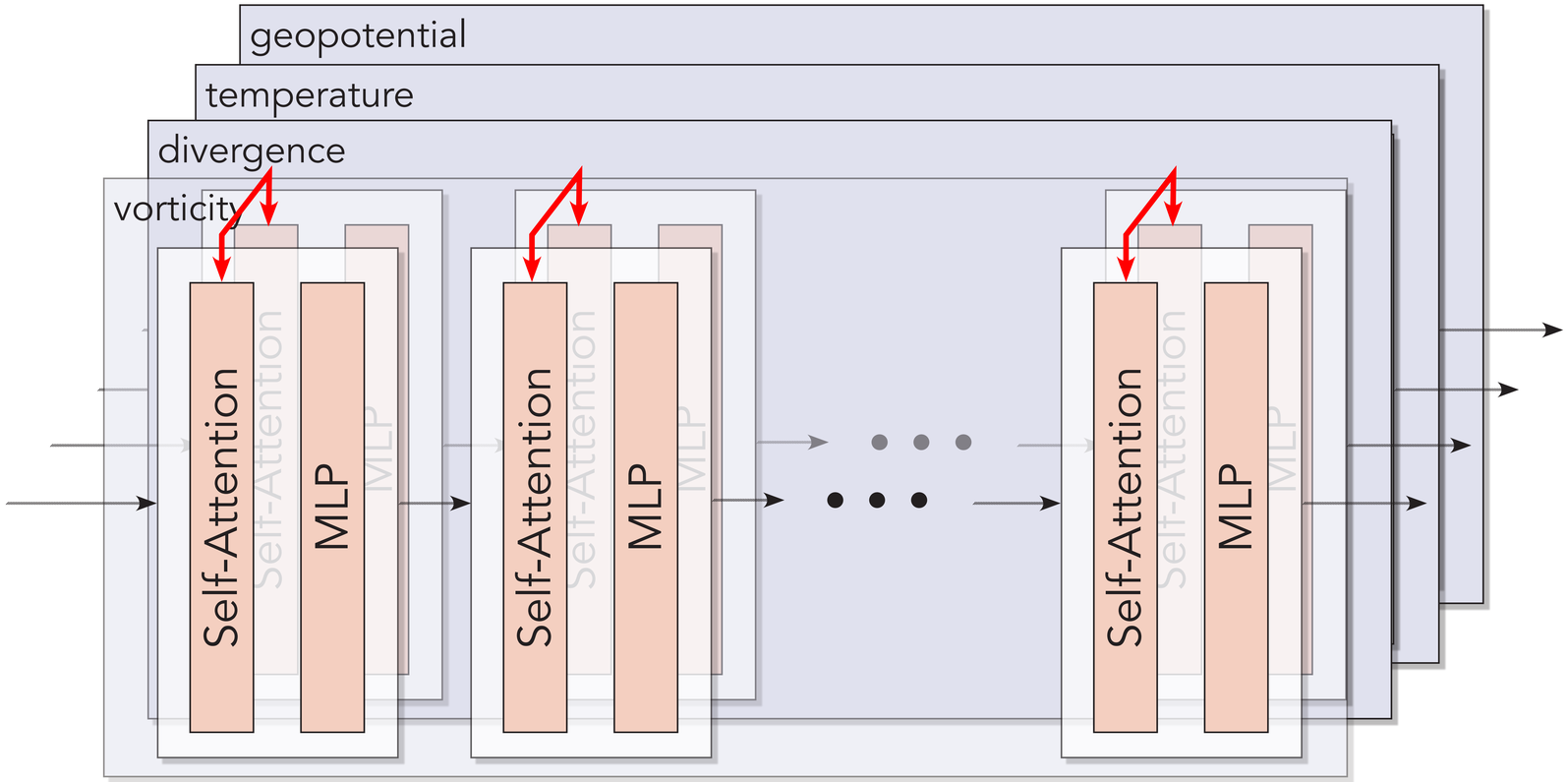
Multiformer

Self attention

$$\sigma\left(Q\,K^T\right)V$$

Cross attention

$$\sigma\left(Q_\zeta\,K_\mu^T\right)V_\mu$$



© Christian Lessig, 2023

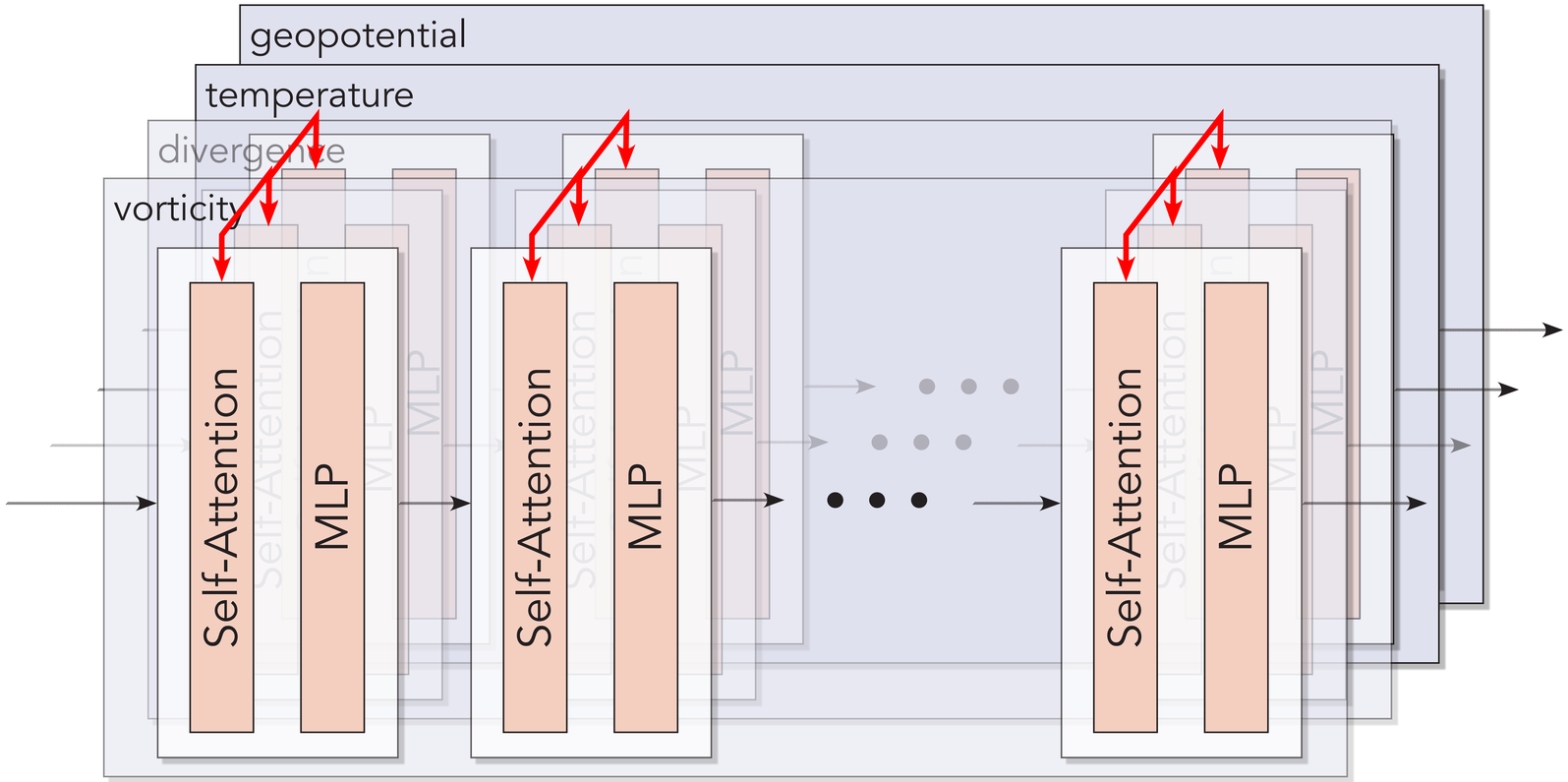
Multiformer

Self attention

$$\sigma\left(Q\,K^T\right)V$$

Cross attention

$$\sigma\left(Q_\zeta\,K_\mu^T\right)V_\mu$$



© Christian Lessig, 2023

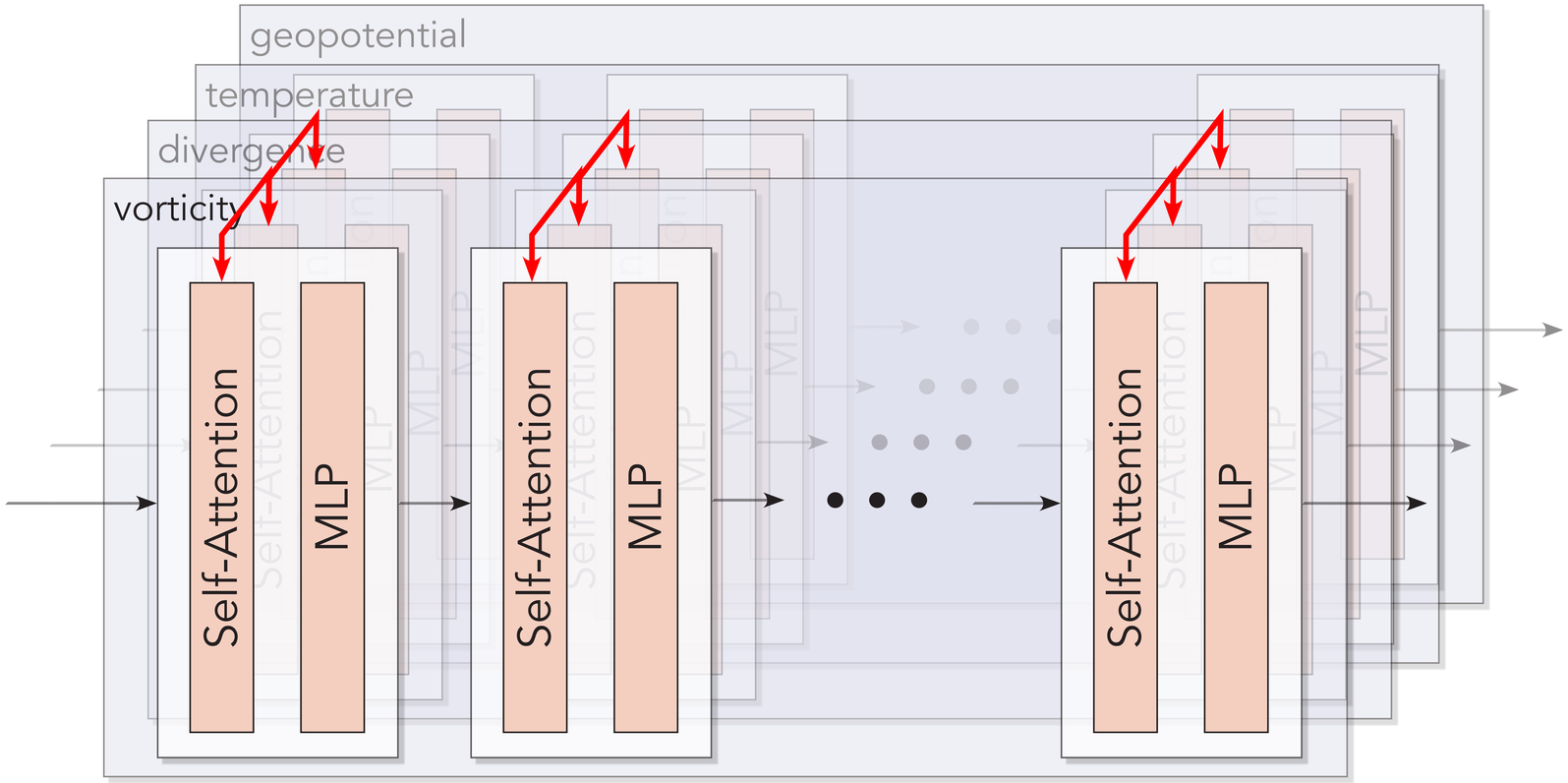
Multiformer

Self attention

$$\sigma\left(Q\,K^T\right)V$$

Cross attention

$$\sigma\left(Q_\zeta\,K_\mu^T\right)V_\mu$$



© Christian Lessig, 2023

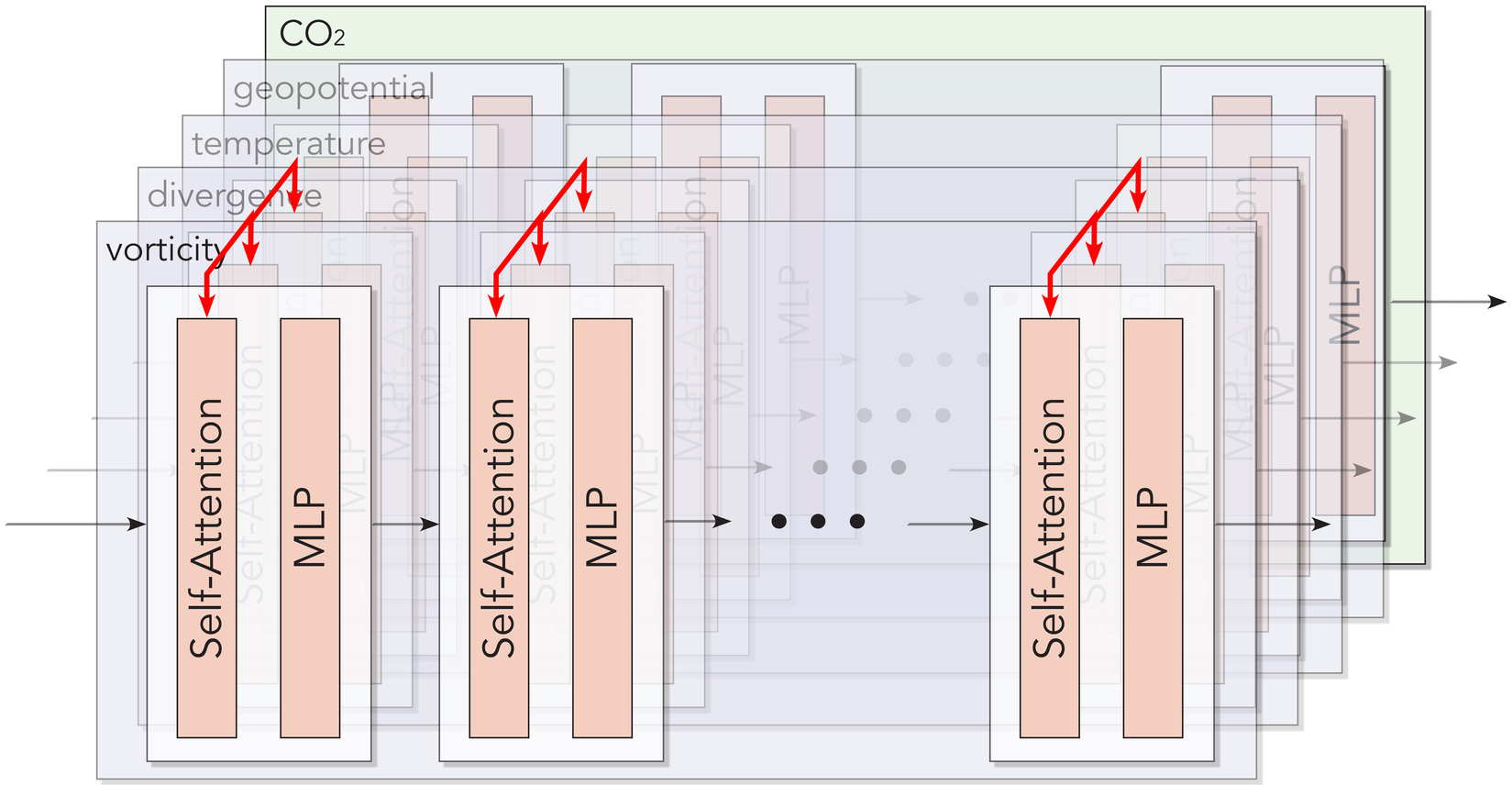
Multiformer

Self attention

$$\sigma\left(Q\,K^T\right)V$$

Cross attention

$$\sigma\left(Q_\zeta\,K_\mu^T\right)V_\mu$$



© Christian Lessig, 2023

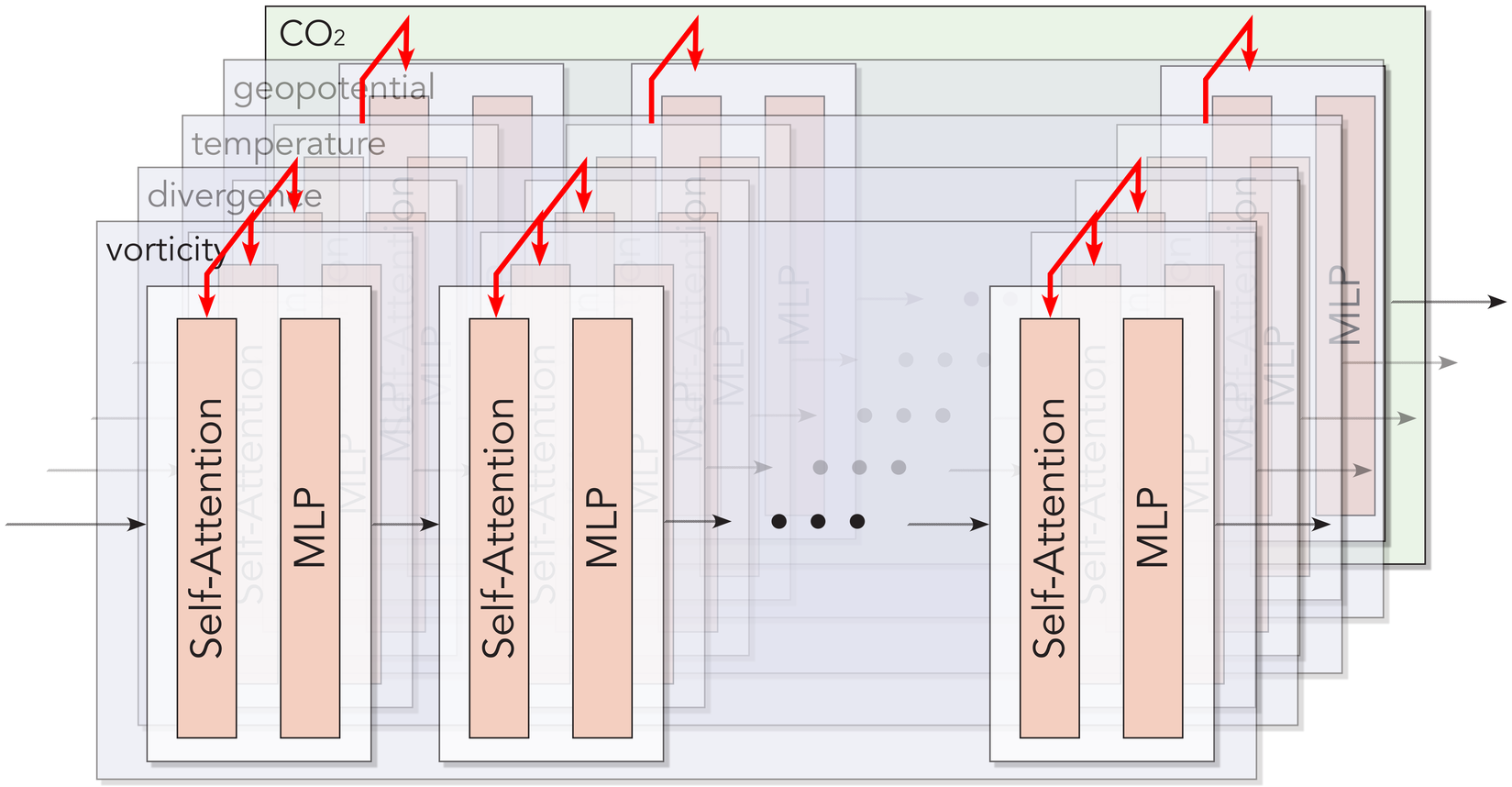
Multiformer

Self attention

$$\sigma\left(Q\,K^T\right)V$$

Cross attention

$$\sigma\left(Q_\zeta\,K_\mu^T\right)V_\mu$$



© Christian Lessig, 2023

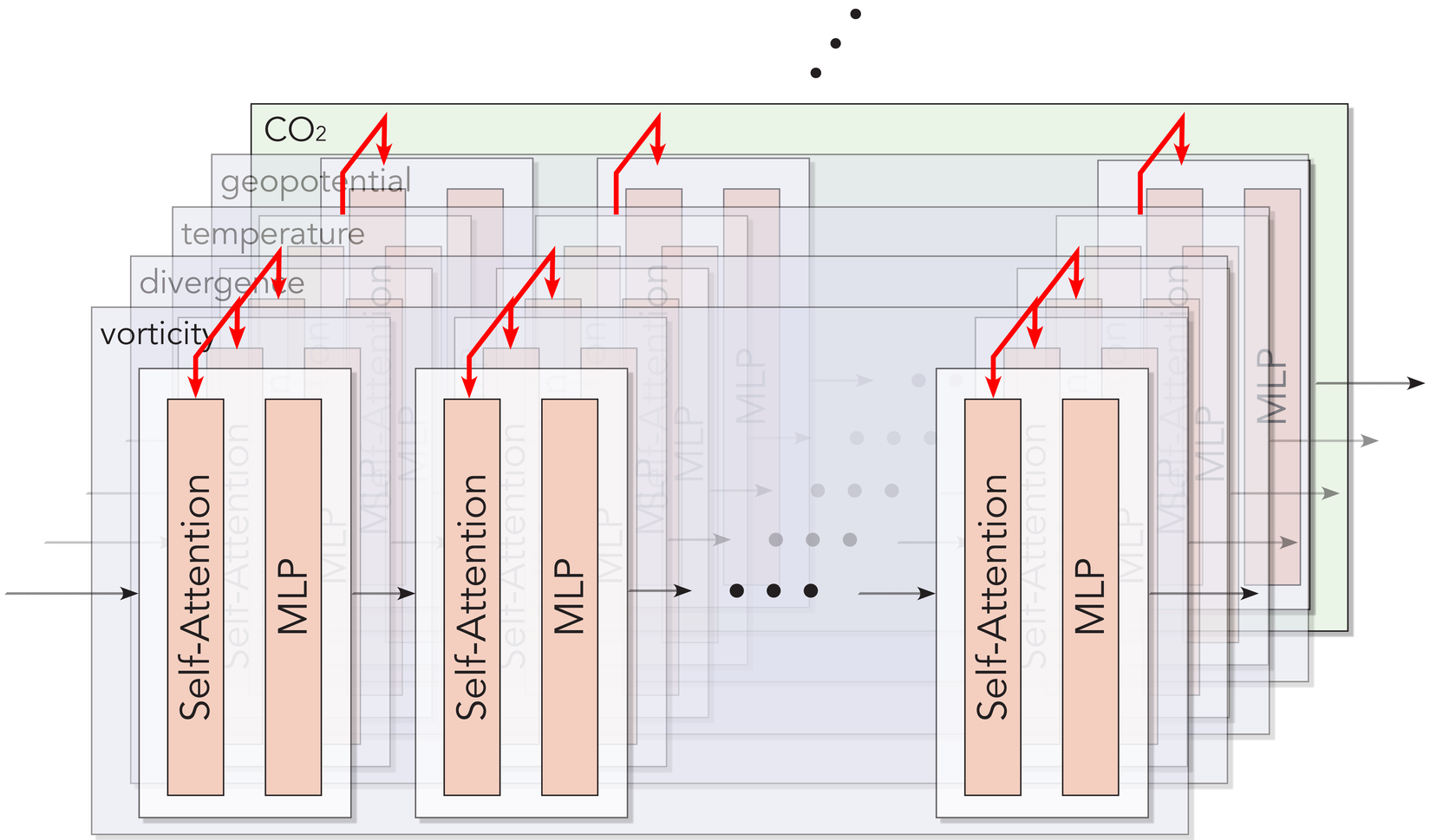
Multiformer

Self attention

$$\sigma\left(Q\,K^T\right)V$$

Cross attention

$$\sigma\left(Q_\zeta\,K_\mu^T\right)V_\mu$$

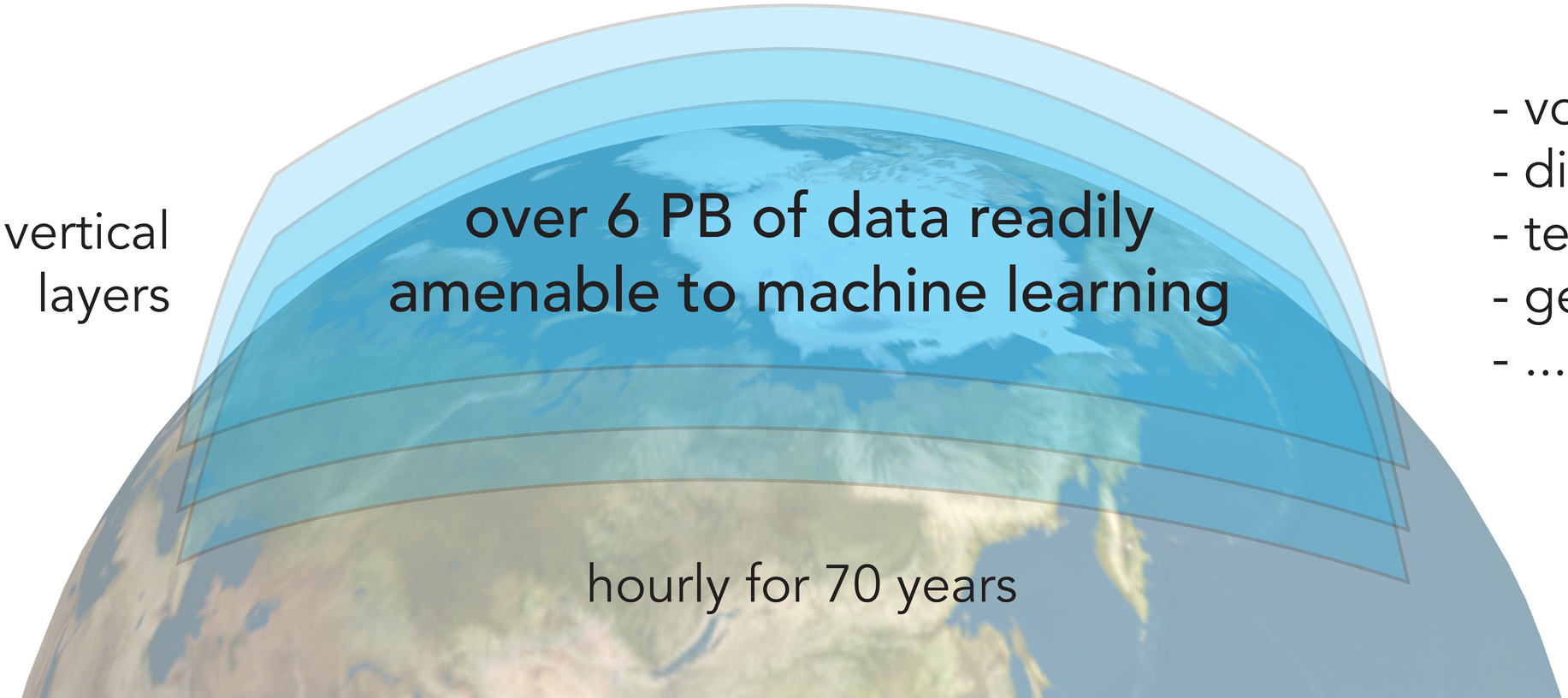


© Christian Lessig, 2023

Data: ERA5 reanalysis

721x1440 horizontal grid (0.25 degree)

137 vertical layers



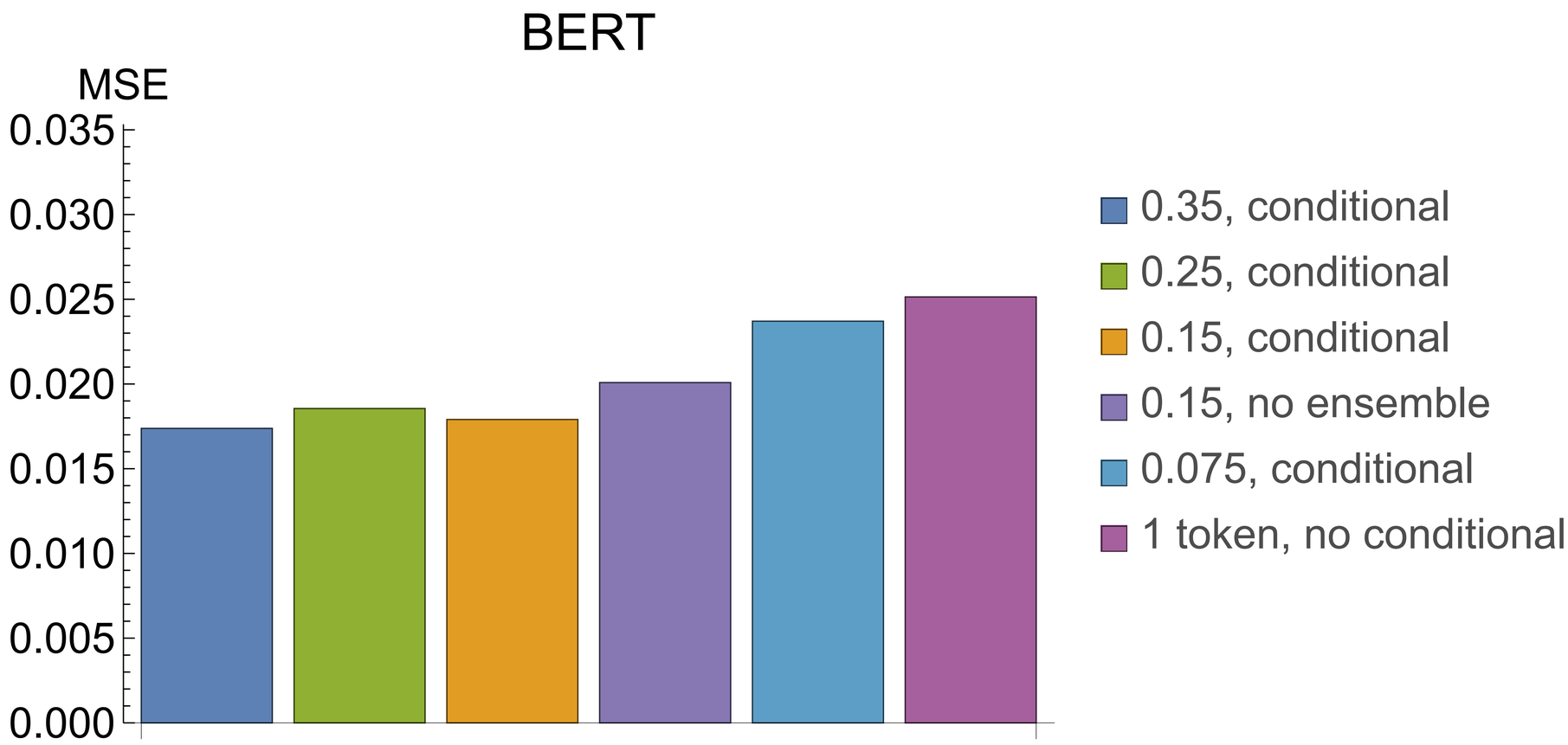
over 6 PB of data readily amenable to machine learning

- vorticity
- divergence
- temperature
- geopotential
- ...

hourly for 70 years

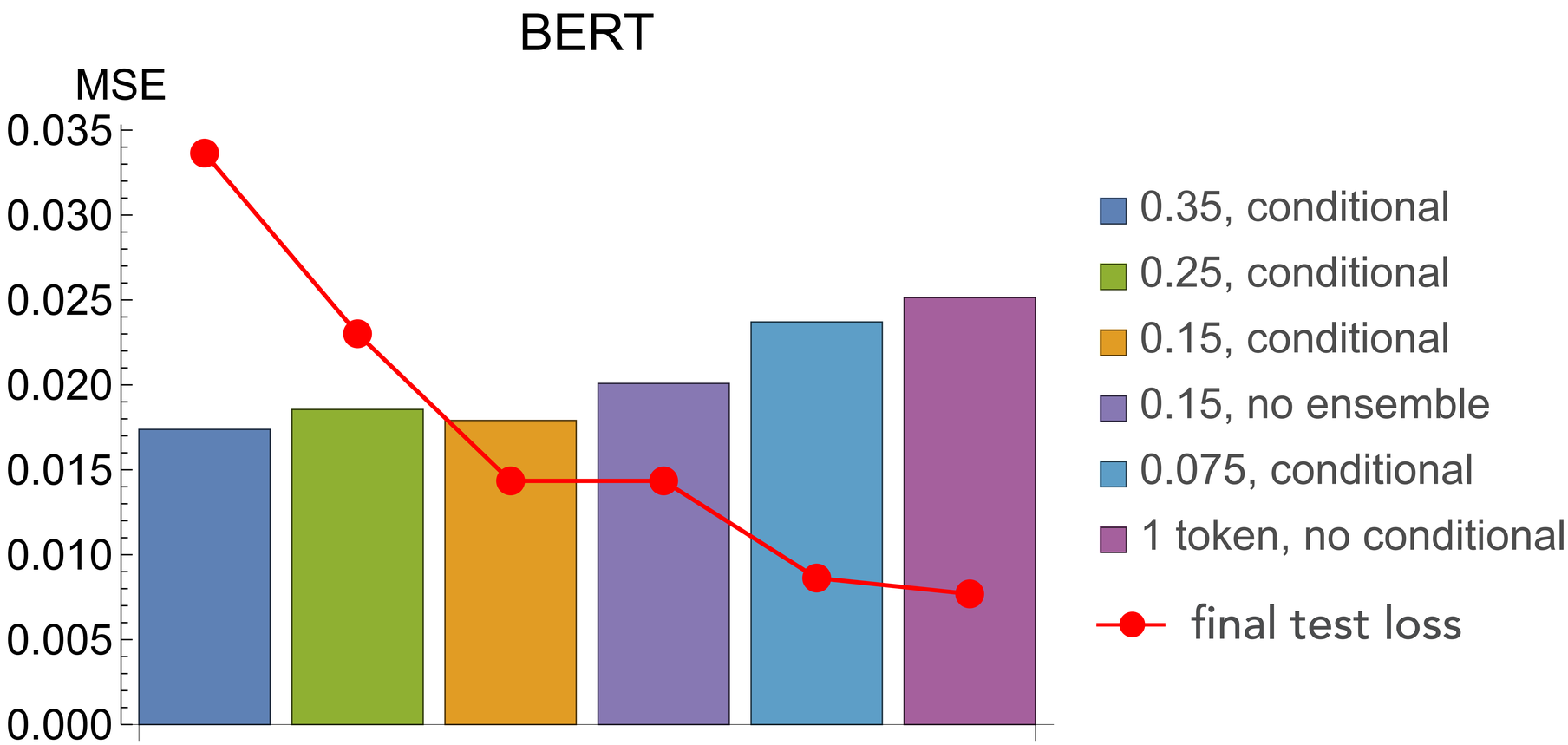
© Christian Lessig, 2023

Zero shot forecasting



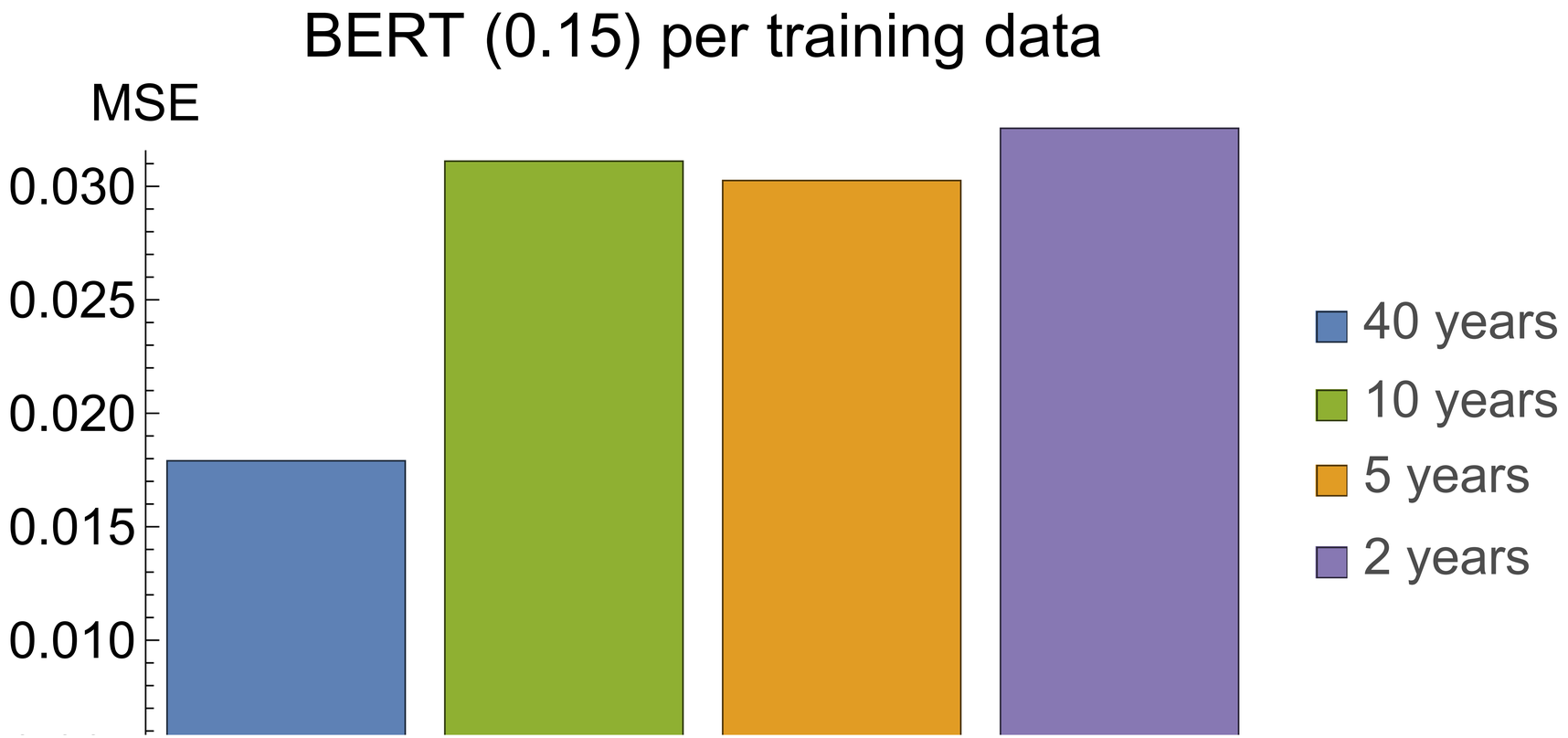
© Christian Lessig, 2023

Zero shot forecasting



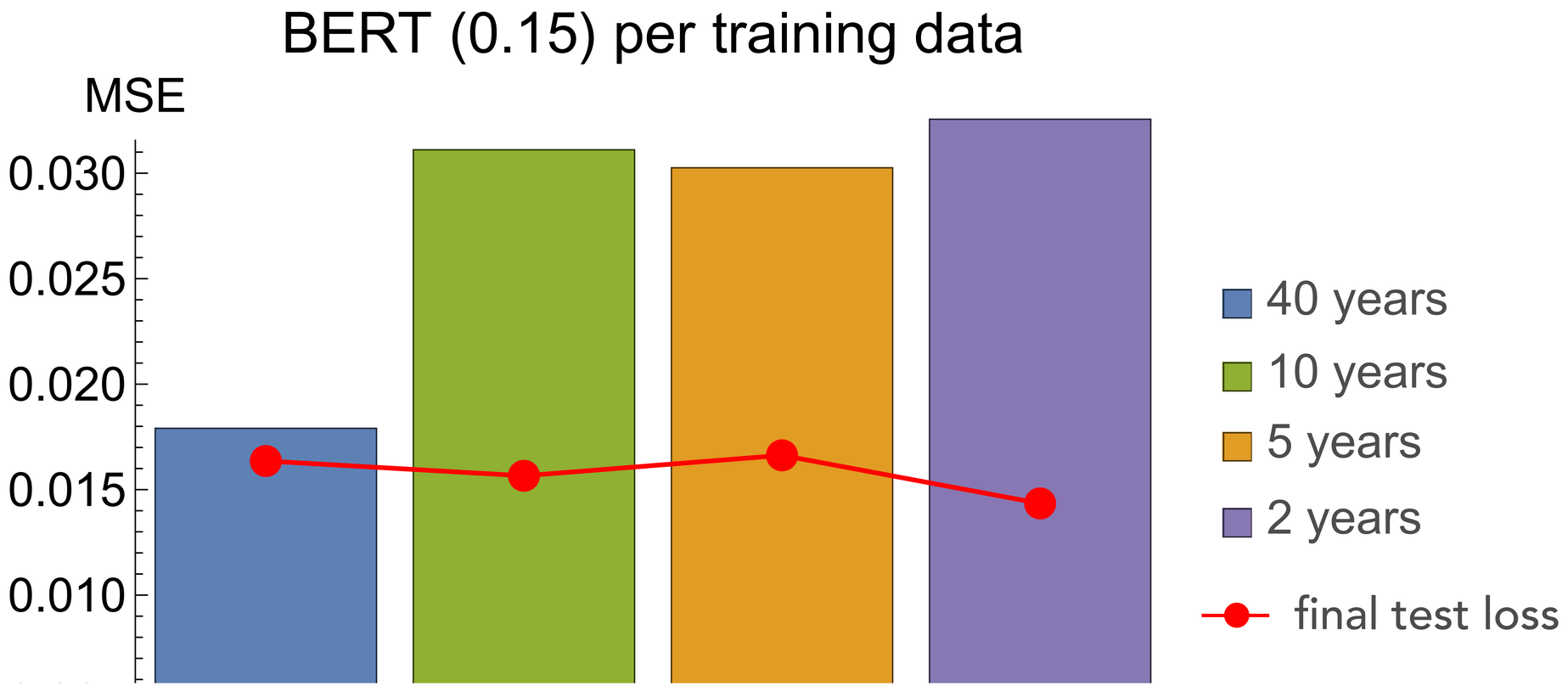
© Christian Lessig, 2023

Zero shot forecasting



© Christian Lessig, 2023

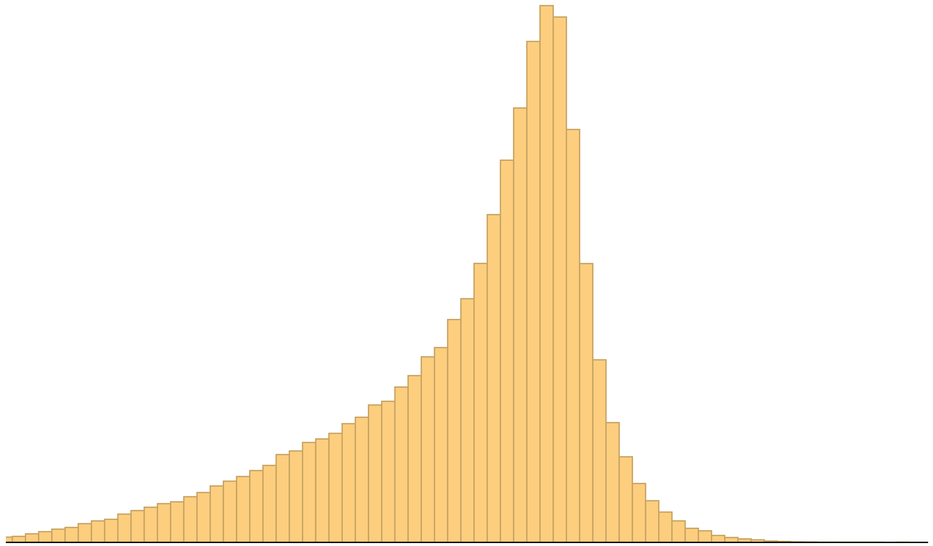
Zero shot forecasting



© Christian Lessig, 2023

Zero shot forecasting

Cape town: historgram of vorticity

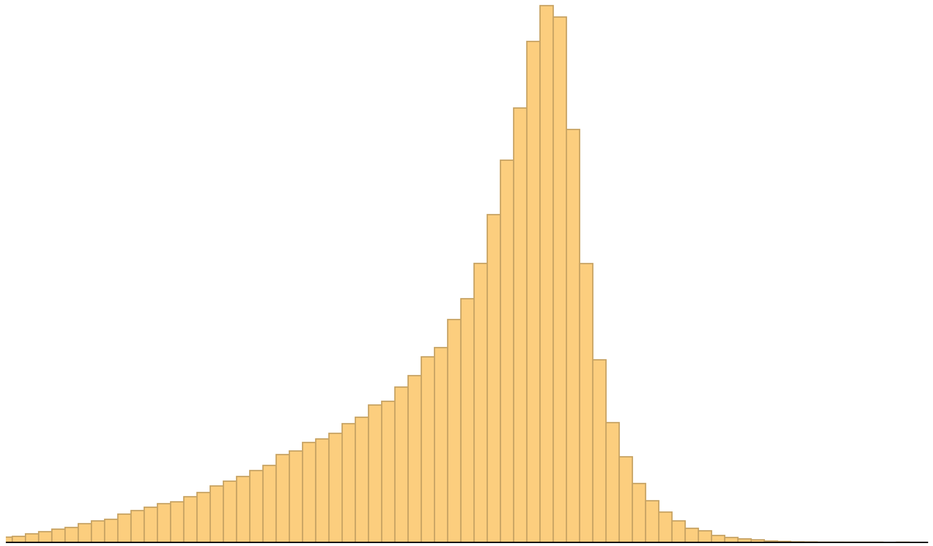


ERA5

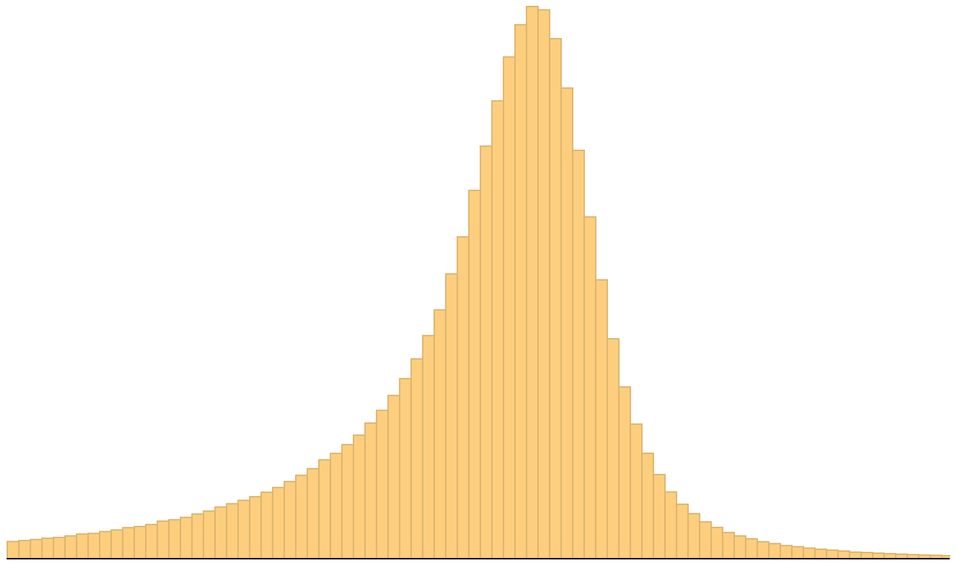
© Christian Lessig, 2023

Zero shot forecasting

Cape town: historgram of vorticity



ERA5



predictions

© Christian Lessig, 2023