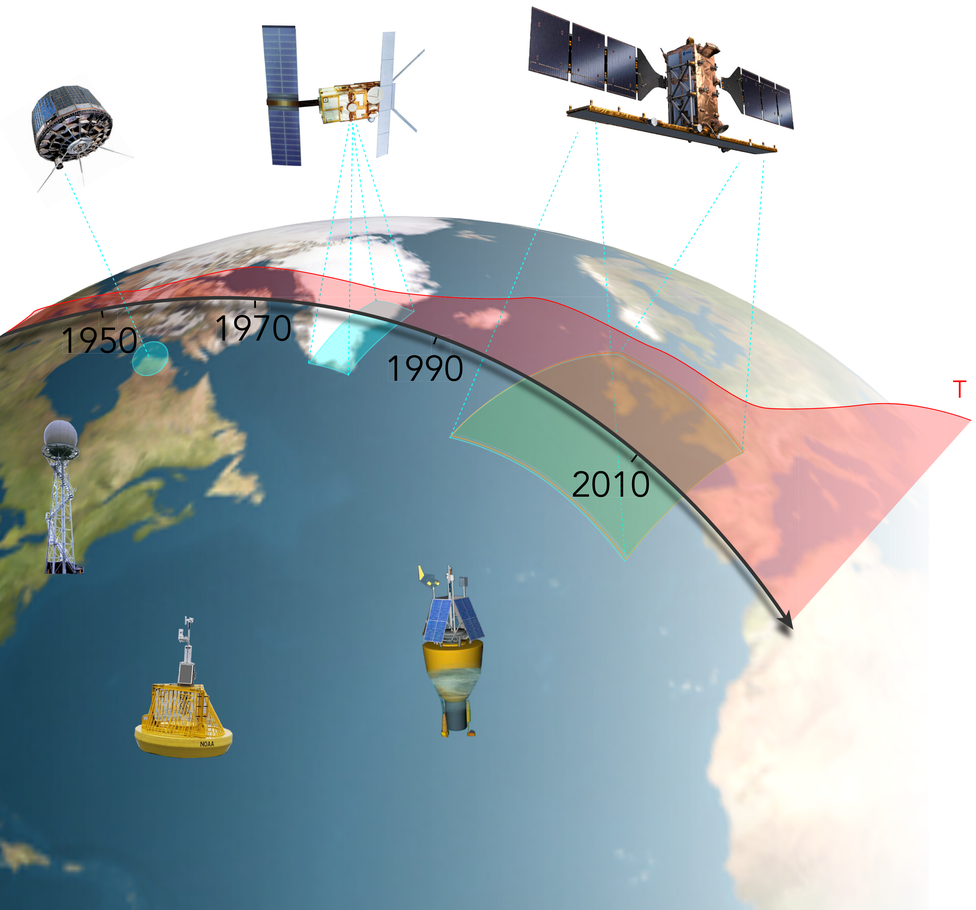
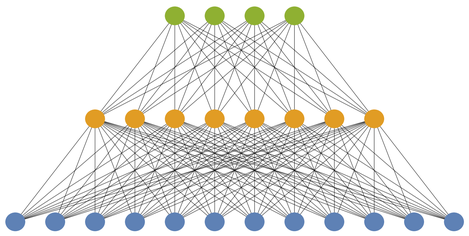


AtmoRep

Large scale representation learning of atmospheric dynamics

Christian Lessig, Ilaria Luise, Martin Schultz, et al.



address climate change

large scale machine learning

scientific insight

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 (O(PB)) with very large networks (O(10^{11}) parameters)
  - 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 (O(PB)) with very large networks (O(10^{11}) parameters)
  - Useful for downstream applications using tail network, fine-tuning, or **in-context learning**

© Christian Lessig, 2023

Can we perform representation learning in the Earth sciences?

© Christian Lessig, 2023

Representation learning for the Earth sciences?

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

© Christian Lessig, 2023

Representation learning for the Earth sciences?

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

© Christian Lessig, 2023

Representation learning for the Earth sciences?

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

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

© Christian Lessig, 2023

Representation learning for the Earth sciences?

- 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 the Earth sciences?

- Very large amounts of **observational** data
- No complete classical model for system and dynamics
- Chaosticisty in atmospheric dynamics leads to ambiguity
  - There is often not one “correct answer“
  - Large networks learn statistical representations

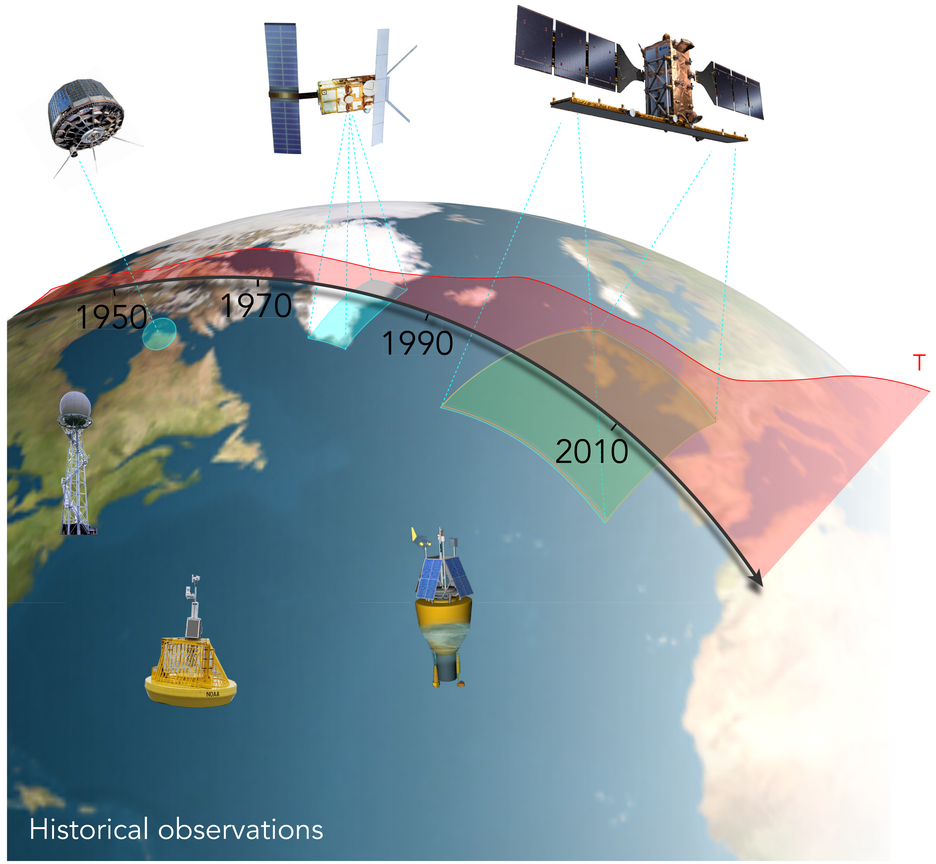
© Christian Lessig, 2023

AtmoRep

Large scale representation learning of atmospheric dynamics

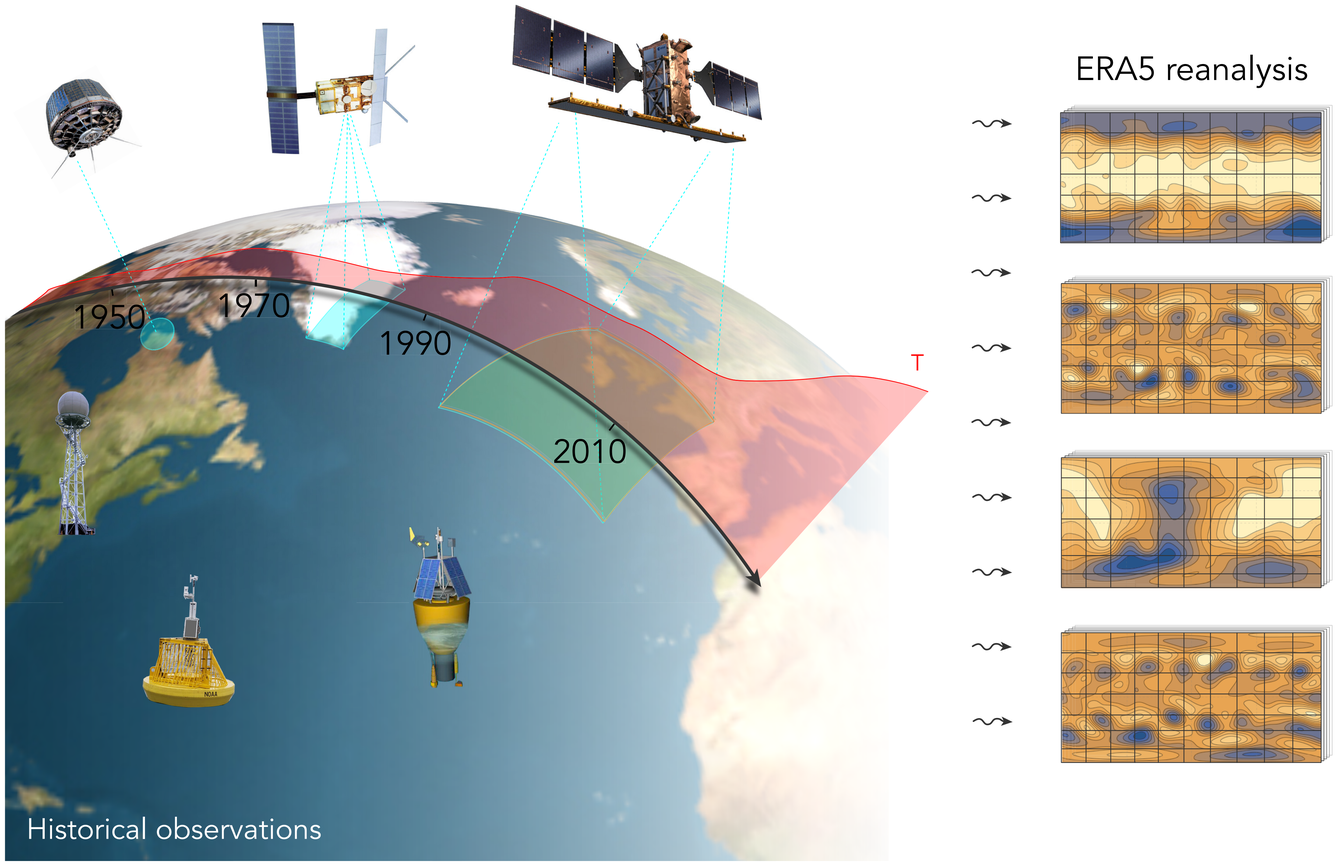
© Christian Lessig, 2023

AtmoRep



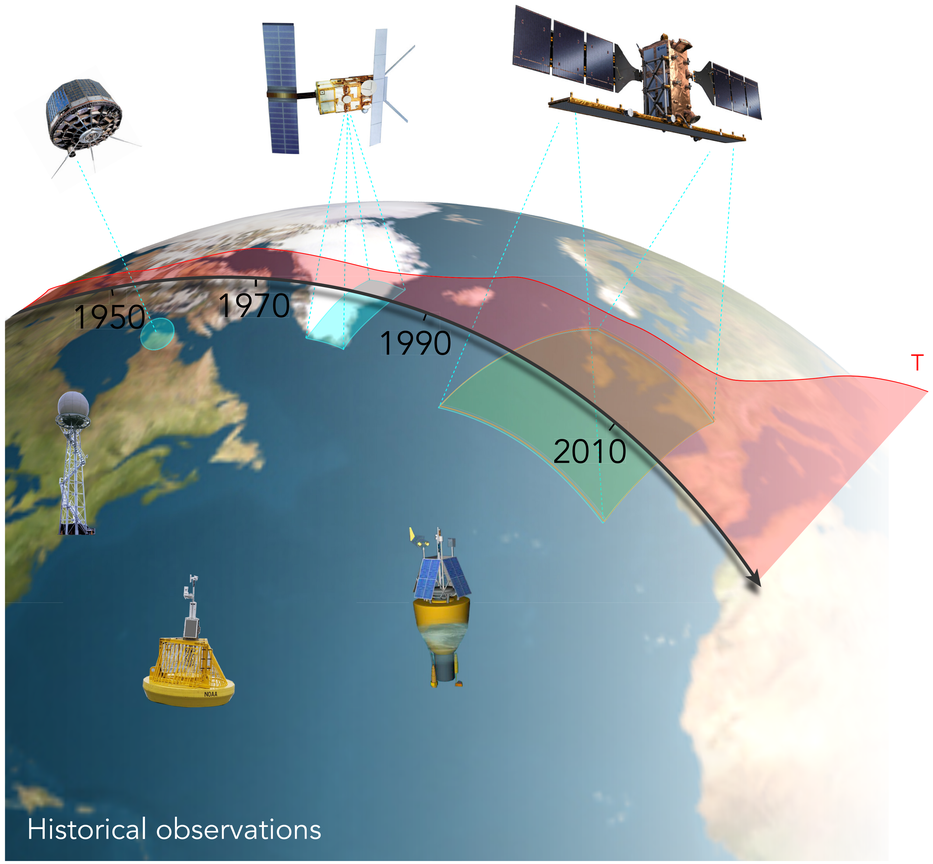
© Christian Lessig, 2023

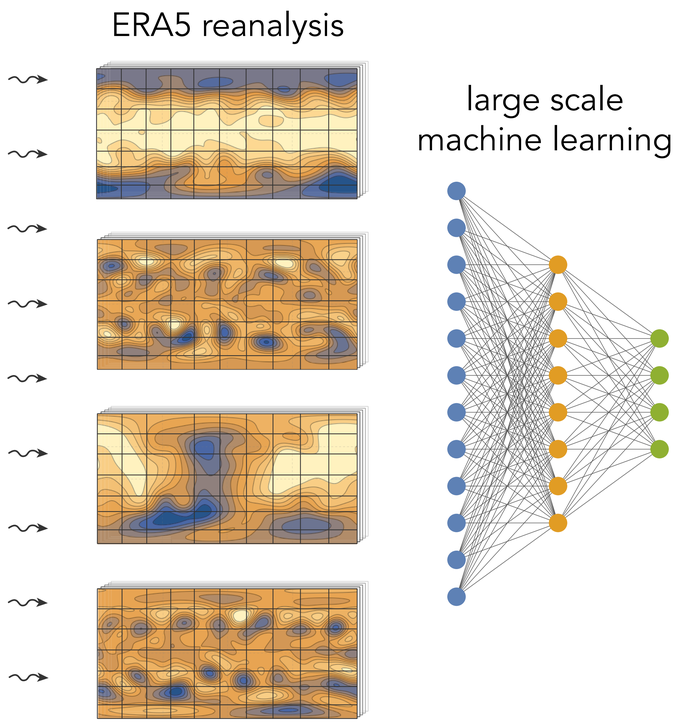
AtmoRep



© Christian Lessig, 2023

AtmoRep

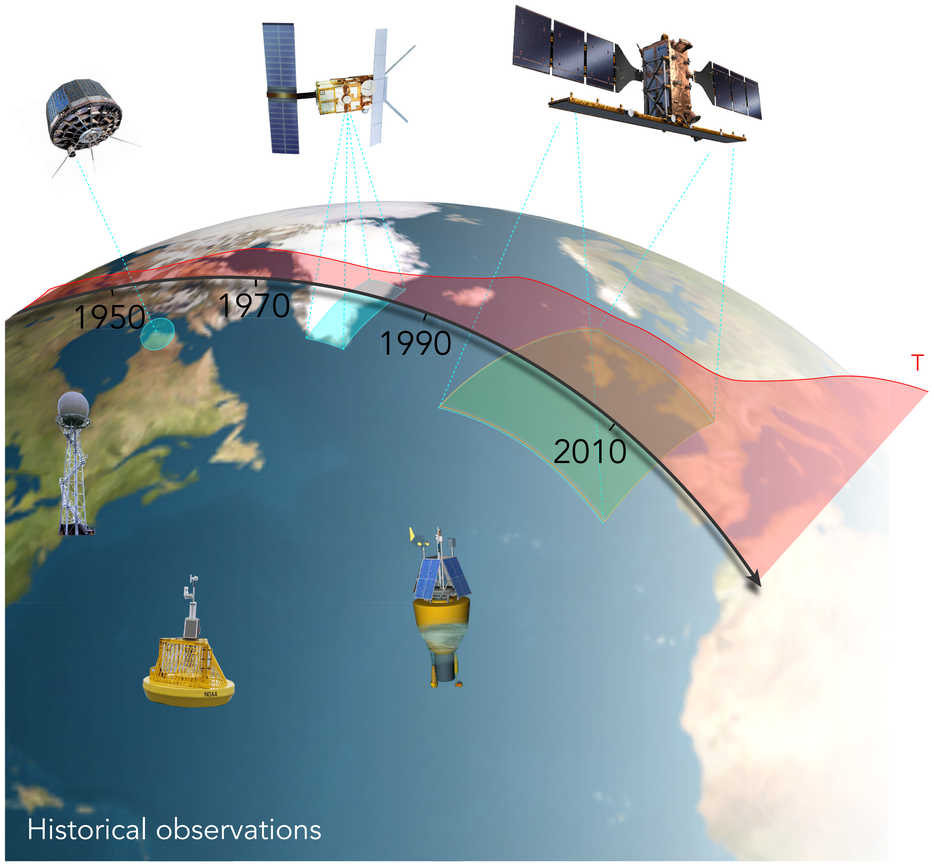


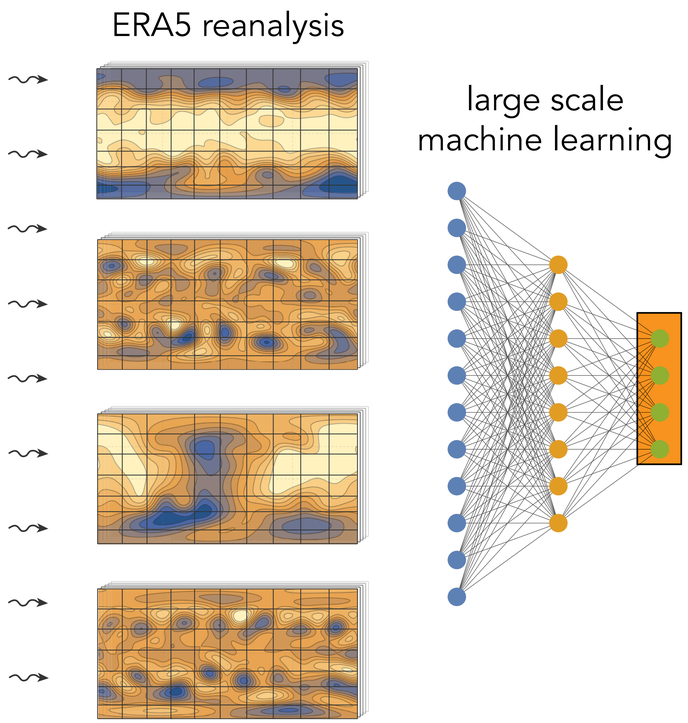


large scale
machine learning

© Christian Lessig, 2023

AtmoRep

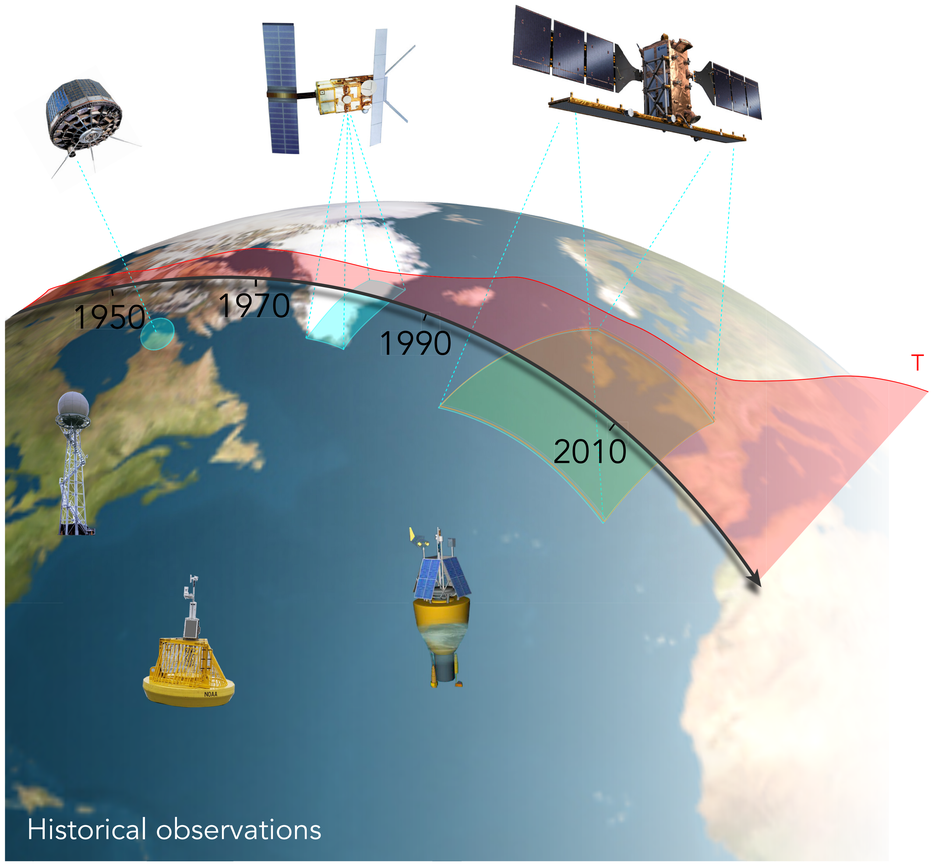


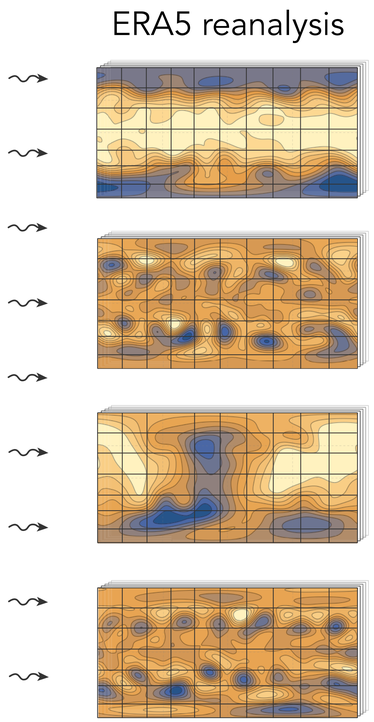


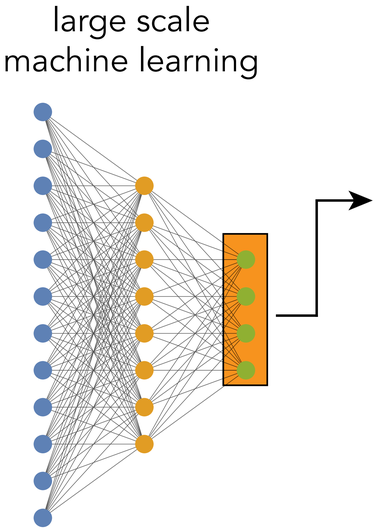
large scale
machine learning

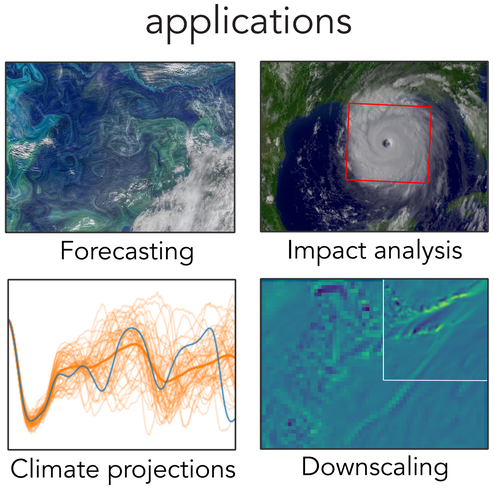
© Christian Lessig, 2023

AtmoRep



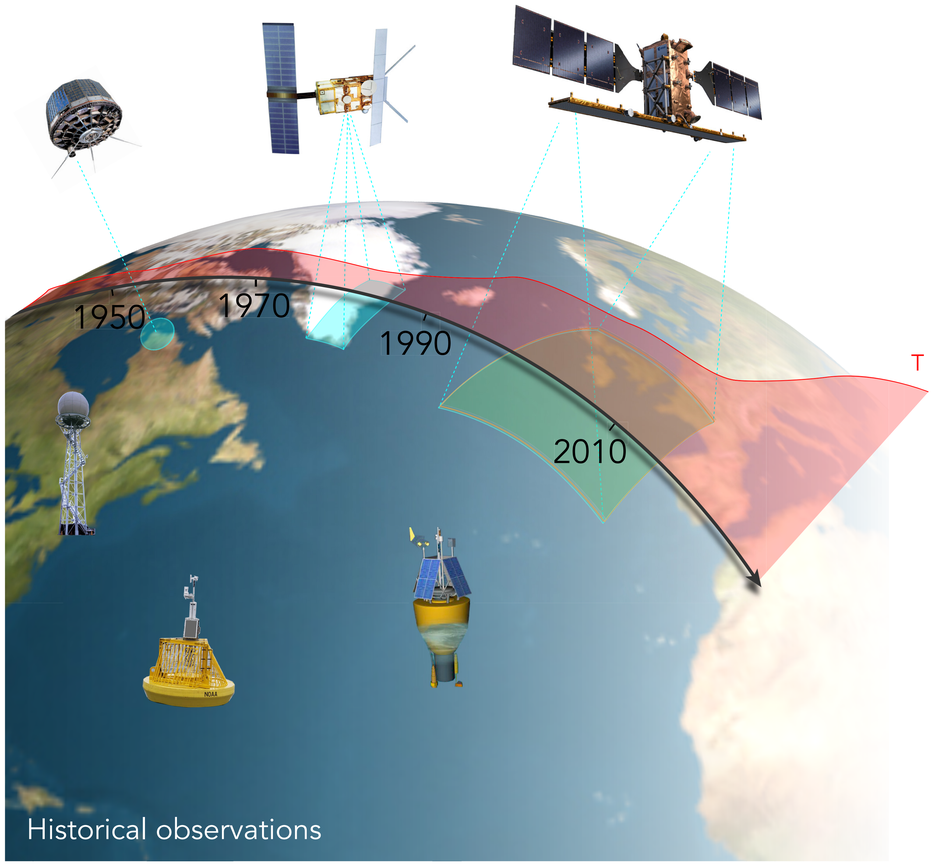


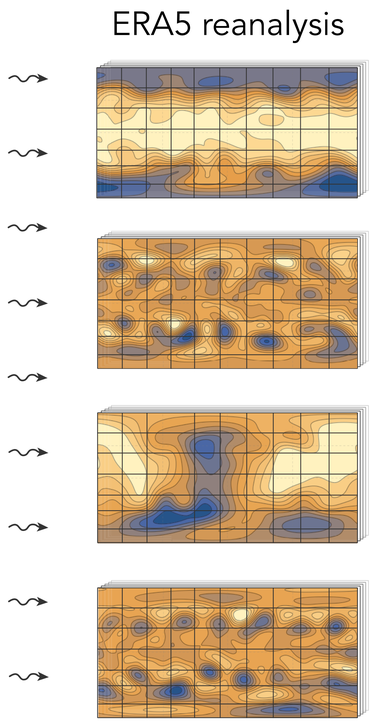


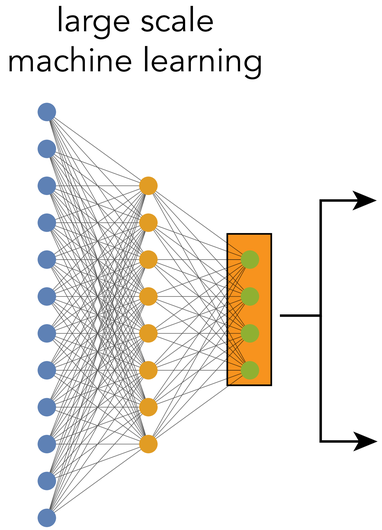


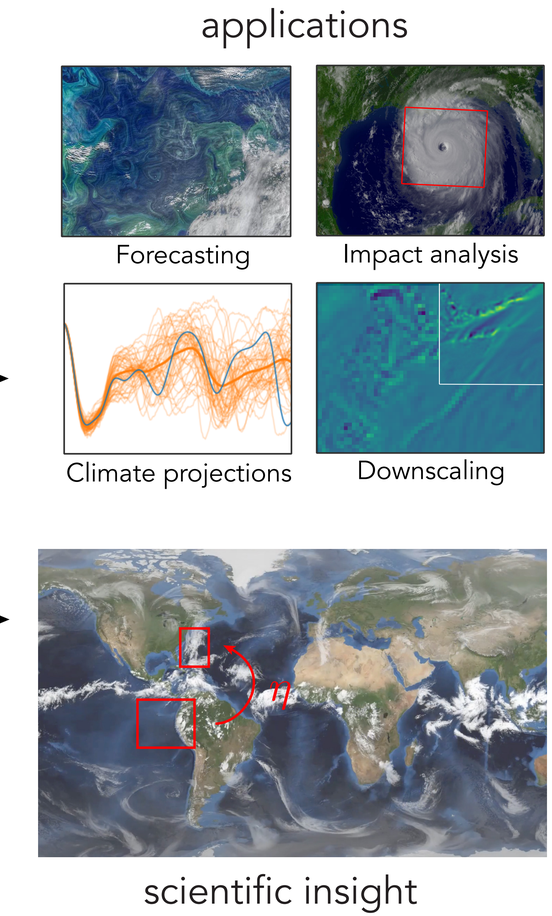
© Christian Lessig, 2023

AtmoRep



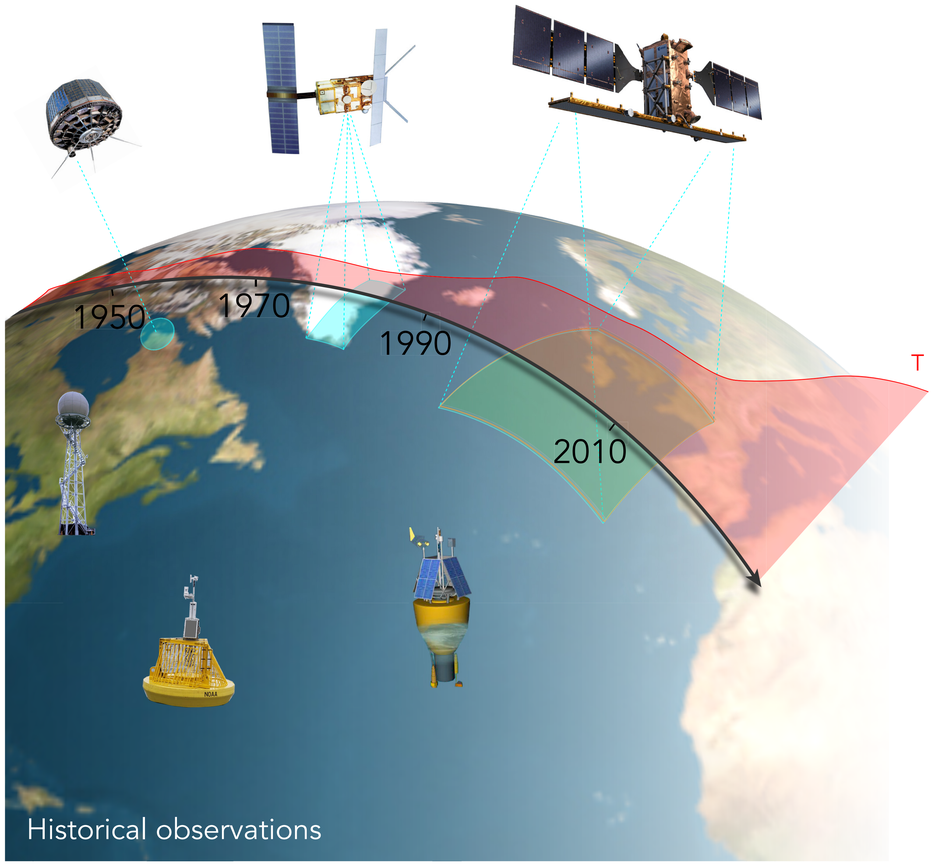


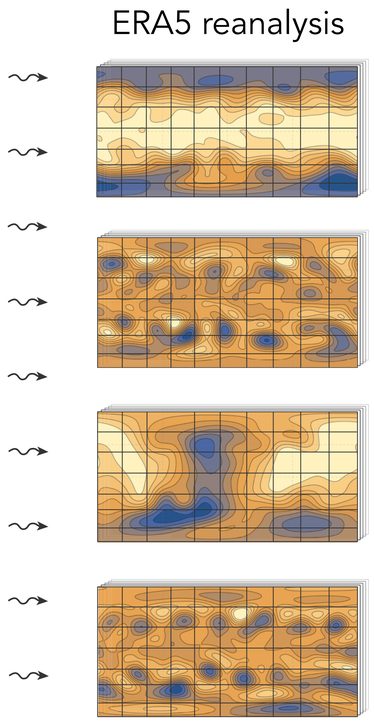


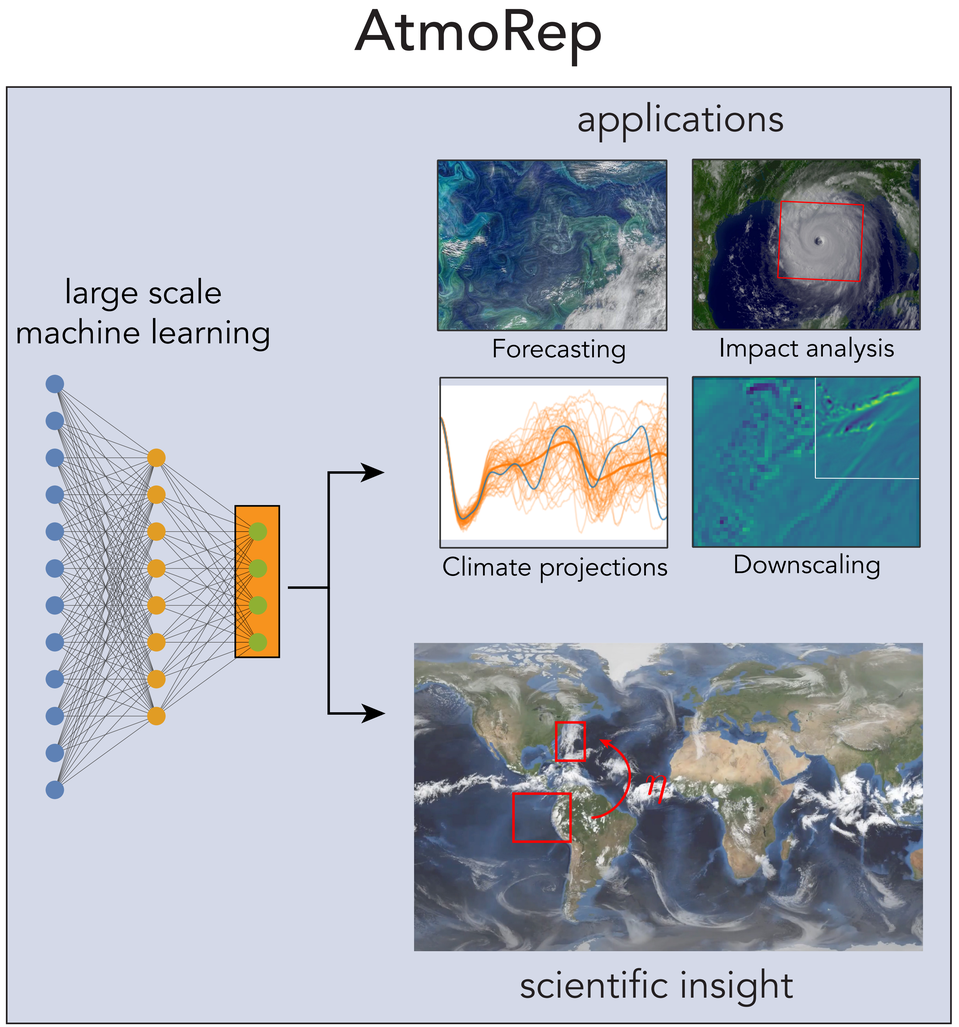


© Christian Lessig, 2023

AtmoRep







© Christian Lessig, 2023

AtmoRep: a theoretical formulation

© Christian Lessig, 2023

AtmoRep: a theoretical formulation

- Atmosphere as abstract stochastical dynamical system:

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

© Christian Lessig, 2023

AtmoRep: 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

AtmoRep: a theoretical formulation

- Atmosphere as abstract stochastical dynamical system:

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

output state

initial condition

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

© Christian Lessig, 2023

AtmoRep: 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

AtmoRep: a theoretical formulation

- Atmosphere as abstract stochastical dynamical system:

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

- Neural network model:

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

© Christian Lessig, 2023

AtmoRep: 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

AtmoRep: 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., Dall-E, diffusion models)

© Christian Lessig, 2023

AtmoRep: 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., Dall-E, diffusion models)

- forecasting
- downscaling
- model correction
- ...

© Christian Lessig, 2023

AtmoRep: a theoretical formulation

- Neural network model as factorization of $p_\theta\left(y|x,\alpha\right)$

© Christian Lessig, 2023

AtmoRep: a theoretical formulation

- Neural network model as factorization of $p_\theta\left(y|x,\alpha\right)$
- Intrinsically statistical/probabilistic formulation
  - Fits naturally the statistical/chaotic nature of the atmosphere

© Christian Lessig, 2023

AtmoRep: a theoretical formulation

- Neural network model as factorization of $p_\theta\left(y|x,\alpha\right)$
- Intrinsically statistical/probabilistic formulation
  - Fits naturally the statistical/chaotic nature of the atmosphere
- Loss derivation
  - Expectation maximization, ELBO, ...
  - Probabilistic bound on skill of network

© 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 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 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" the output state y without learning

© Christian Lessig, 2023

AtmoRep: in-context learning

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

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

positional encoding
for output

© Christian Lessig, 2023

AtmoRep: in-context learning

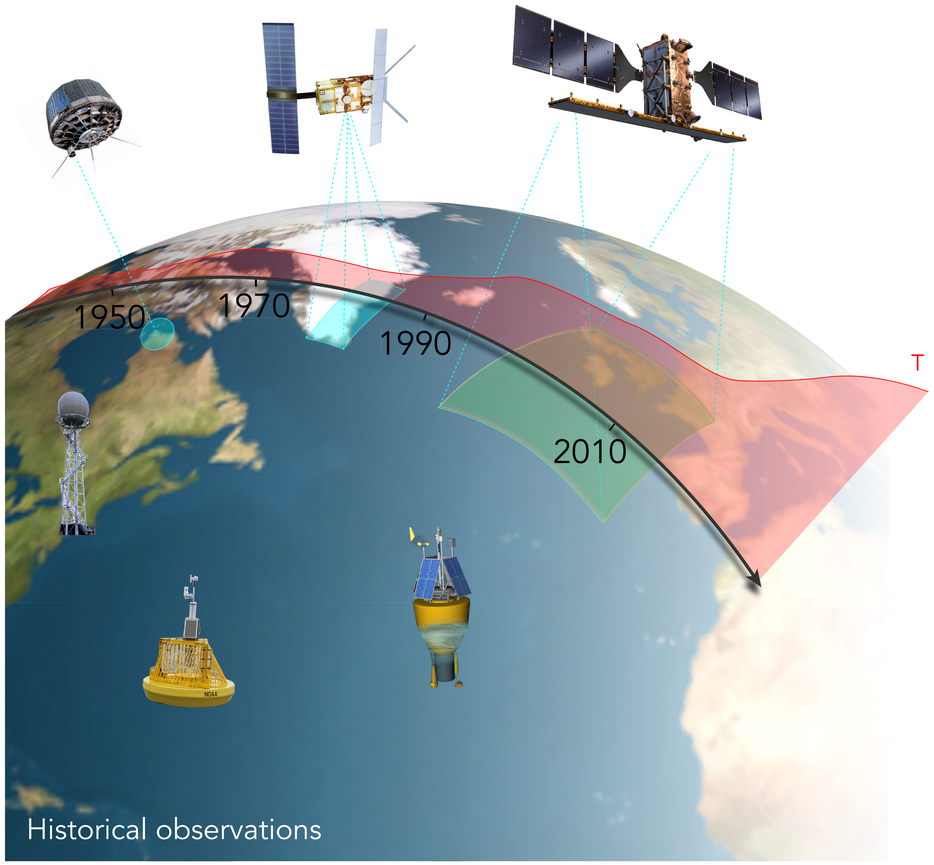
- The model $p_\theta\left(y|x,\alpha\right)$ implies that what we want to "control" the output state y without learning

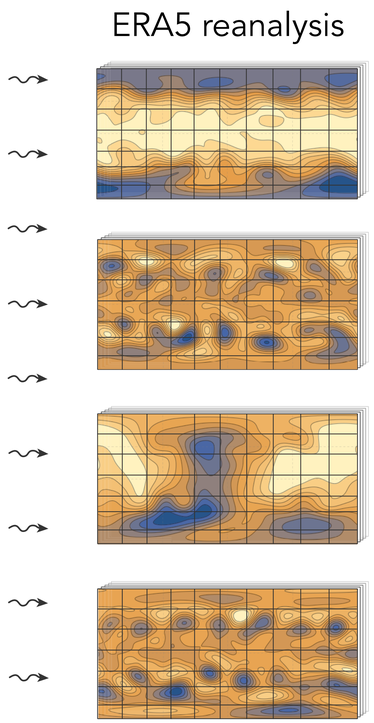
$$p_\theta\left(y|x,\alpha,\tau\right)$$

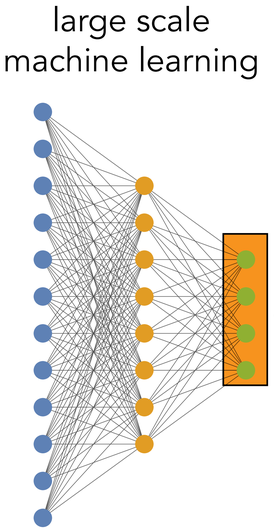
  - Spatial and temporal location, resolution, quality
  - Few shot: "explain" τ to the network

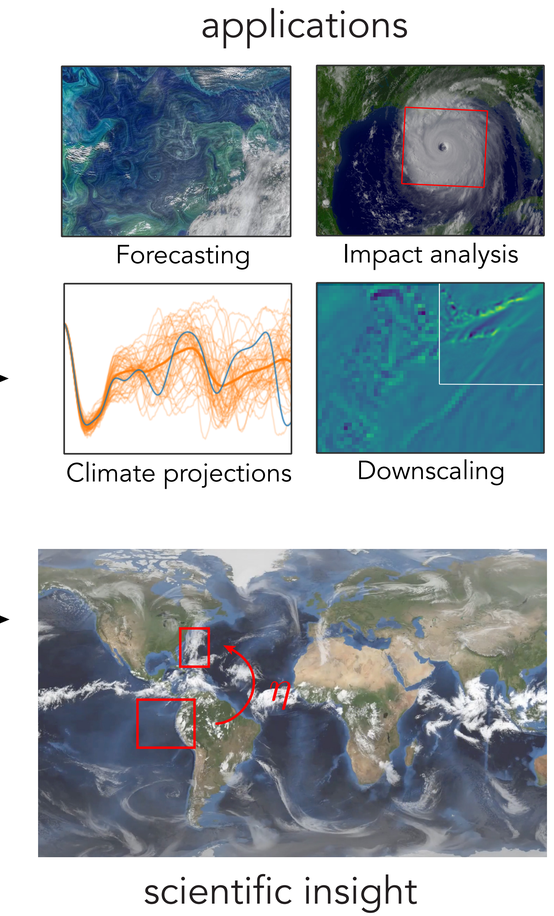
© Christian Lessig, 2023

AtmoRep



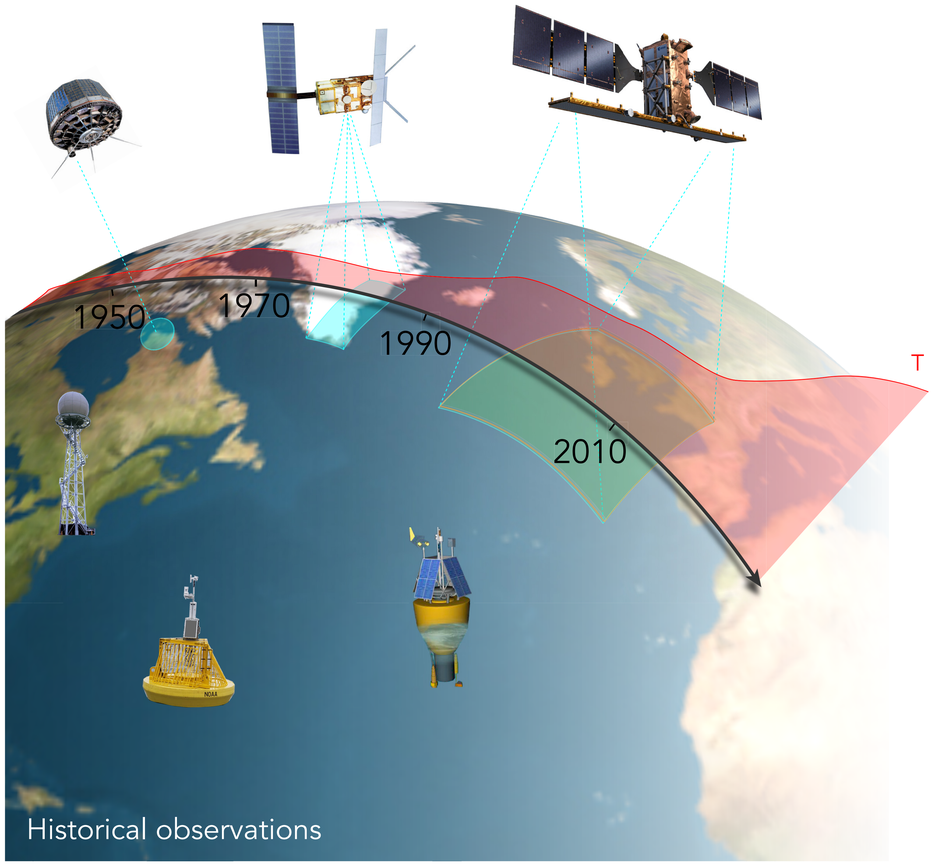




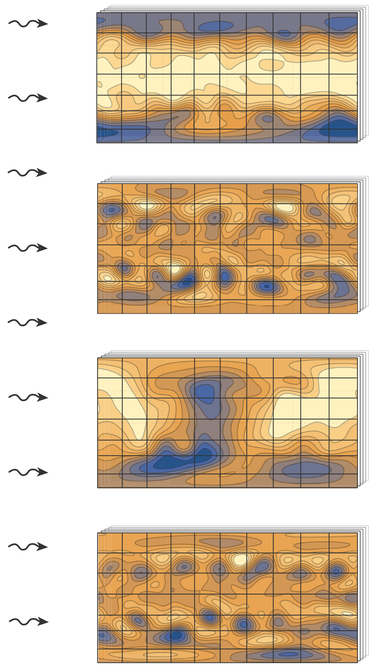


© Christian Lessig, 2023

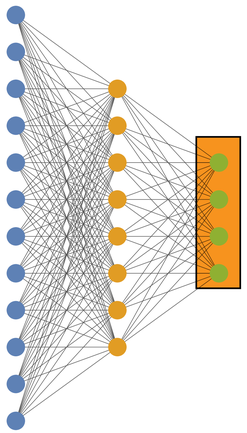
AtmoRep



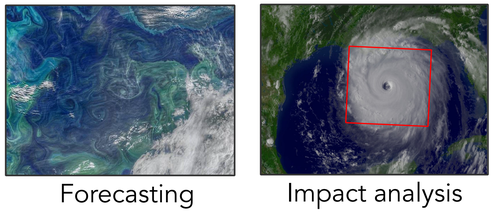
ERA5 reanalysis

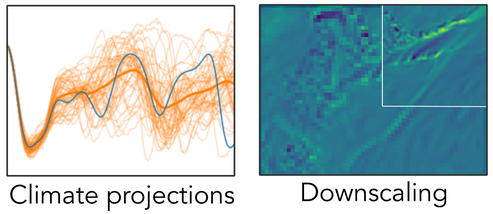


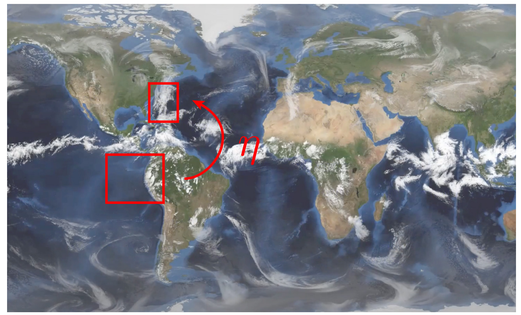
$p_\theta\left(y|x,\alpha\right)$



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







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

© Christian Lessig, 2023

AtmoRep network architecture

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

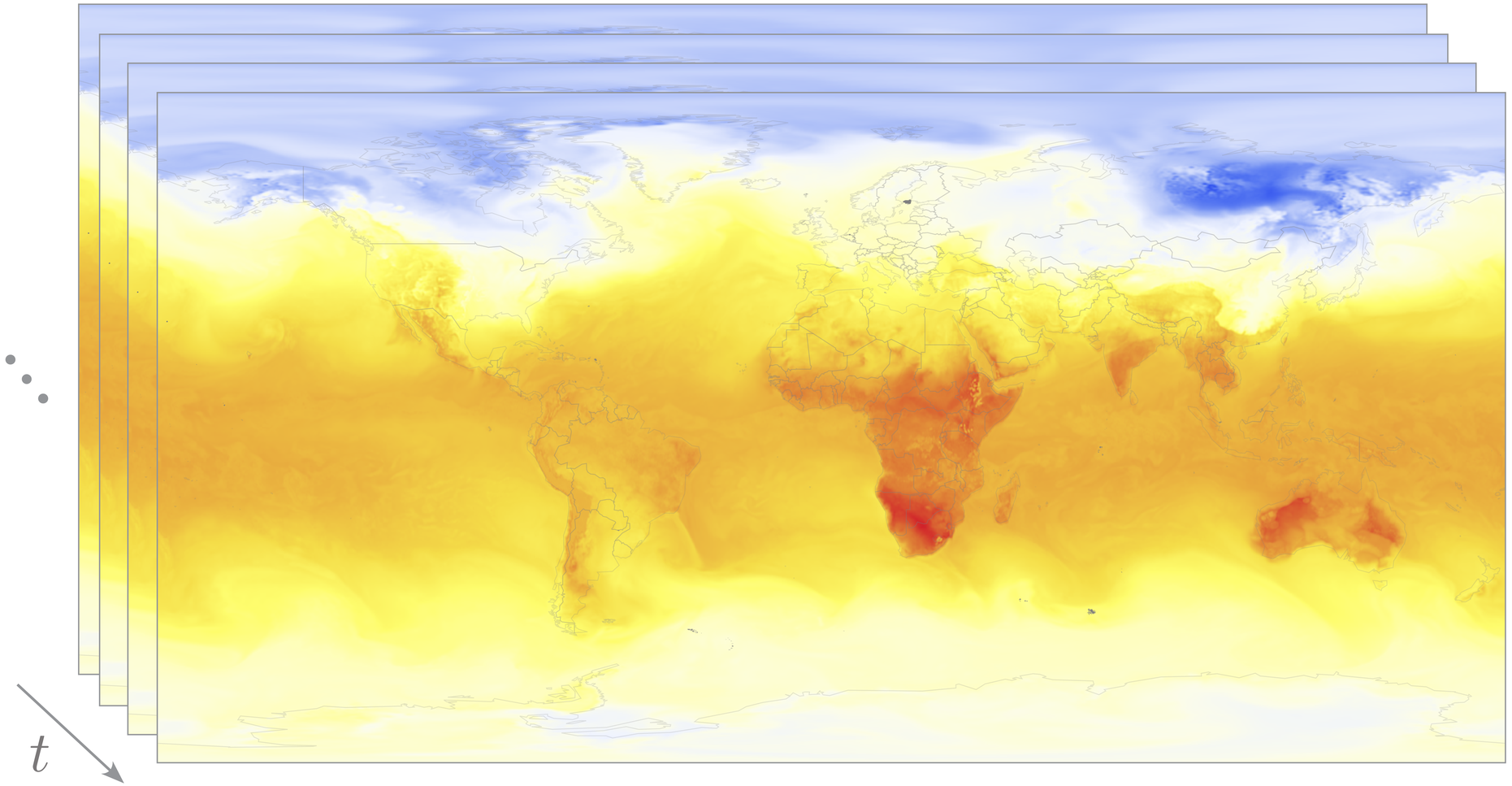
© Christian Lessig, 2023

AtmoRep network architecture

- Network is local in space-time

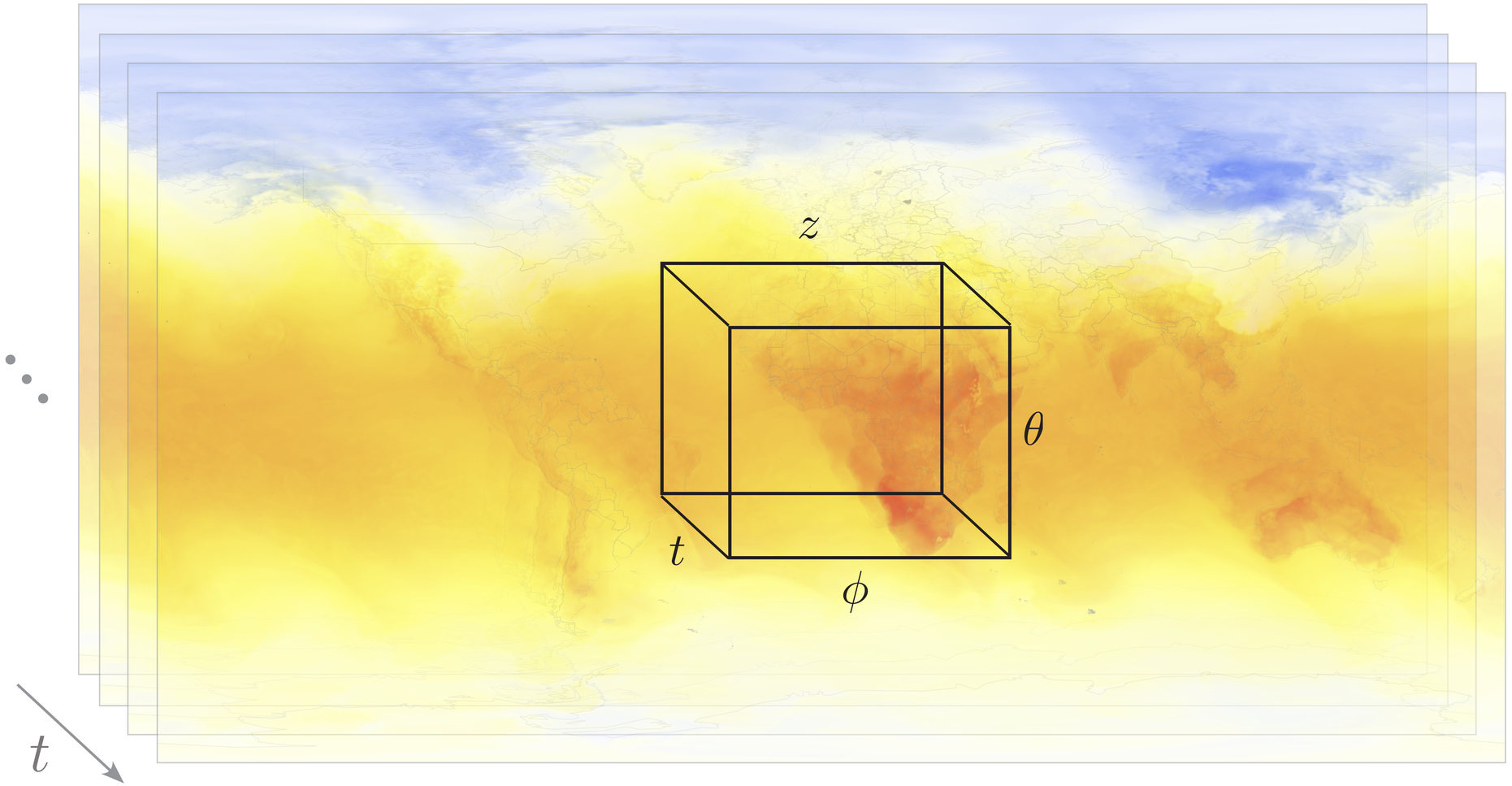
© Christian Lessig, 2023

AtmoRep network architecture



© Christian Lessig, 2023

AtmoRep network architecture



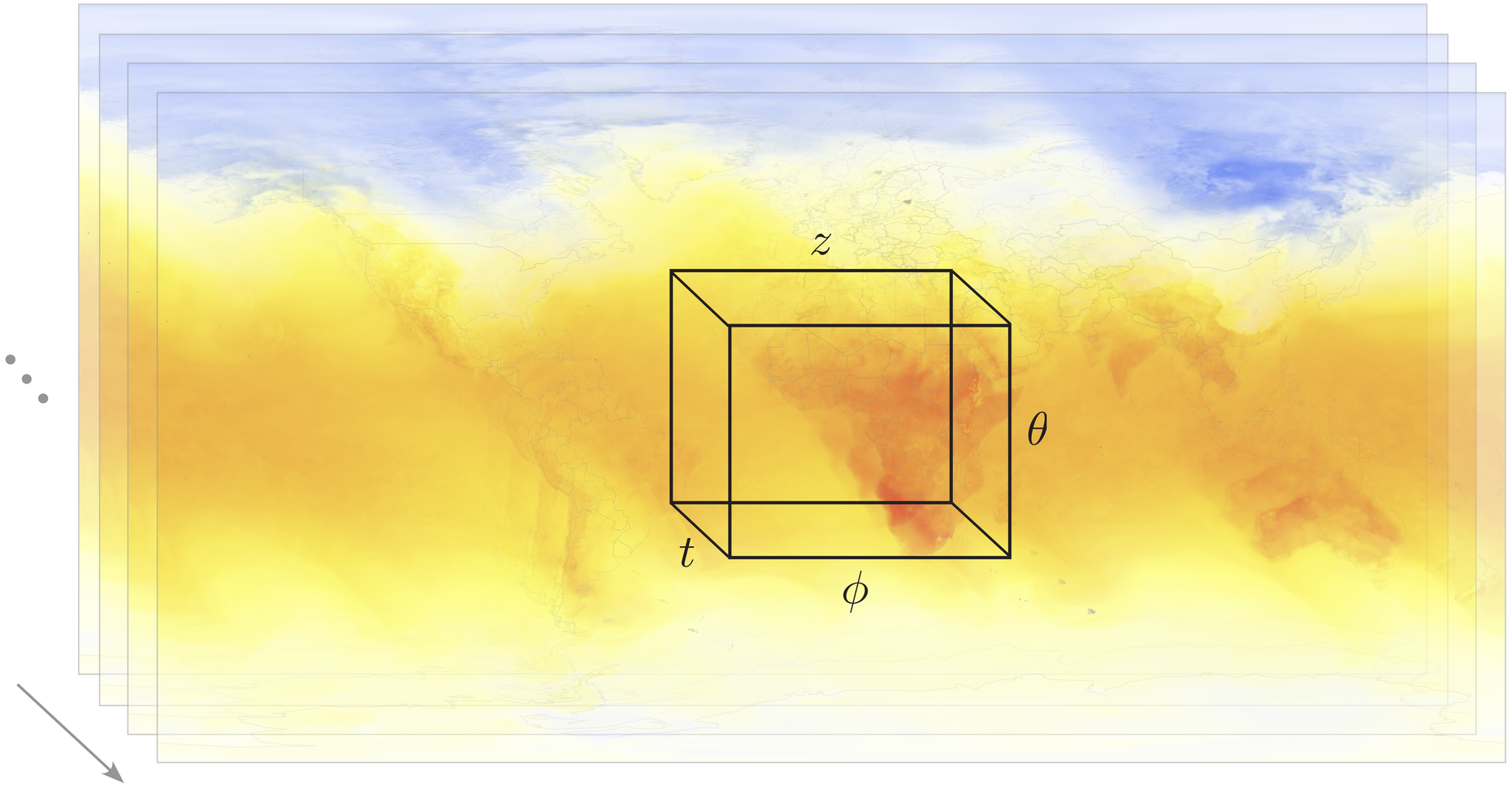
© Christian Lessig, 2023

AtmoRep network architecture

- Network is local in space-time
  - Physics of dynamics are universally valid
  - Local particularities can be learned by providing time + space position as auxiliary information

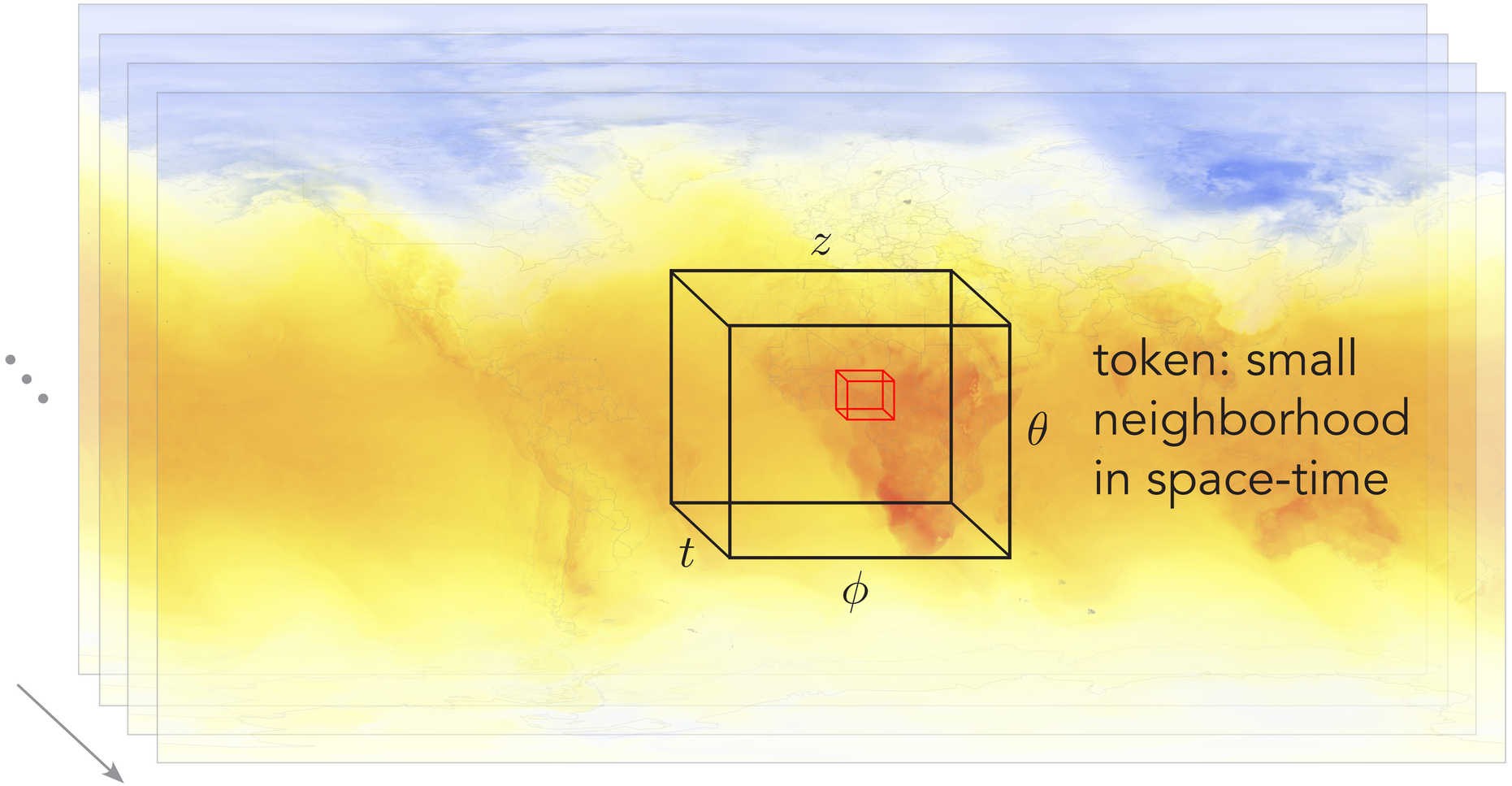
© Christian Lessig, 2023

What is a token?



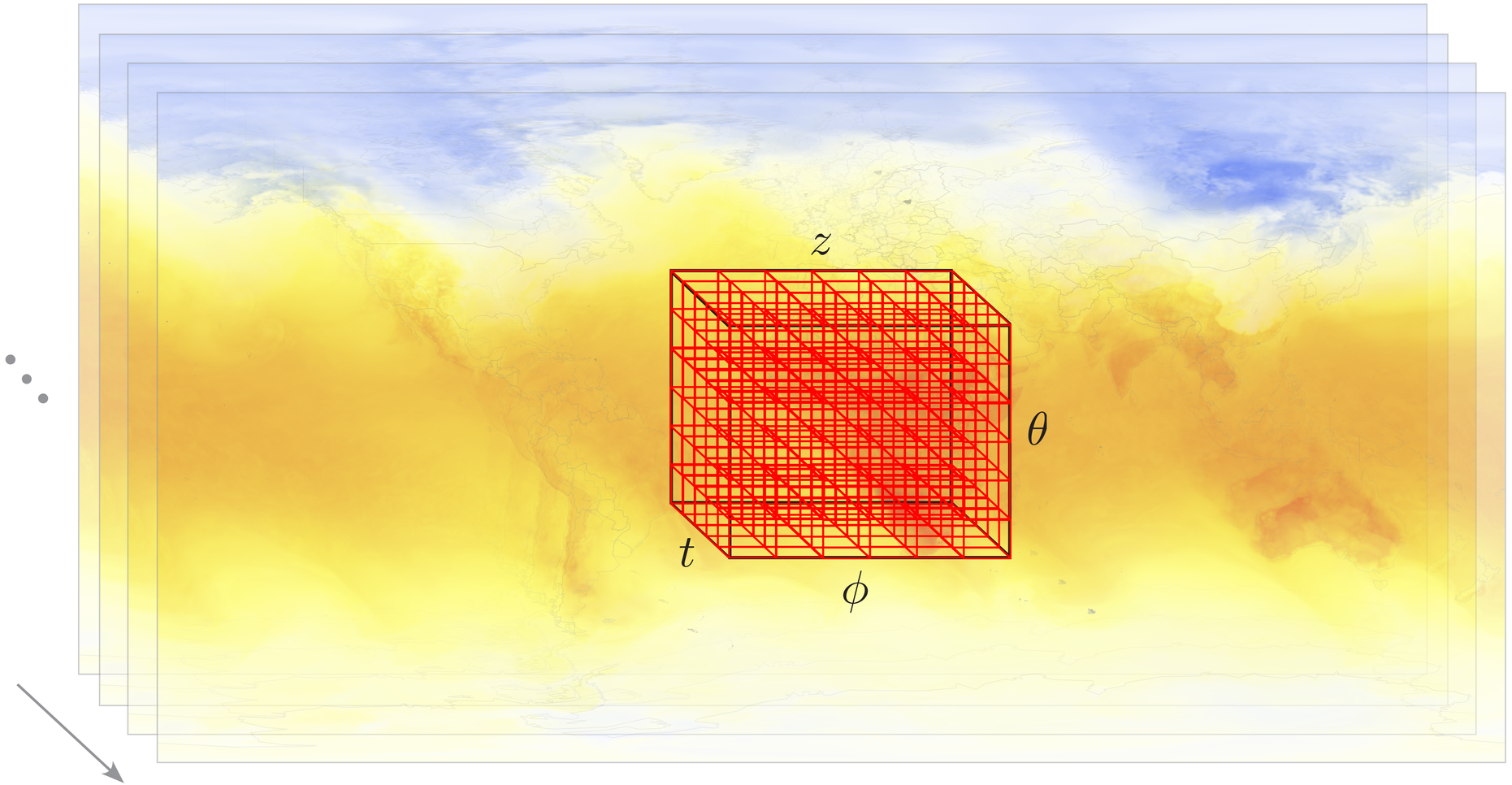
© Christian Lessig, 2023

What is a token?



© Christian Lessig, 2023

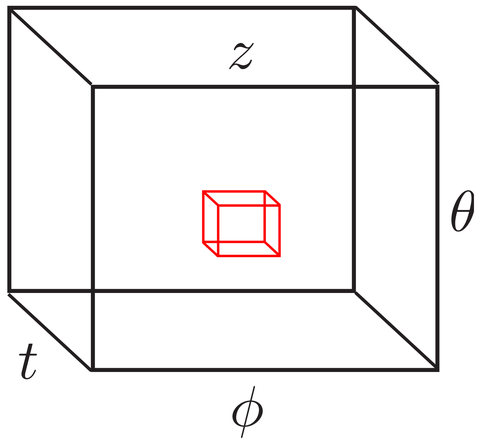
What is a token?



© Christian Lessig, 2023

What is a token?

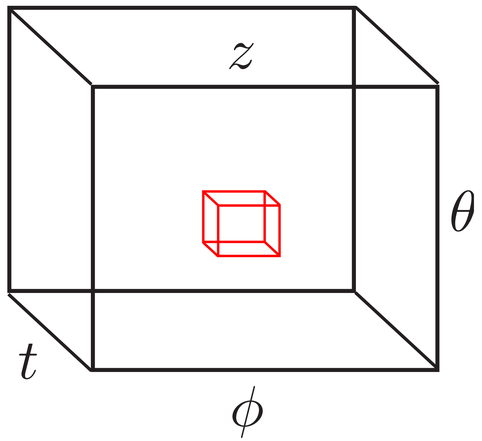
- Token is small neighborhood in space-time
  - Small for token attention / inter-action to be informative
  - Big enough so token has rich internal strurcture



© Christian Lessig, 2023

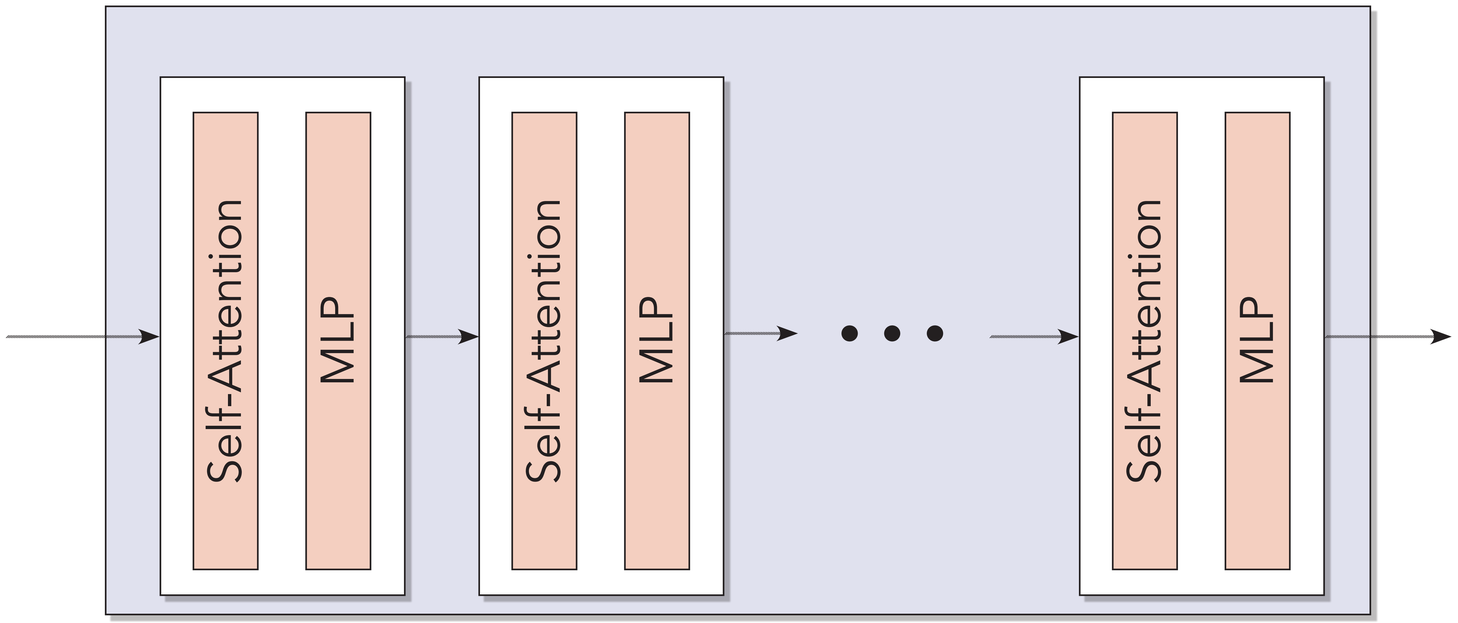
What is a token?

- 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

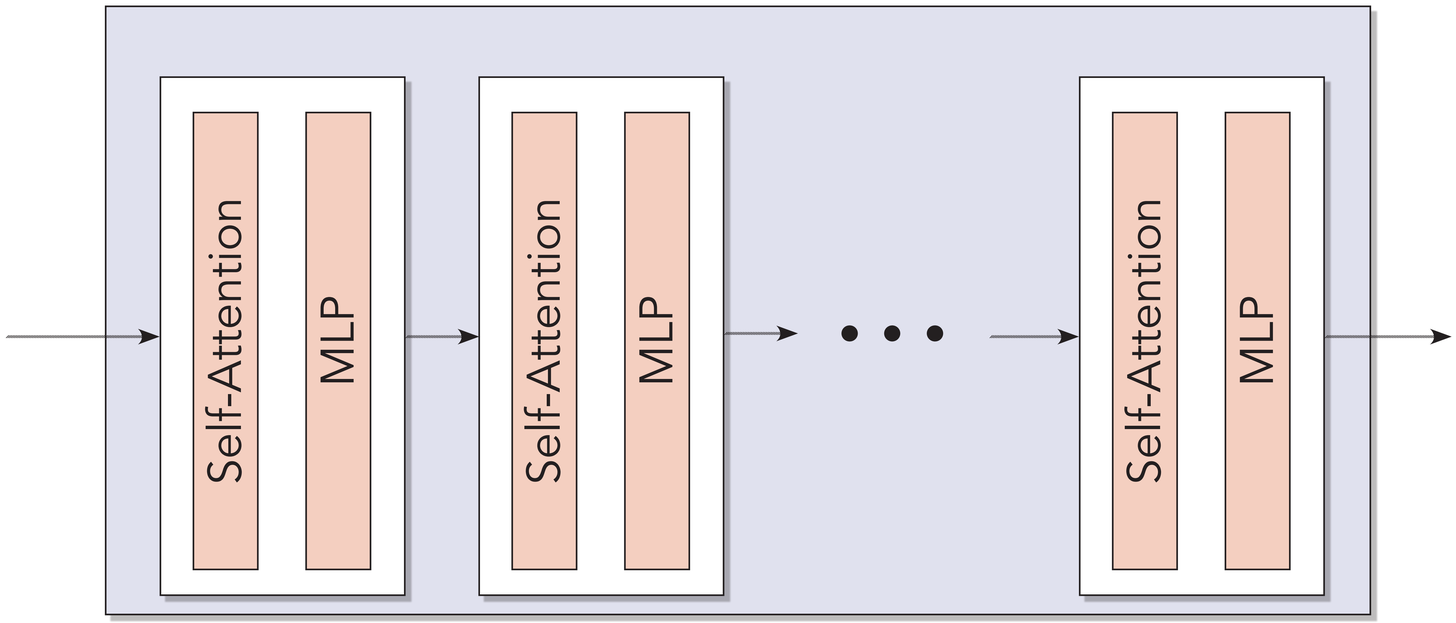


© Christian Lessig, 2023

Multiformer

Self
attention

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

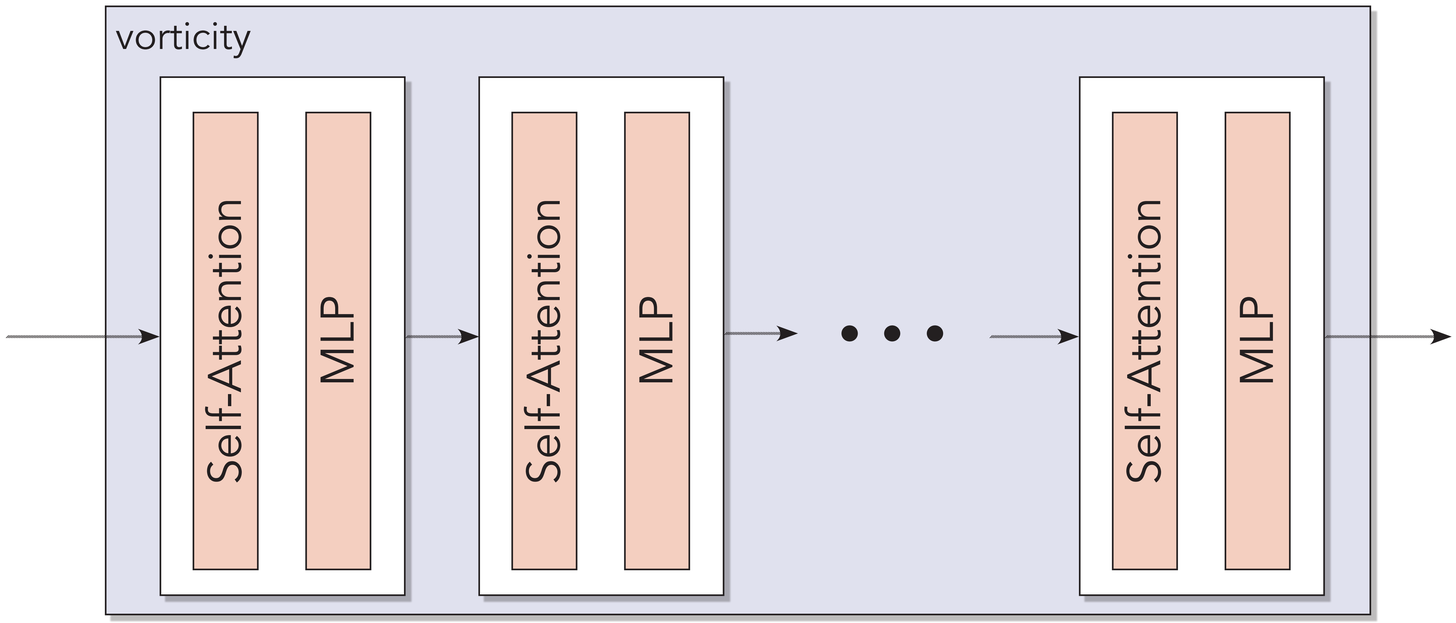


© Christian Lessig, 2023

Multiformer

Self
attention

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

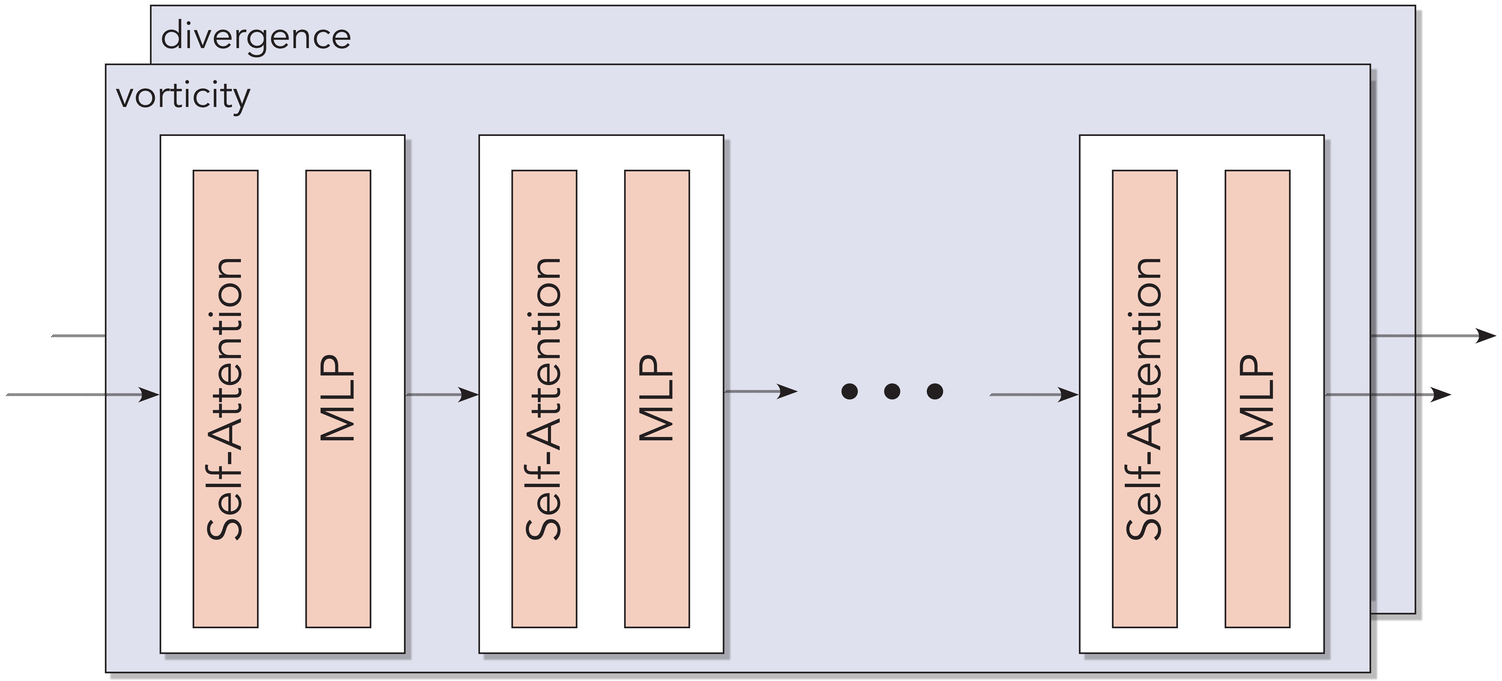


© Christian Lessig, 2023

Multiformer

Self
attention

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

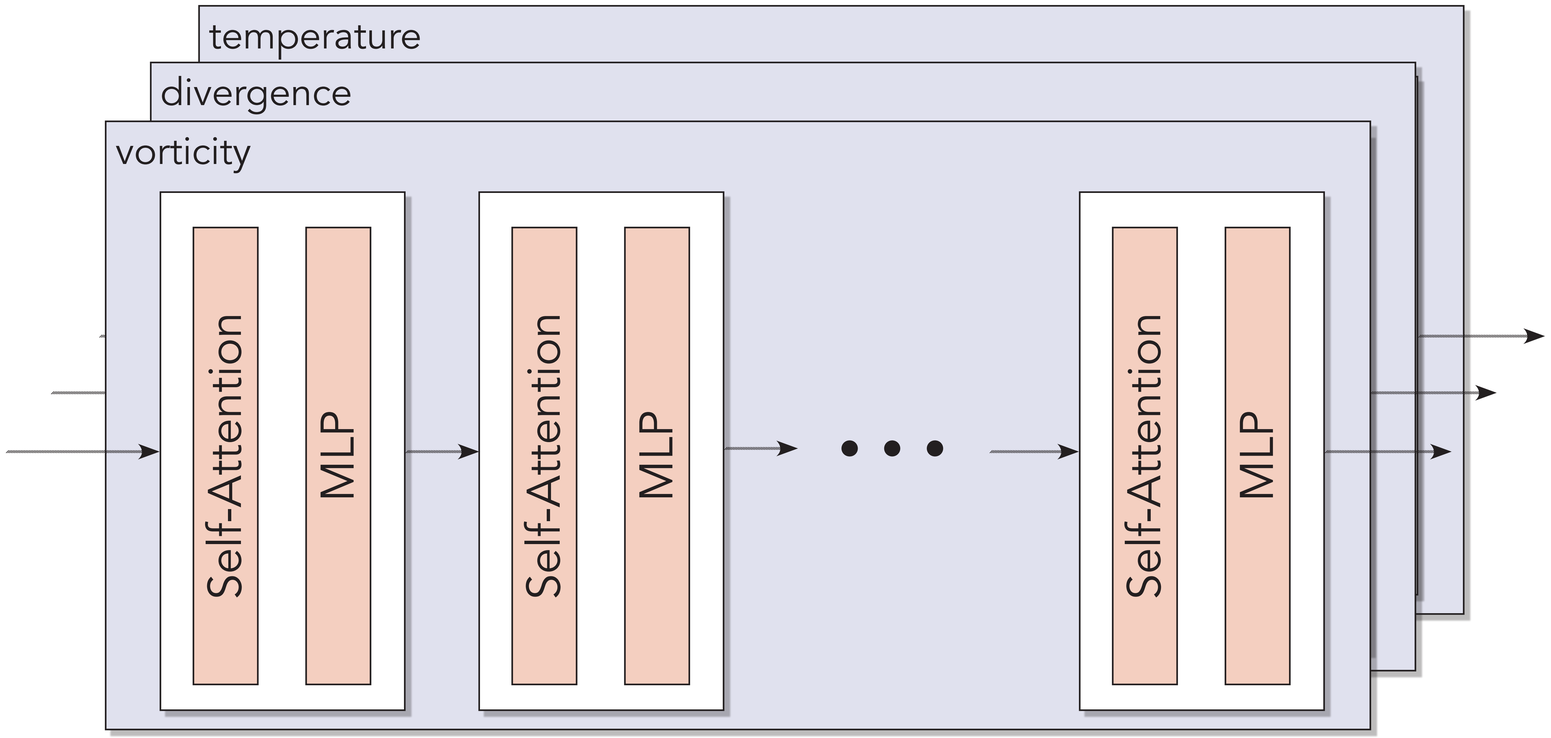


© Christian Lessig, 2023

Multiformer

Self
attention

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

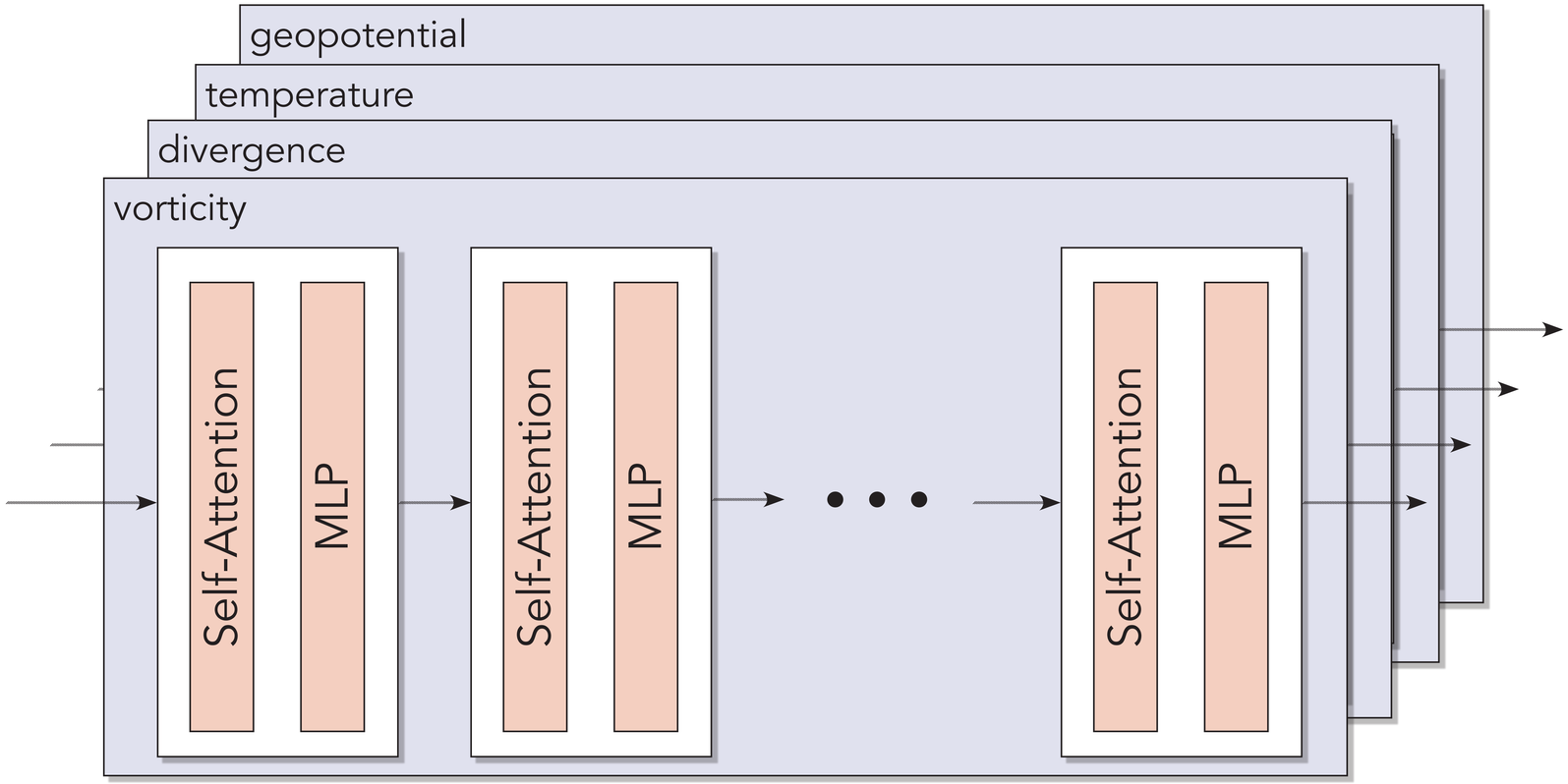


© Christian Lessig, 2023

Multiformer

Self
attention

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

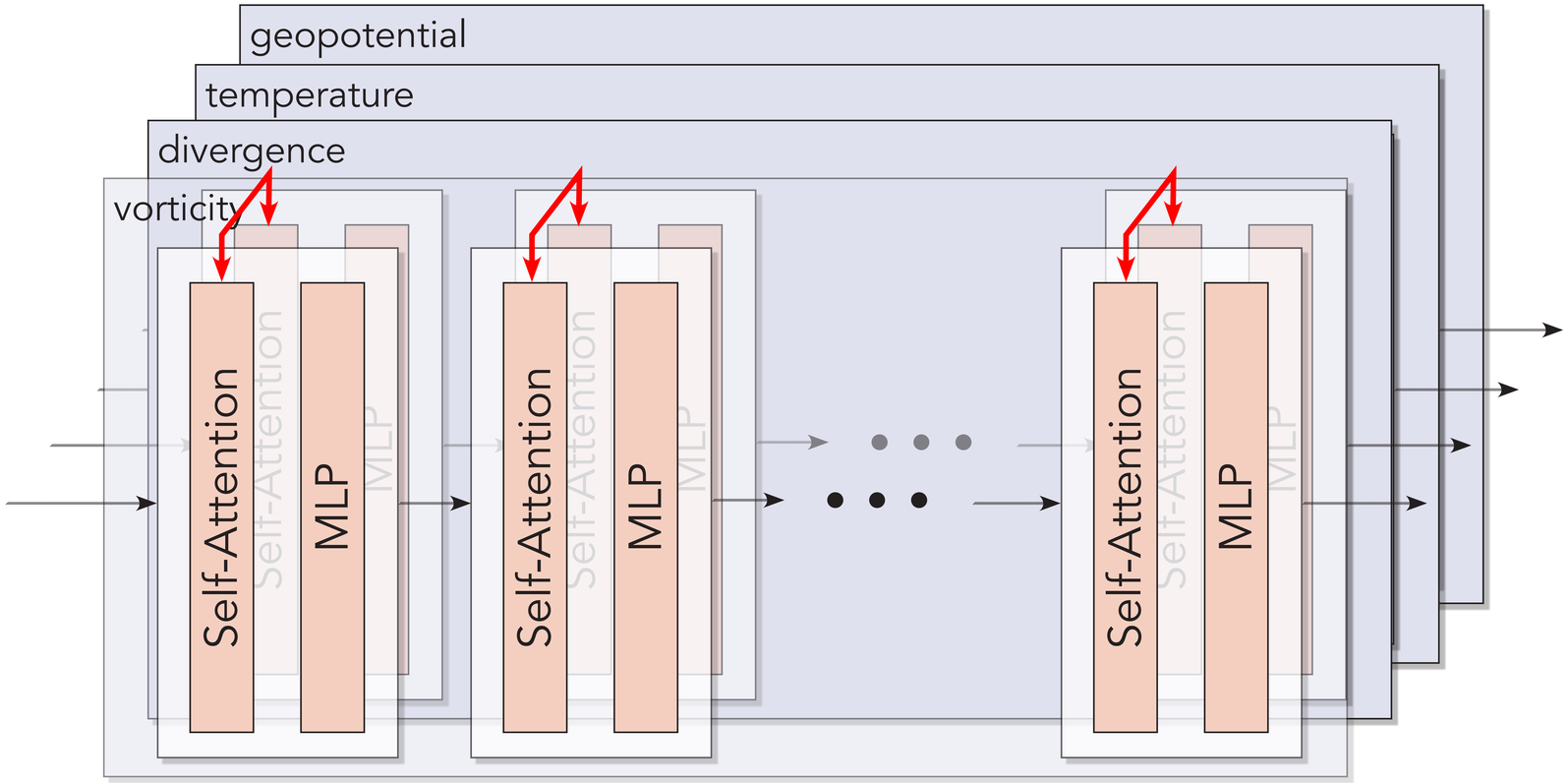


© Christian Lessig, 2023

Multiformer

Self
attention

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



© Christian Lessig, 2023

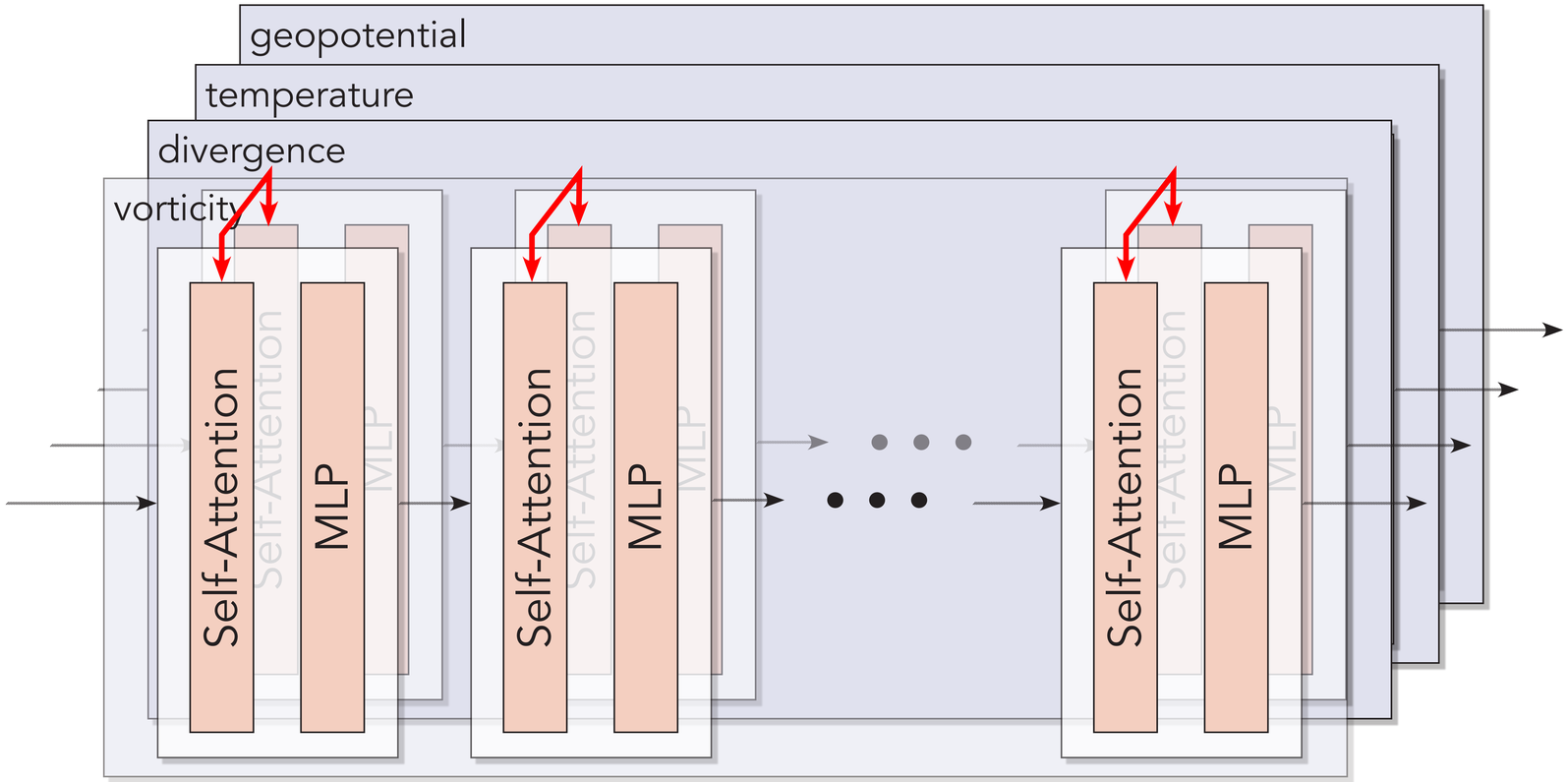
Multiformer

Self attention

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

Cross attention

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



© Christian Lessig, 2023

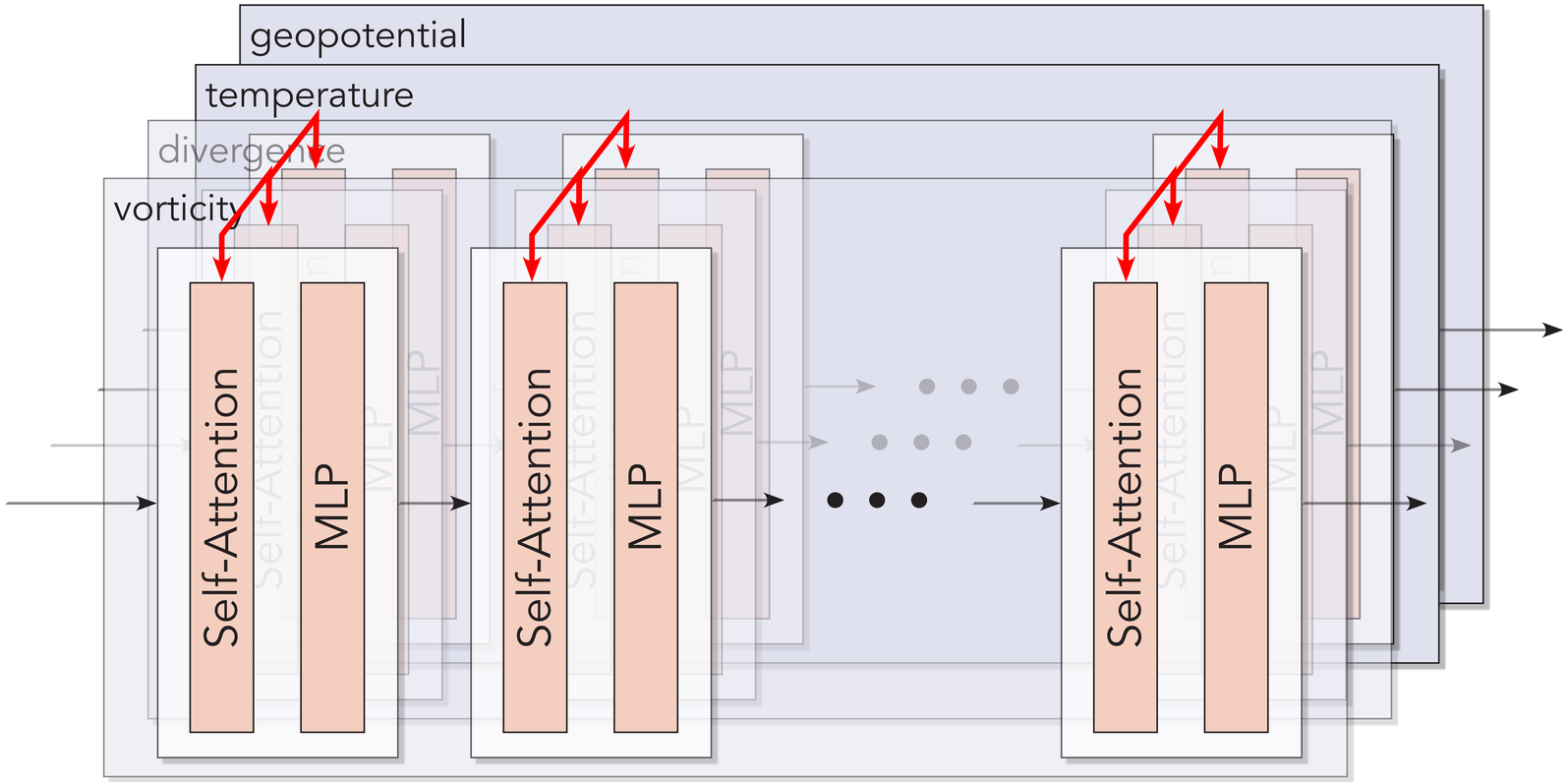
Multiformer

Self attention

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

Cross attention

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



© Christian Lessig, 2023

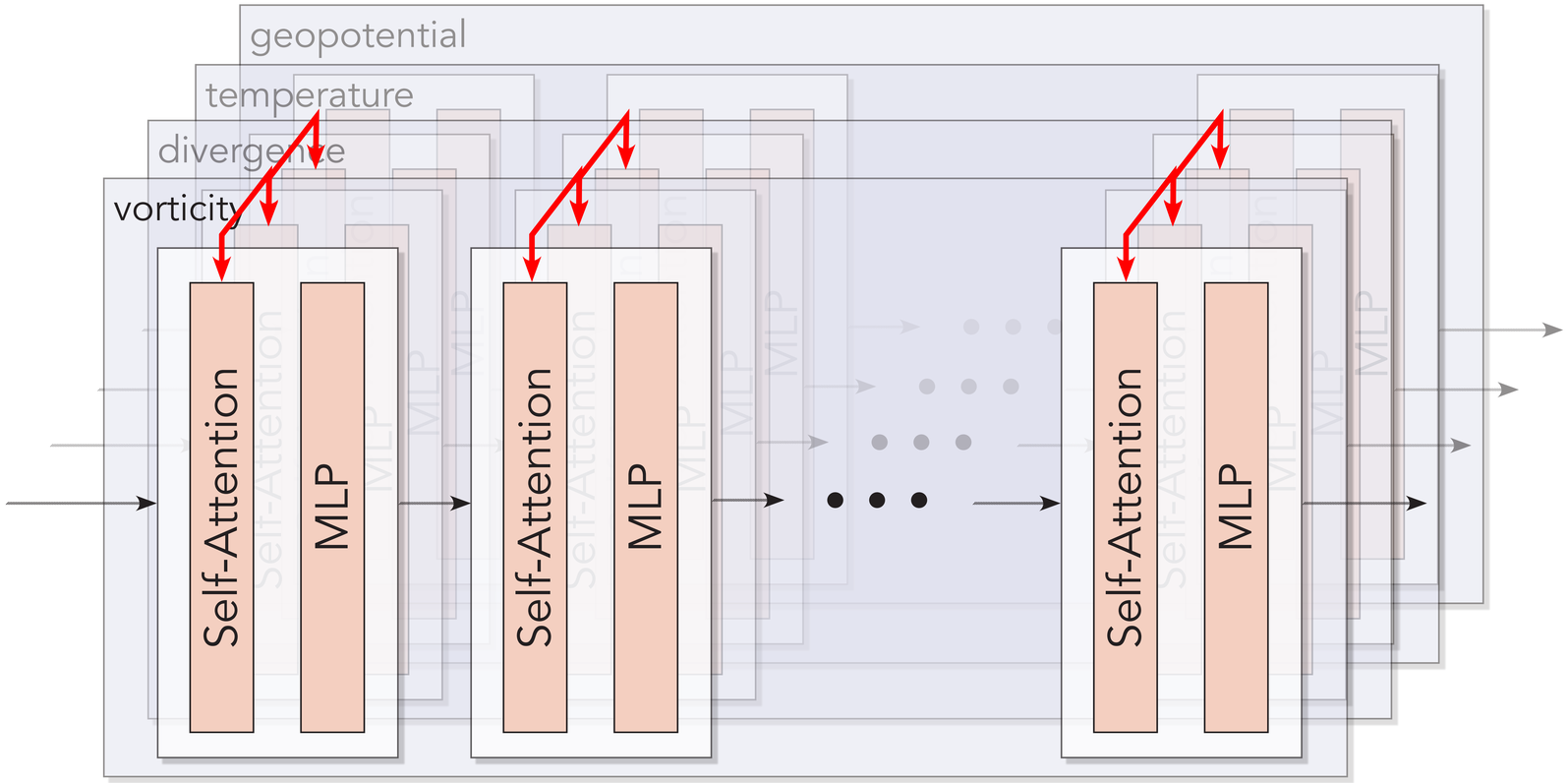
Multiformer

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

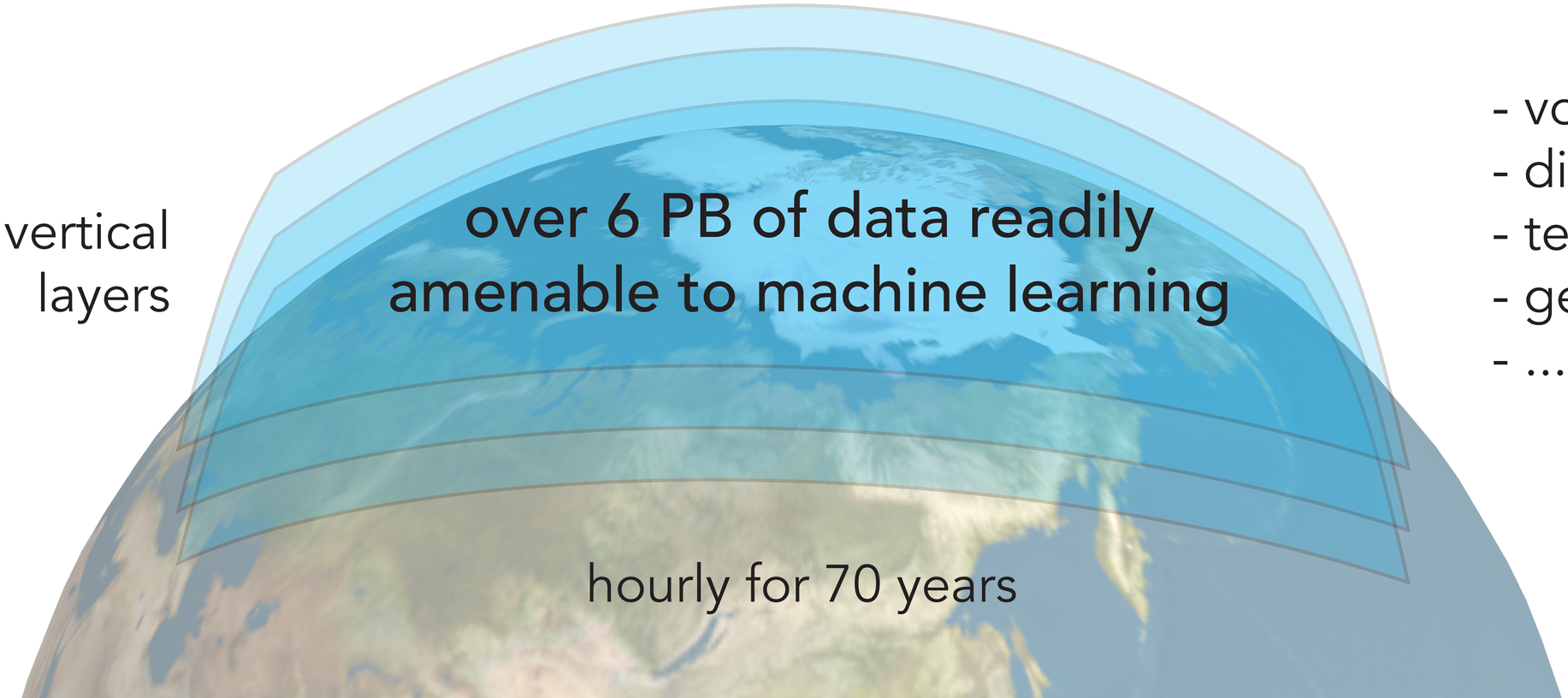
- Different physical fields with different properties have separate latent spaces (and transformations for these)
- Individual fields can be pre-trained independently
- Plug-and-play of fields
  - Fields can be added/removed with limited (or no) computational effort
- Cross-attention allows for explicit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

© 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

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

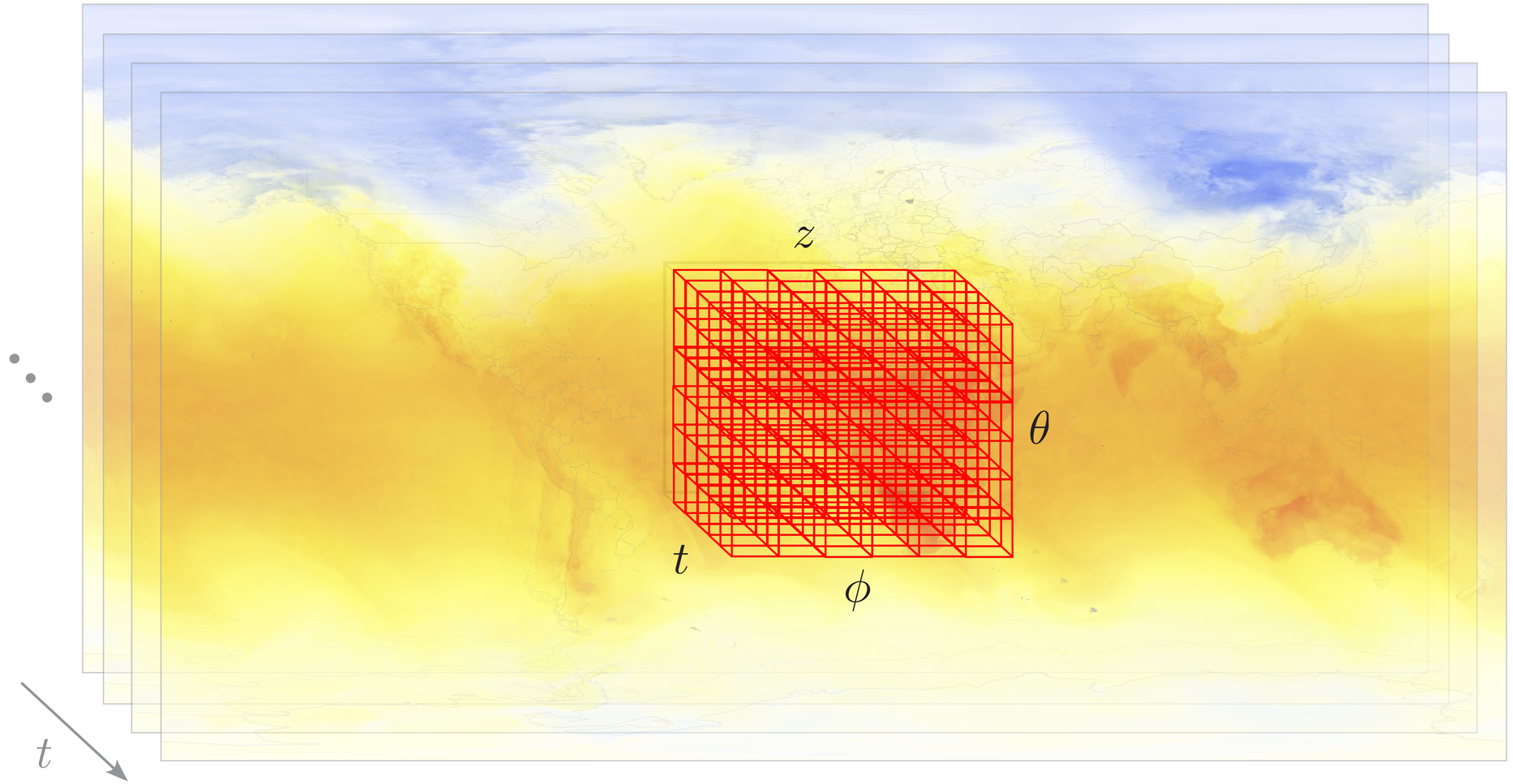
© Christian Lessig, 2023

Spatio-temporal BERT

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

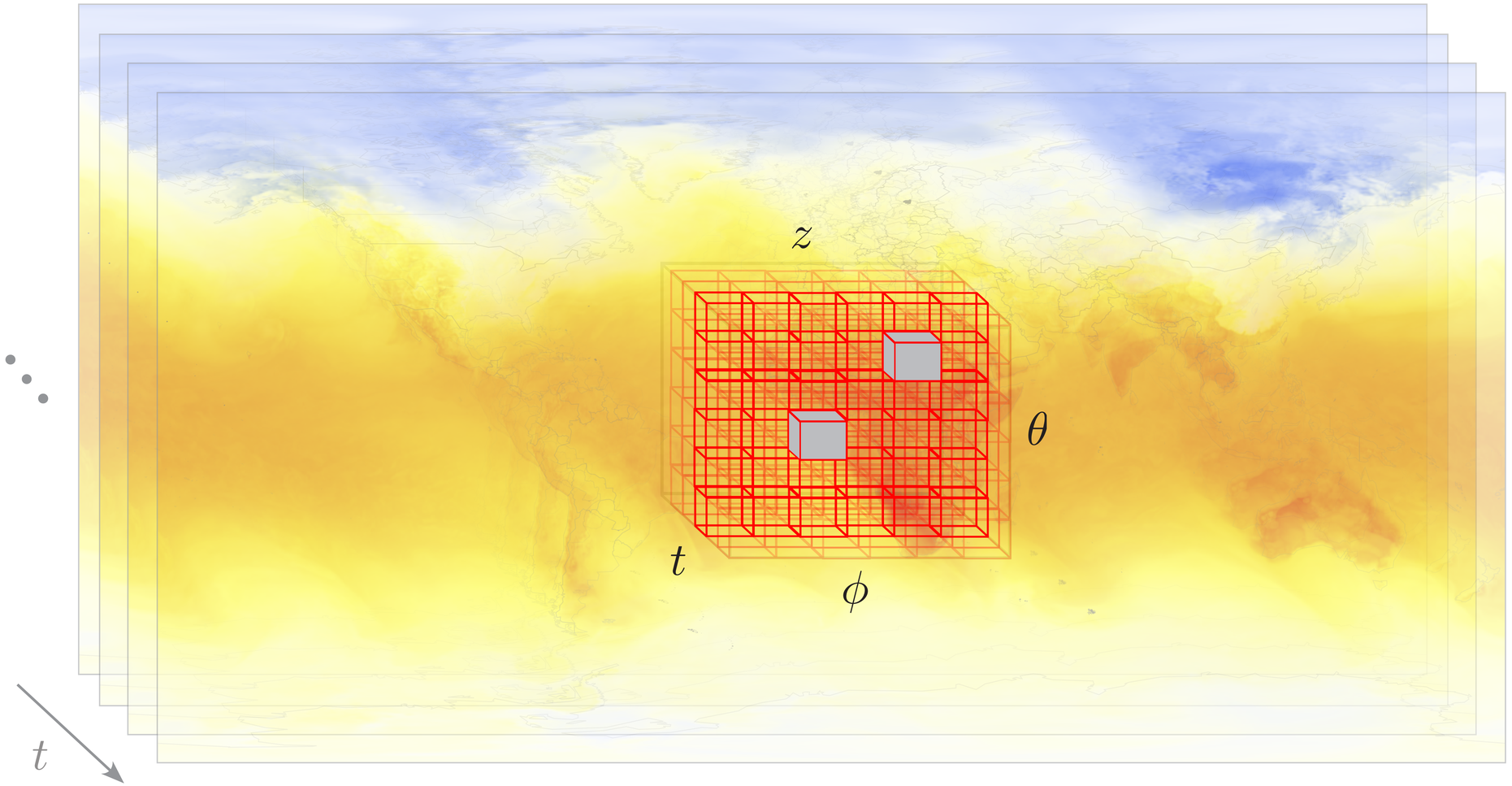
© Christian Lessig, 2023

Spatio-temporal BERT



© Christian Lessig, 2023

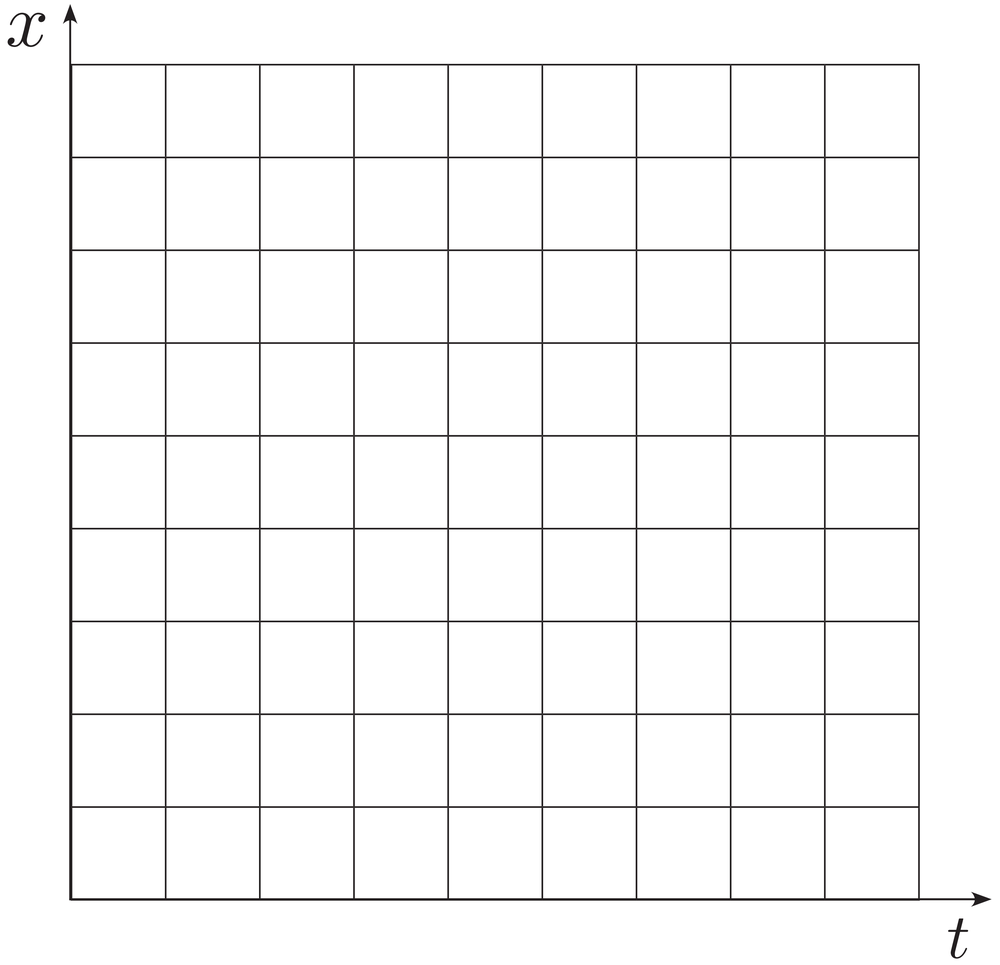
Spatio-temporal BERT



© Christian Lessig, 2023

Spatio-temporal BERT

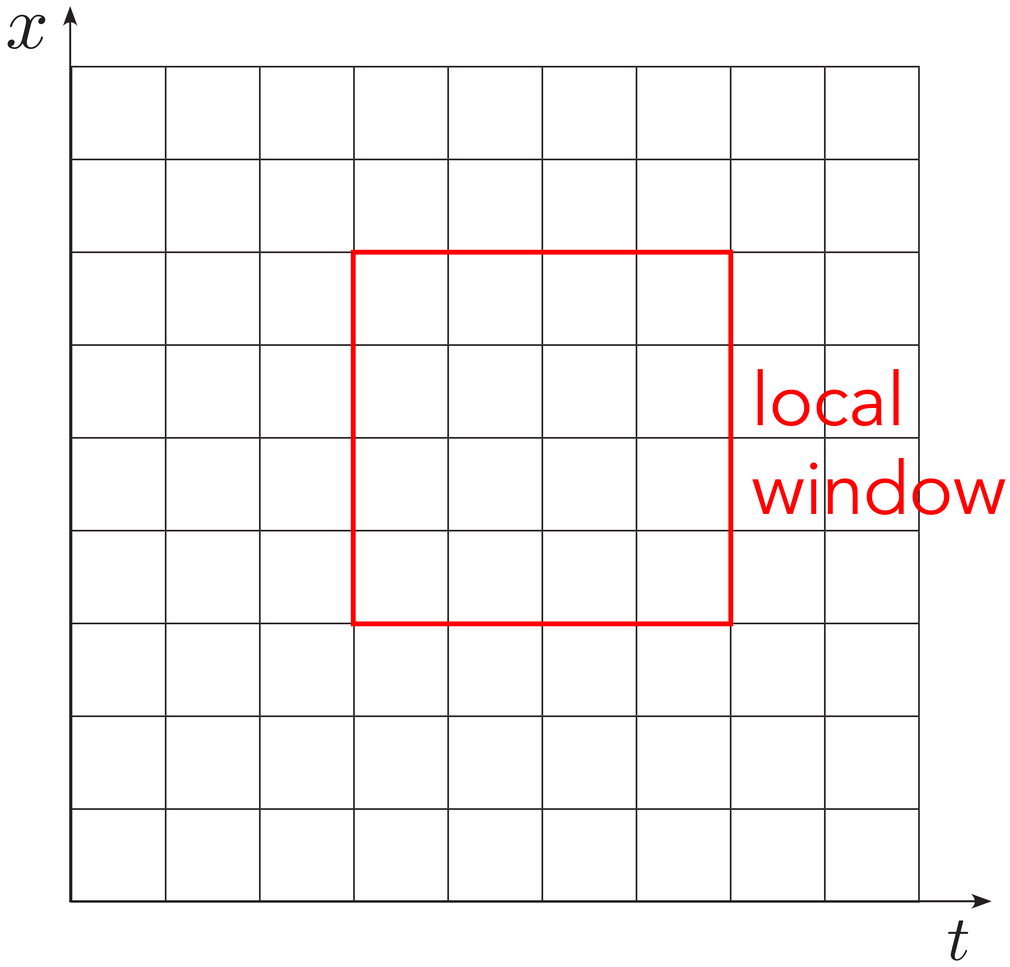
Flatland
view



© Christian Lessig, 2023

Spatio-temporal BERT

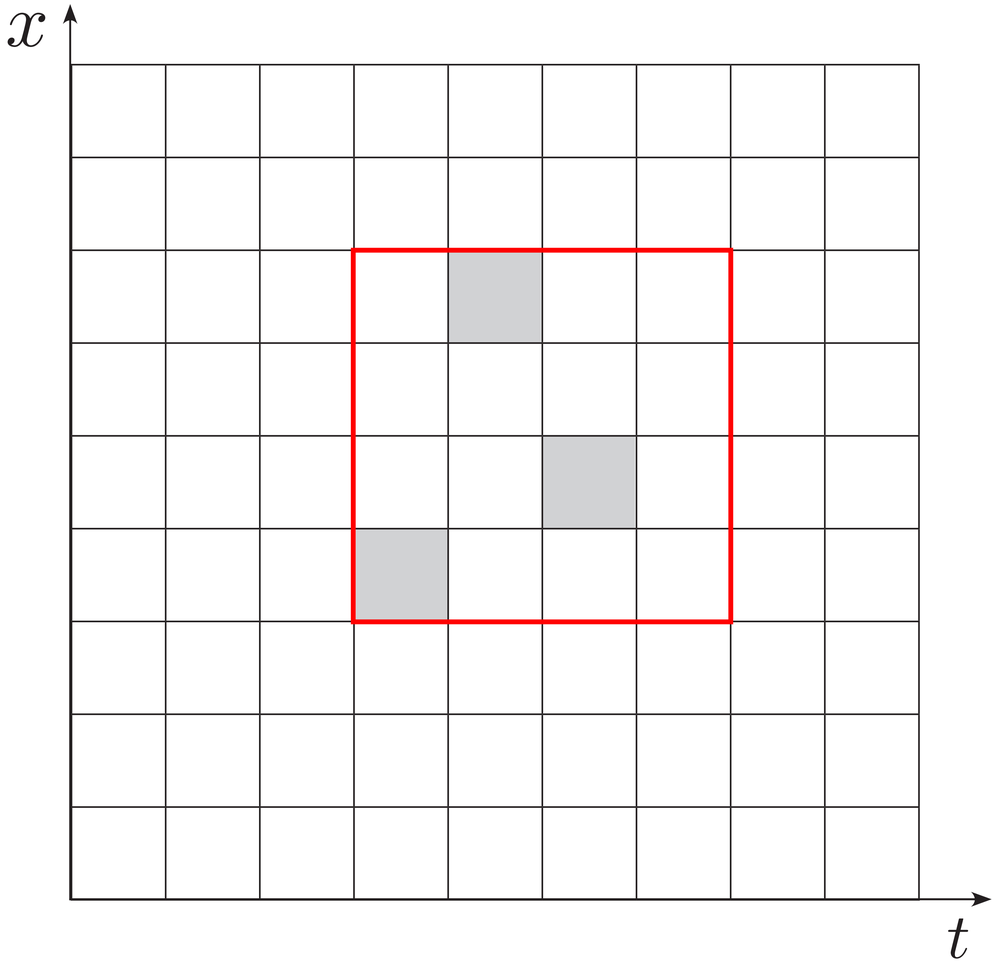
Flatland
view



© 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

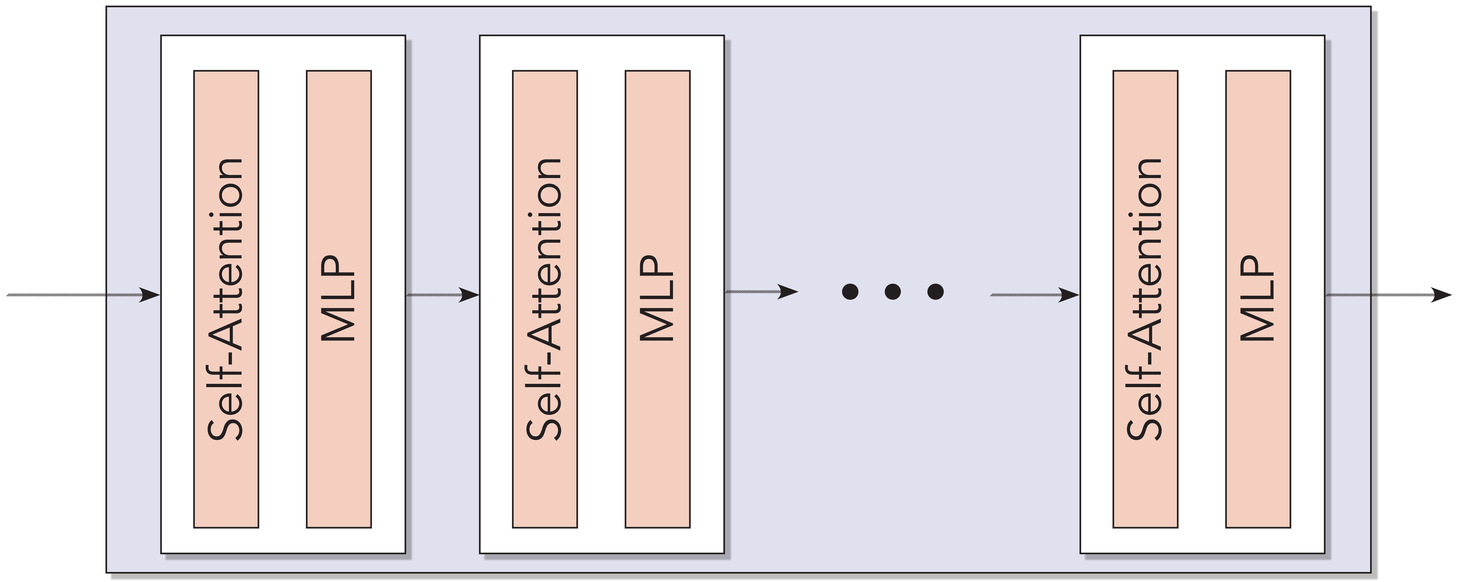
© Christian Lessig, 2023

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

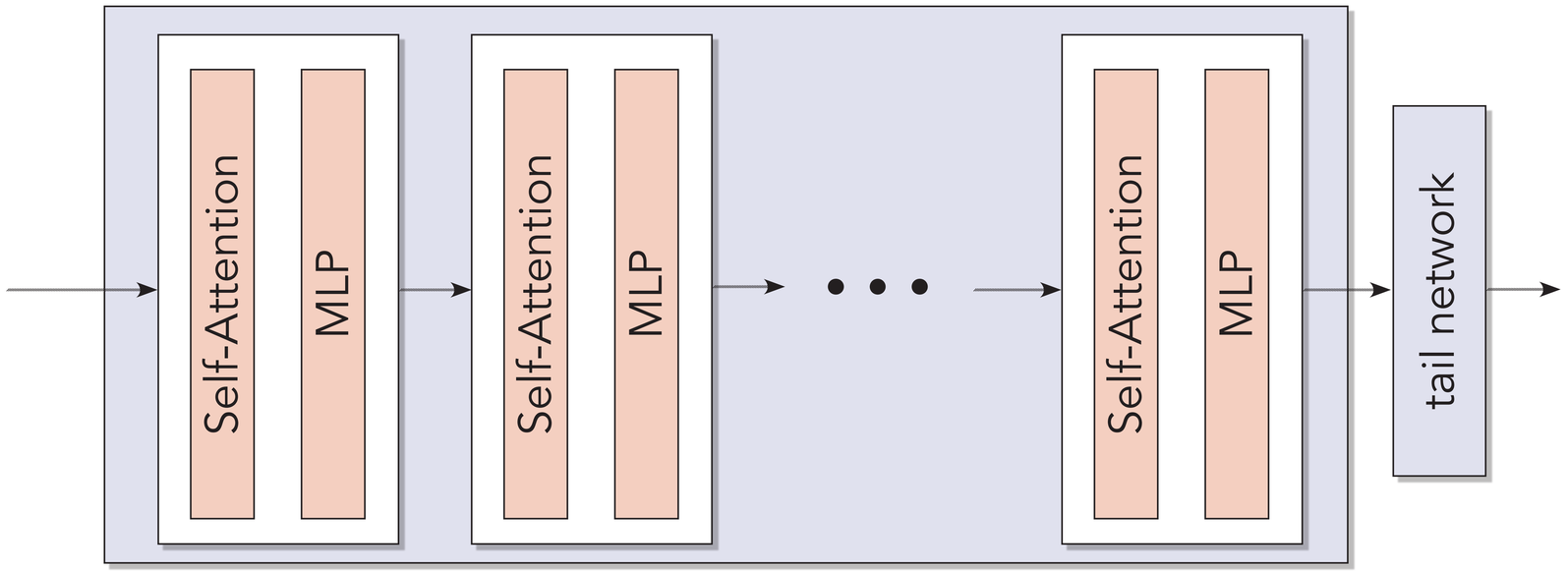
© Christian Lessig, 2023

Statistical loss



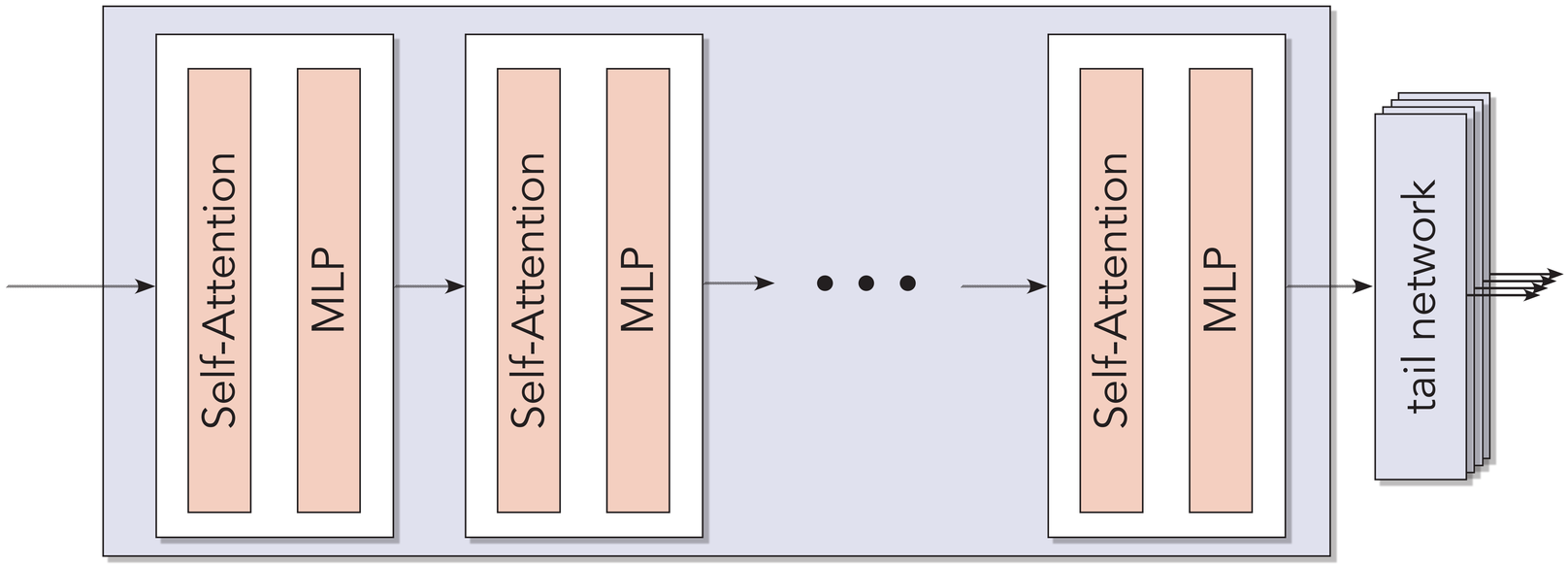
© Christian Lessig, 2023

Statistical loss



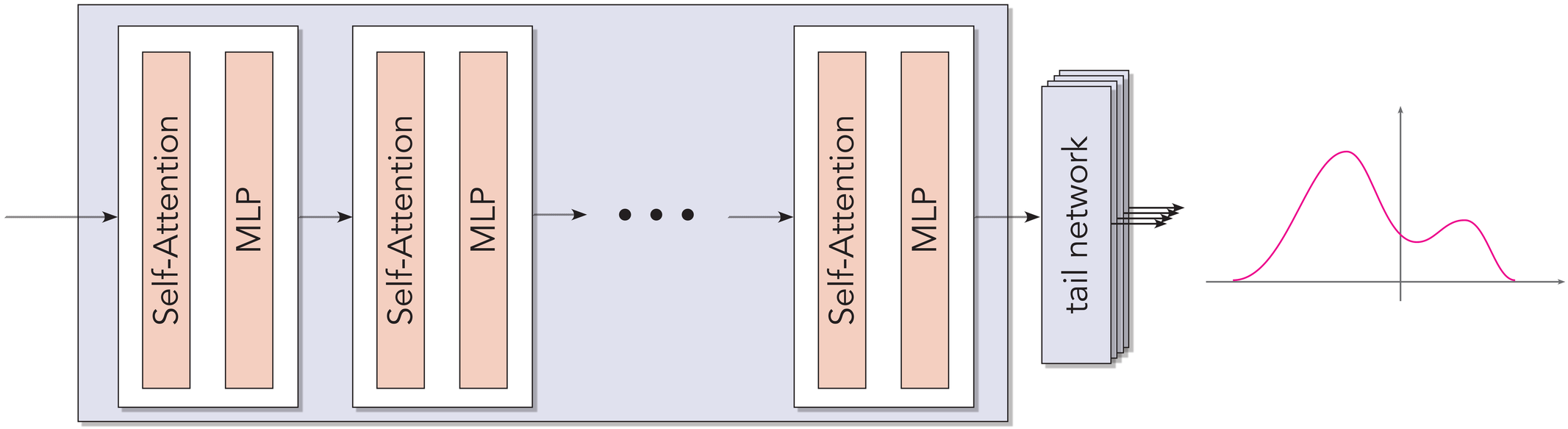
© Christian Lessig, 2023

Statistical loss



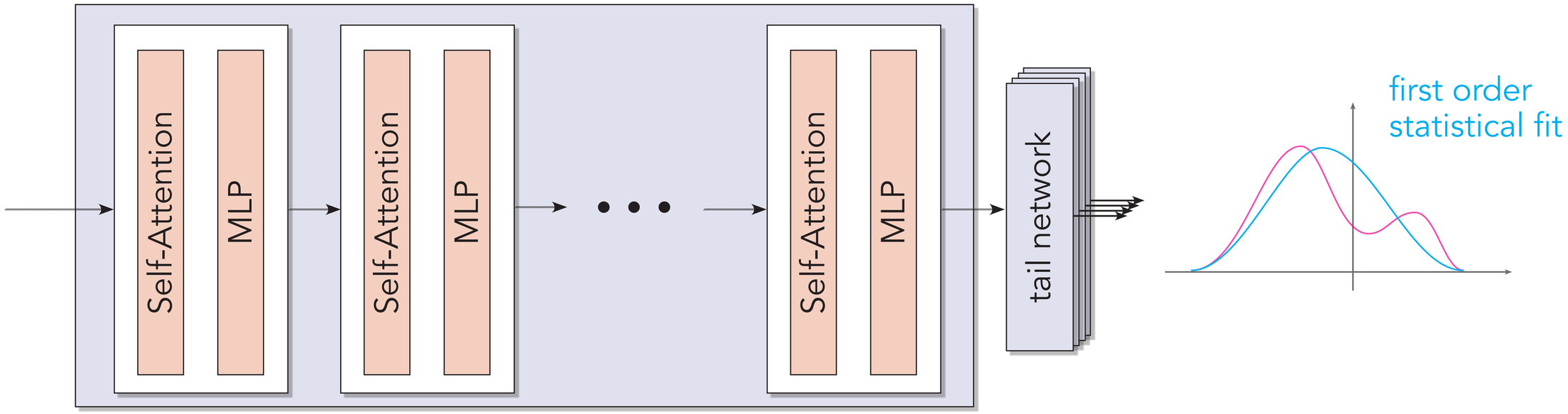
© Christian Lessig, 2023

Statistical loss



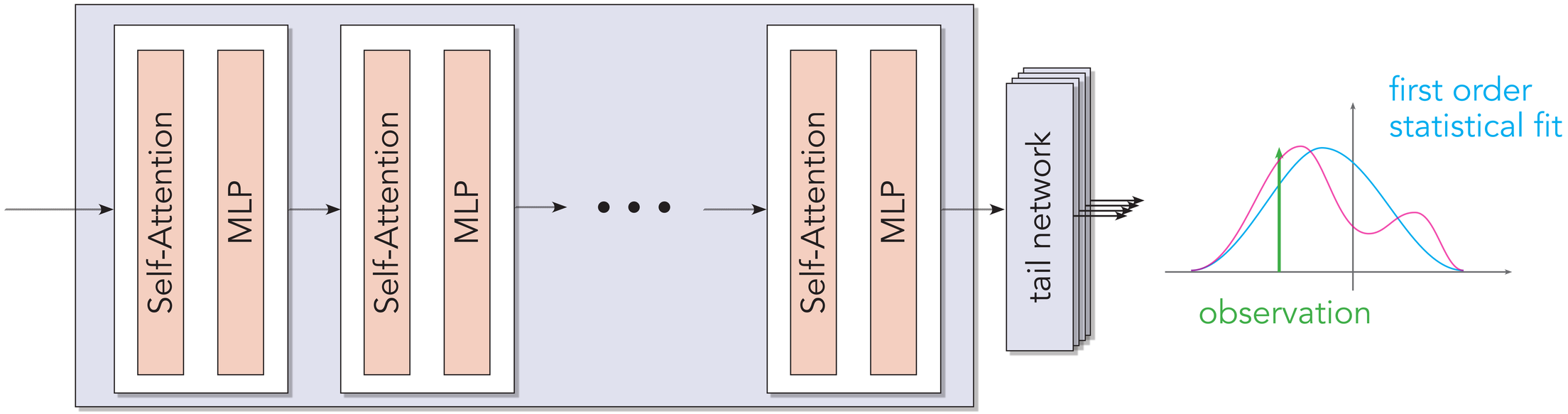
© Christian Lessig, 2023

Statistical loss



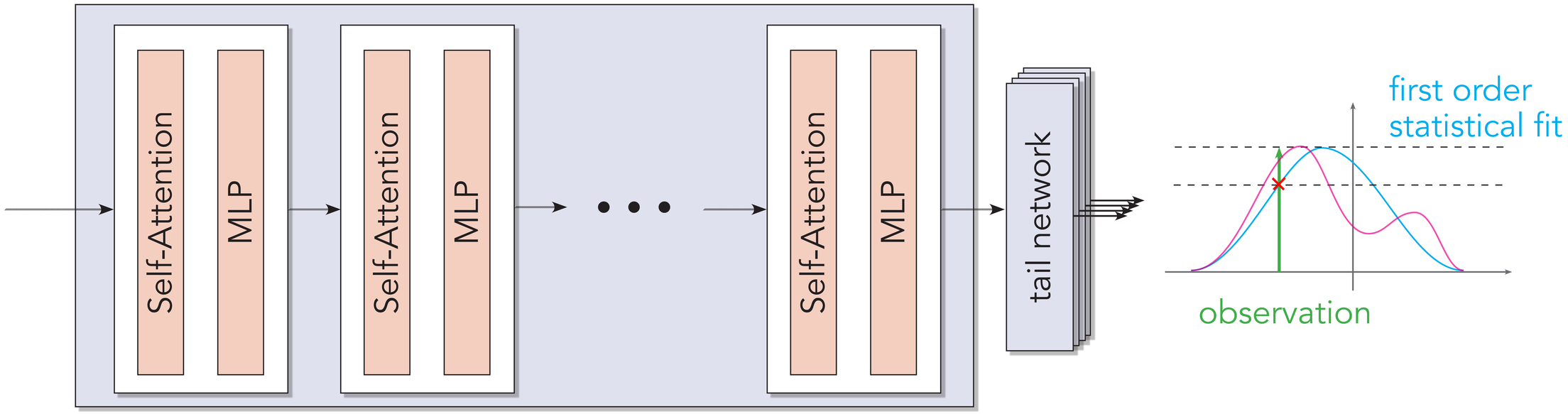
© Christian Lessig, 2023

Statistical loss



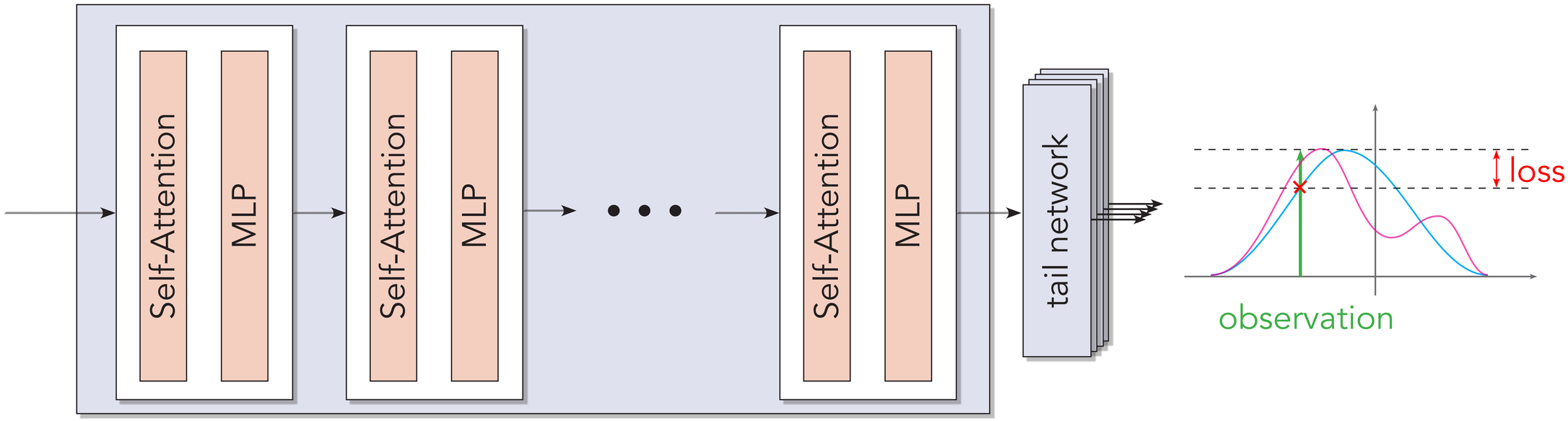
© Christian Lessig, 2023

Statistical loss



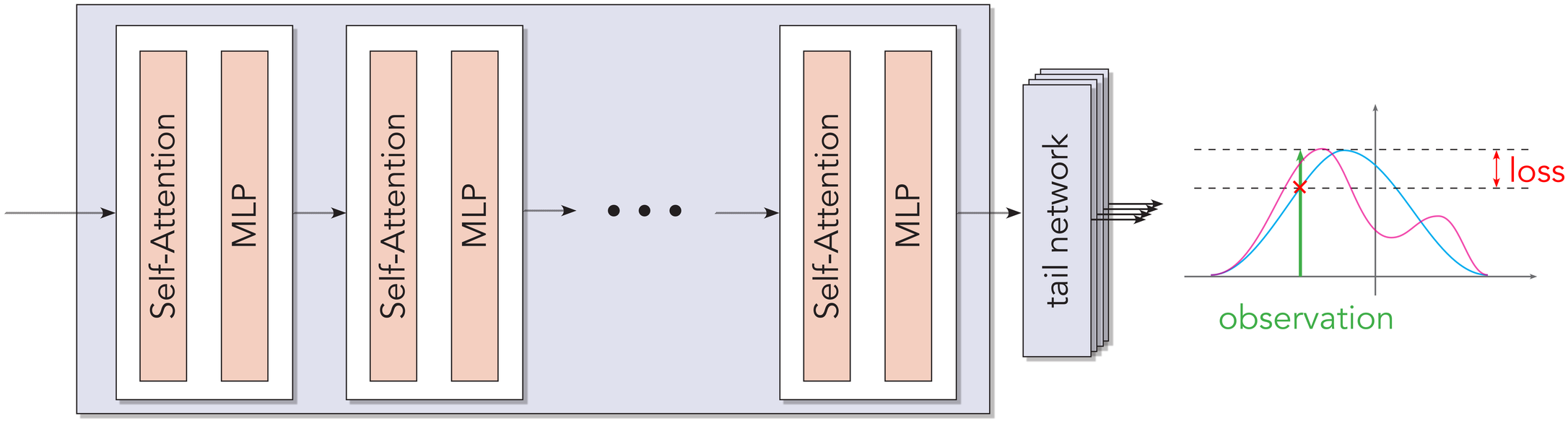
© Christian Lessig, 2023

Statistical loss



© Christian Lessig, 2023

Statistical loss



end-to-end training encourages statistical representation

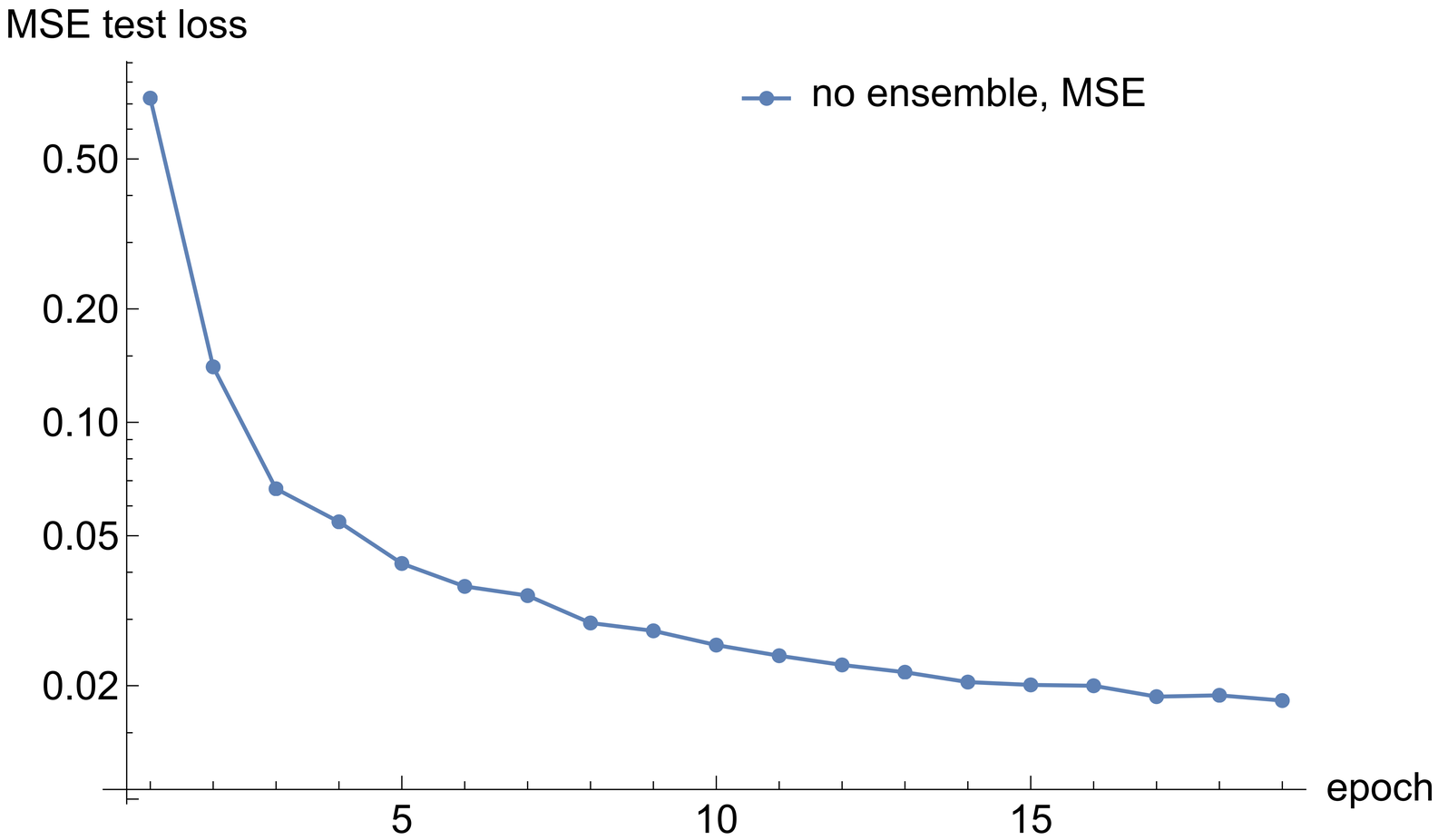
© Christian Lessig, 2023

Statistical loss: experiments

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

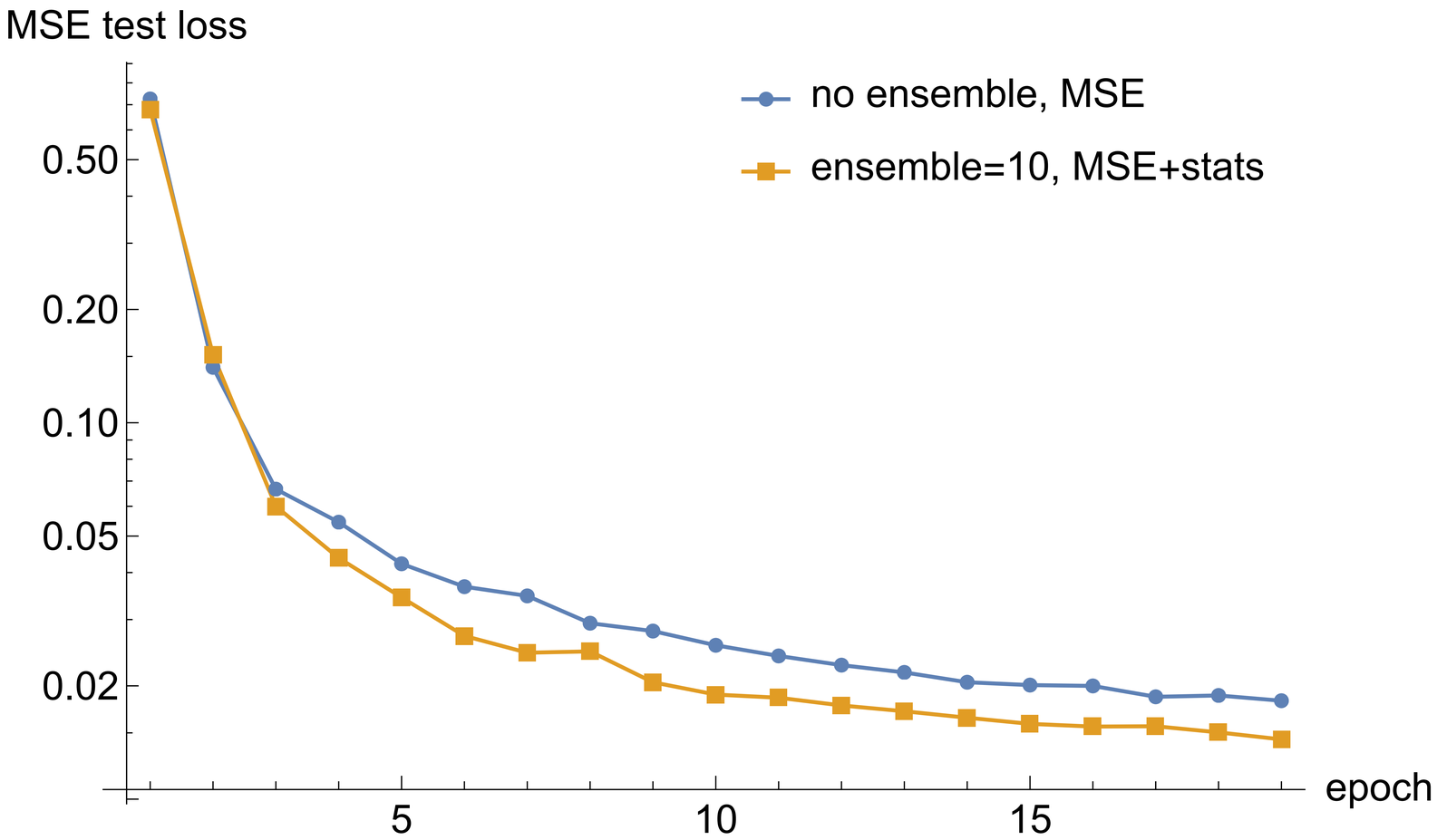
© Christian Lessig, 2023

Statistical loss



© Christian Lessig, 2023

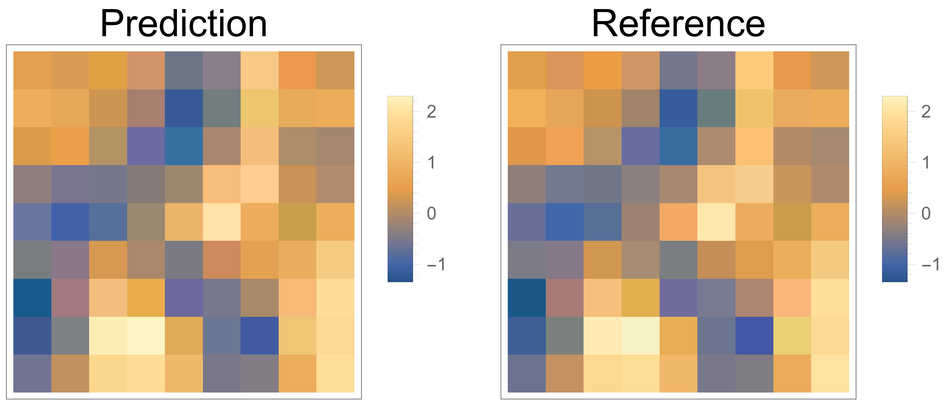
Statistical loss

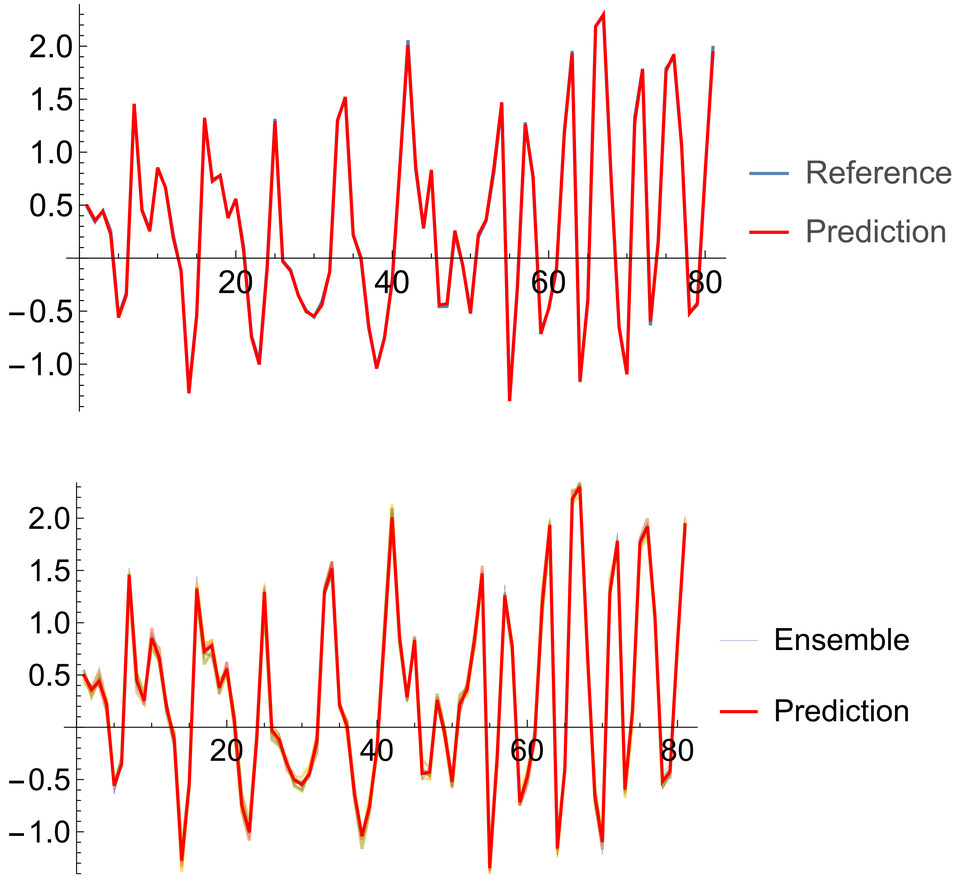


© Christian Lessig, 2023

Statistical loss

- Predictions:

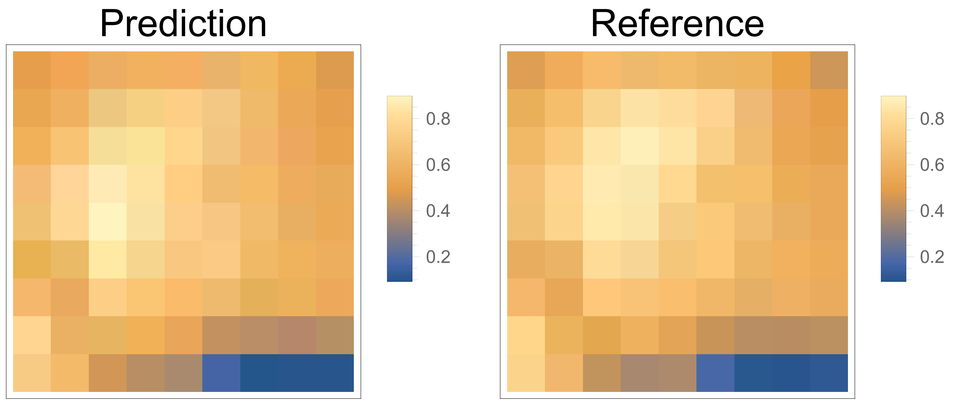


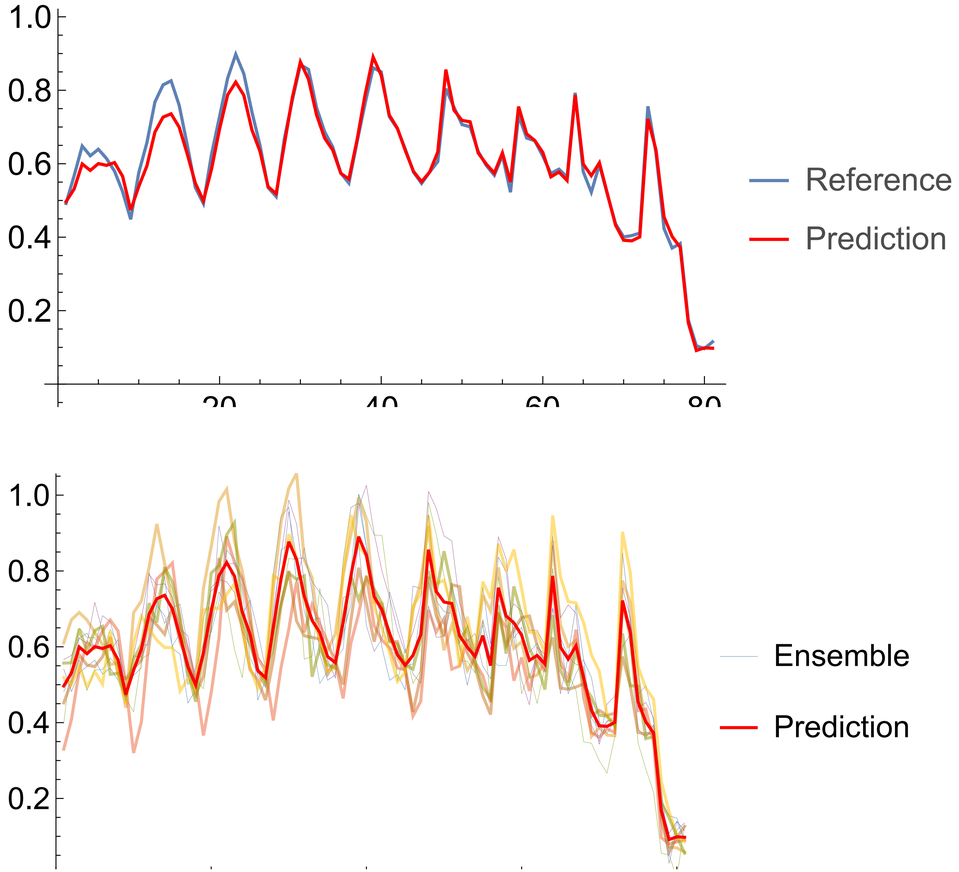


© Christian Lessig, 2023

Statistical loss

- Predictions:

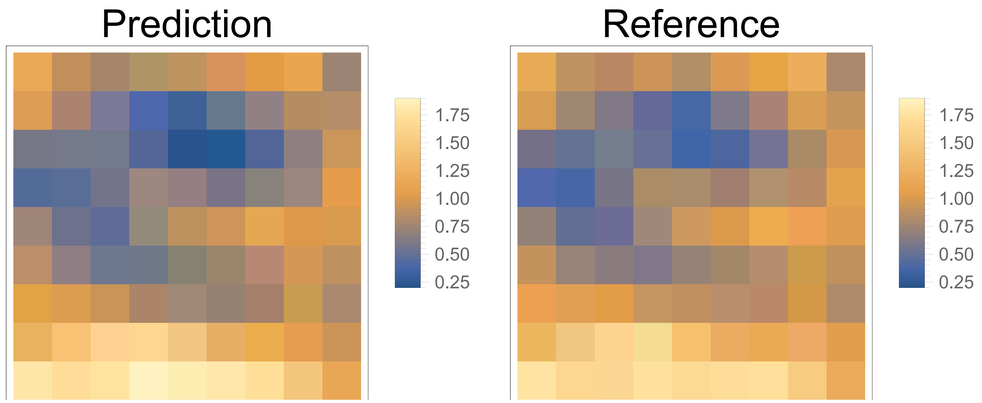


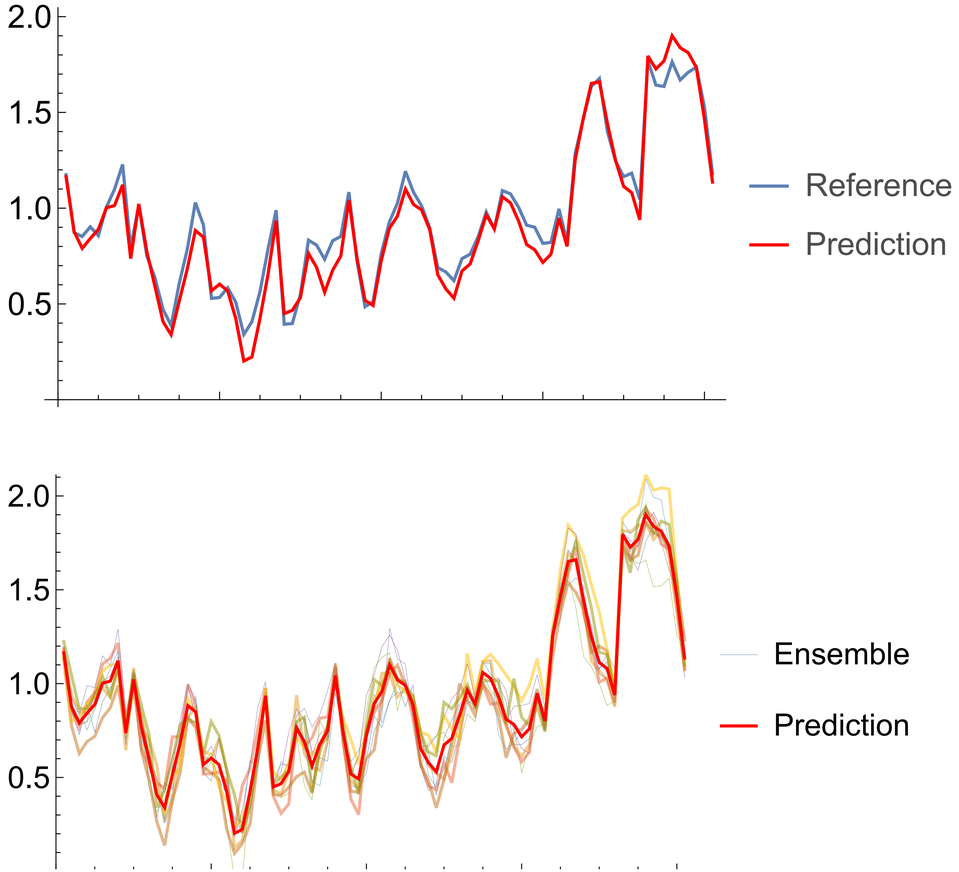


© Christian Lessig, 2023

Statistical loss

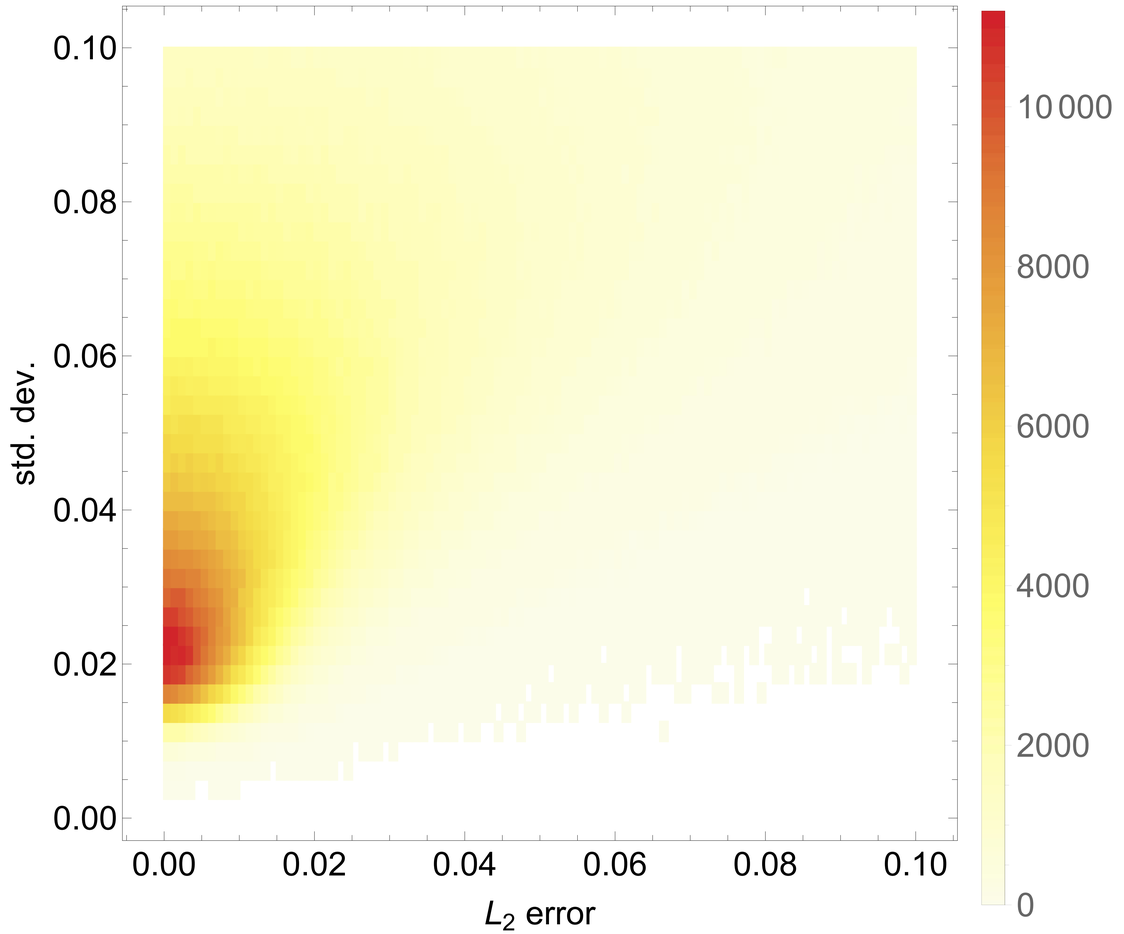
- Predictions:





© Christian Lessig, 2023

Statistical loss



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

© Christian Lessig, 2023

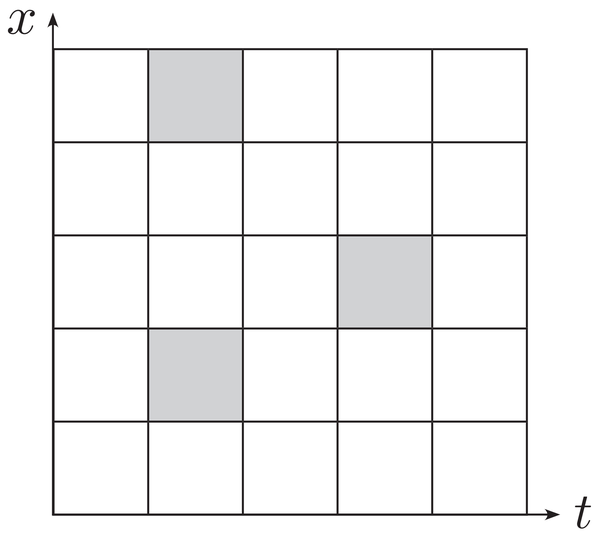
Zero shot evaluation

- Evaluate performance on representation network *as is*

© Christian Lessig, 2023

Zero shot evaluation

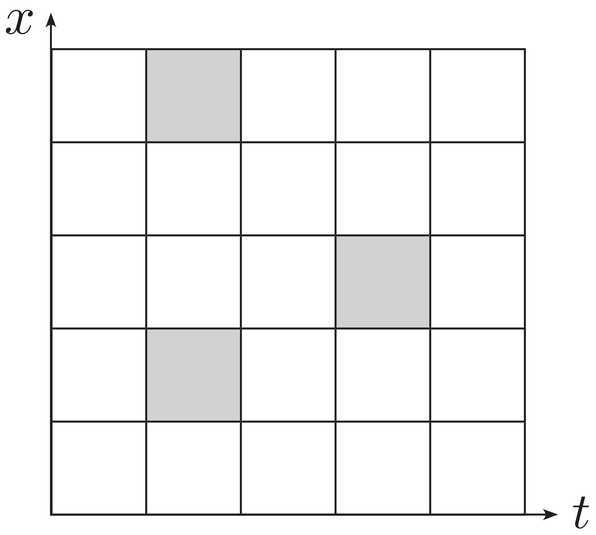
BERT



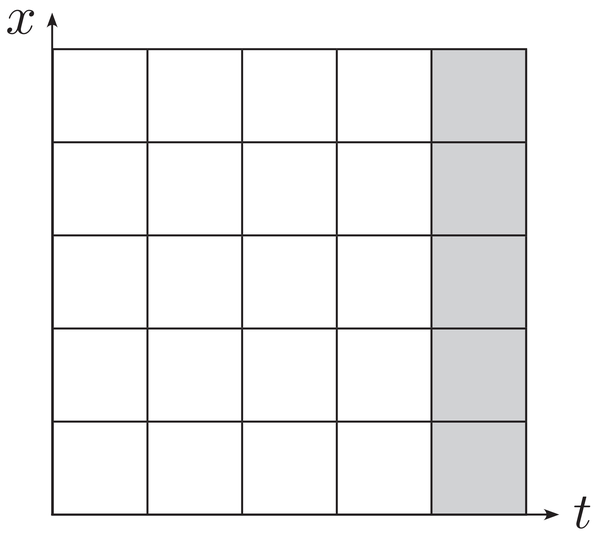
© Christian Lessig, 2023

Zero shot evaluation

BERT



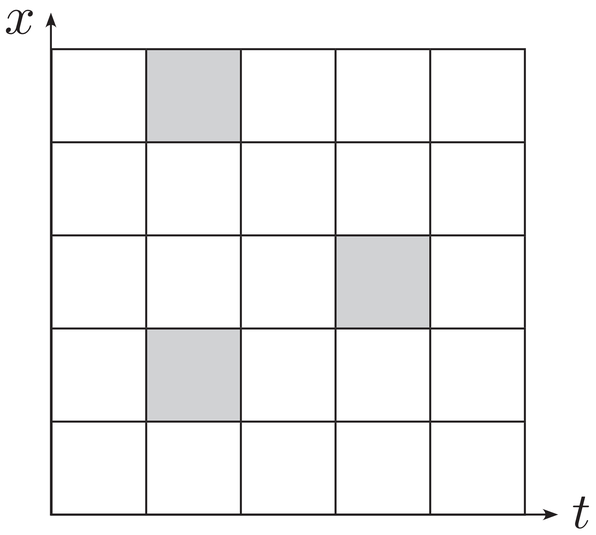
BERT-Forecast



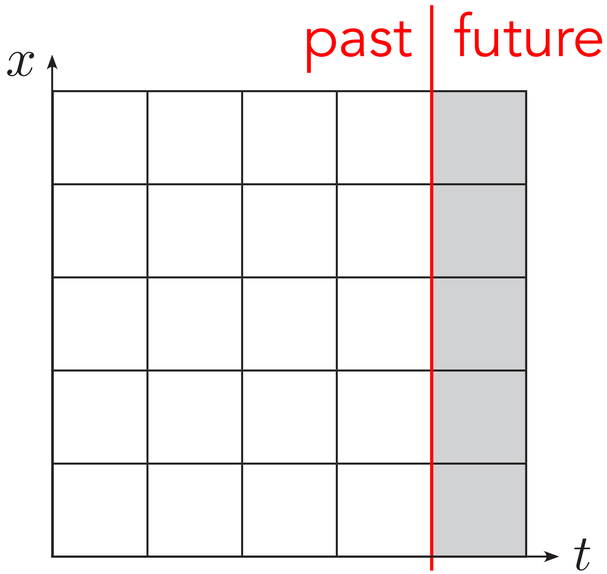
© Christian Lessig, 2023

Zero shot evaluation

BERT

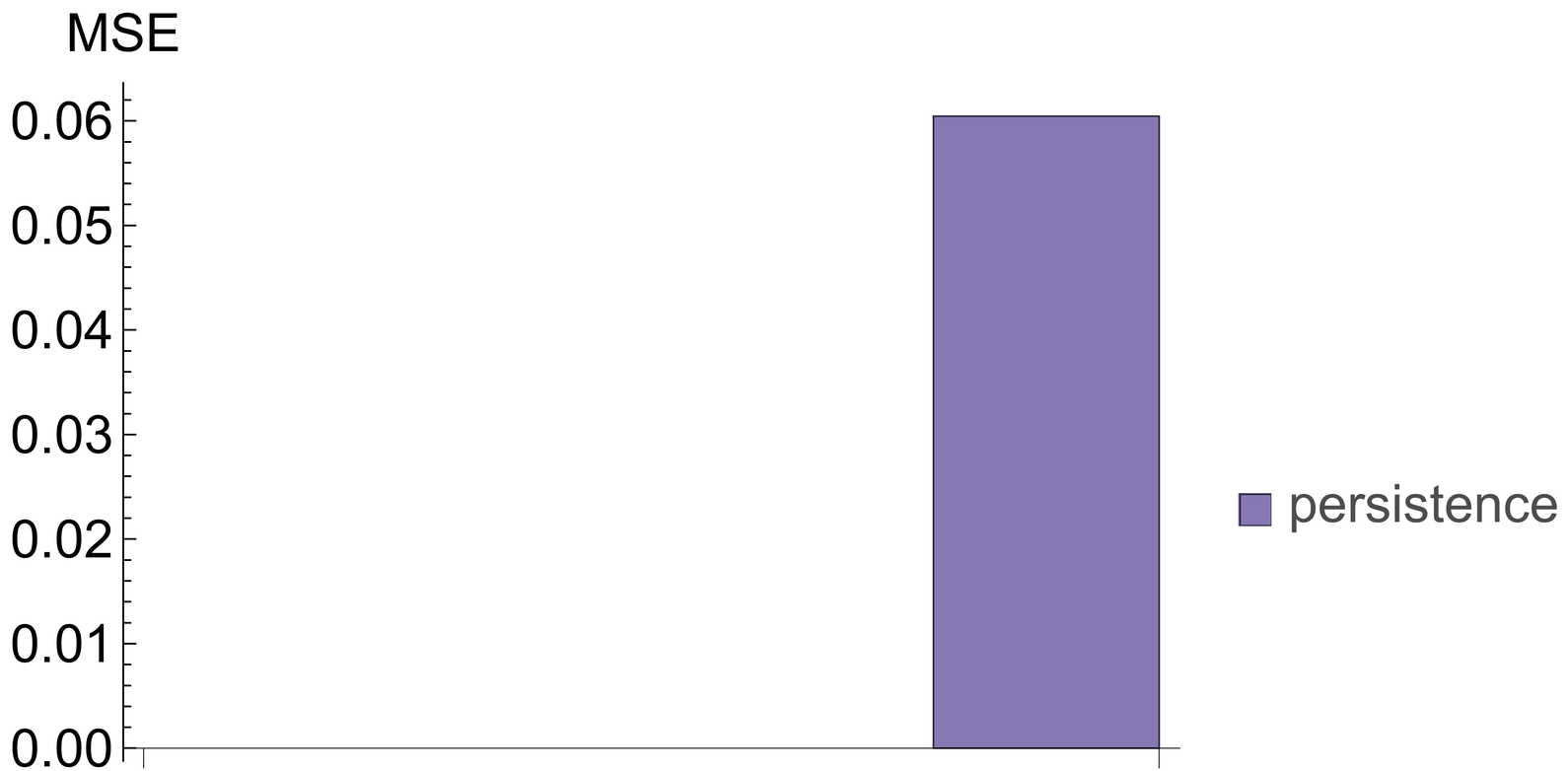


BERT-Forecast



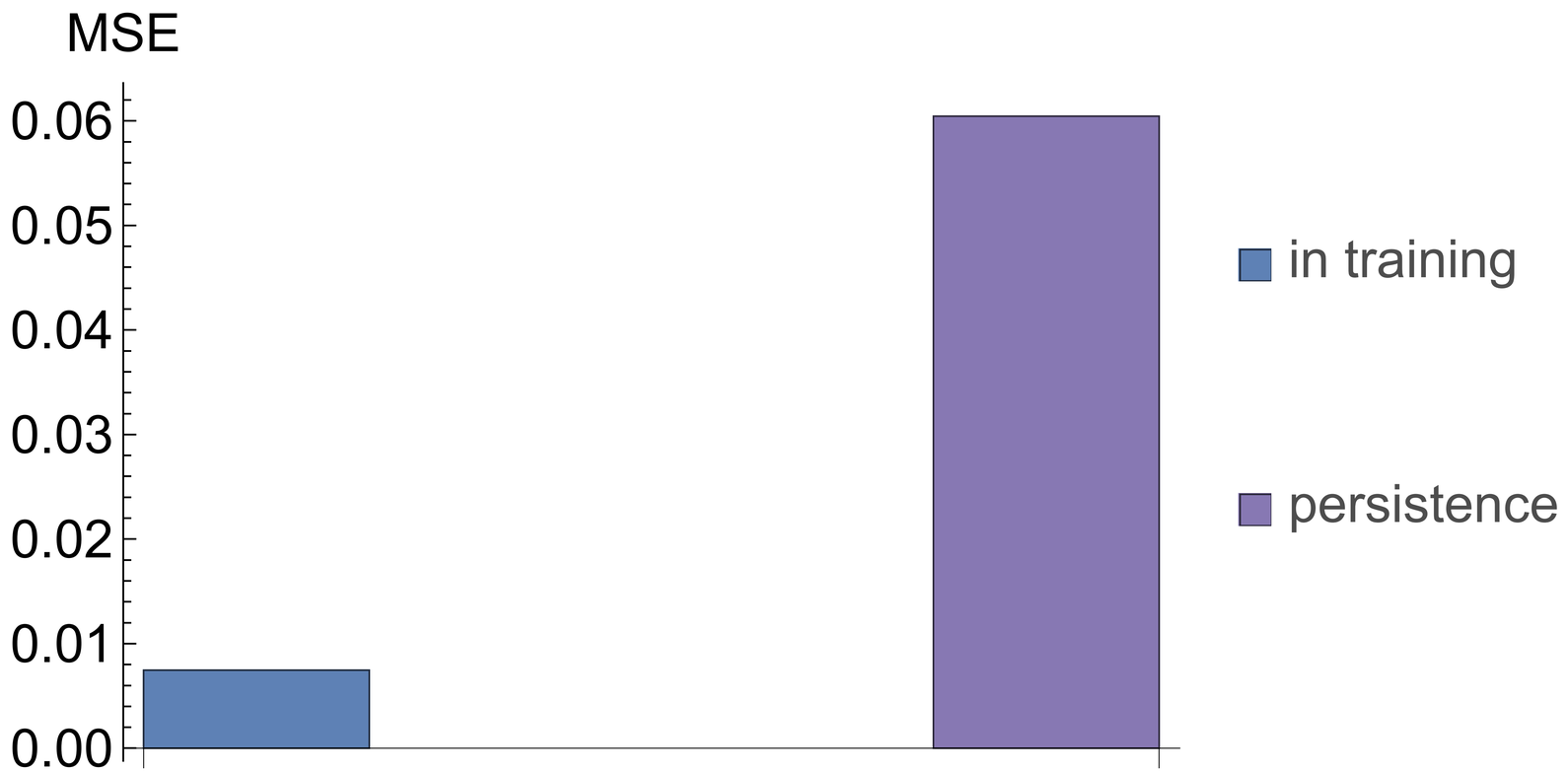
© Christian Lessig, 2023

Zero shot performance



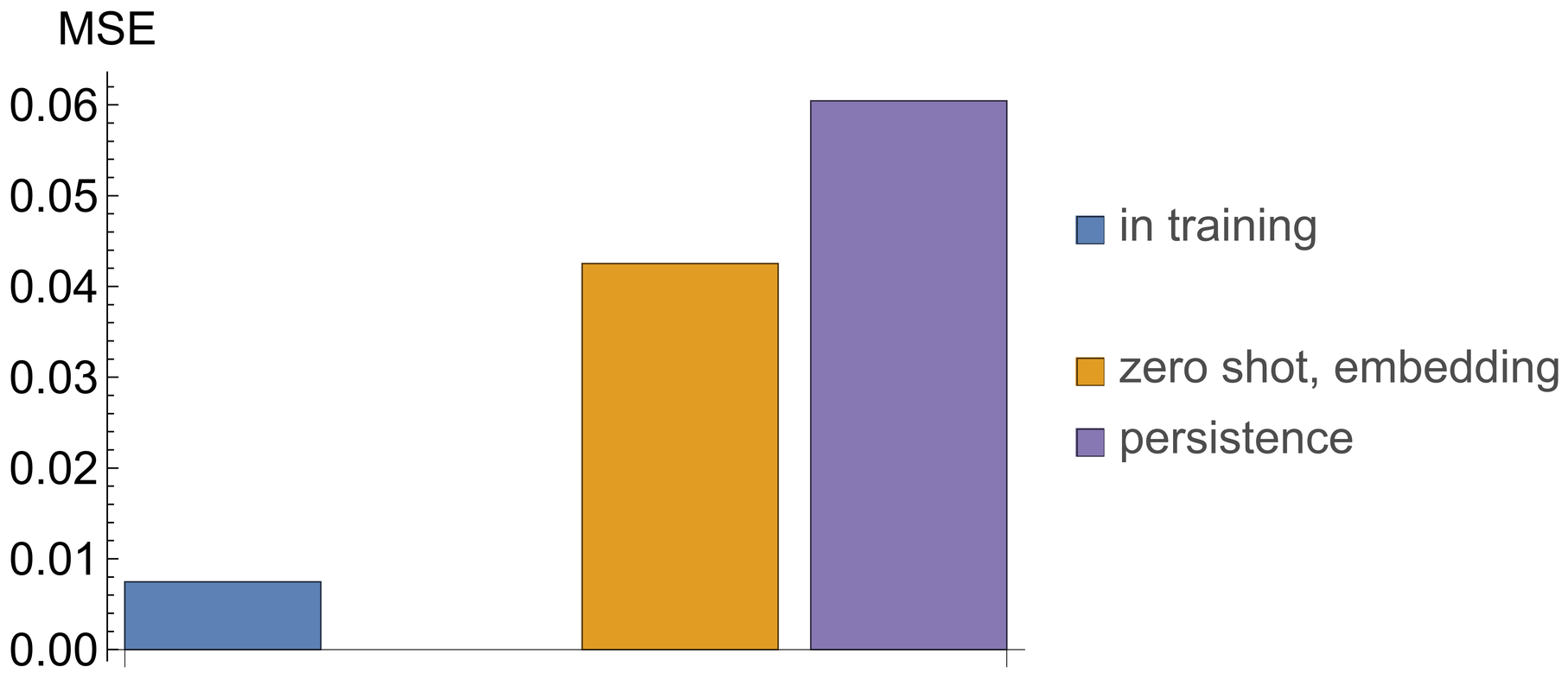
© Christian Lessig, 2023

Zero shot performance



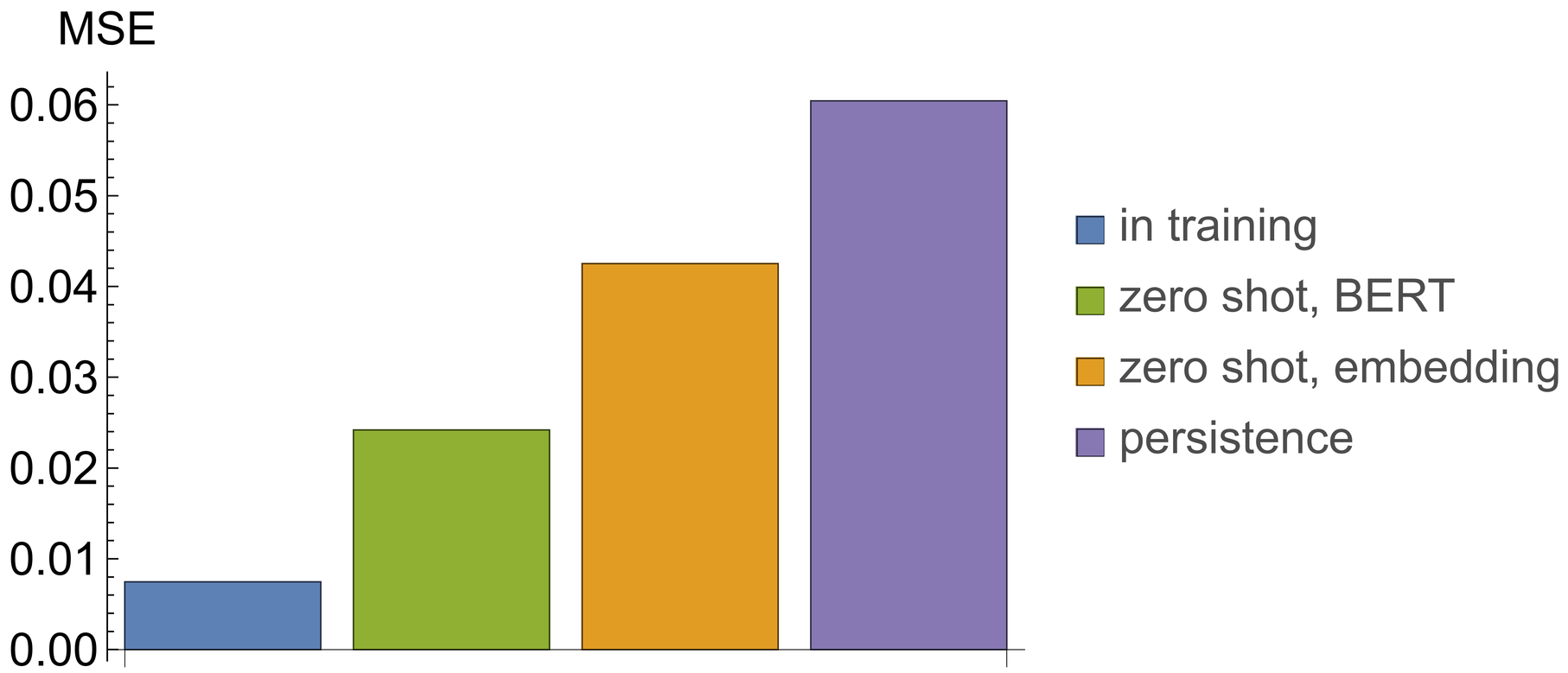
© Christian Lessig, 2023

Zero shot performance



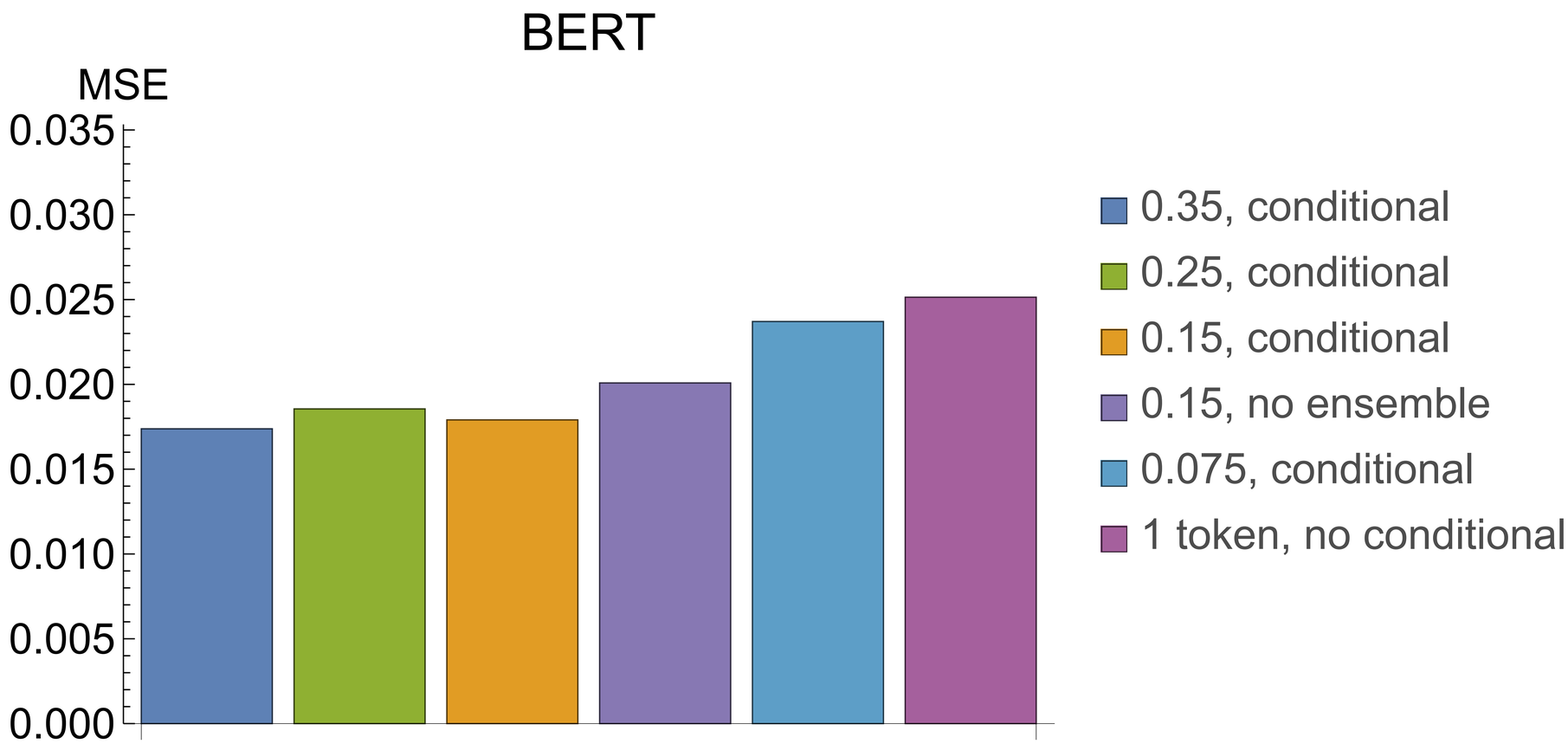
© Christian Lessig, 2023

Zero shot performance



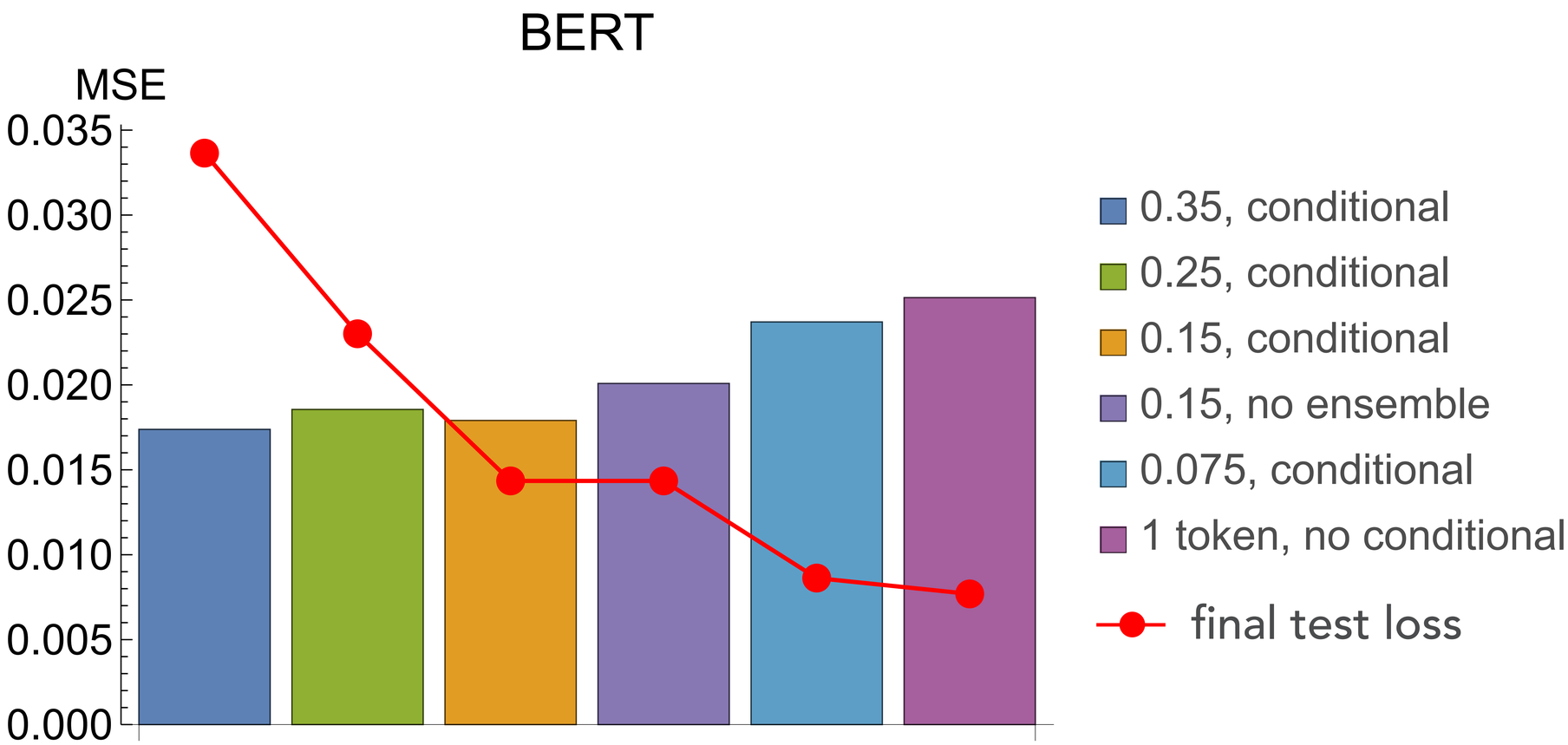
© Christian Lessig, 2023

Zero shot performance



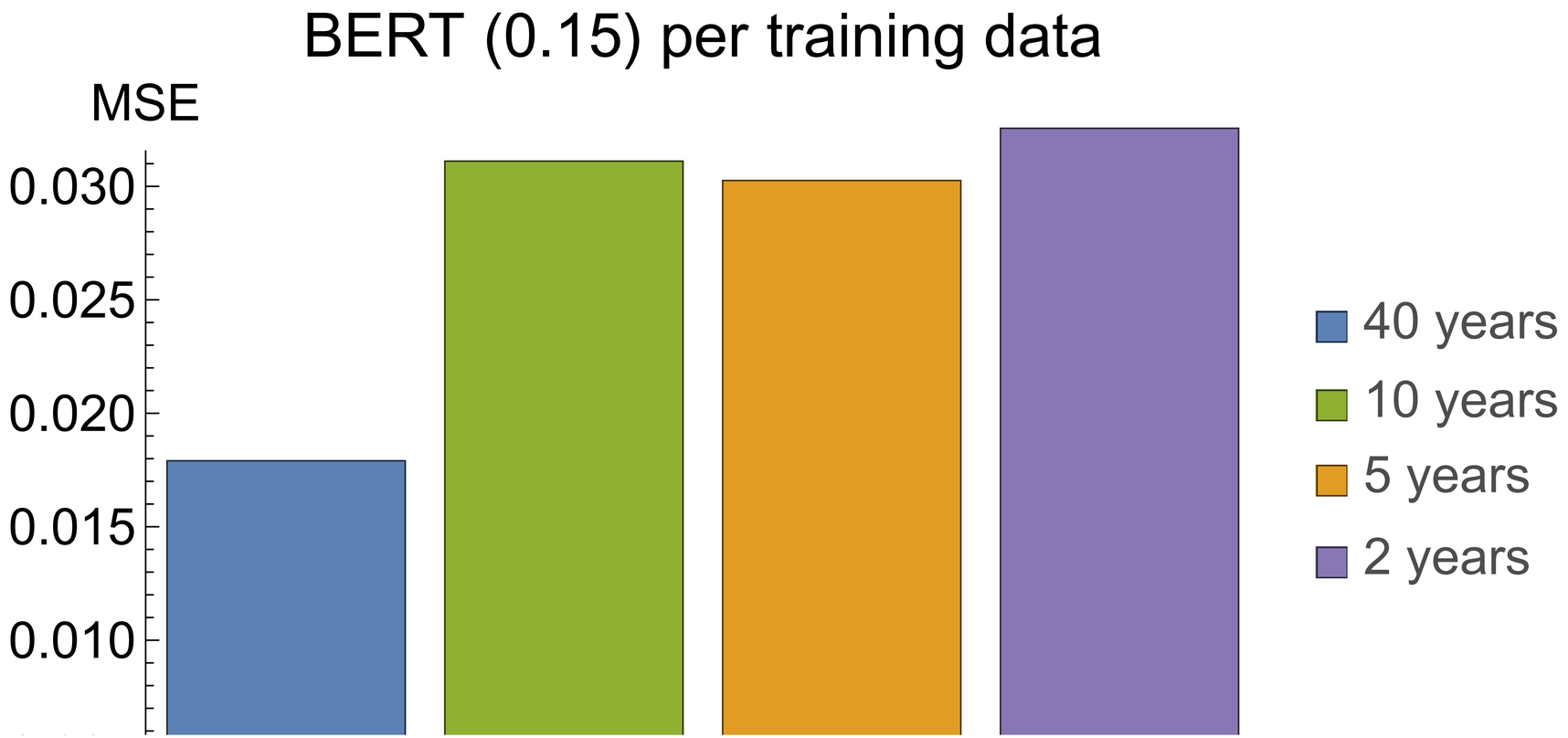
© Christian Lessig, 2023

Zero shot performance



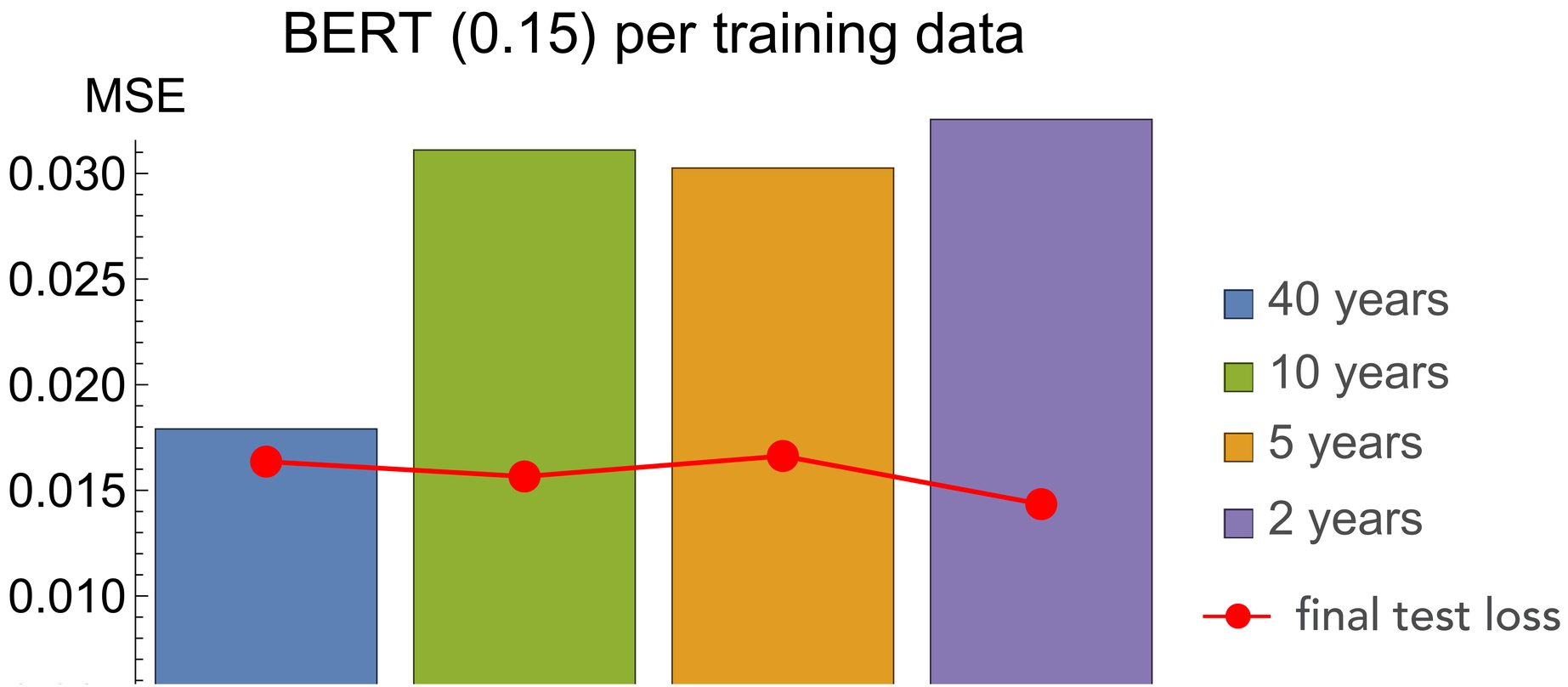
© Christian Lessig, 2023

Zero shot performance



© Christian Lessig, 2023

Zero shot performance



© Christian Lessig, 2023

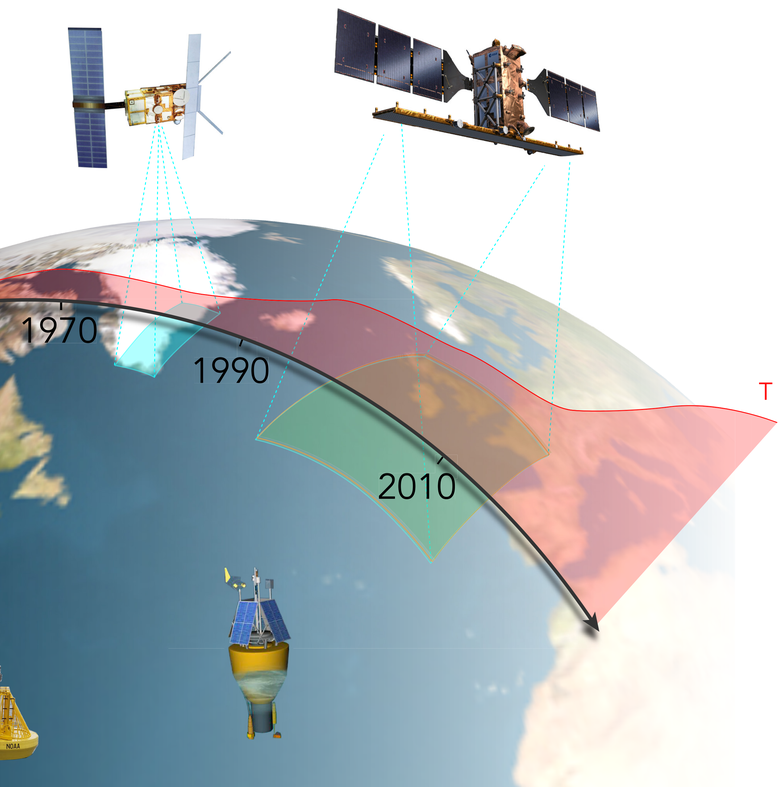
AtmoRep: longer term objectives

- Weather forecasting
- Climate projections
- Coupled Earth system
- Scientific model
- Training/fine-tuning on direct observational data

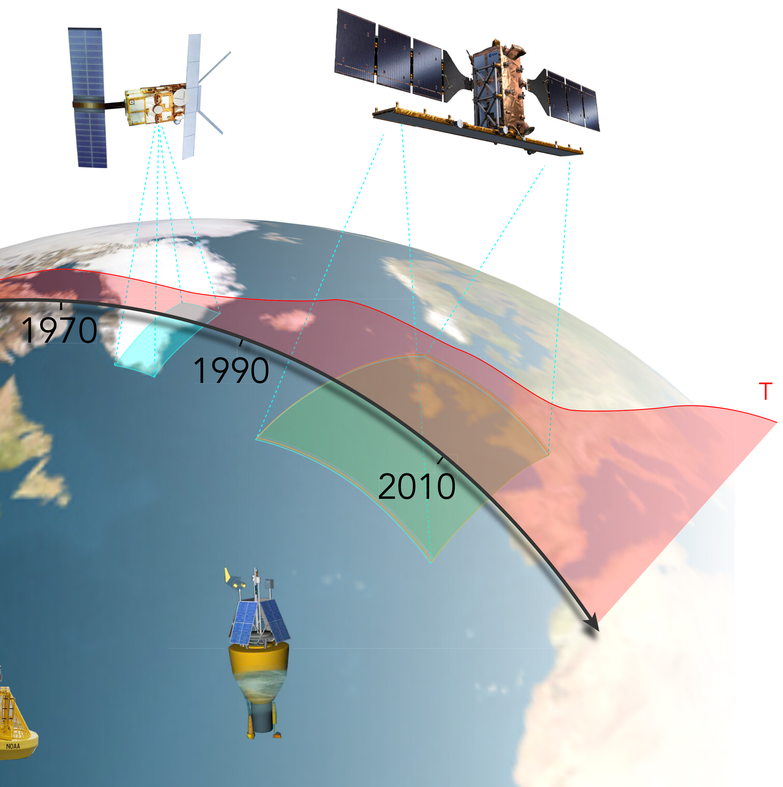
© Christian Lessig, 2023

Current / next steps

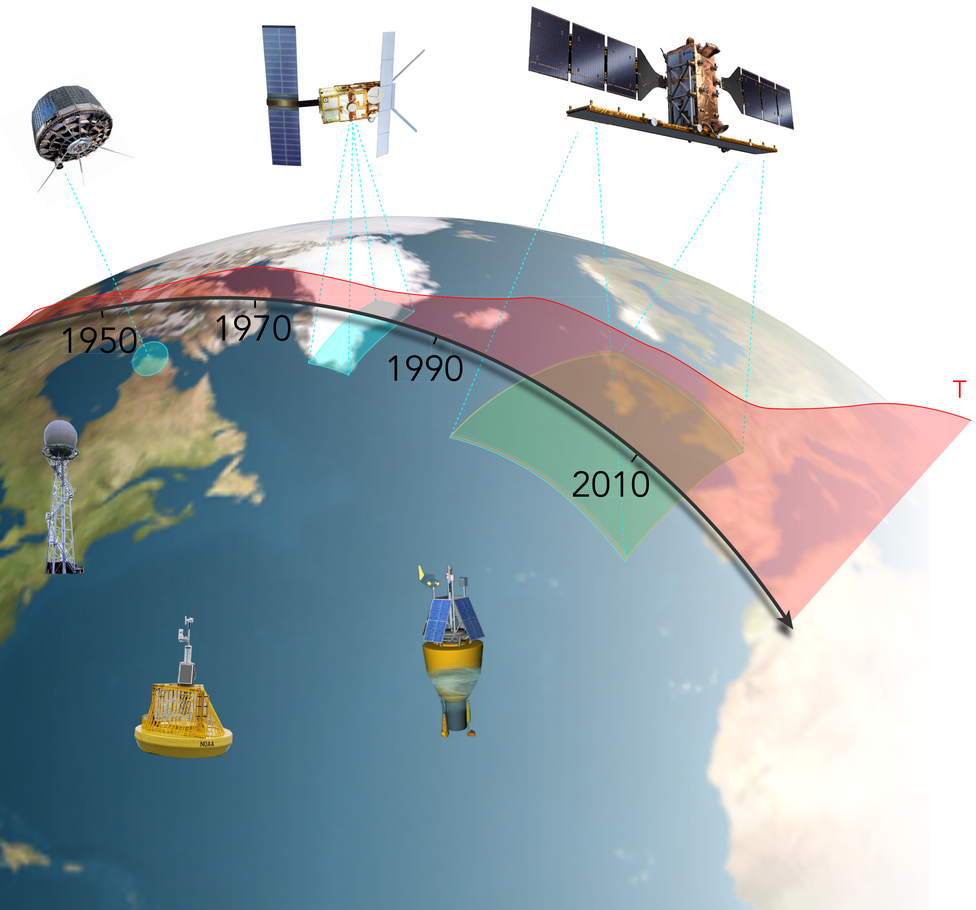
- Complete representation learning model
  - Scale data and network size
  - Different training tasks and protocols



Current / next steps

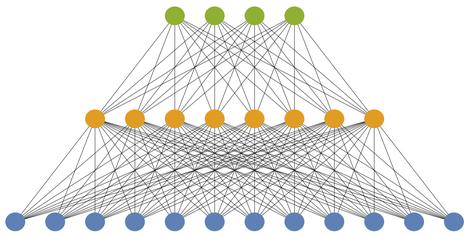


- Complete representation learning model
- Downstream applications
  - Weather forecasting
  - Downscaling
  - Model correction
  - ...



AtmoRep

Large scale representation learning of atmospheric dynamics

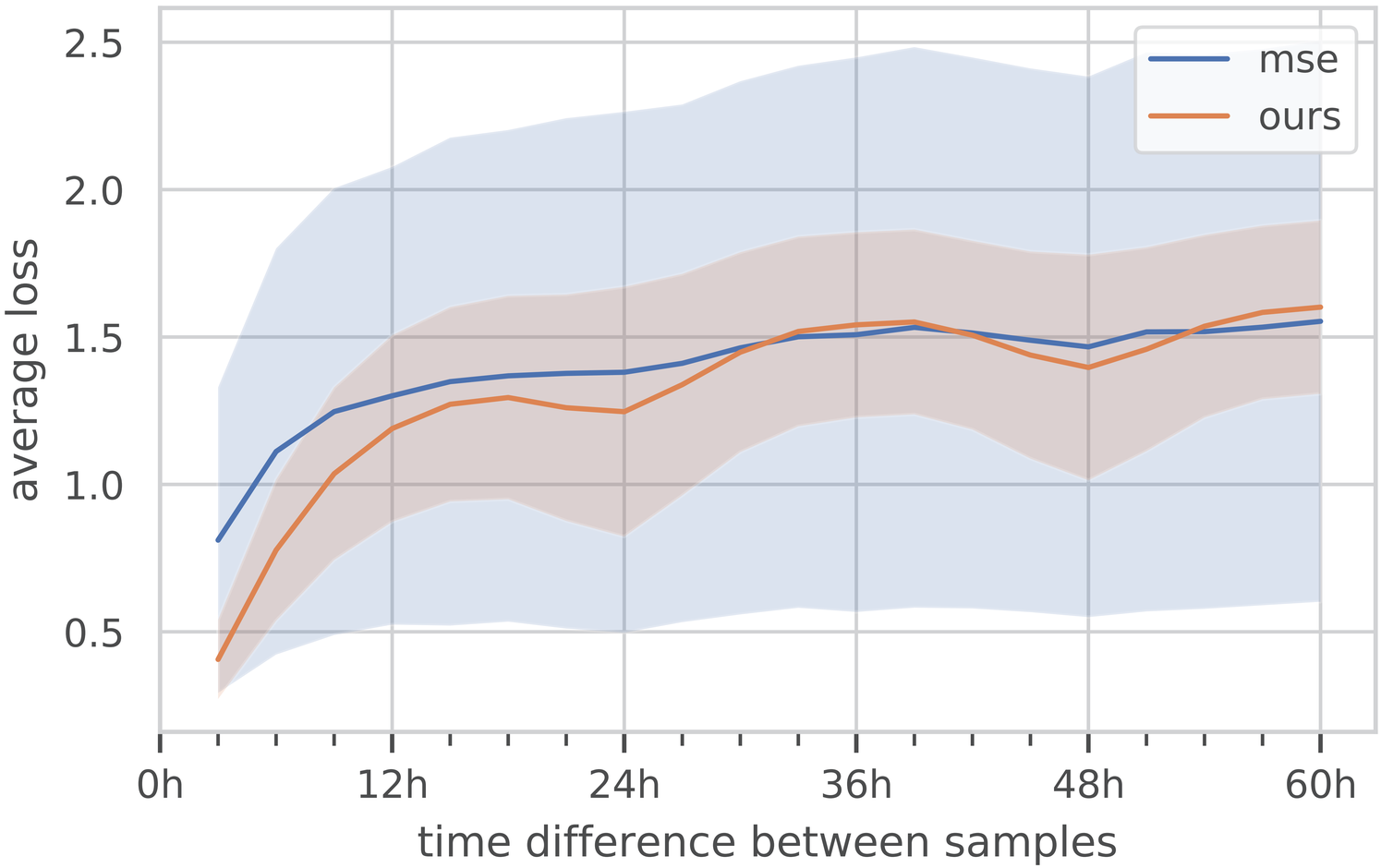


address climate change

large scale machine learning

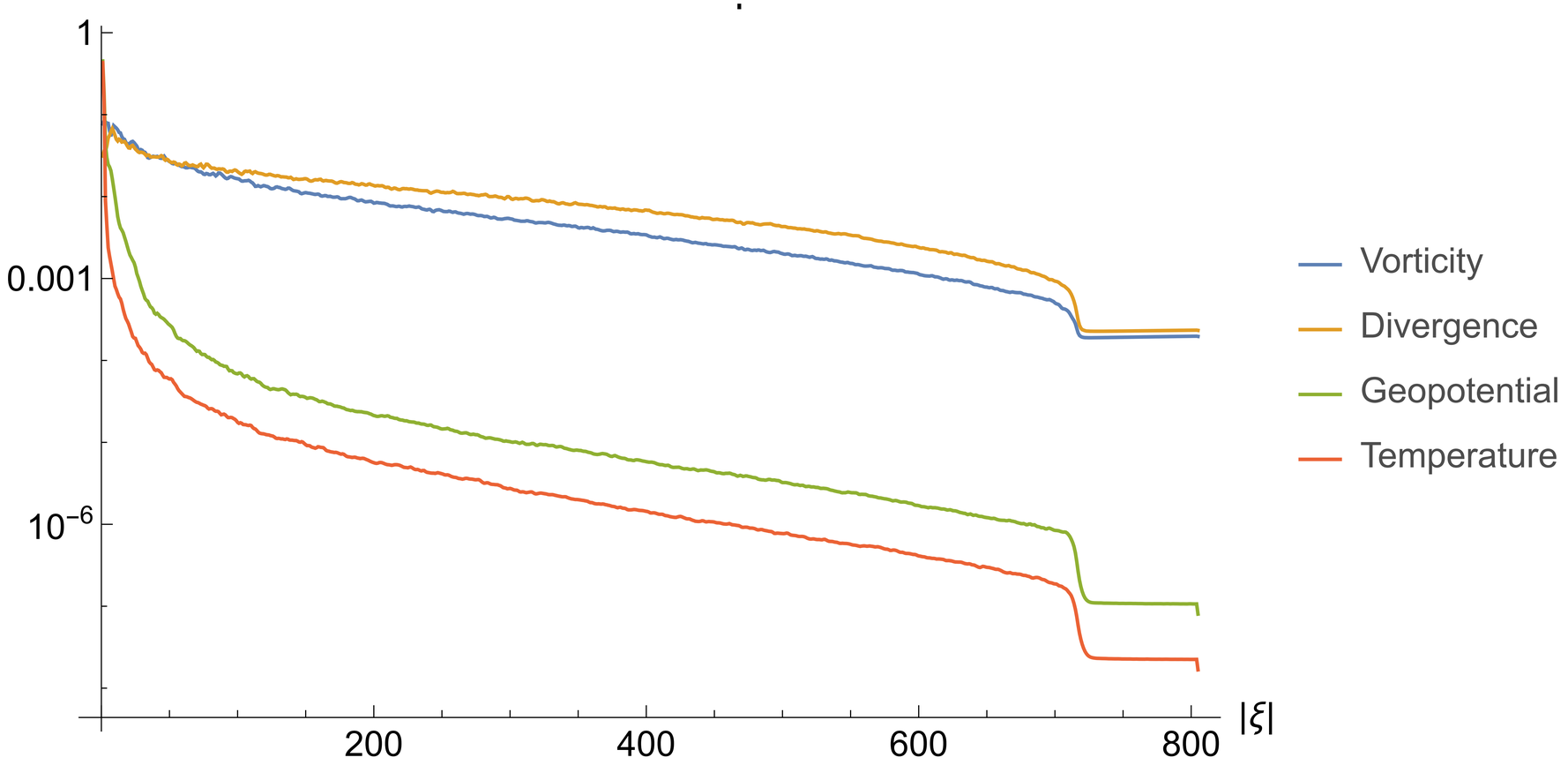
scientific insight

AtmoDist: evaluation



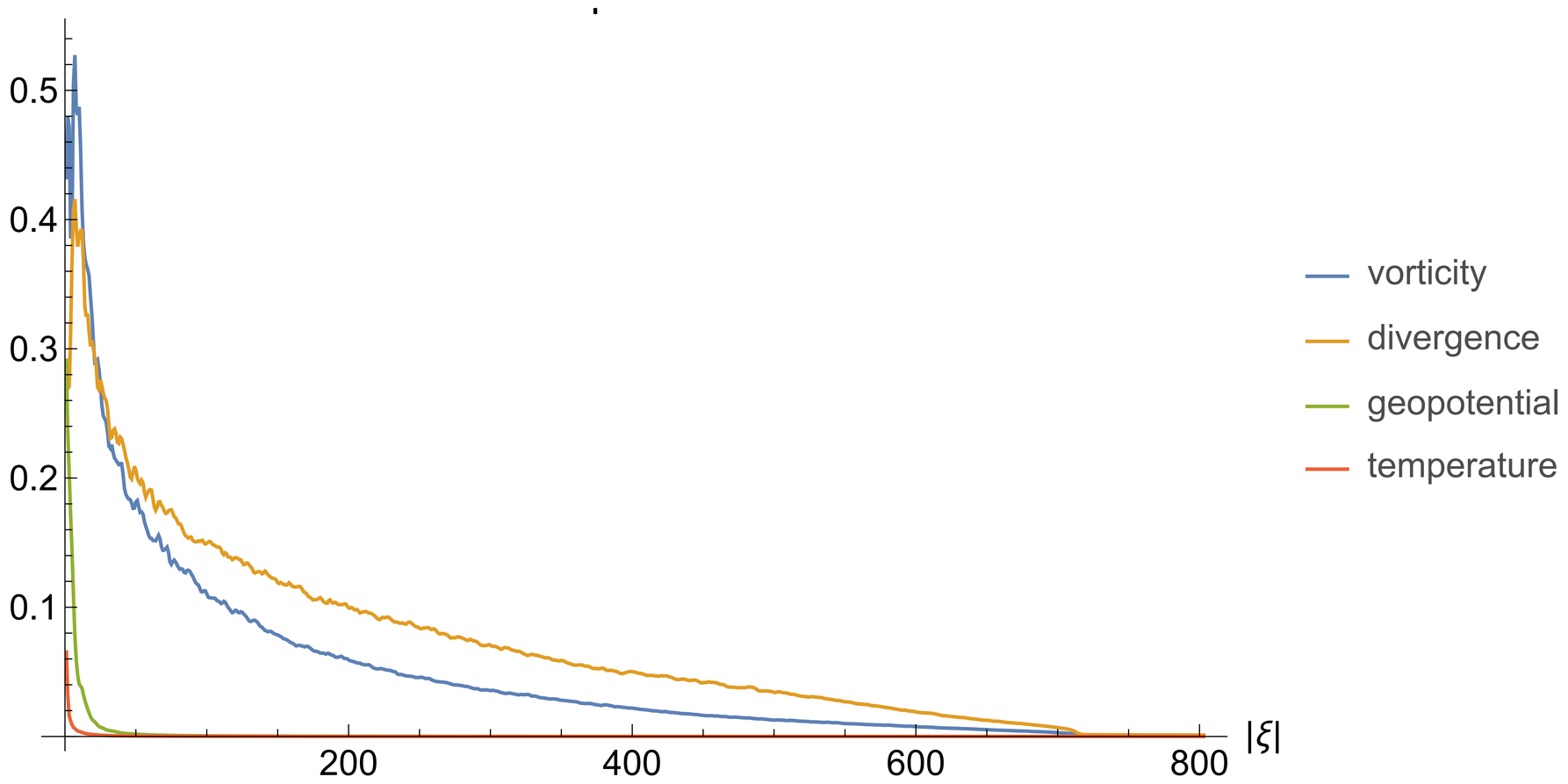
© Christian Lessig, 2023

AtmoRep data



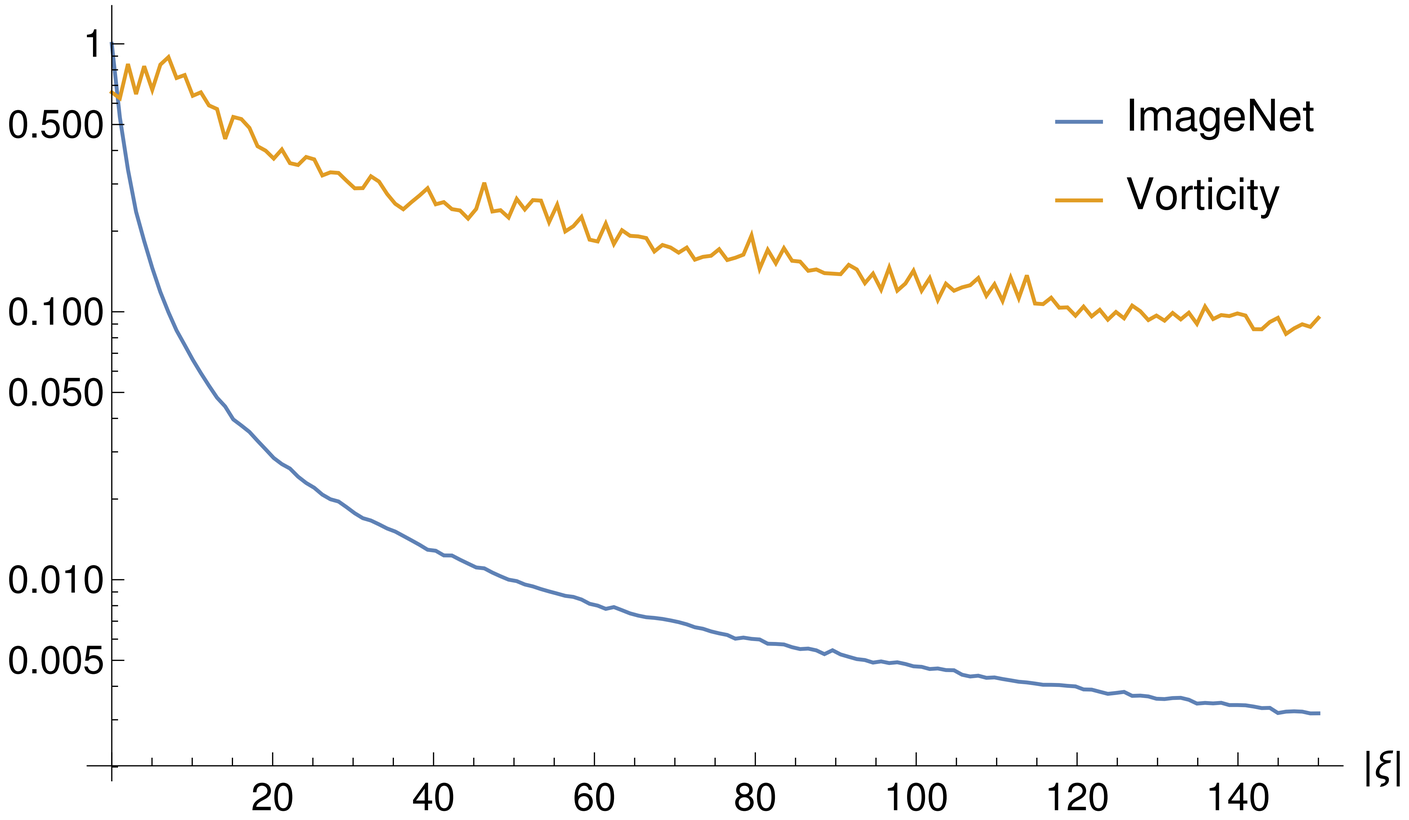
© Christian Lessig, 2023

AtmoRep data



© Christian Lessig, 2023

ERA5 versus ImageNet



© Christian Lessig, 2023

ERA5 versus ImageNet

stream function
velocity potential



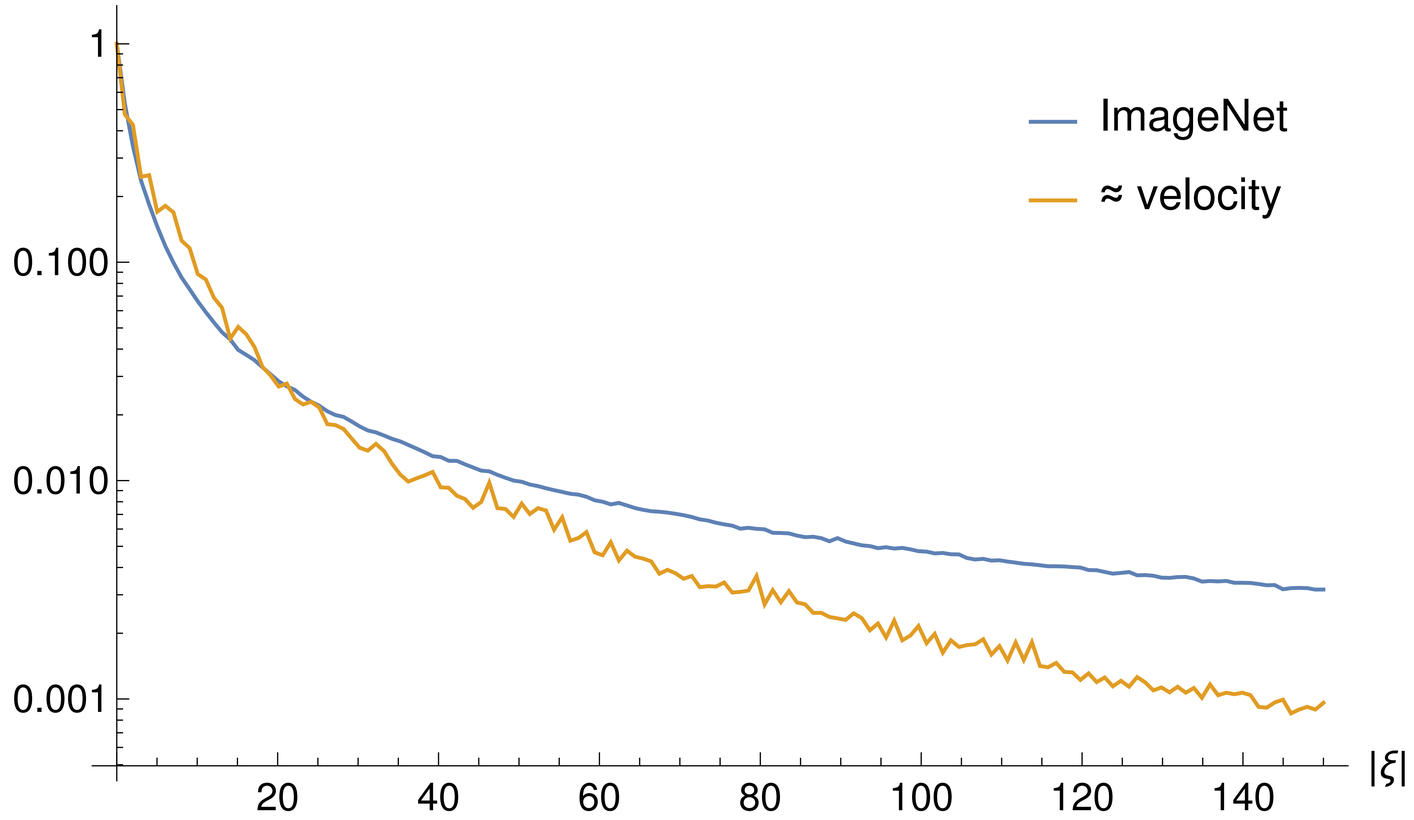
velocity
vector field



vorticity
divergence

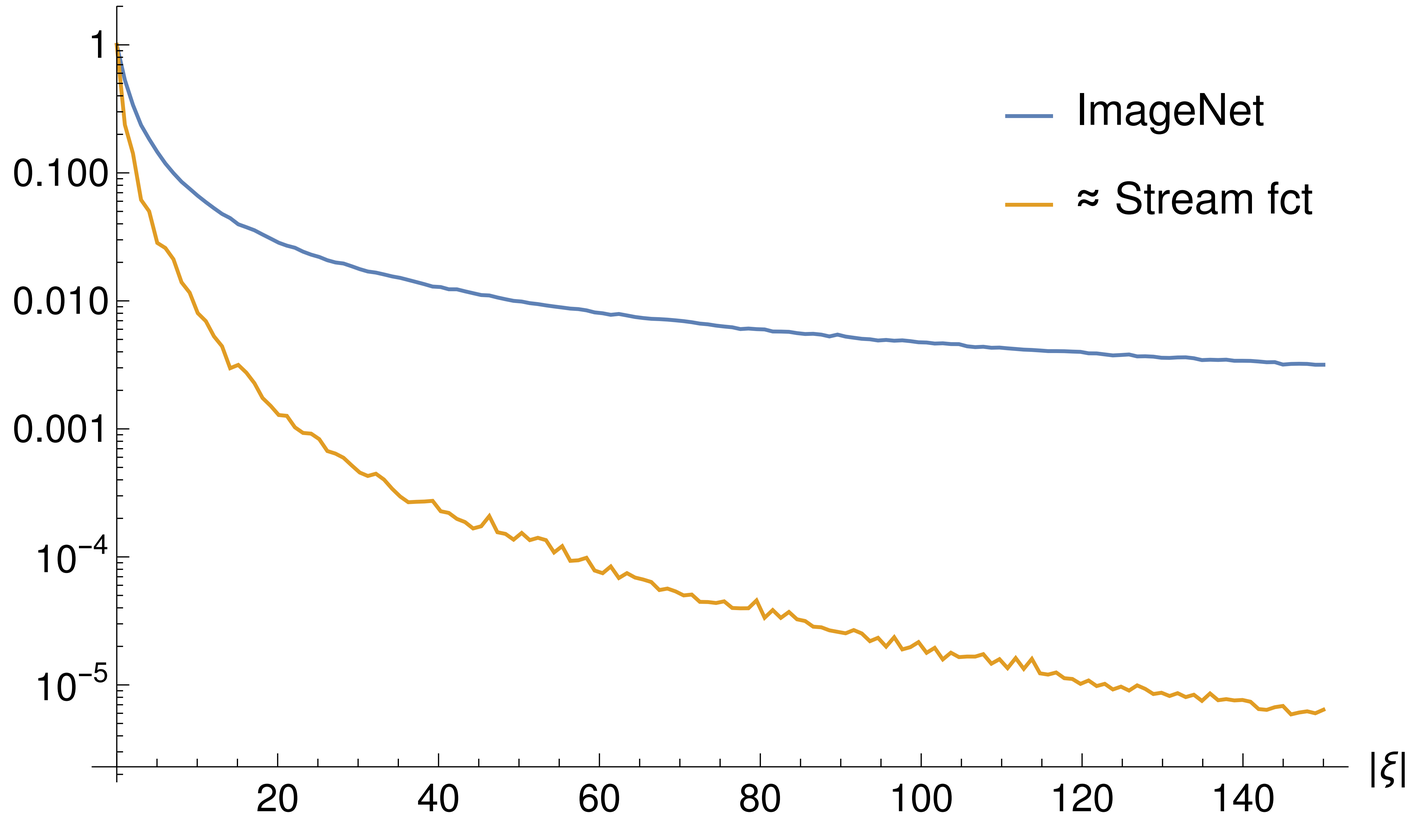
© Christian Lessig, 2023

ERA5 versus ImageNet



© Christian Lessig, 2023

ERA5 versus ImageNet



© Christian Lessig, 2023

ERA5 versus ImageNet

stream function
velocity potential



velocity
vector field



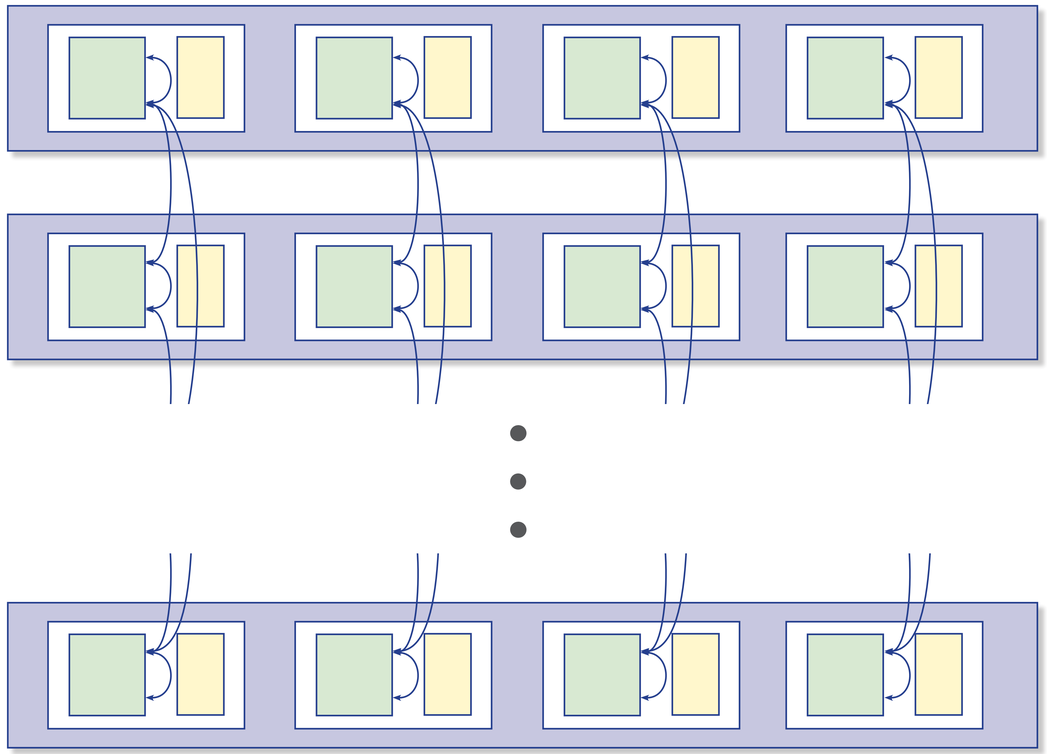
vorticity
divergence

Work with velocity with norm that emphasizes small scale detail

© Christian Lessig, 2023

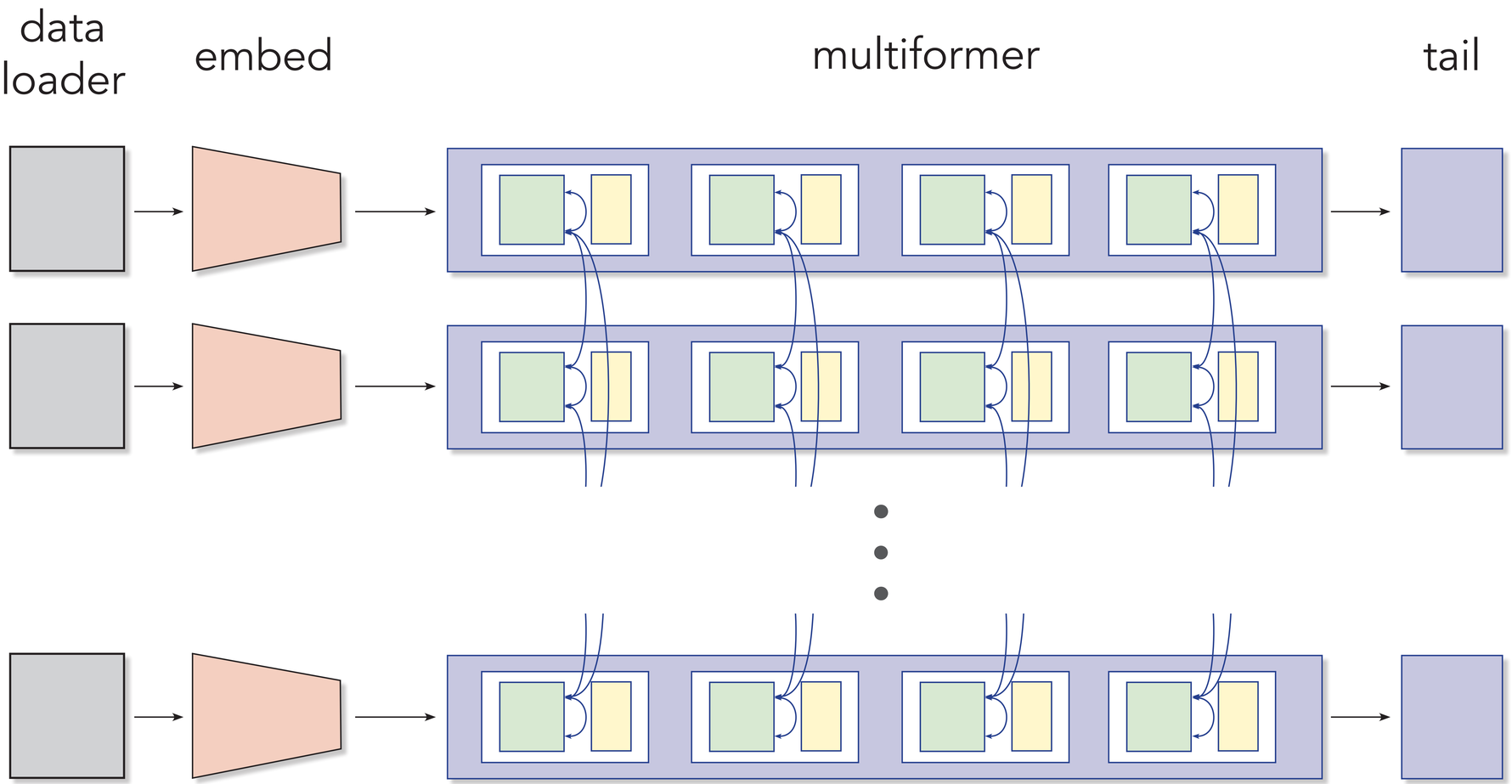
Embedding of tokens

multiformer



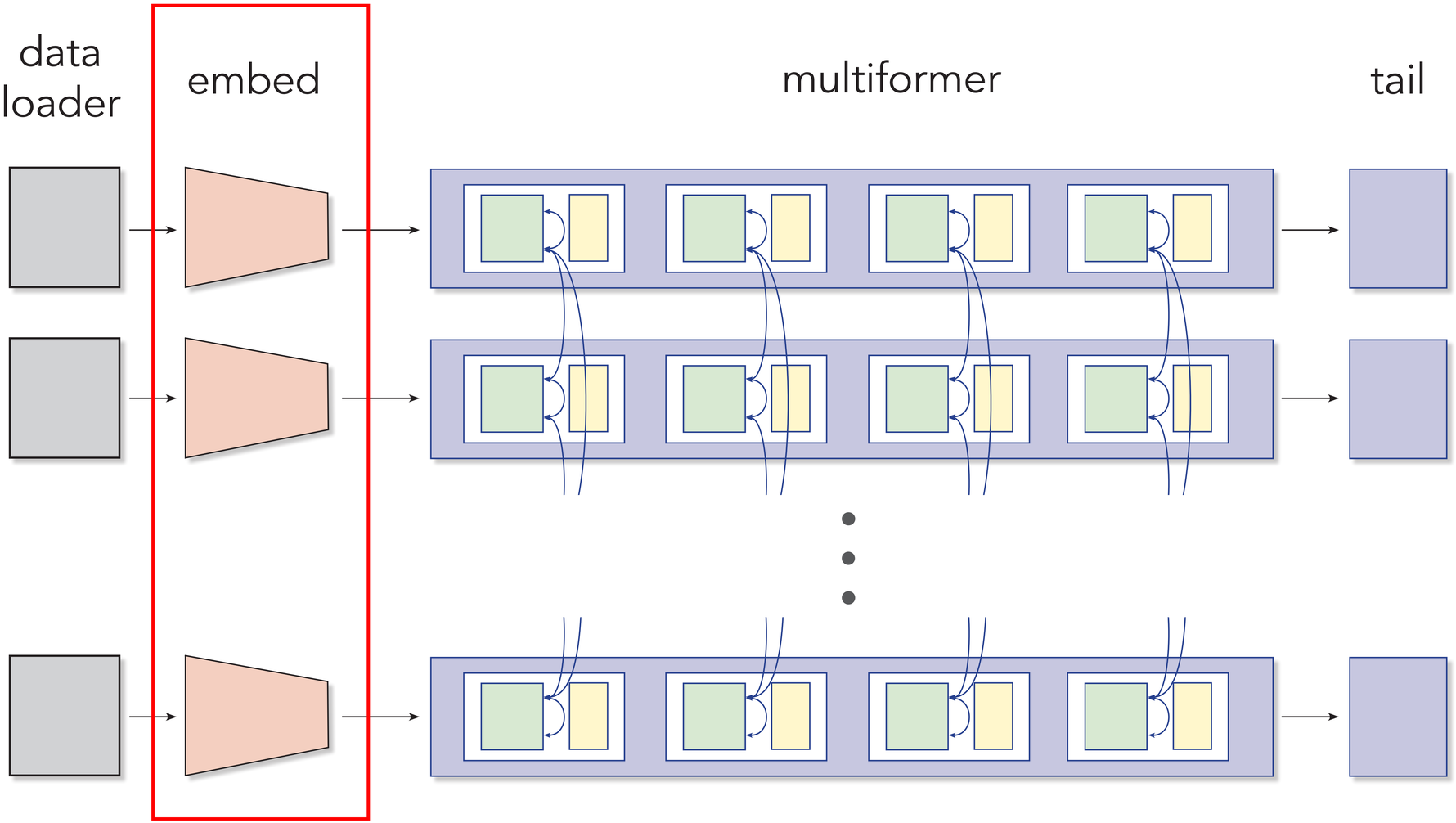
© Christian Lessig, 2023

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

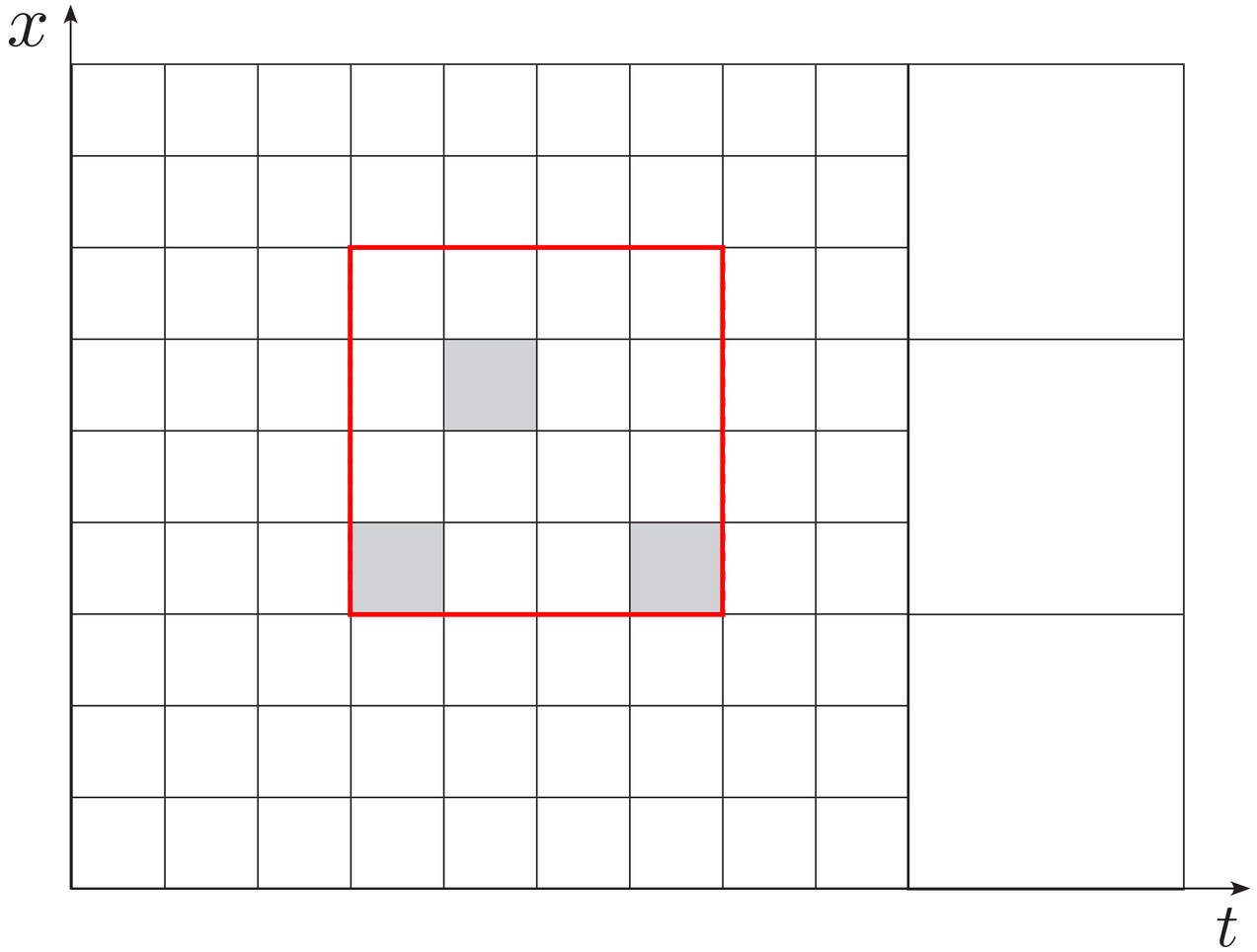
© Christian Lessig, 2023

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

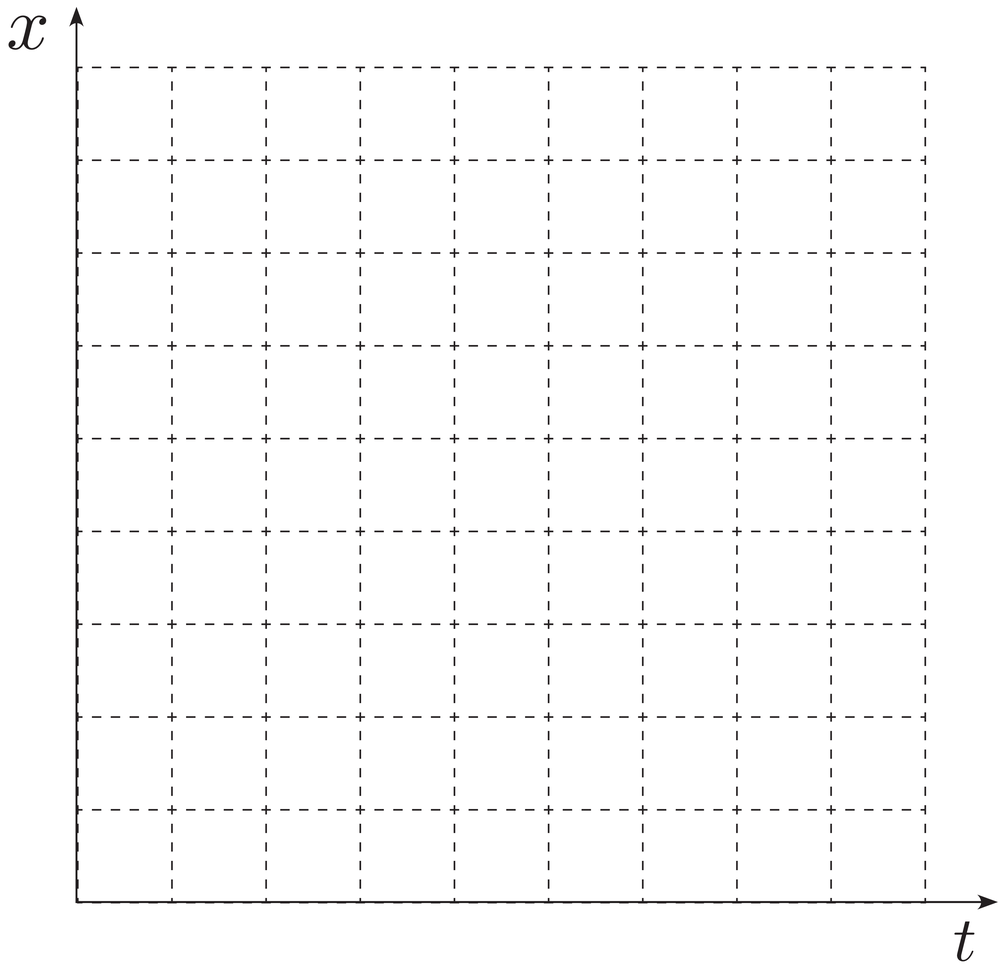
Embedding of tokens

- 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

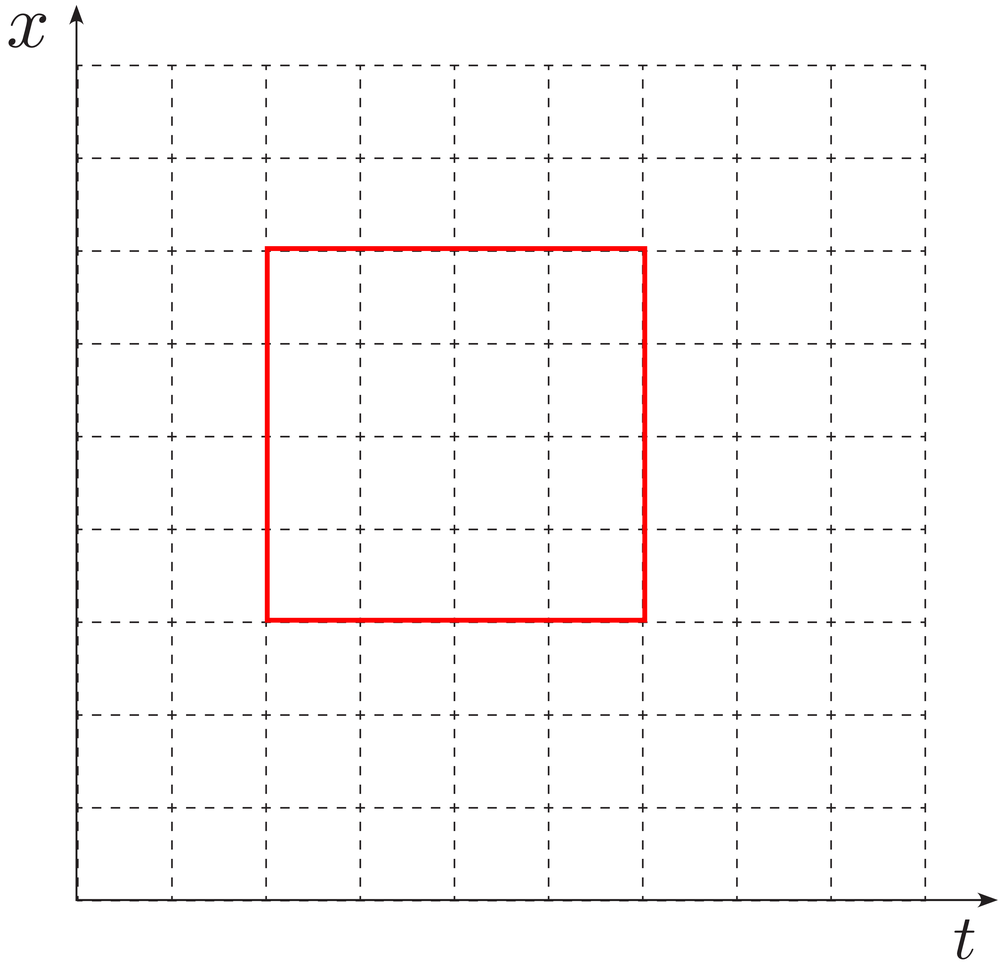
© Christian Lessig, 2023

Embedding of tokens



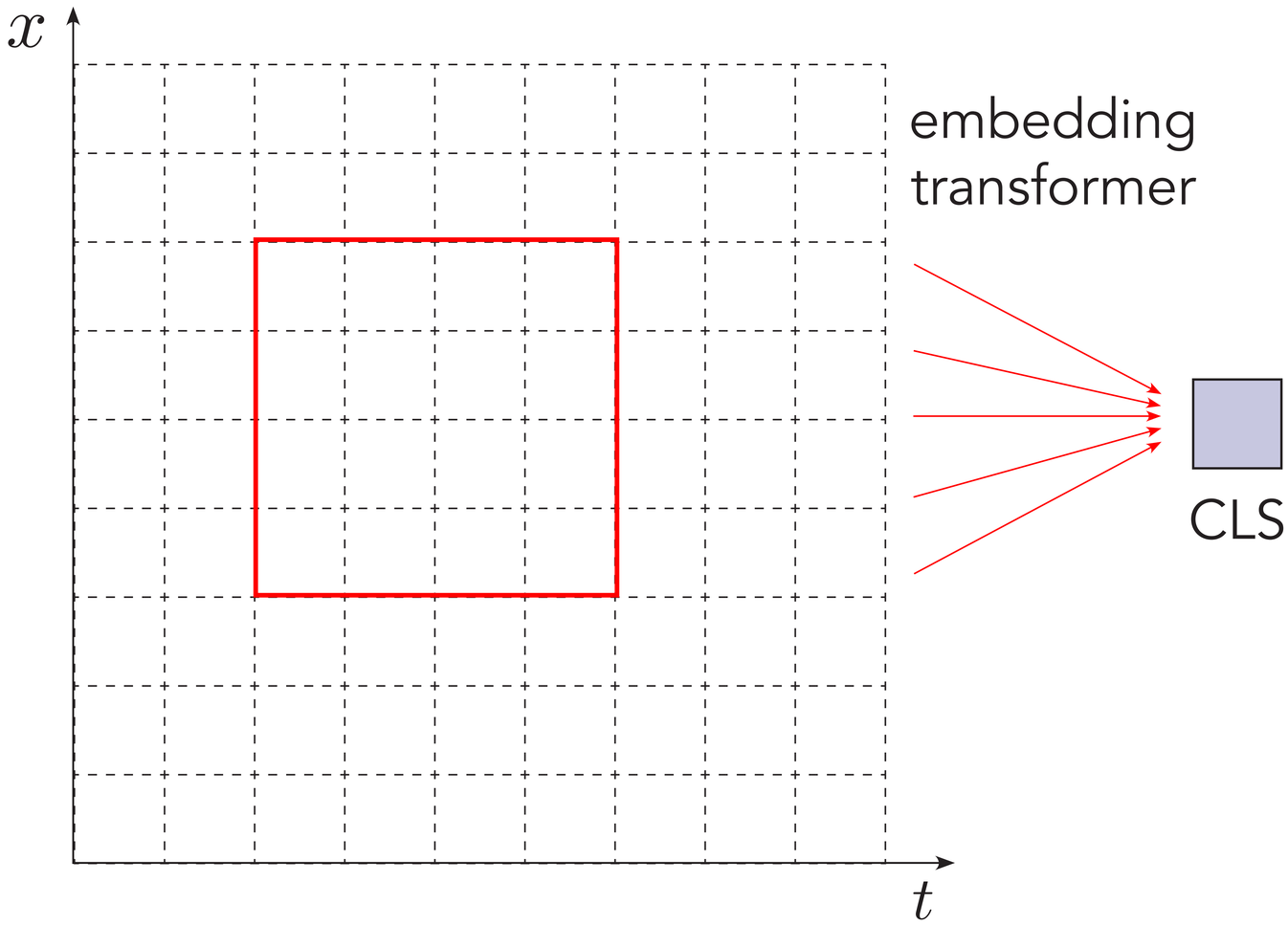
© Christian Lessig, 2023

Embedding of tokens



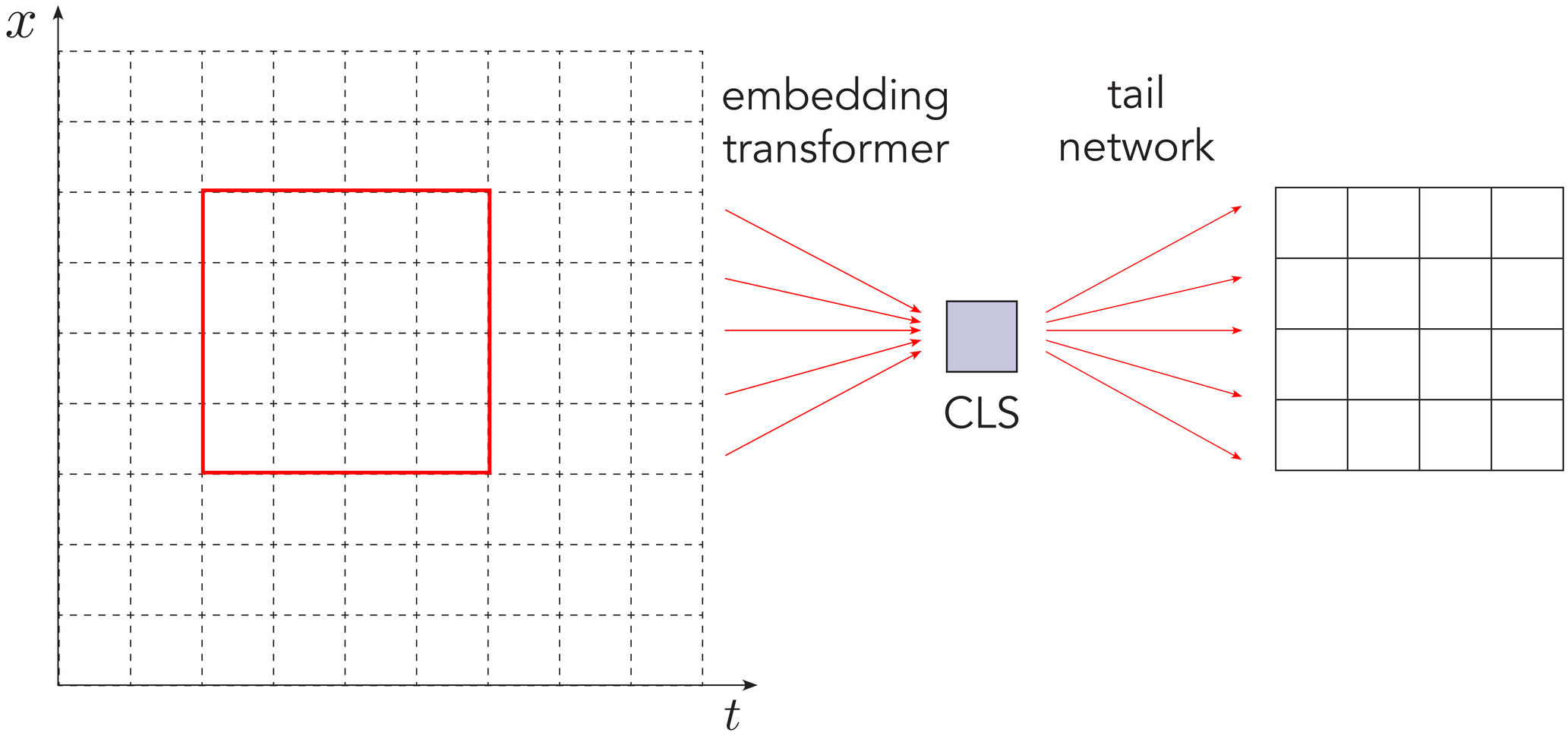
© Christian Lessig, 2023

Embedding of tokens



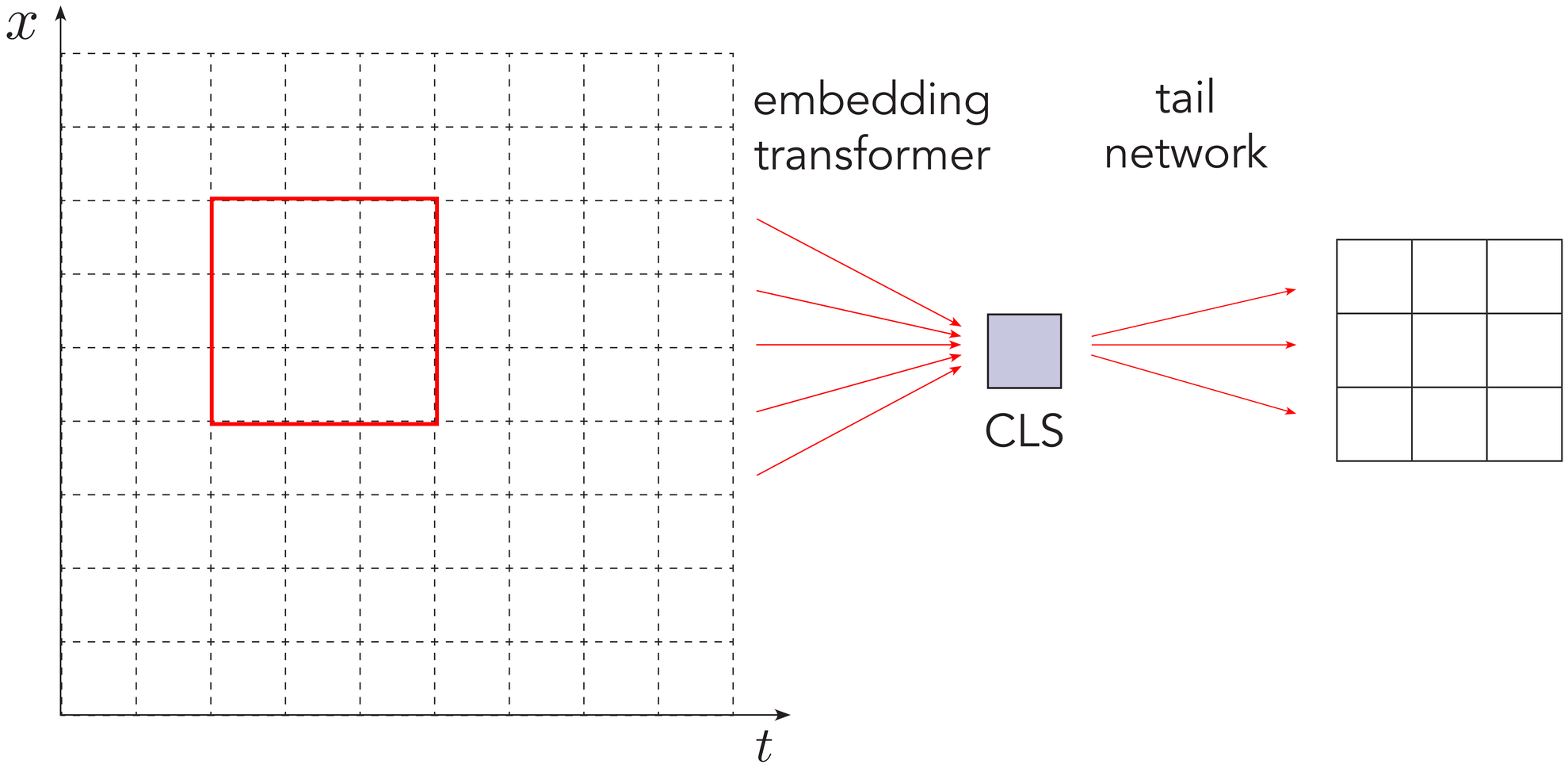
© Christian Lessig, 2023

Embedding of tokens



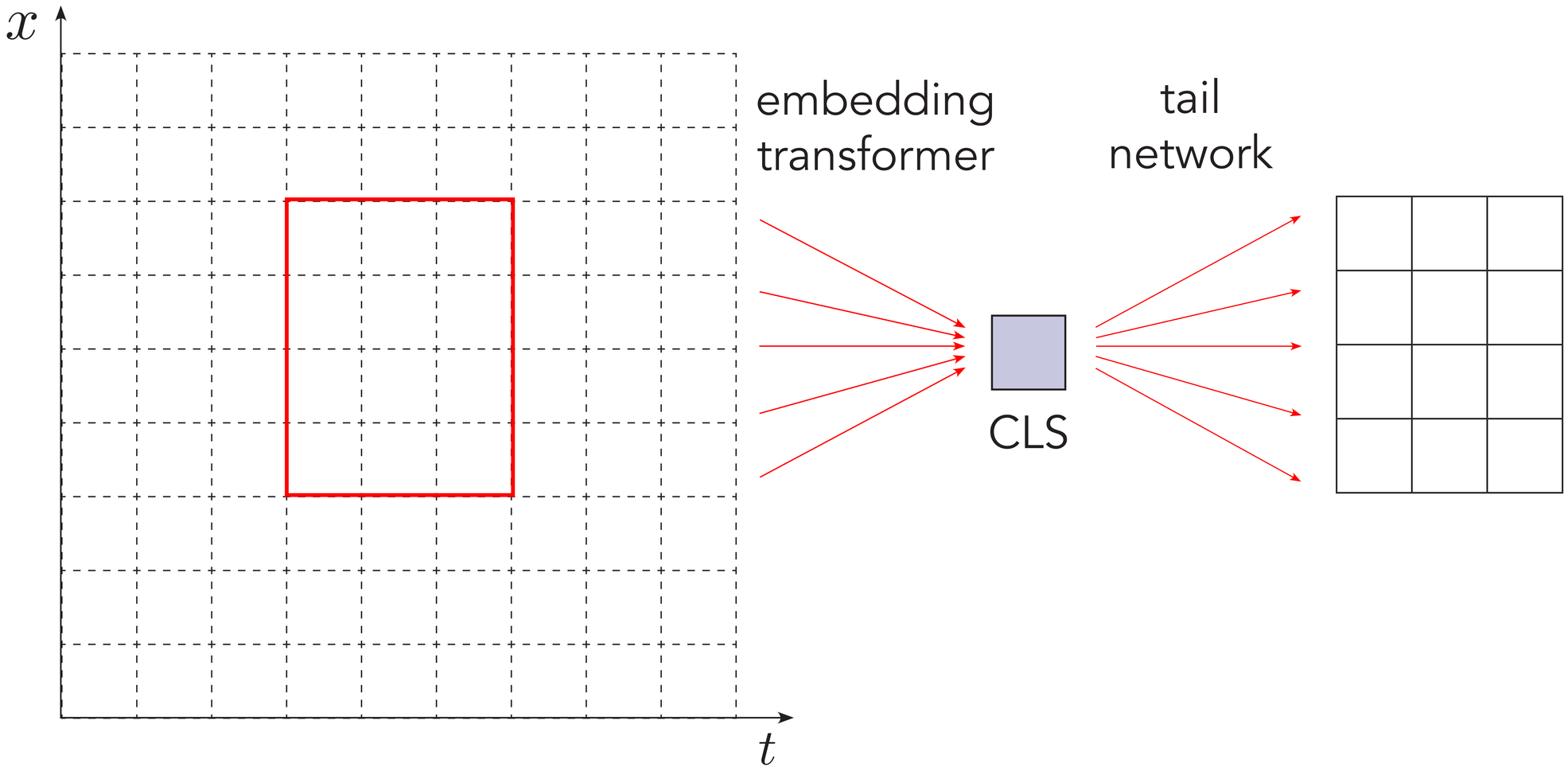
© Christian Lessig, 2023

Embedding of tokens



© Christian Lessig, 2023

Embedding of tokens



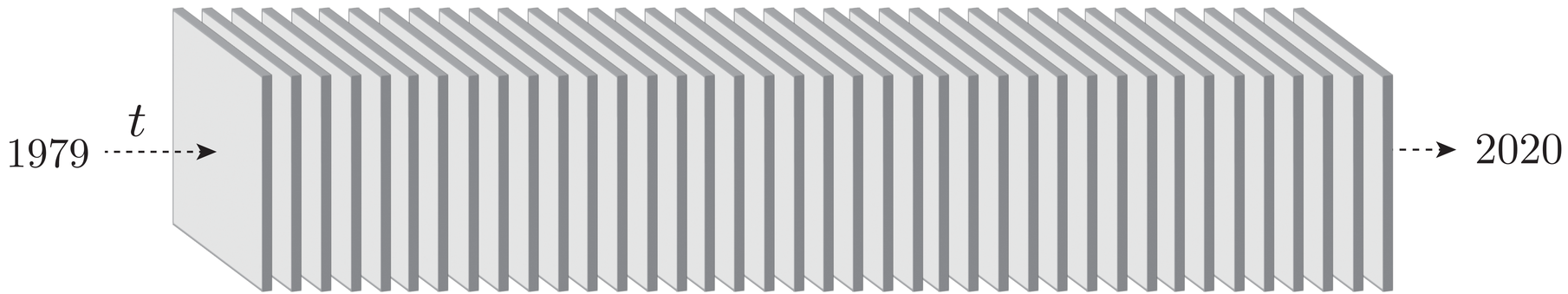
© Christian Lessig, 2023

Training

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

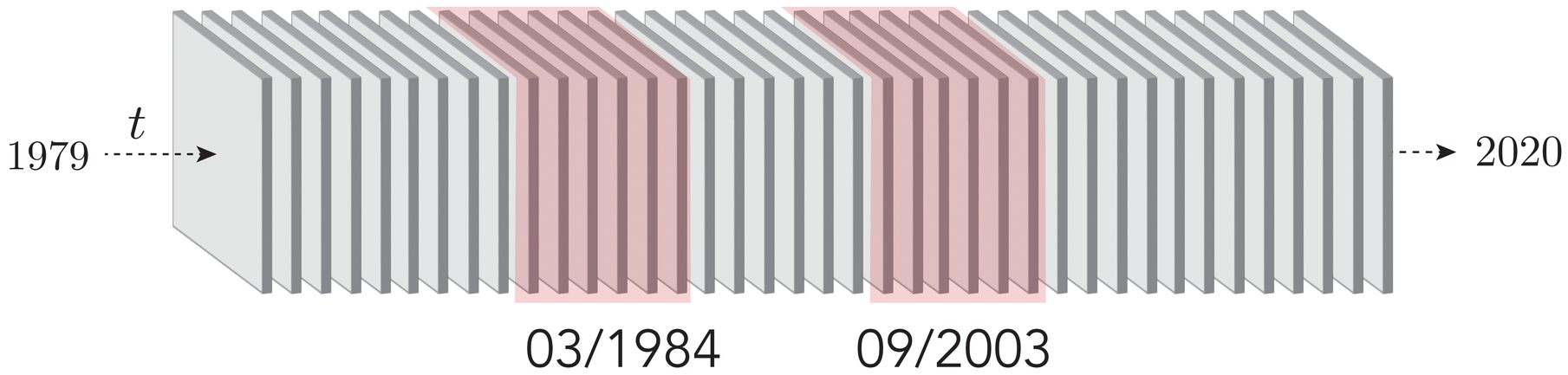
© Christian Lessig, 2023

Training



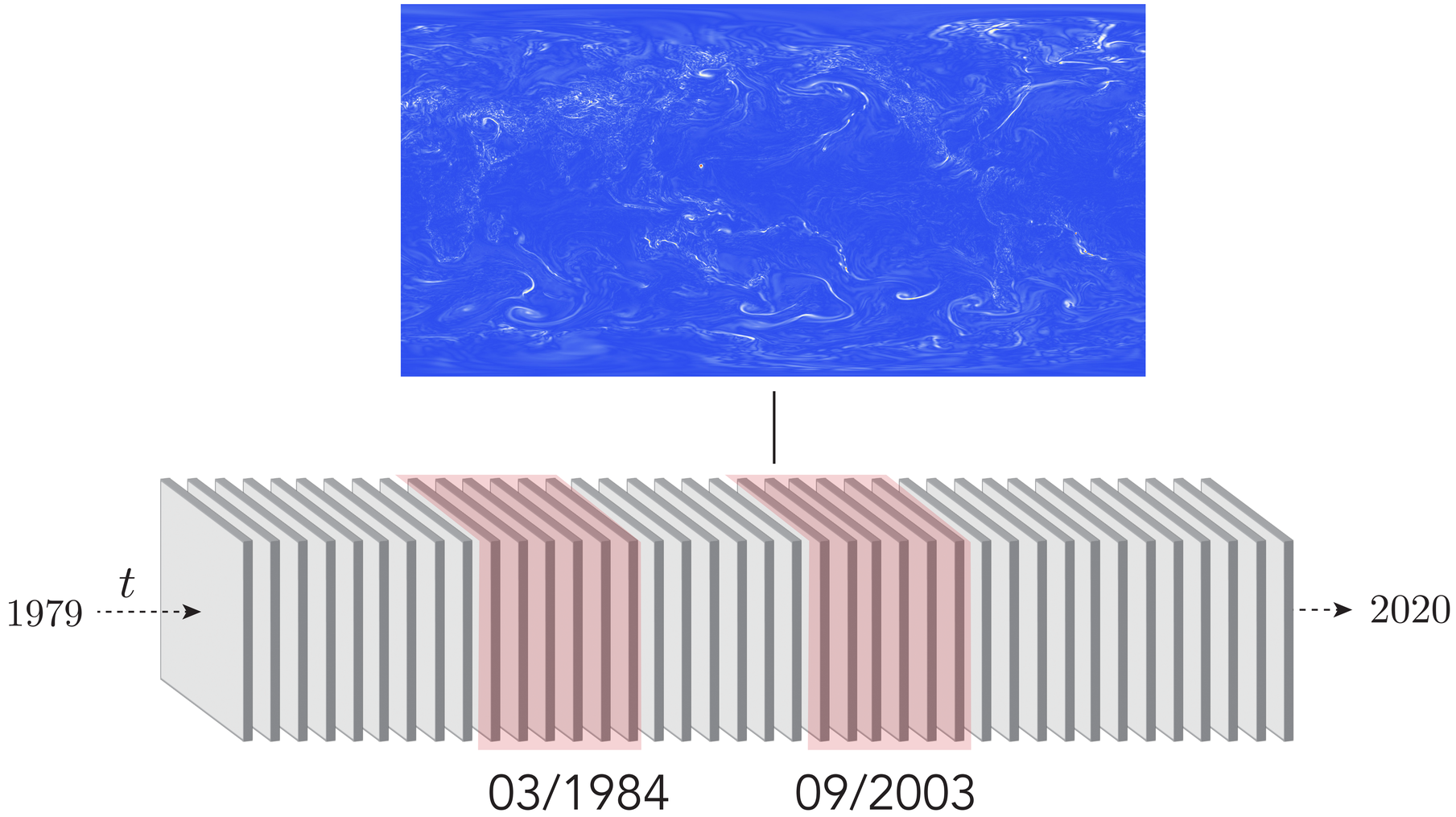
© Christian Lessig, 2023

Training



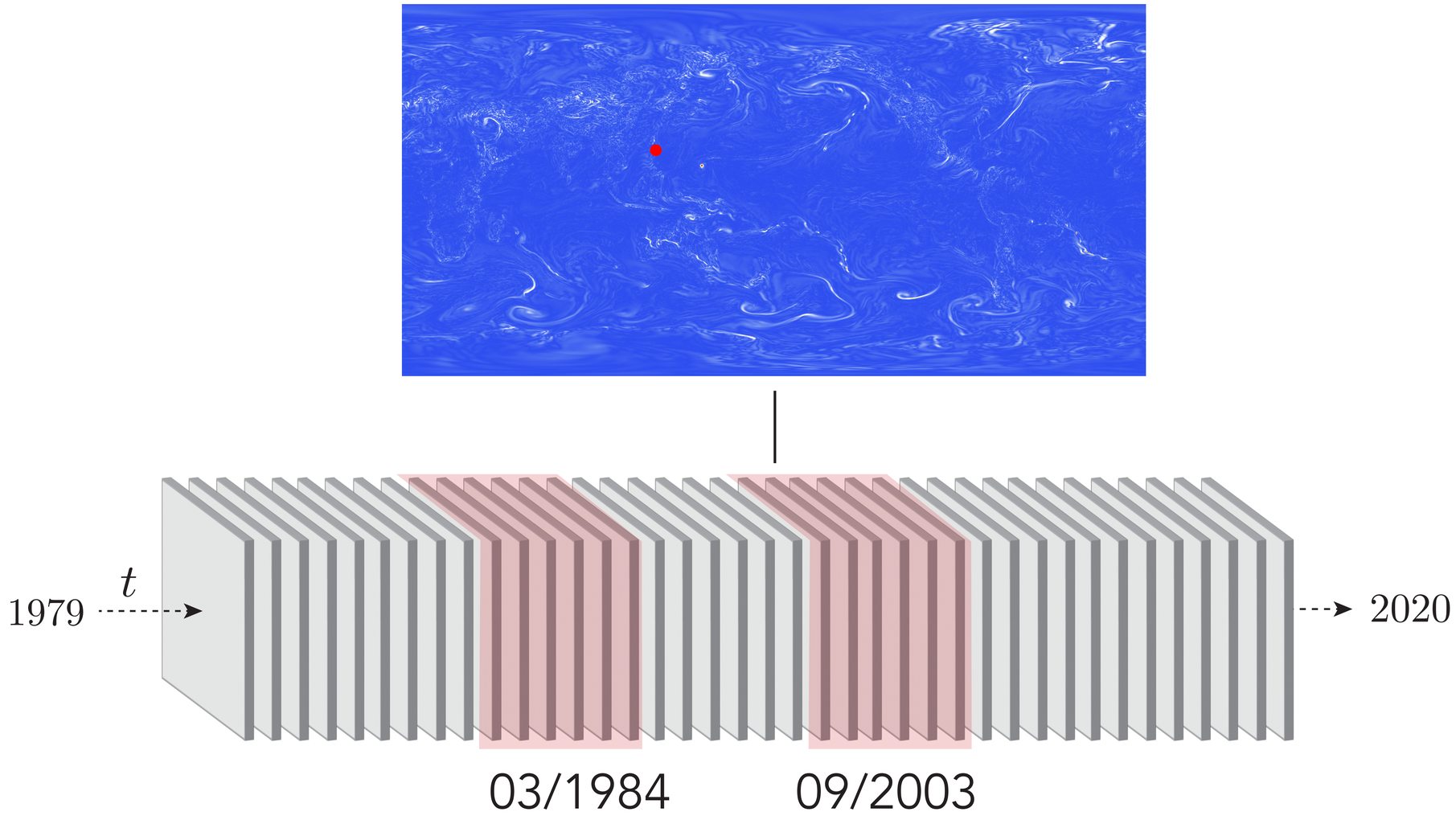
© Christian Lessig, 2023

Training



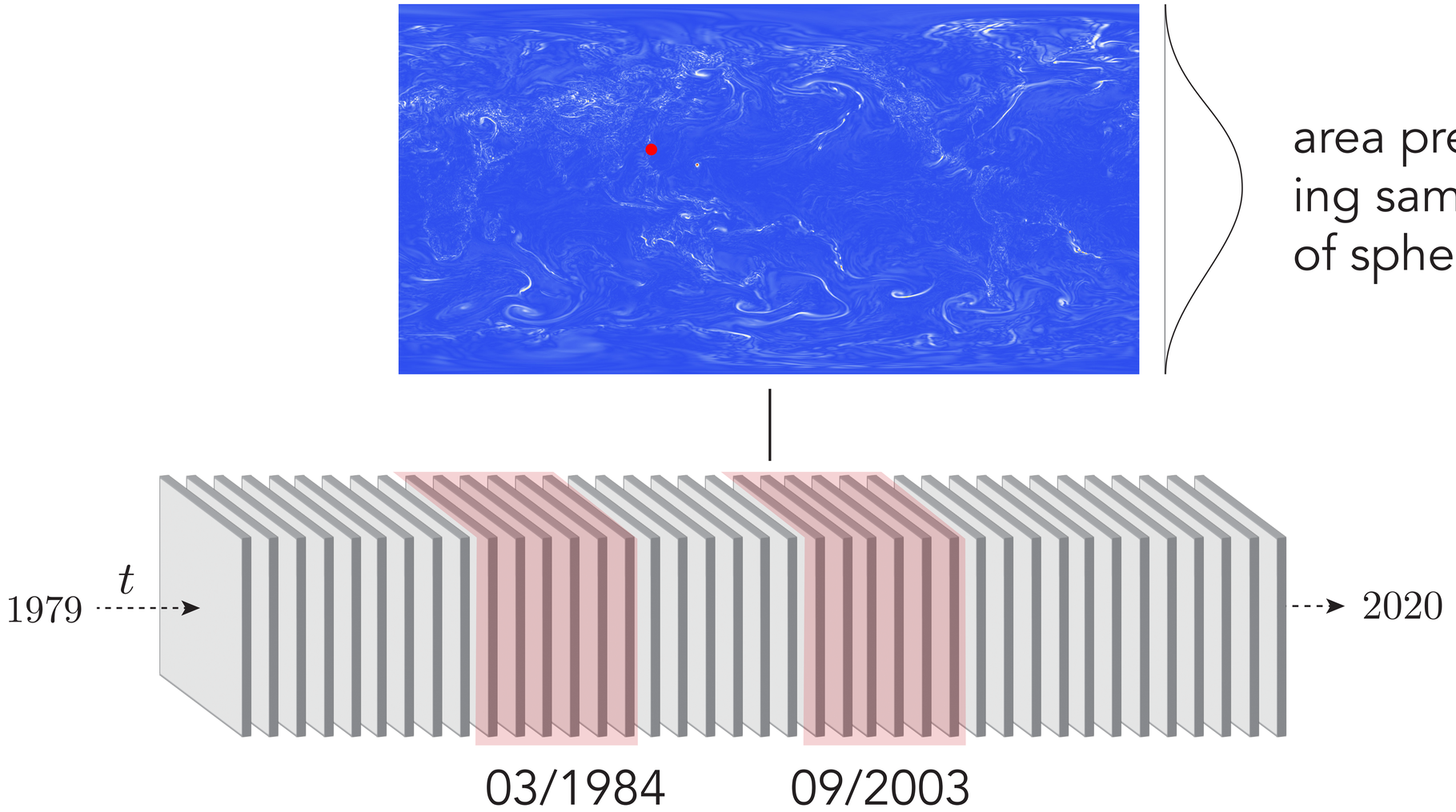
© Christian Lessig, 2023

Training



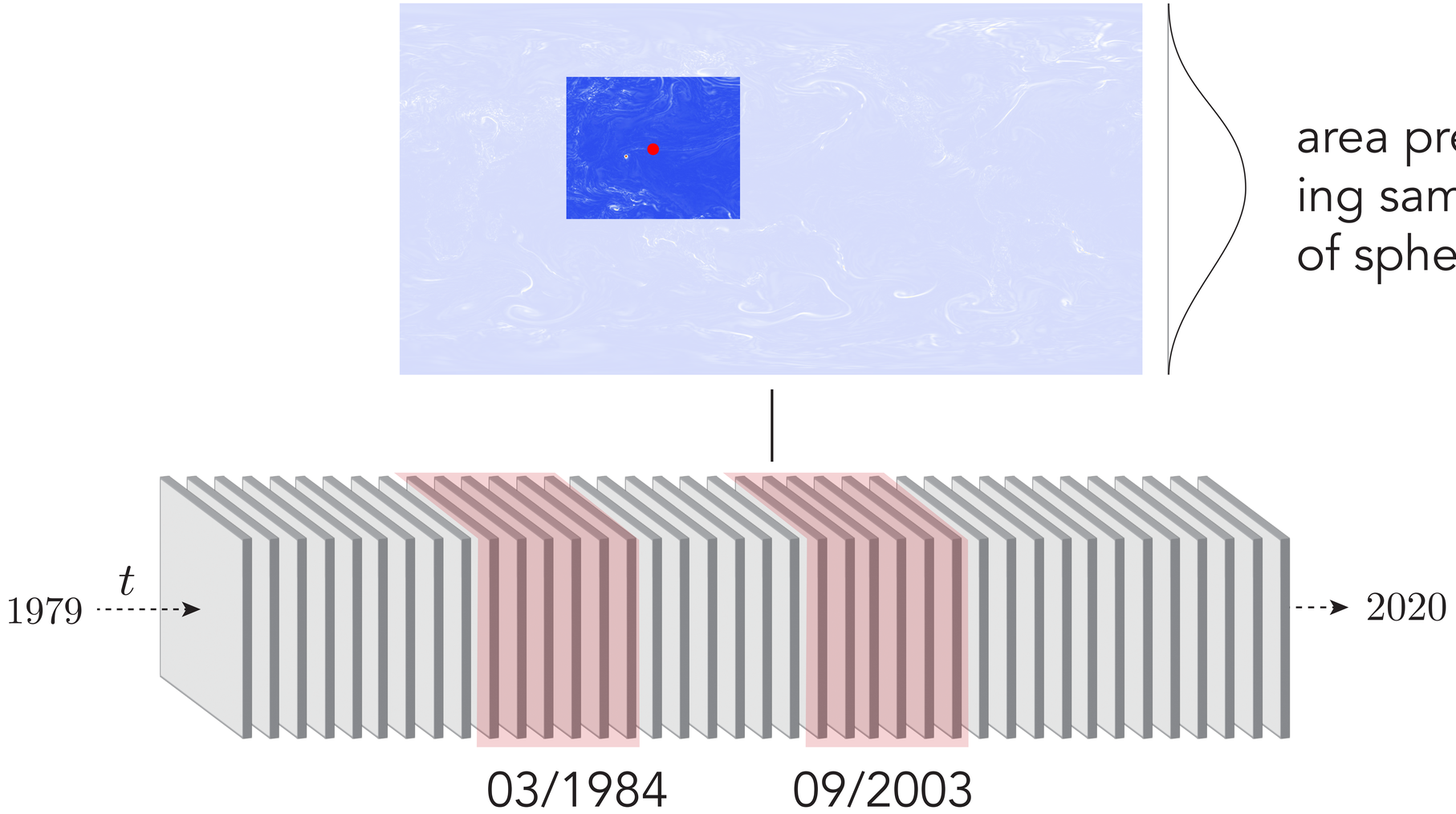
© Christian Lessig, 2023

Training



© Christian Lessig, 2023

Training



© Christian Lessig, 2023

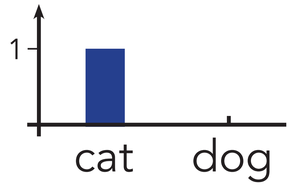
Statistical loss



© Christian Lessig, 2023

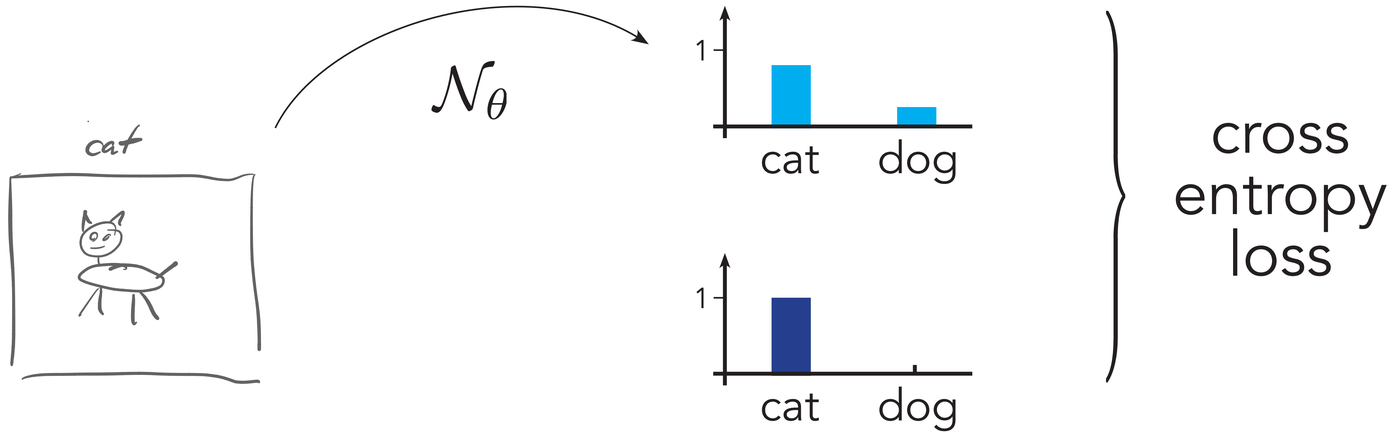
Statistical loss





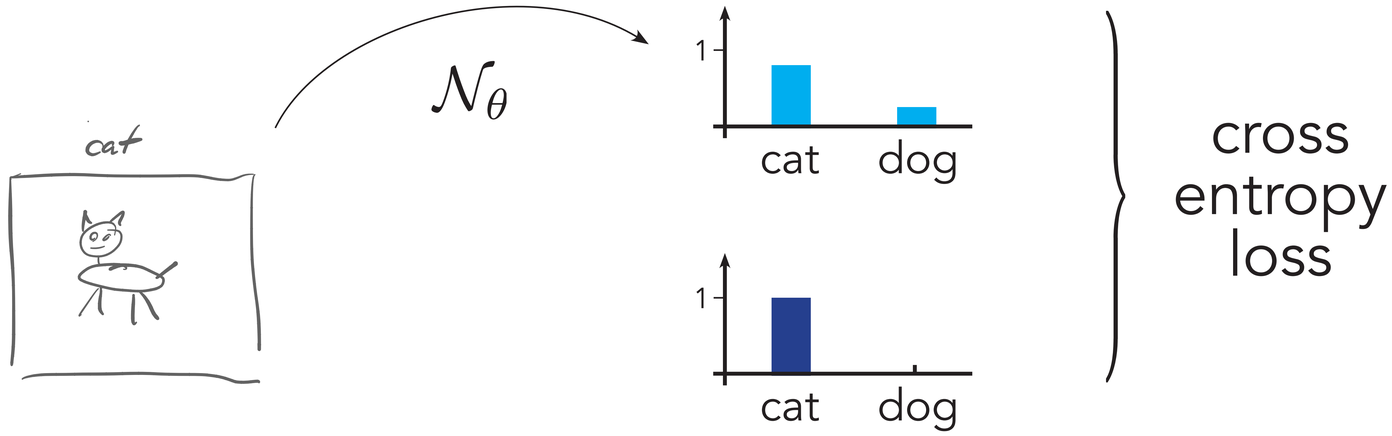
© Christian Lessig, 2023

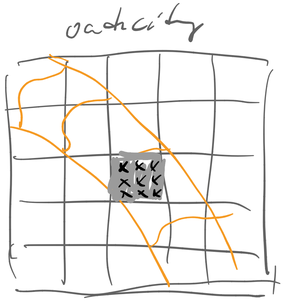
Statistical loss



© Christian Lessig, 2023

Statistical loss



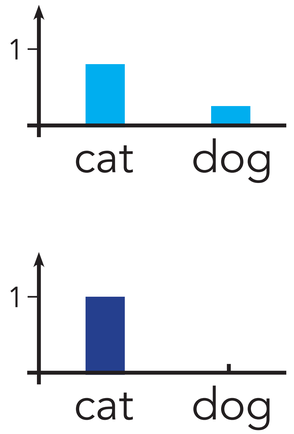


© Christian Lessig, 2023

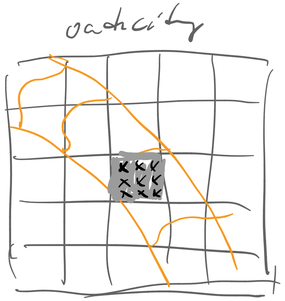
Statistical loss



$\mathcal{N}_\theta$



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:

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

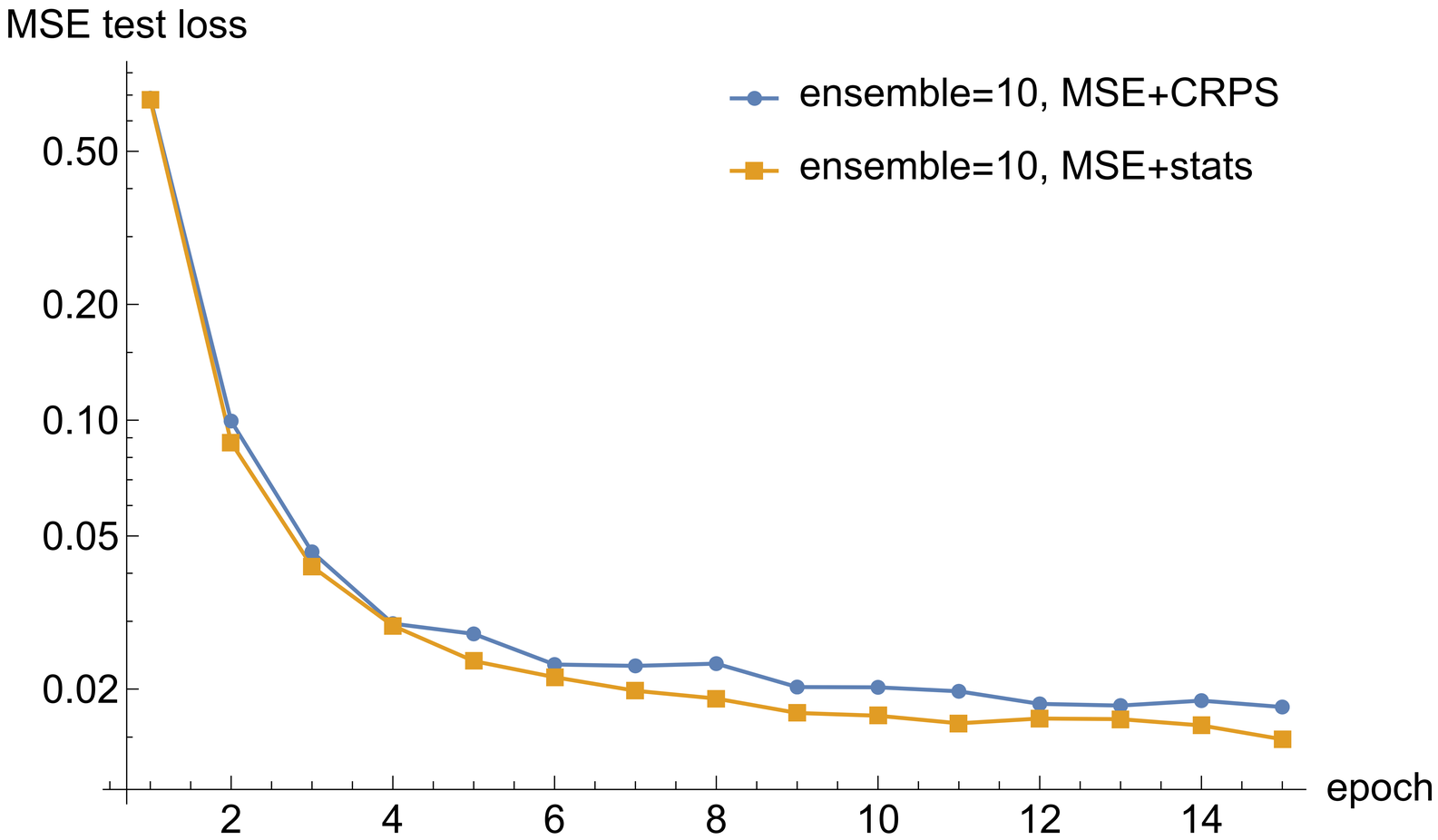
- CRPS:[1]

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

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

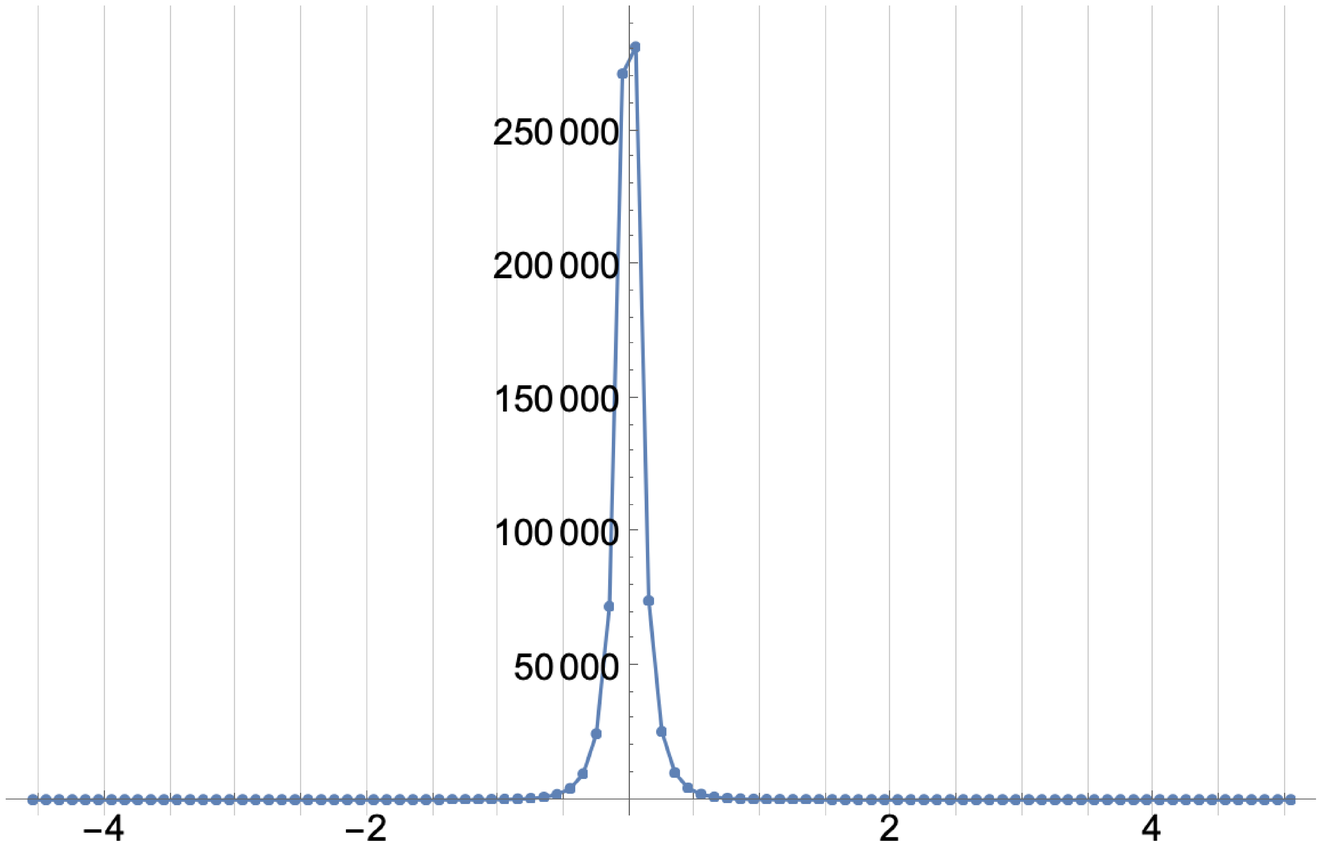
© Christian Lessig, 2023

Statistical loss



© Christian Lessig, 2023

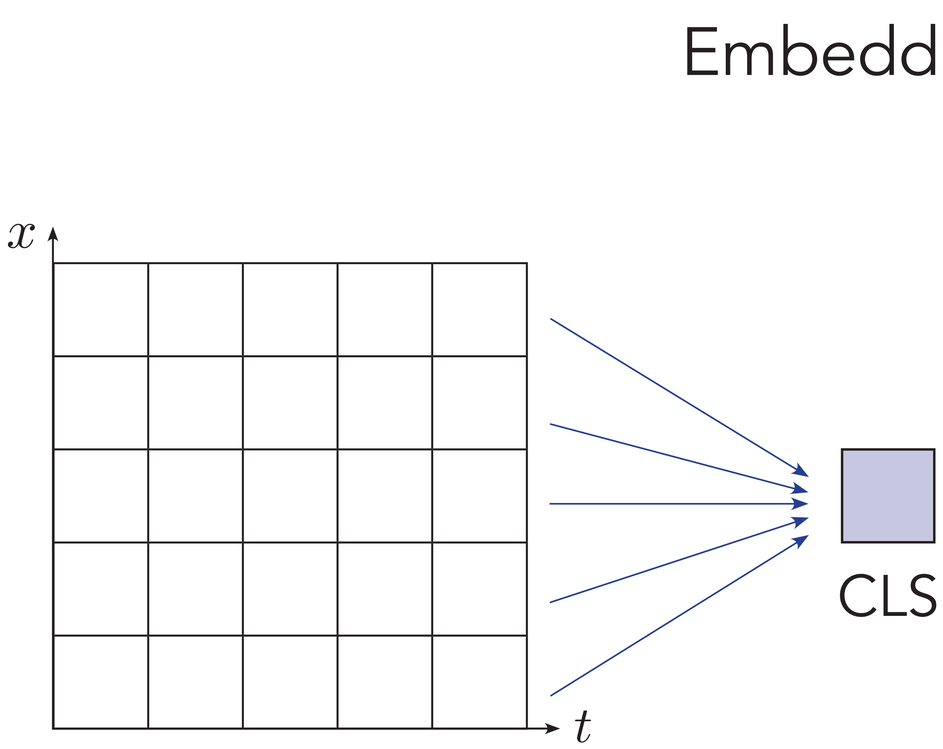
Statistical loss



Histogram of ensemble errors

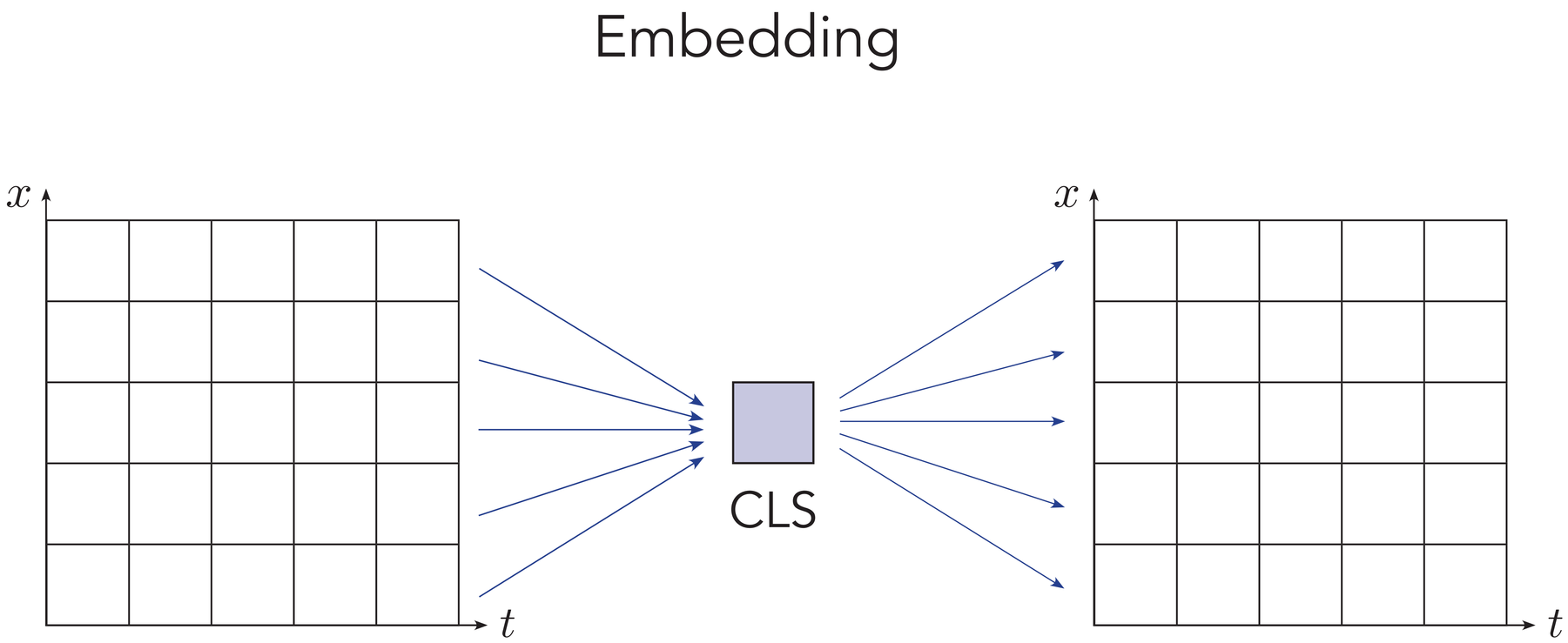
© Christian Lessig, 2023

Zero shot evaluation



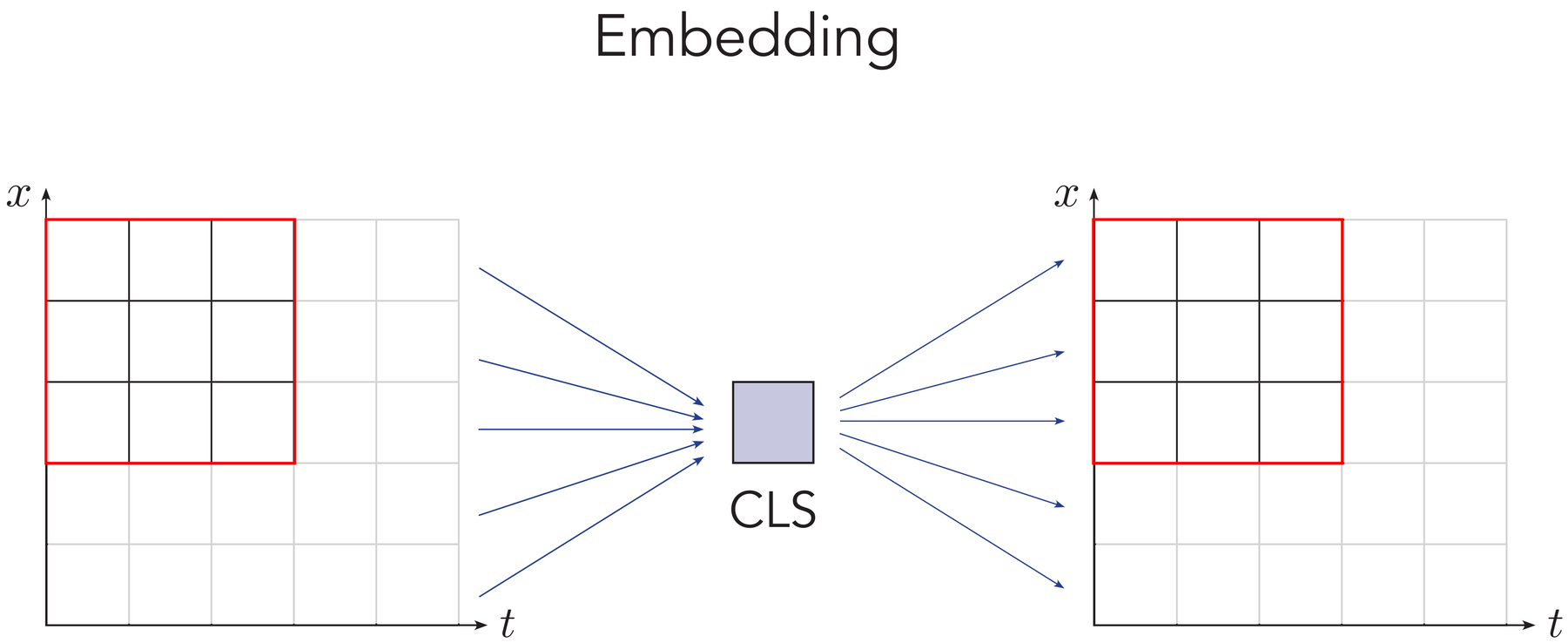
© Christian Lessig, 2023

Zero shot evaluation



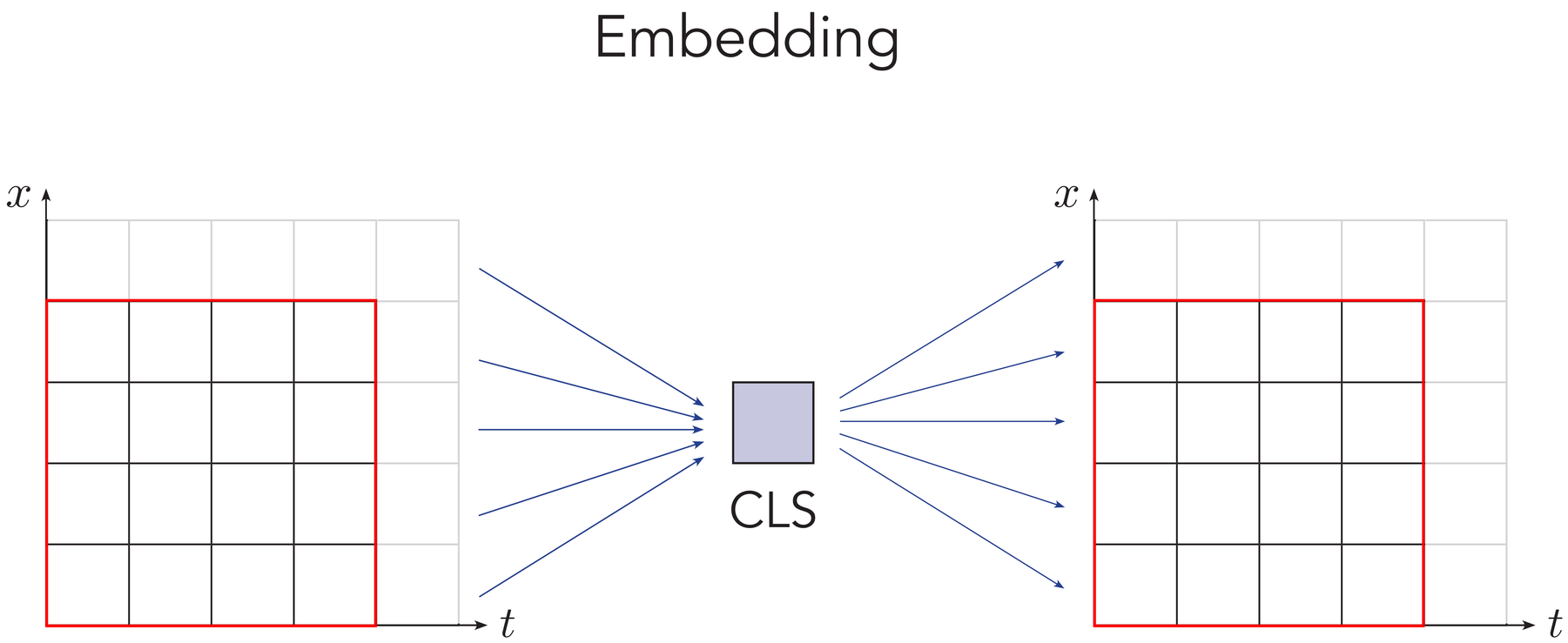
© Christian Lessig, 2023

Zero shot evaluation



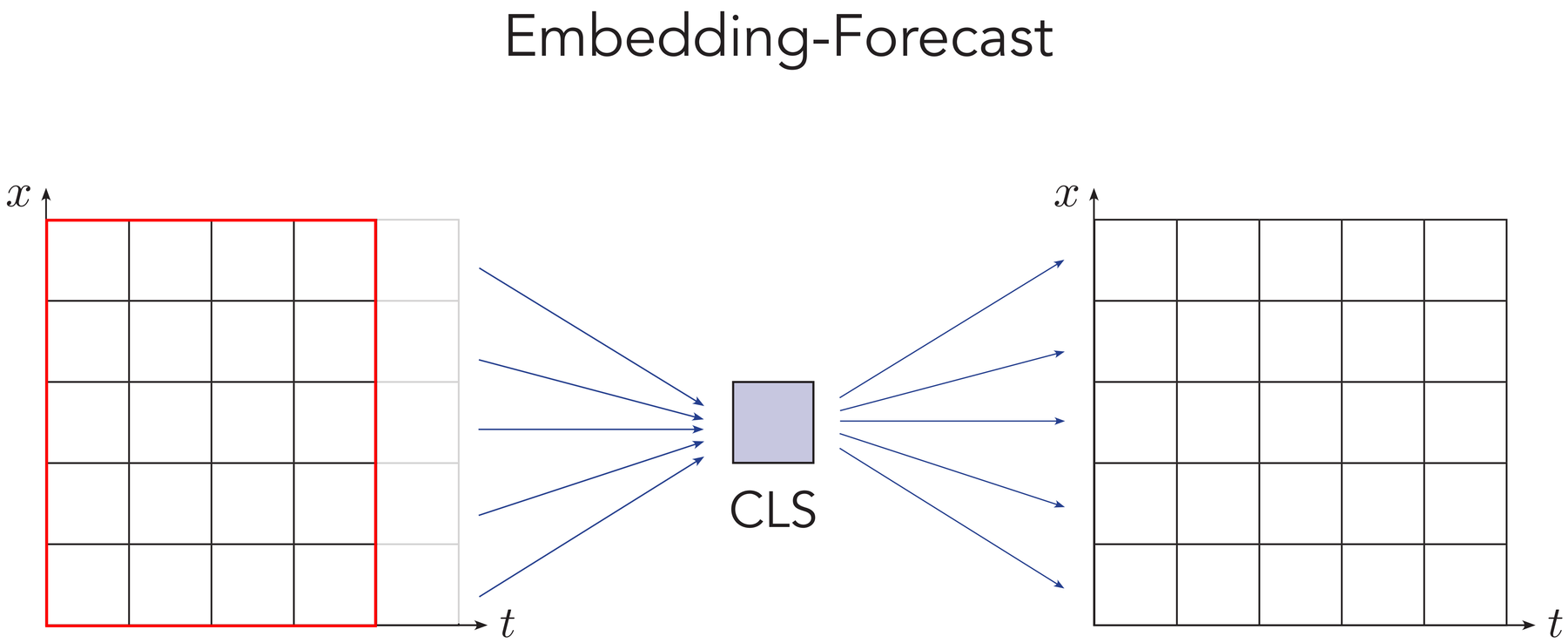
© Christian Lessig, 2023

Zero shot evaluation



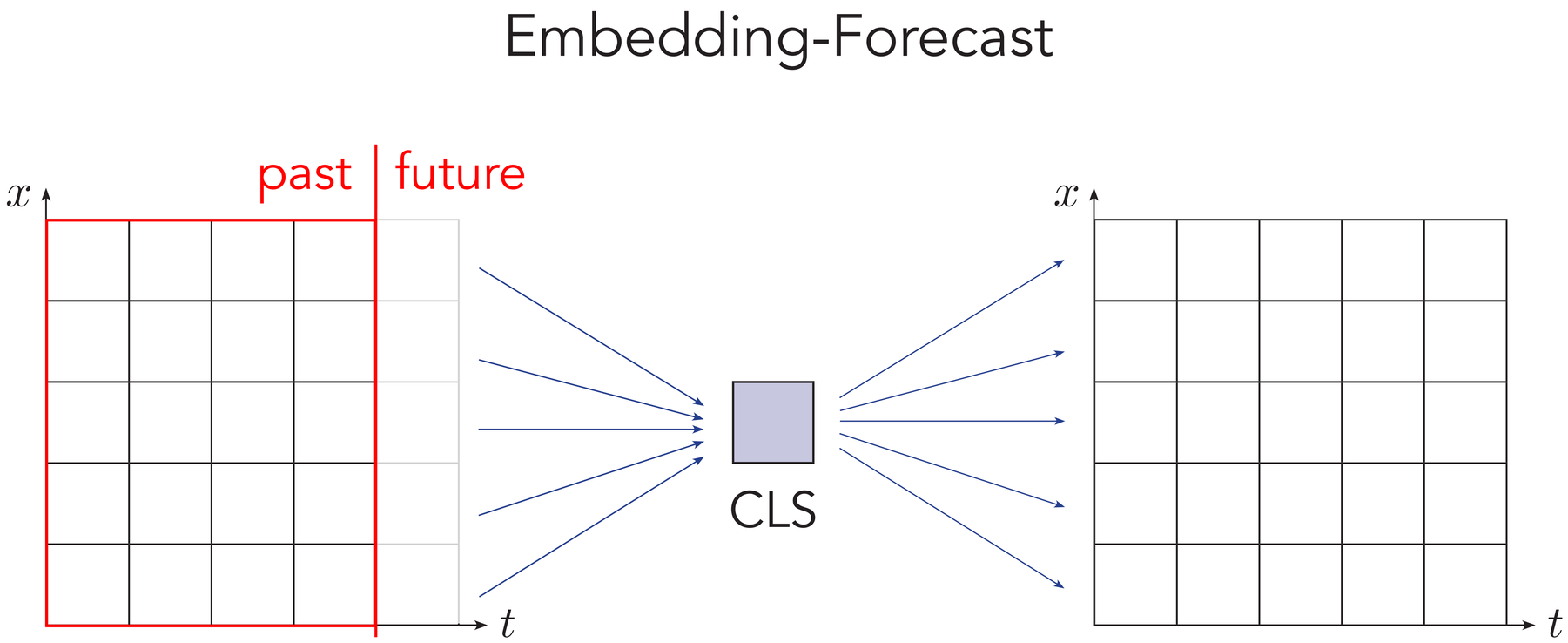
© Christian Lessig, 2023

Zero shot evaluation



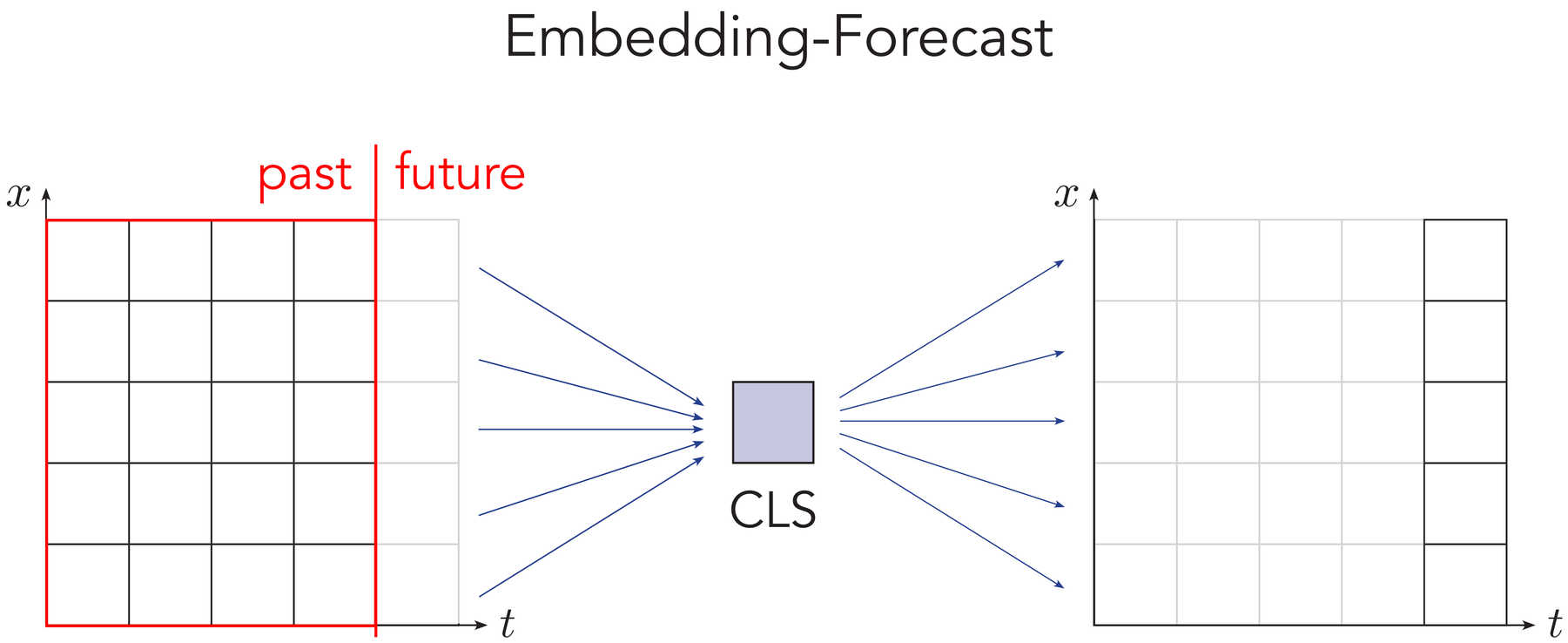
© Christian Lessig, 2023

Zero shot evaluation



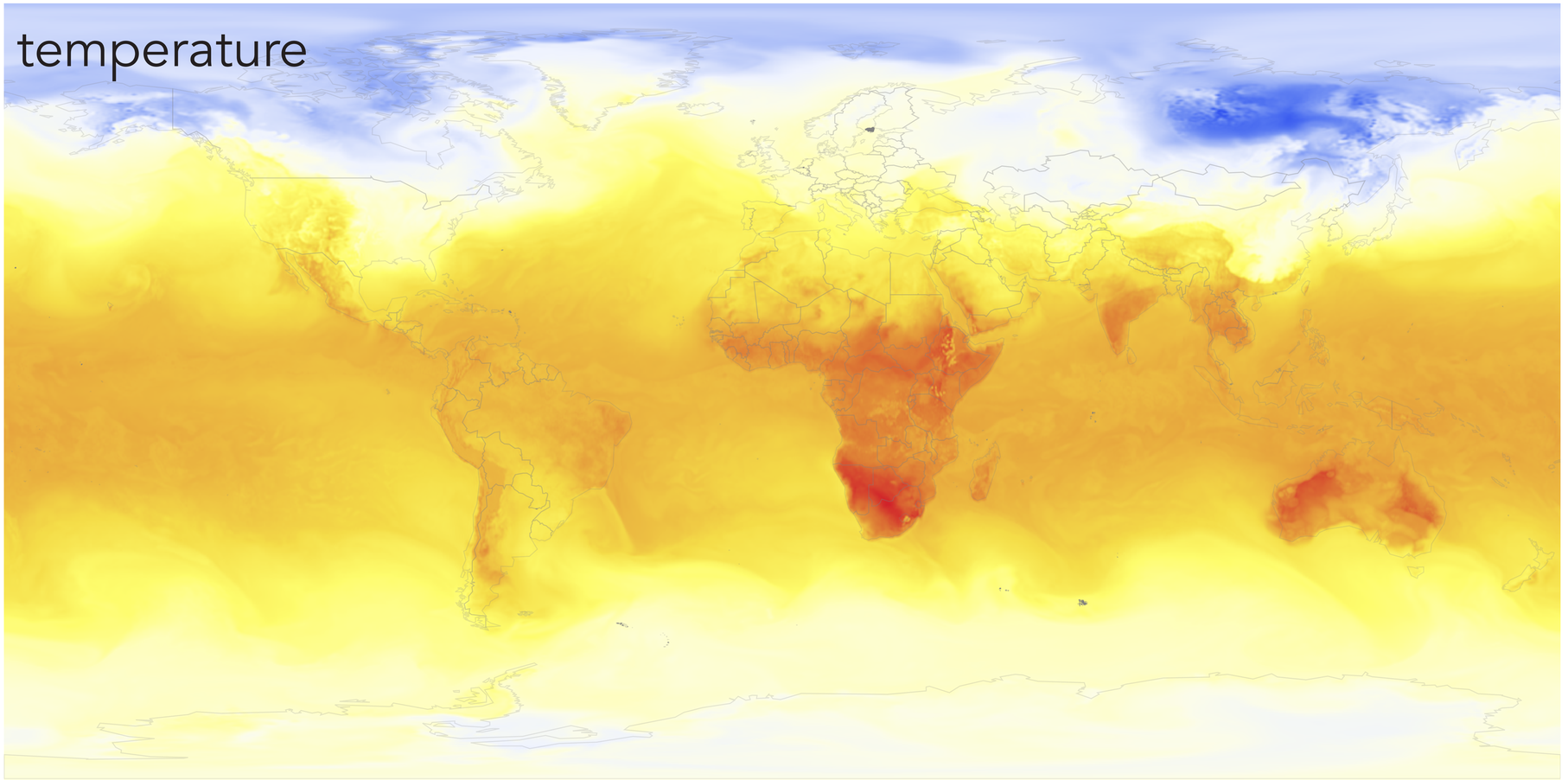
© Christian Lessig, 2023

Zero shot evaluation



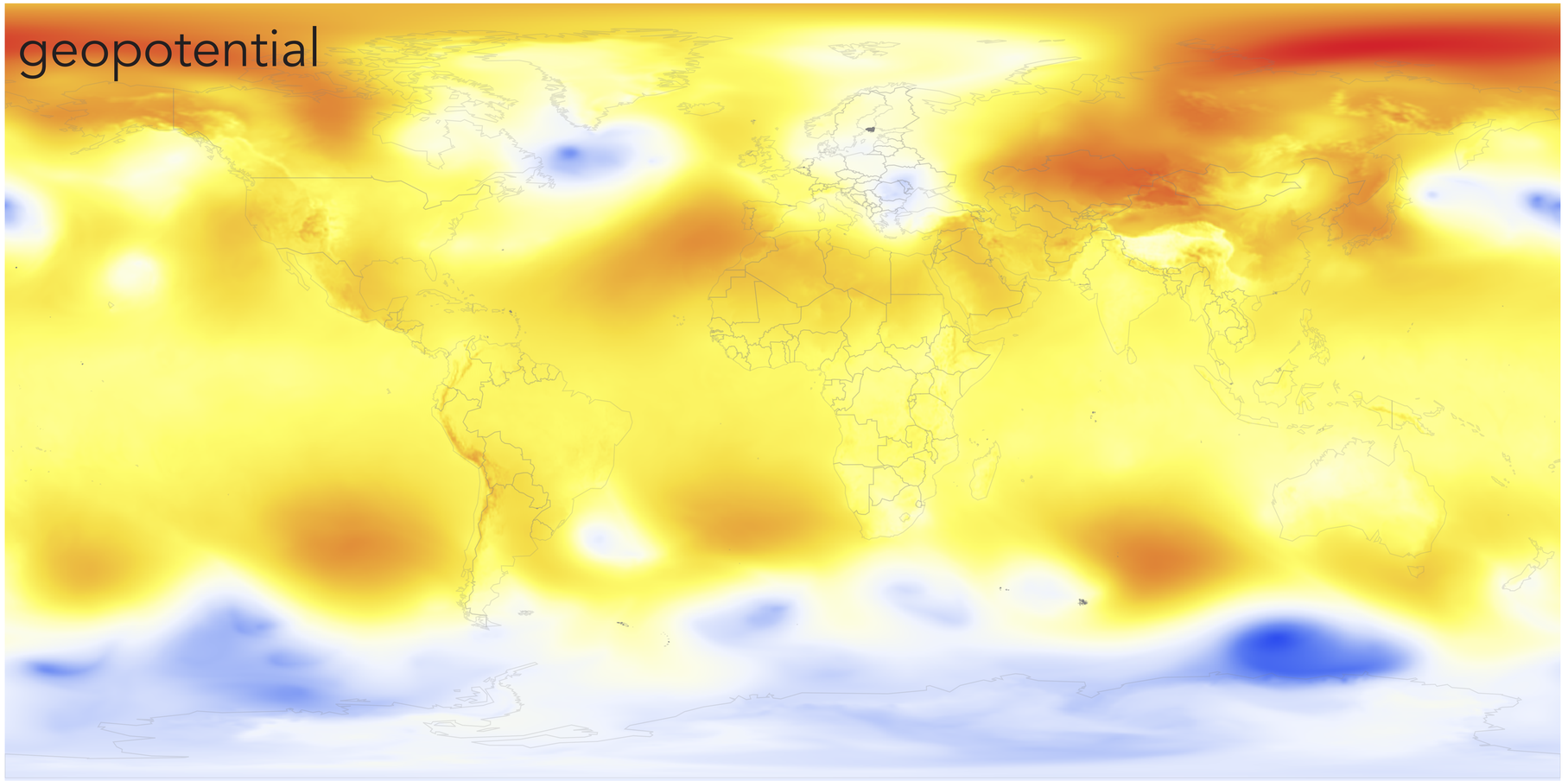
© Christian Lessig, 2023

AtmoRep data



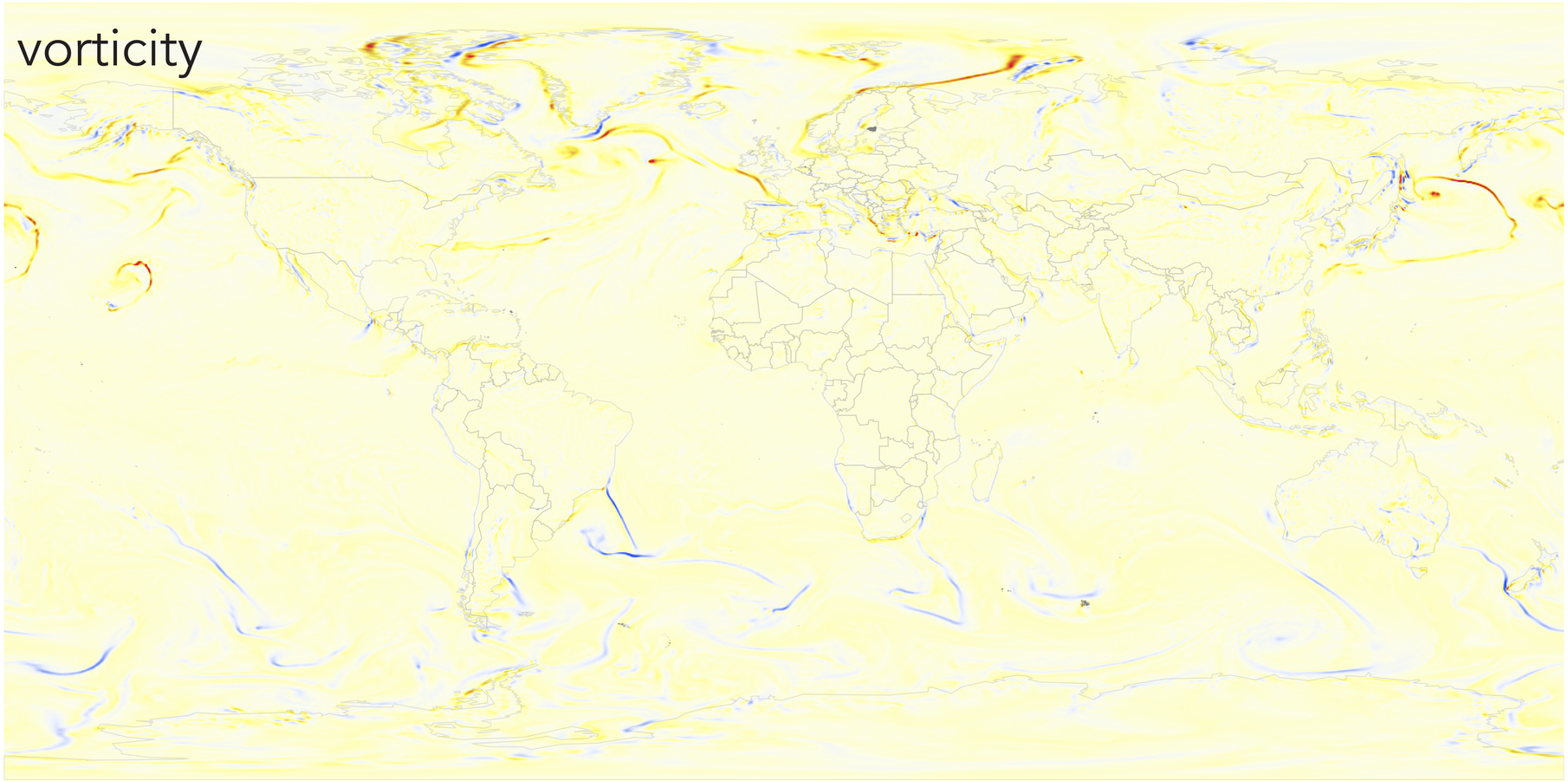
© Christian Lessig, 2023

AtmoRep data



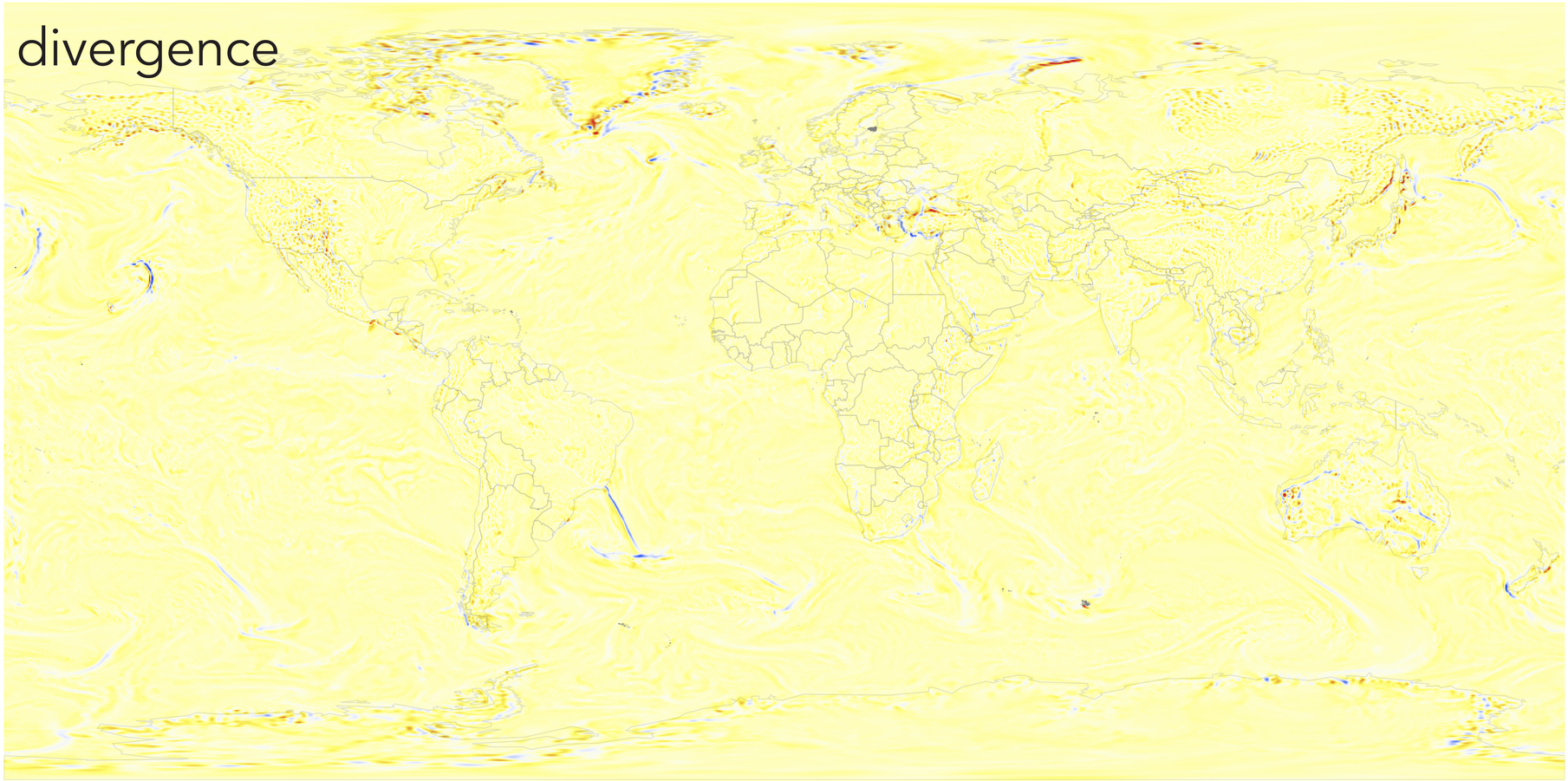
© Christian Lessig, 2023

AtmoRep data



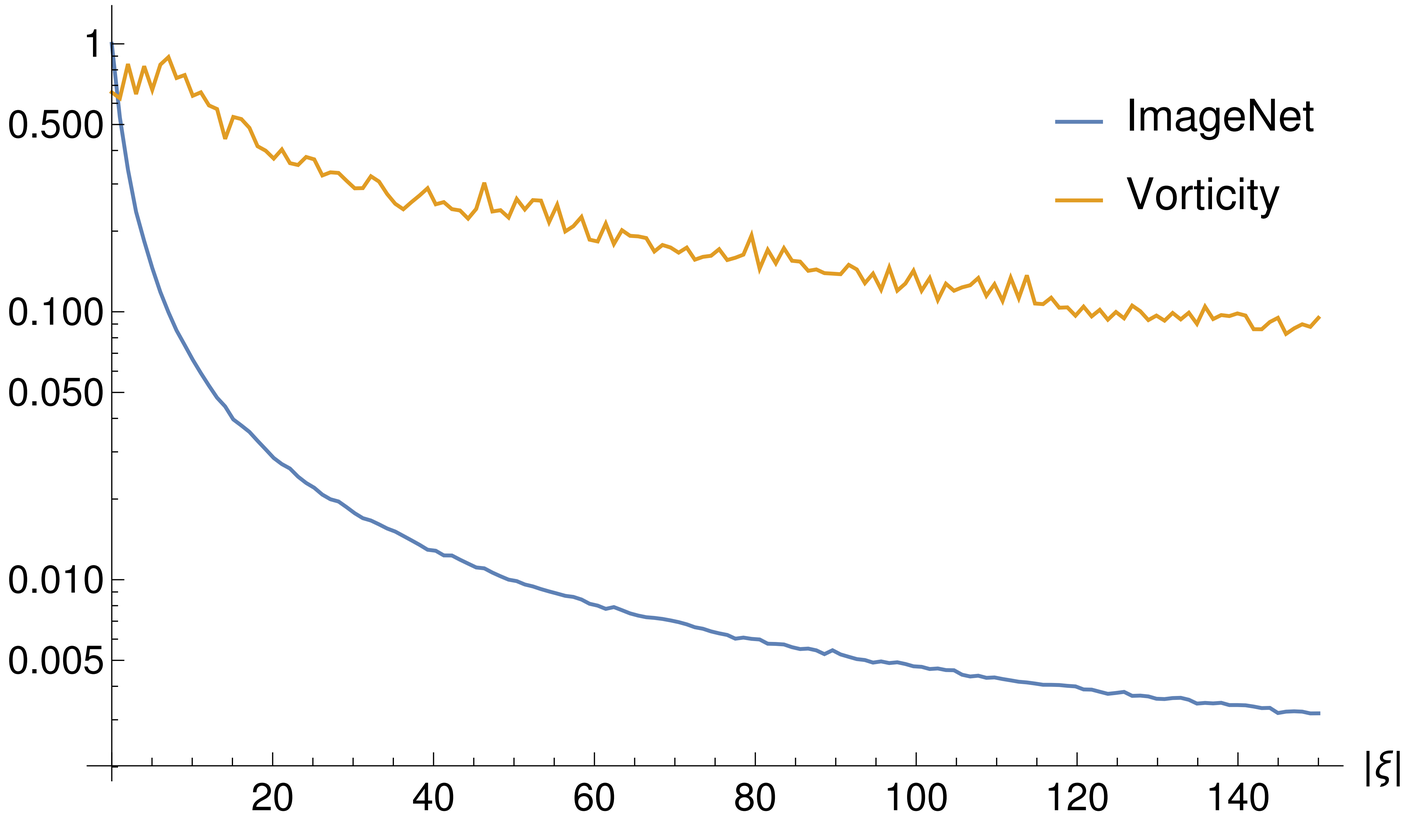
© Christian Lessig, 2023

AtmoRep data



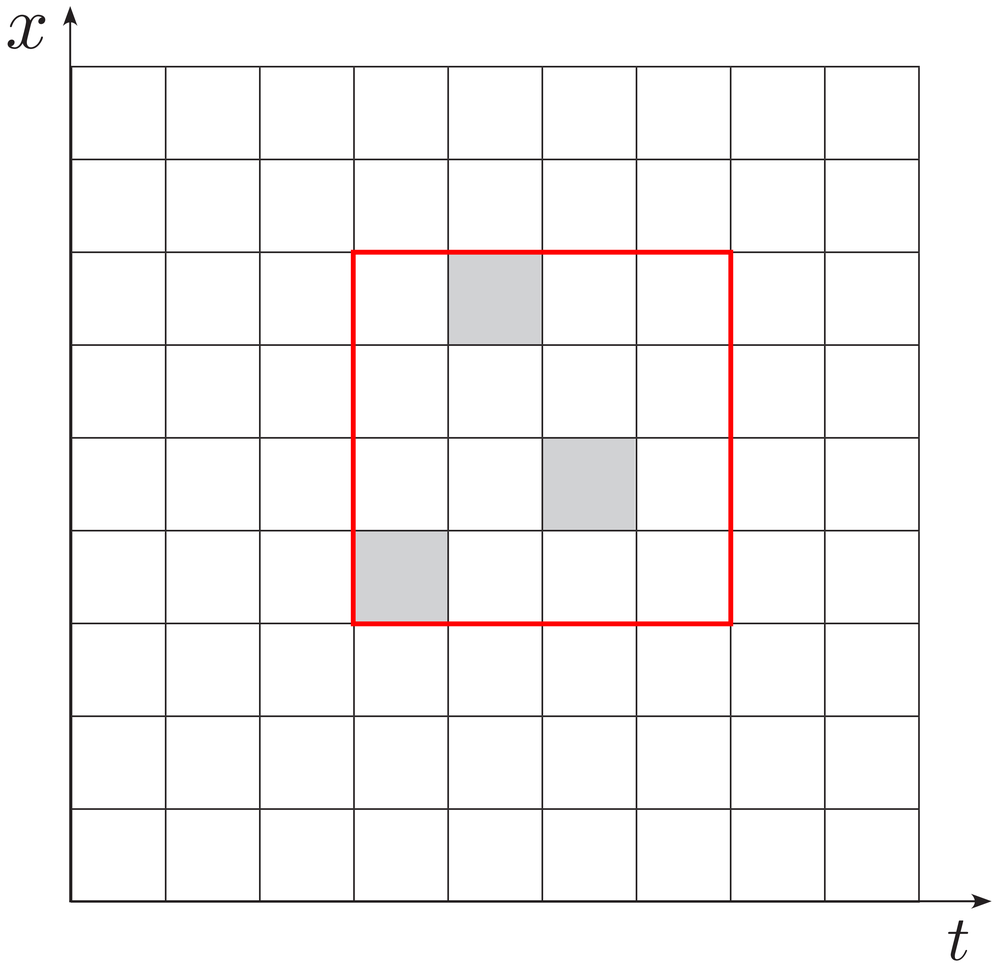
© Christian Lessig, 2023

AtmoRep data



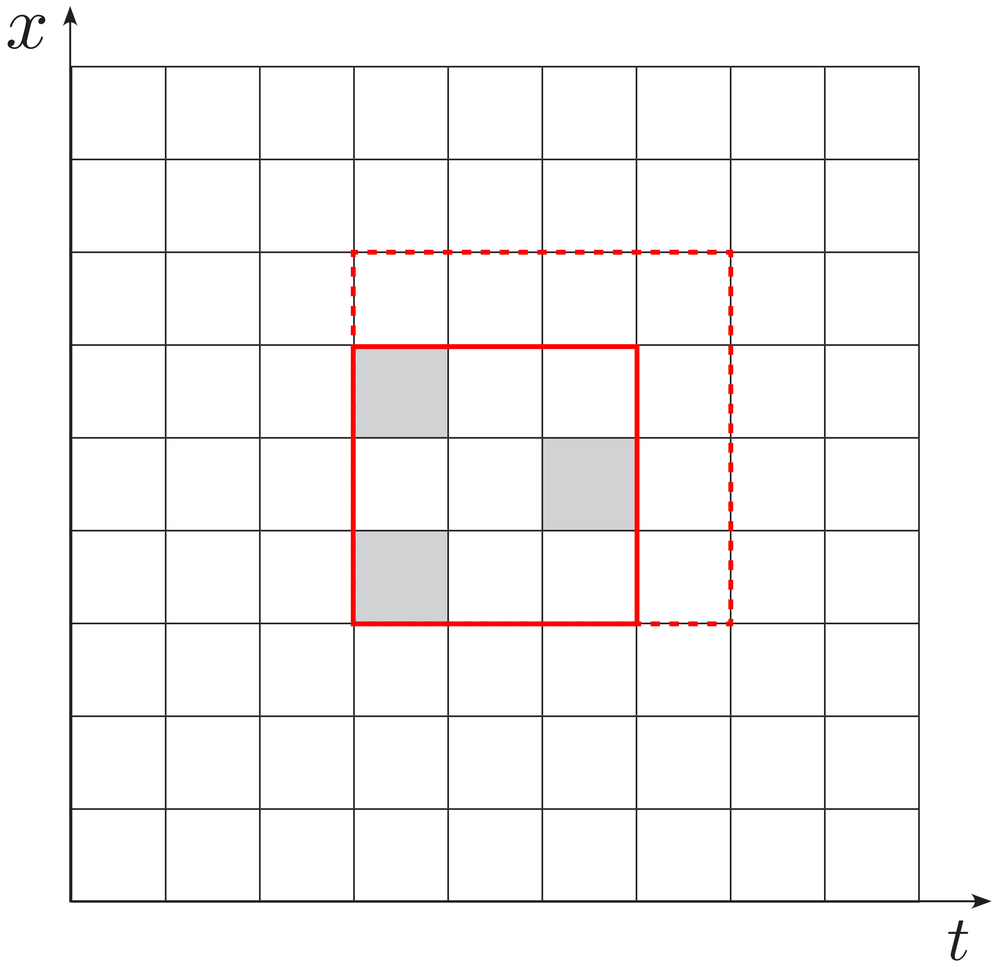
© Christian Lessig, 2023

Spatio-temporal BERT



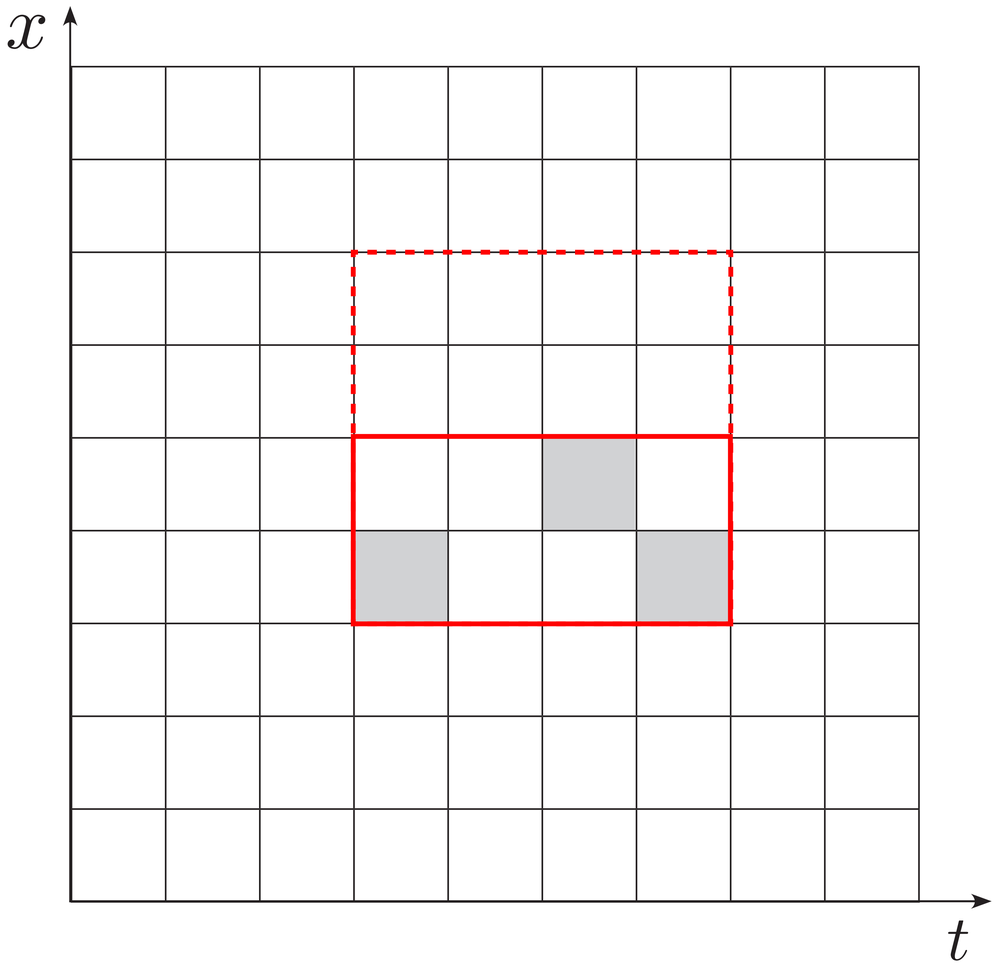
© Christian Lessig, 2023

Spatio-temporal BERT



© Christian Lessig, 2023

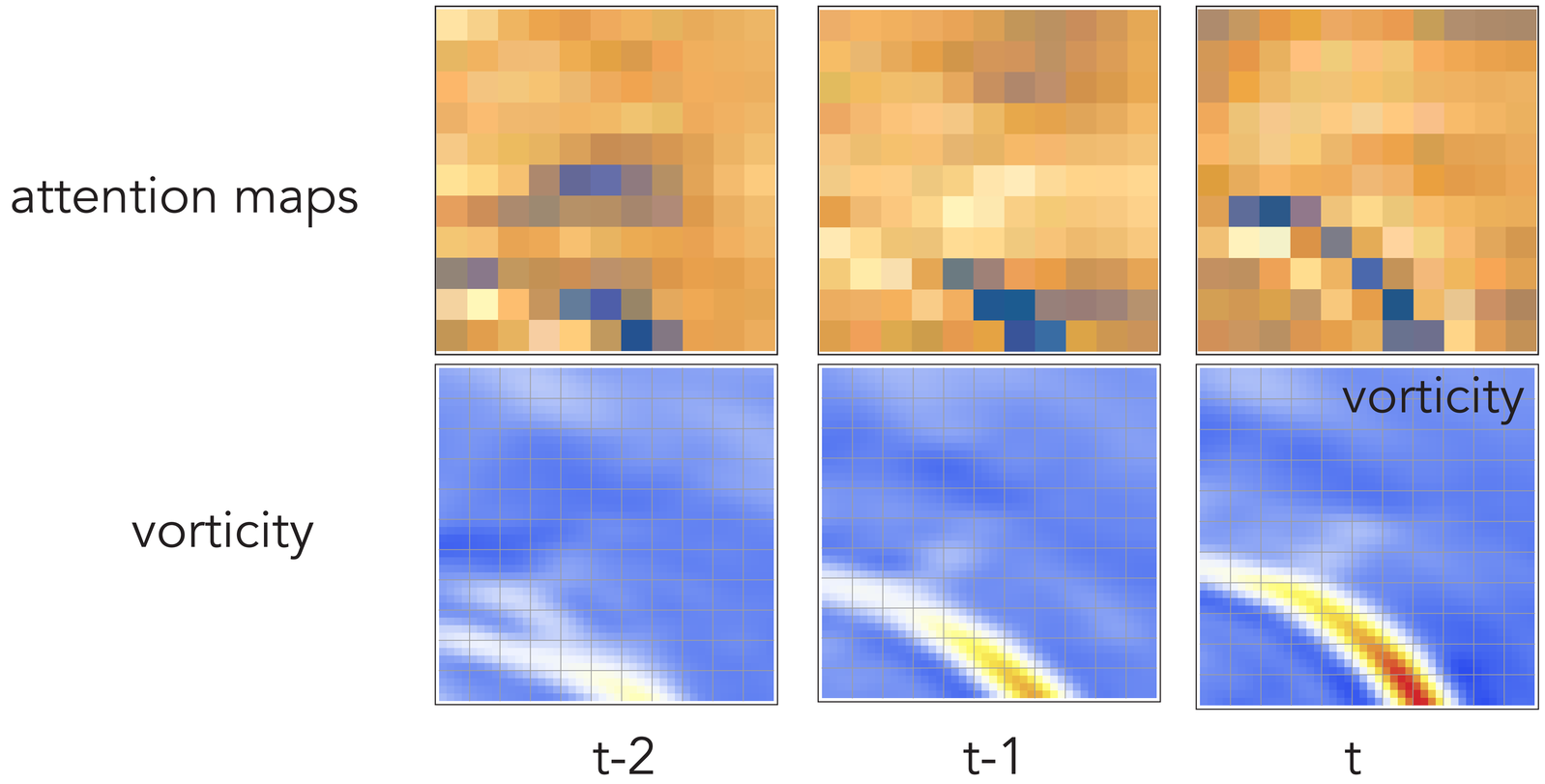
Spatio-temporal BERT



© Christian Lessig, 2023

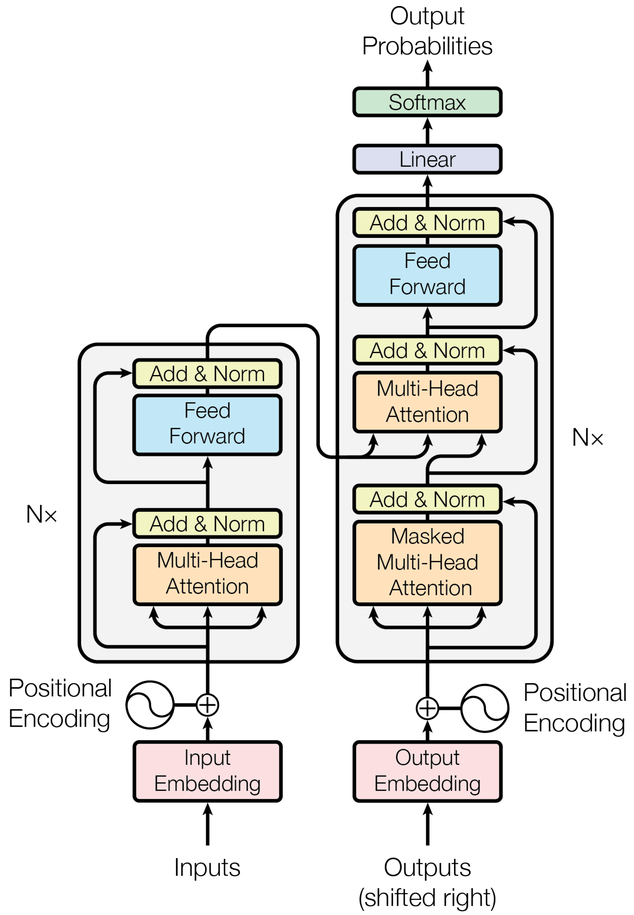
Statistical loss

- Attention maps:



© Christian Lessig, 2023

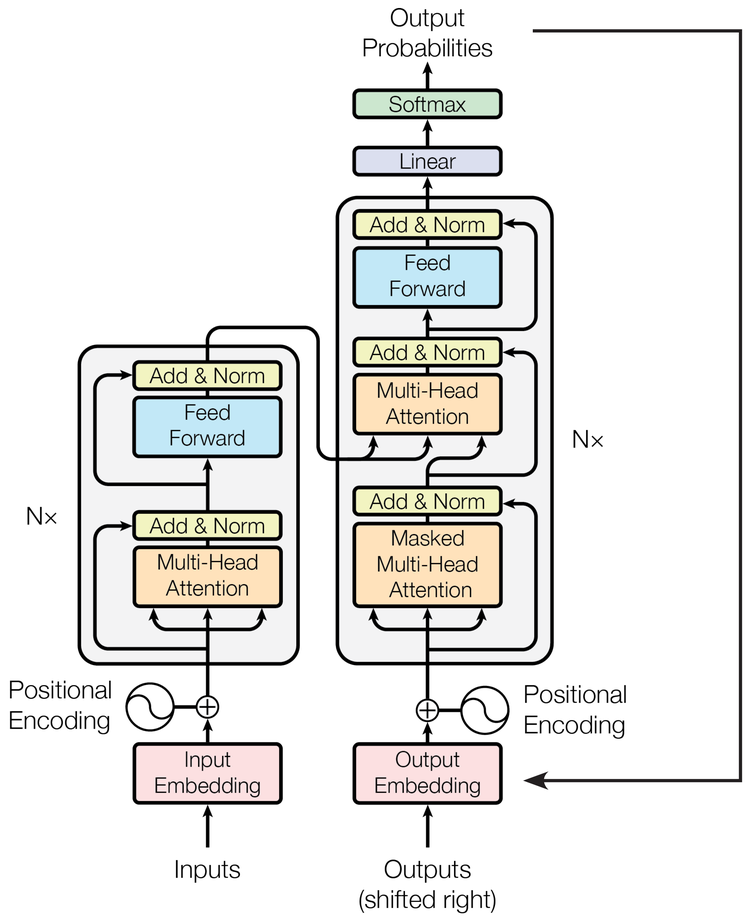
Forecasting / projections



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

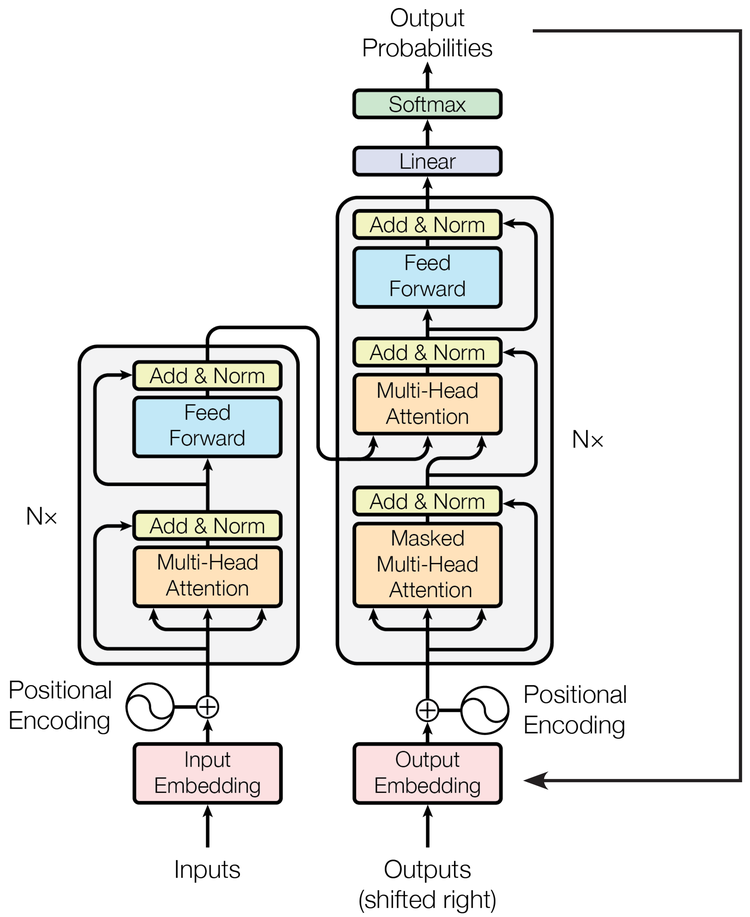


autoregressive, generative modeling

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

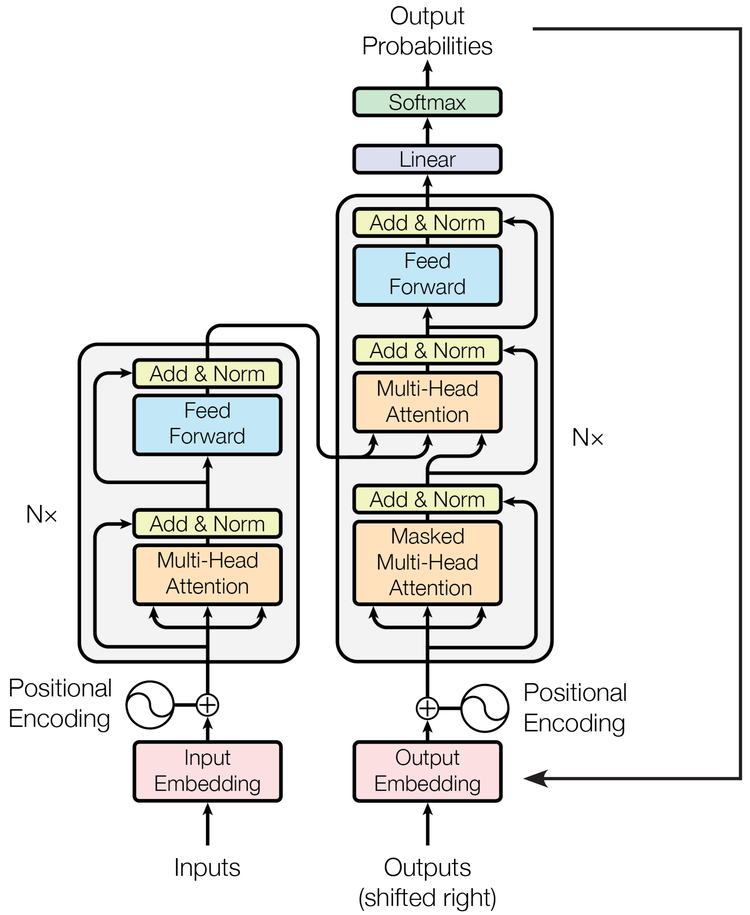
akin to time stepping loop (roll out) for forecasting/projections

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

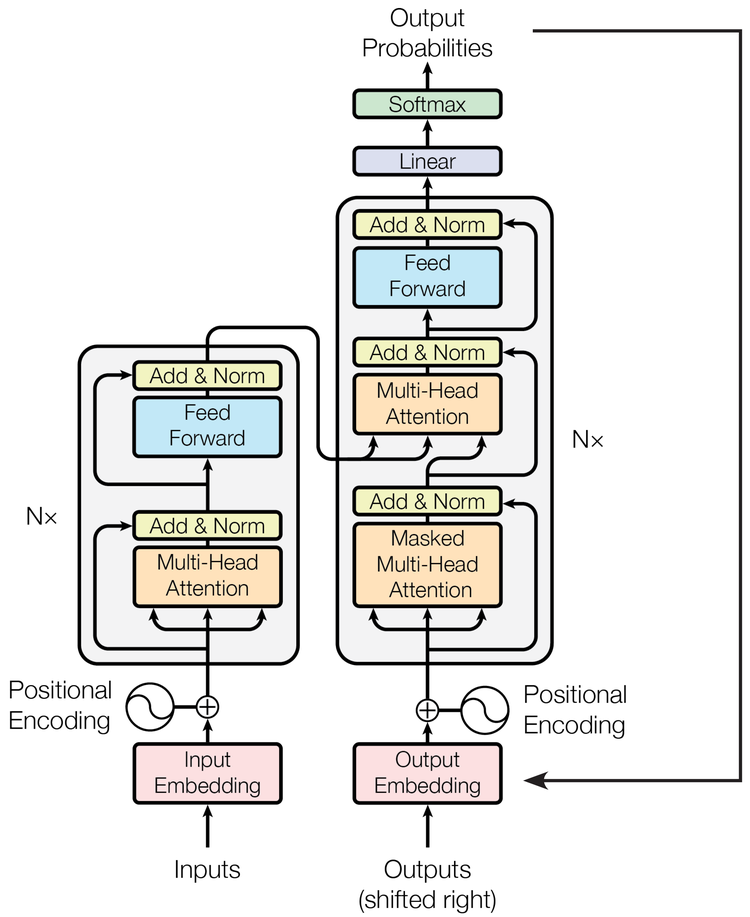
A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017.

© Christian Lessig, 2023

Forecasting / projections

coarse scale/simple classical model

slow climate variables



autoregressive, generative modeling

akin to time stepping loop (roll out) for 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