Alexander von Humboldt Foundation
ERC
AFOSR
DFG
DAAD
Spanish Ministry

Control and Machine Learning

Enrique Zuazua

FAU & AvH, Erlangen, Germany

CoDeFeL

CONTROL FOR DEEP AND FEDERATED LEARNING

2016

DYNAMIC CONTROL

Digital Twins & LLM (2024)

First demonstration of predictive control of fusion plasma by digital twin

by National Institutes of Natural Sciences

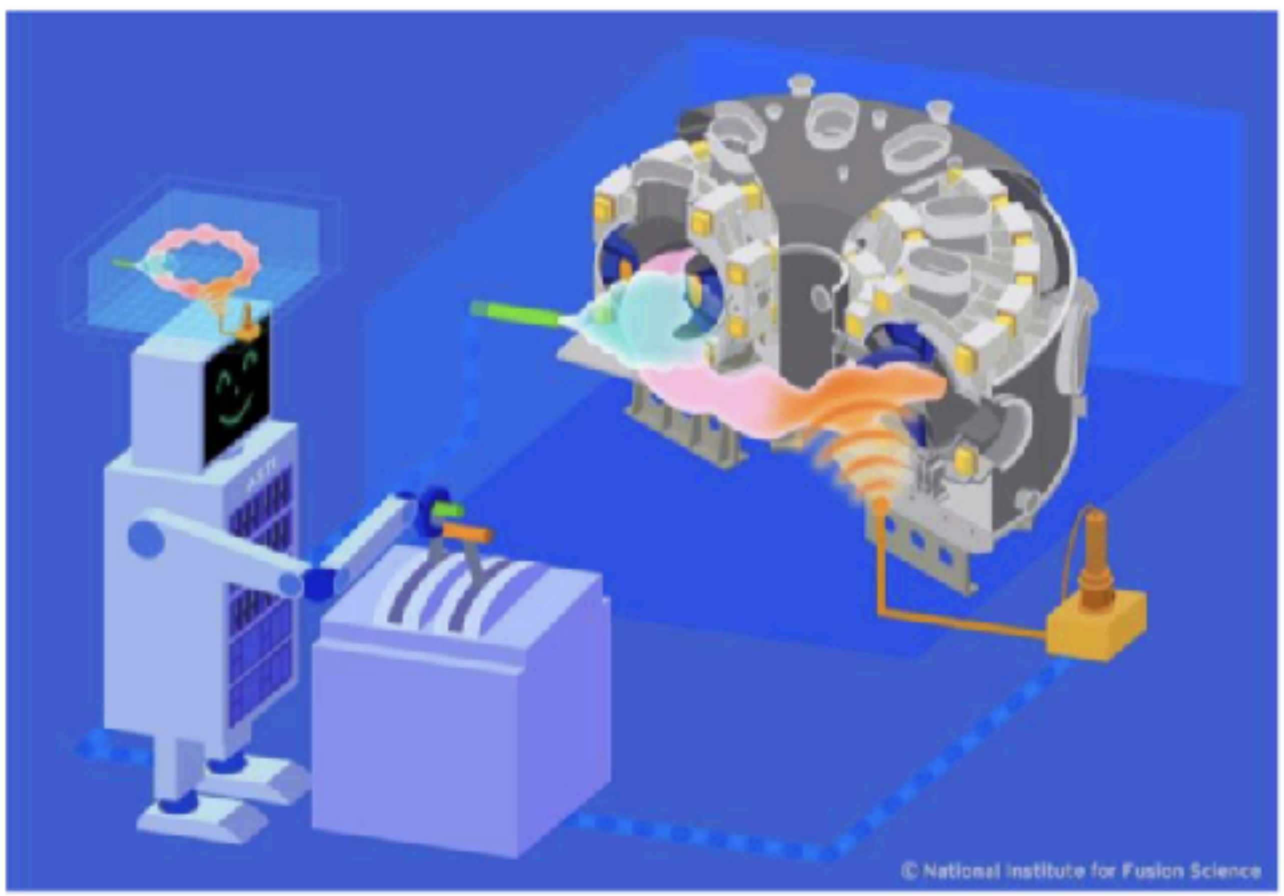
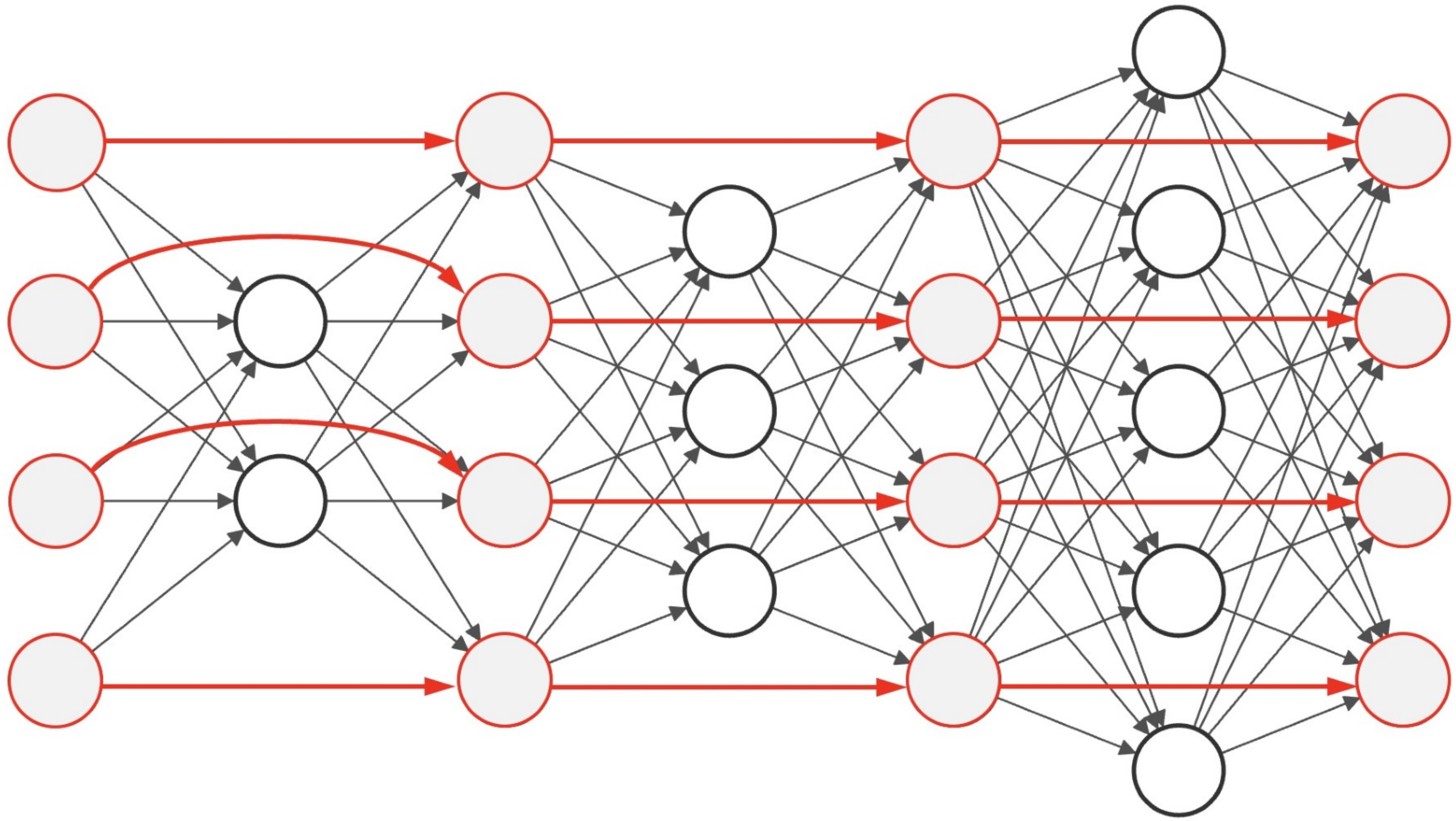


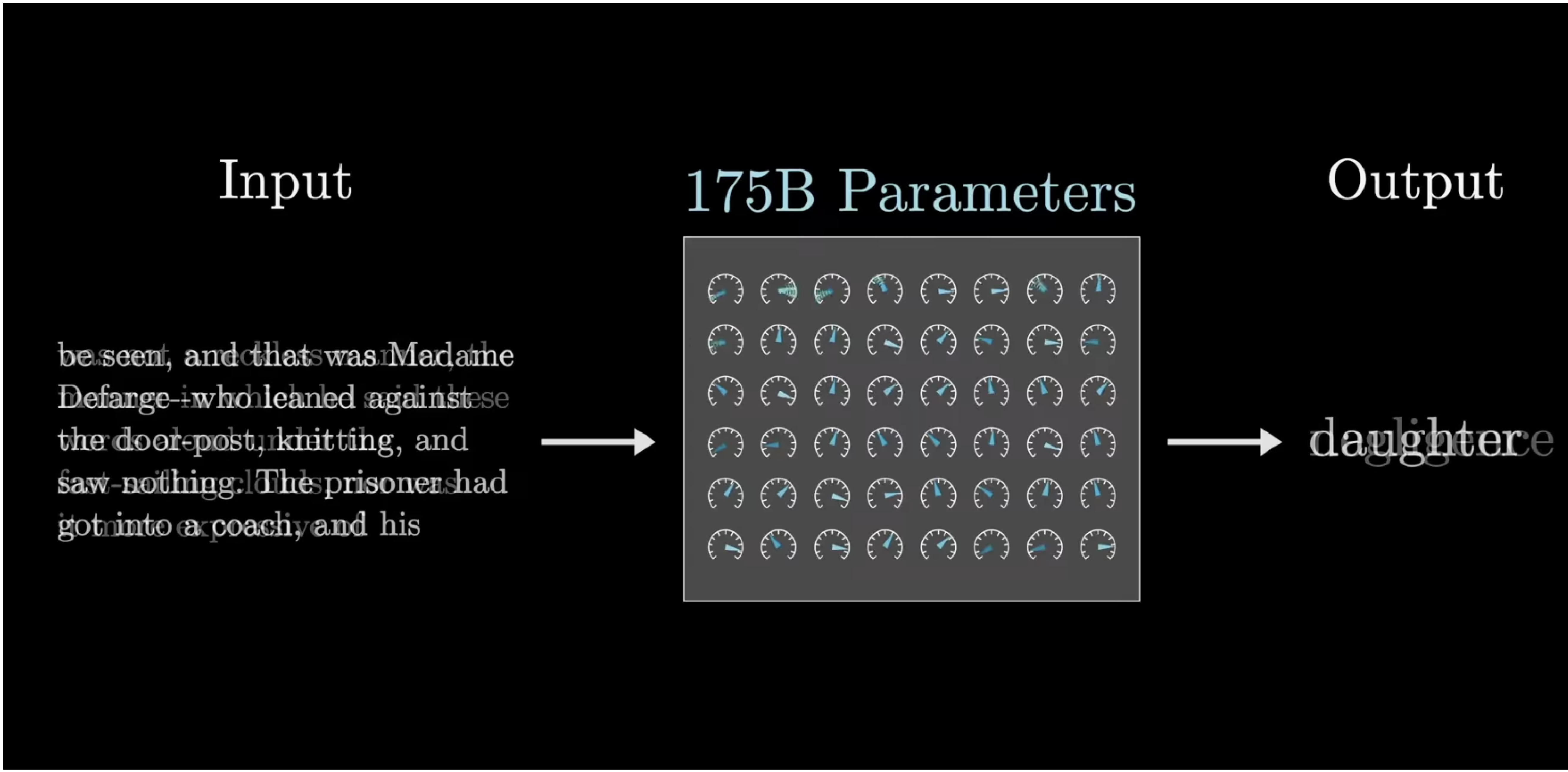
Image of digital twin control, in which real plasma is controlled by virtual plasm

Standard computational practice

$$\underbrace{\frac{1}{N}\sum_{i=1}^{N}\sum_{k}\text{loss}\left(x_i^k, y^{(i)}\right)}_{\text{empirical risk } := E(x(\cdot))} + \alpha \sum_{j,k} \left\| (\mathbf{a}_j^k, \mathbf{w}_j^k, b_j^k) \right\|^2$$



Complexity



Input
175B Parameters
Output
be seen, and that was Madame
Defarge--who leaned against
the door-post, knitting, and
saw nothing. The prisoner had
got into a coach, and his
daughter

Math. Control Signals Systems (1989) 2: 303–314

Mathematics of Control, Signals, and Systems

Approximation by Superpositions of a Sigmoidal Function*

G. Cybenko†

$$\sum_{j=1}^{N} \alpha_j \sigma(y_j^{\mathrm{T}} x + \theta_j), \tag{1}$$

where $y_j \in \mathbb{R}^n$ and $\alpha_j, \theta \in \mathbb{R}$ are fixed. (y^{T} is the transpose of y so that $y^{\mathrm{T}}x$ is the inner product of y and x.) Here the univariate function σ depends heavily on the context of the application. Our major concern is with so-called sigmoidal σ's:

$$\sigma(t) \to \begin{cases} 1 & \text{as} \quad t \to +\infty, \\ 0 & \text{as} \quad t \to -\infty. \end{cases}$$

Two different questions

?1

How does it work?

Does it actually work? Convergence? Error estimates?

Why it works relatively well?

Can traditional applied mathematics contribute to explain the theoretical foundations of this success?

?2

What can Applied Maths learn from these new tools?

Merging: PDE+D(ata)

Digital Twins: Where Data, Mathematics, Models, and Decisions Collide

?1

How does it work?

Supervised learning

Goal: Find an approximation of a function $f_\rho : \mathbb{R}^d \to \mathbb{R}^m$ from a dataset

$$\{\vec{x}_i, \vec{y}_i\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R}^m$$

drawn from an unknown probability measure ρ on $\mathbb{R}^d \times \mathbb{R}^m$.

Classification: match points (images) to respective labels (cat, dog).

This is typically done by **training a neural network**. We will do it through the **simultaneous or ensemble control of Neural ODEs.**

Neural differential equations

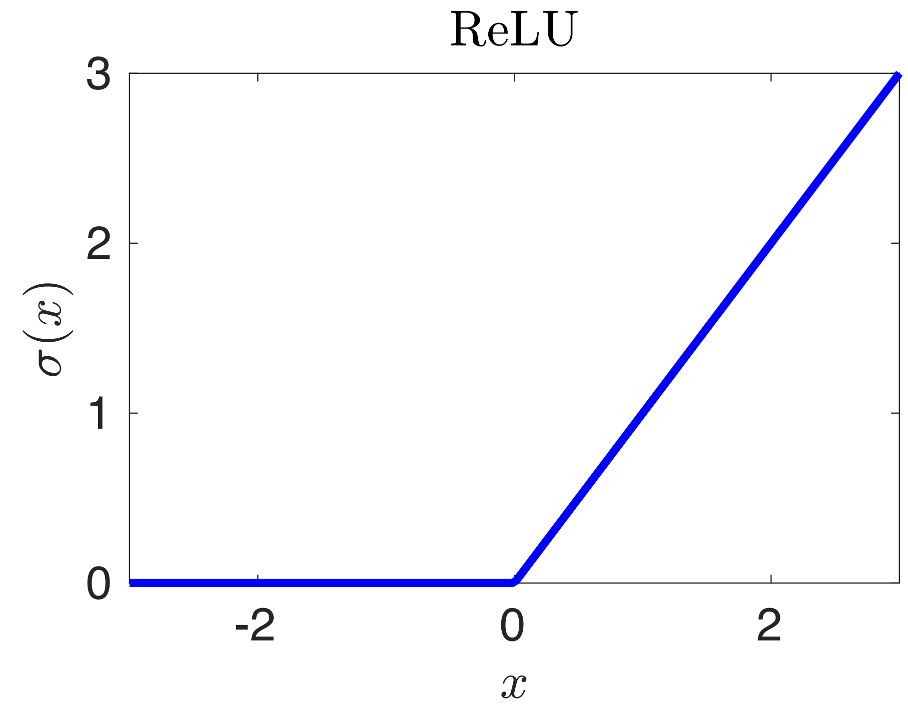
$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{w}(t)\,\sigma(\mathbf{a}(t)\cdot\mathbf{x}(t) + b(t))$$

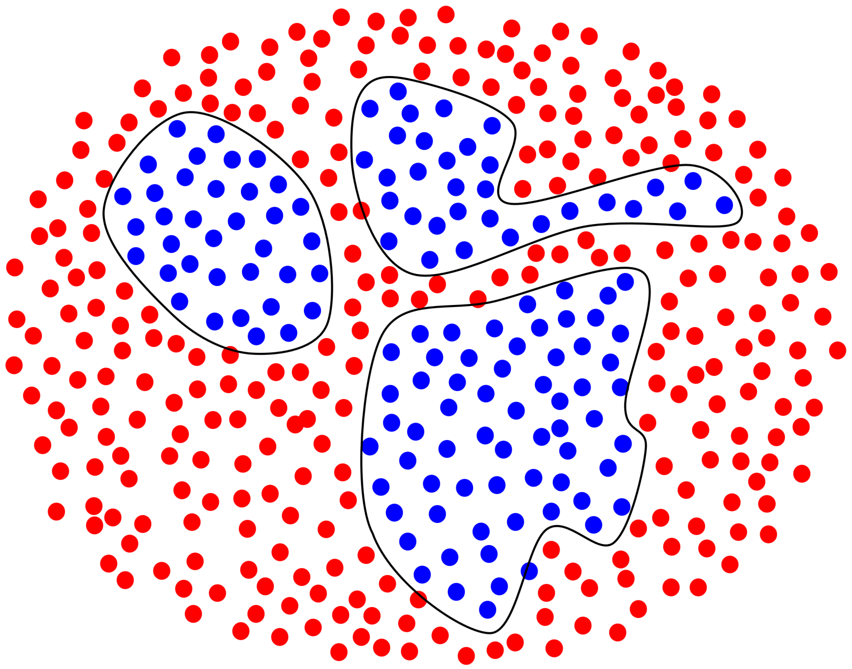
$$\Updownarrow$$

$$\mathbf{x}^{k+1} = \mathbf{x}^k + h\,\mathbf{w}^k\sigma(\mathbf{a}^k\cdot\mathbf{x}^k + b^k)$$

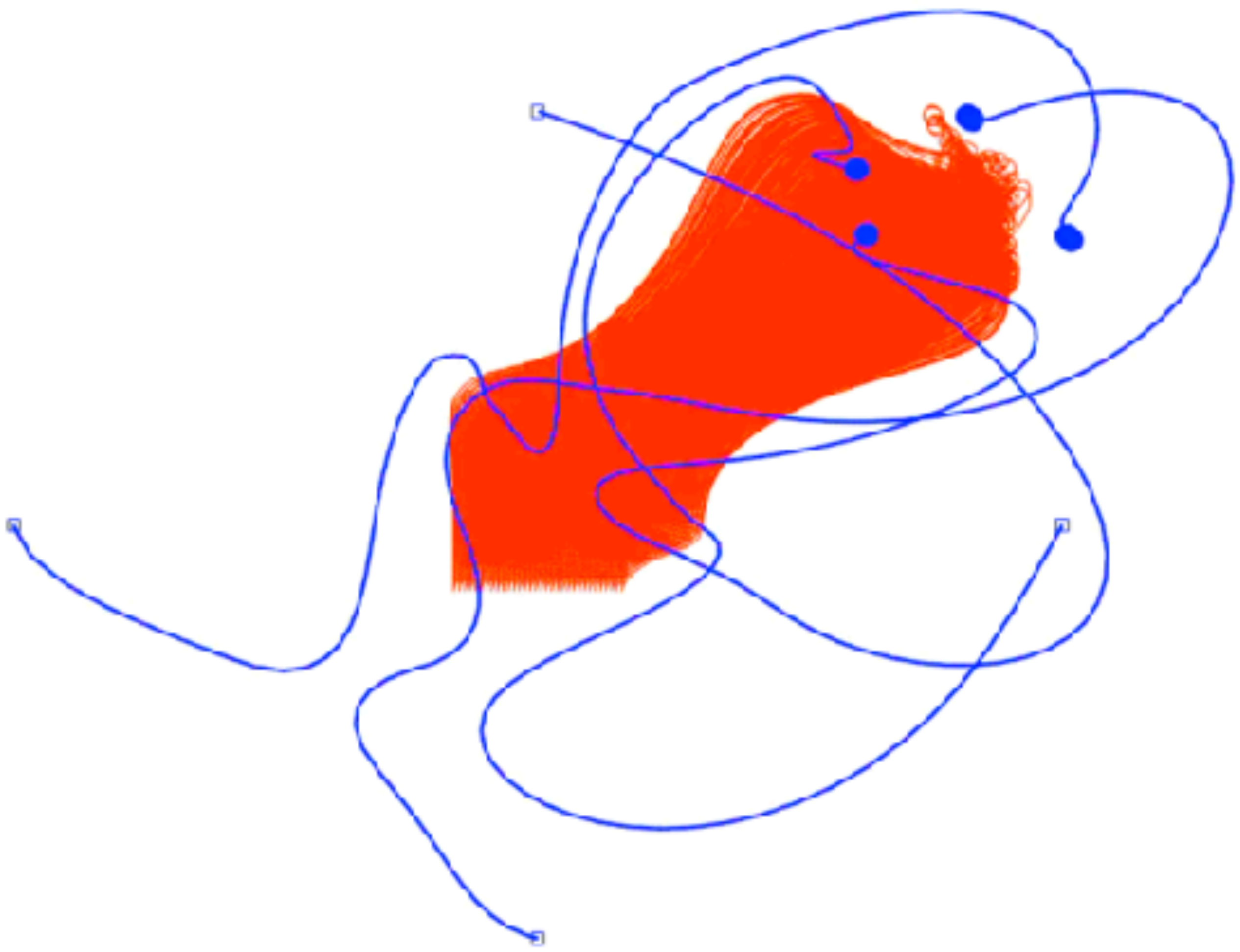
$$\Updownarrow$$

$$f(x) \sim \sum_{j=1}^{K} \mathbf{w}_j\sigma(\mathbf{a}_j\cdot x + b_j)$$





Two neighbouring fields



Control: Dogs-Sheep

Supervised Learning

ResNets / Neural ODEs in action (Borjan Geshkovski)

$$\dot{\mathbf{x}}(t) = \mathbf{w}(t)\,\sigma(\mathbf{a}(t)\cdot\mathbf{x}(t) + b(t))$$

[1] K. He, X Zhang, S. Ren, J Sun, 2016: Deep residual learning for image recognition
[2] E. Weinan, 2017. A proposal on machine learning via dynamical systems.
[3] R. Chen, Y. Rubanova, J. Bettencourt, D. Duvenaud, 2018.
[4] E. Sontag, H. Sussmann, 1997.

Classification by simultaneous or ensemble control of Neural ODEs

Theorem (Classification, Domènec Ruiz-Balet & EZ, SIREV, 2023)

In dimension $d \geq 2$, in any time horizon $[0, T]$, a finite number of arbitrary items can be driven to pre-assigned open subsets of the Euclidean space, corresponding to its labels, by piece-wise constant controls.

Generative Neural Transport

Neural ODEs $\dot{\mathbf{x}}(t) = \mathbf{w}(t)\,\sigma(\mathbf{a}(t)\cdot\mathbf{x}(t)+b(t))$, interpreted as the characteristics of the transport equation:

$$\partial_t \rho + \mathrm{div}_x[\,\underbrace{(\mathbf{w}(t)\,\sigma(\mathbf{a}(t)\cdot x + b(t))}_{V(x,t)}\rho] = 0$$

allow transporting atomic measures and constitute a tool for generative transport.

1

[1] Related results for smooth sigmoids using Lie brackets: A. Agrachev and A. Sarychev, arXiv:2008.12702, (2020); Li, Q., Lin, T., & Shen, Z. (2022), JEMS.

What is the ResNet doing? Basic control actions

$$\dot{\mathbf{x}}(t) = \mathbf{w}(t)\,\sigma(\mathbf{a}(t)\cdot\mathbf{x}(t) + b(t))$$

Control functions $(\mathbf{w}, \mathbf{a}, \mathbf{b}) \longrightarrow$ Piecewise constant.
Each time discontinuity $\sim$ change of layer.

- $\mathbf{a}(t), b(t)$ define a hyperplane $H(\mathbf{x}) = \mathbf{a}(t)\cdot\mathbf{x}(t) + b(t) = 0$ in $\mathbb{R}^d$.
- $\sigma(z) = \max\{z, 0\}$ "activates" the halfspace $H(\mathbf{x}) > 0$ and "freezes" $H(\mathbf{x}) \leq 0$.
- $\mathbf{w}(t)$ determines the direction of the field in the active halfspace.

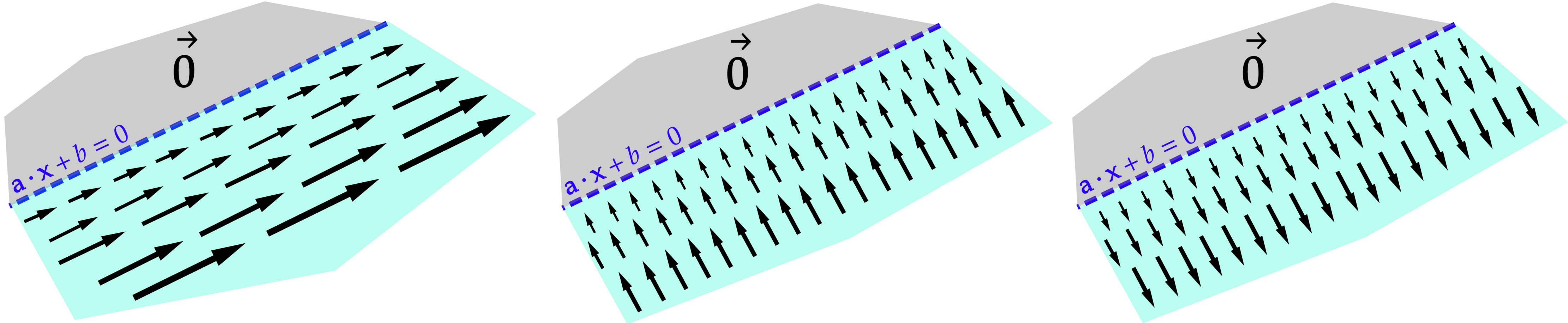
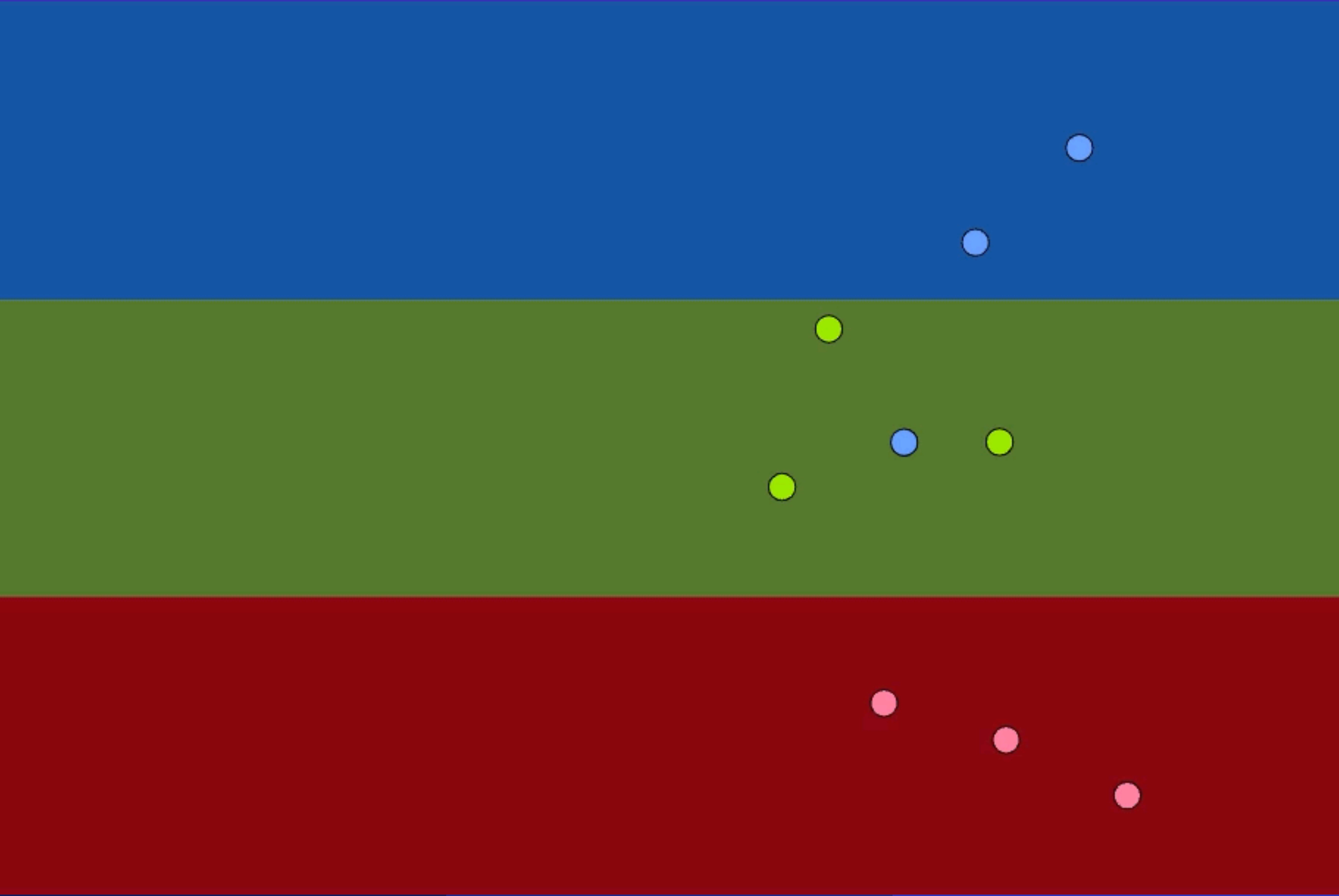


Figure: Parallel (left); Contraction (center); Expansion (right).

Classification by Control of ResNets: One step + Induction



Width versus Depth (A. Álvarez, A. H. Slimane, & E. Z.)

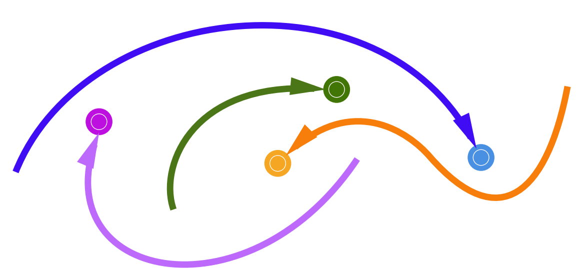
$$\dot{\mathbf{x}}(t) = \sum_{i=1}^{p} \mathbf{w}_i(t)\,\sigma(\mathbf{a}_i(t)\cdot\mathbf{x}(t) + b_i(t))$$

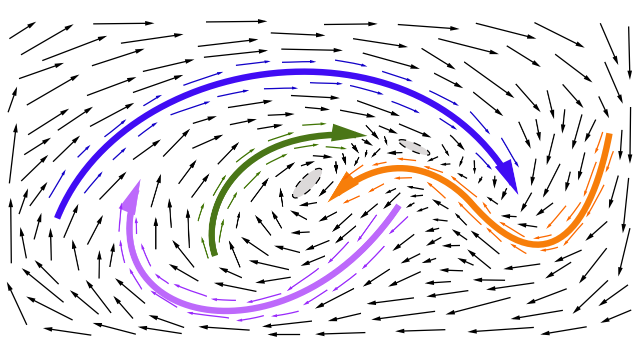
- Increasing the width allows parallelising the consecutive actions of the switching controls and reducing depth:[2]

$$O(N) \to O(1 + N/p) \text{ layers.}$$

- Approximate simultaneous control can be achieved by means of an autonomous, very wide neural field.
  $\to$ Linked to **Turnpike Theory.**

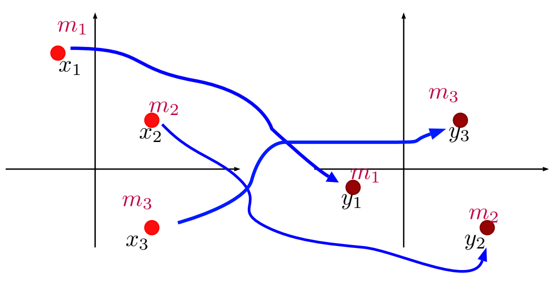
$$\dot{\mathbf{x}}(t) = V(\mathbf{x}(t)) \to V(\mathbf{x}) \sim \sum_{i=1}^{p} \mathbf{w}_i\,\sigma(\mathbf{a}_i\cdot\mathbf{x} + b_i)$$





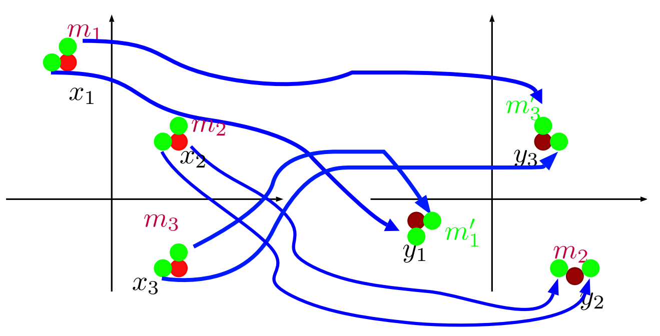
[2] When $d \geq N+1$, the number of layers is $O(N/p)$.

Generative Neural Transport



$$\begin{cases} \dot{\mathbf{x}}(t) = \mathbf{w}(t)\,\sigma(\mathbf{a}(t)\cdot\mathbf{x}(t) + b(t)) \\ \mathbf{x}(0) = \mathbf{x}_0 \end{cases}$$

interpreted as the characteristics of the transport equation:



$$\begin{cases} \partial_t\rho + \mathrm{div}_x[\underbrace{(\mathbf{w}(t)\,\sigma(\mathbf{a}(t)\cdot x + b(t))}_{V(x,t)}\rho] = 0 \\ \rho(0) = \rho^0 \end{cases}$$

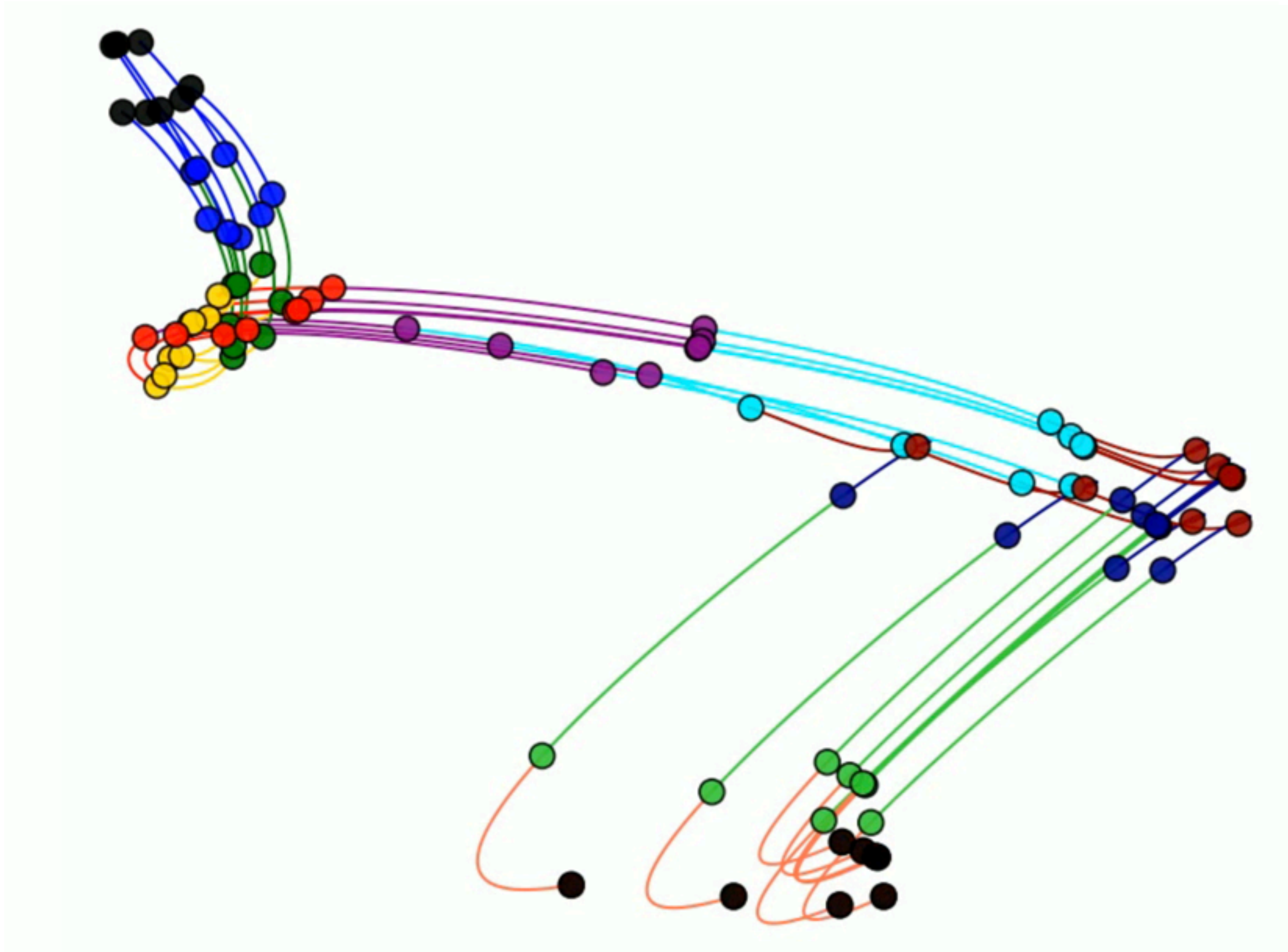
Atomic initial data can be driven to atomic final targets

Theory of Optimal Transport $\rightarrow$ Neural Transport

Chebyshev inequality allows estimating the statistical error of sampling for normalizing flows

Tracking dynamical systems

Joint work with Z. Li, K. Liu and L. Liverani



Semi-autonomous NODEs

A time-independent choice of the parameters leads to a non-autonomous dynamics, with a trivial time-dependence,

$$\dot{\mathbf{x}}(t) = \sum_{i=1}^{p} w_i\, \sigma(a_i^1 \cdot \mathbf{x}(t) + a_i^2 t + b_i)$$

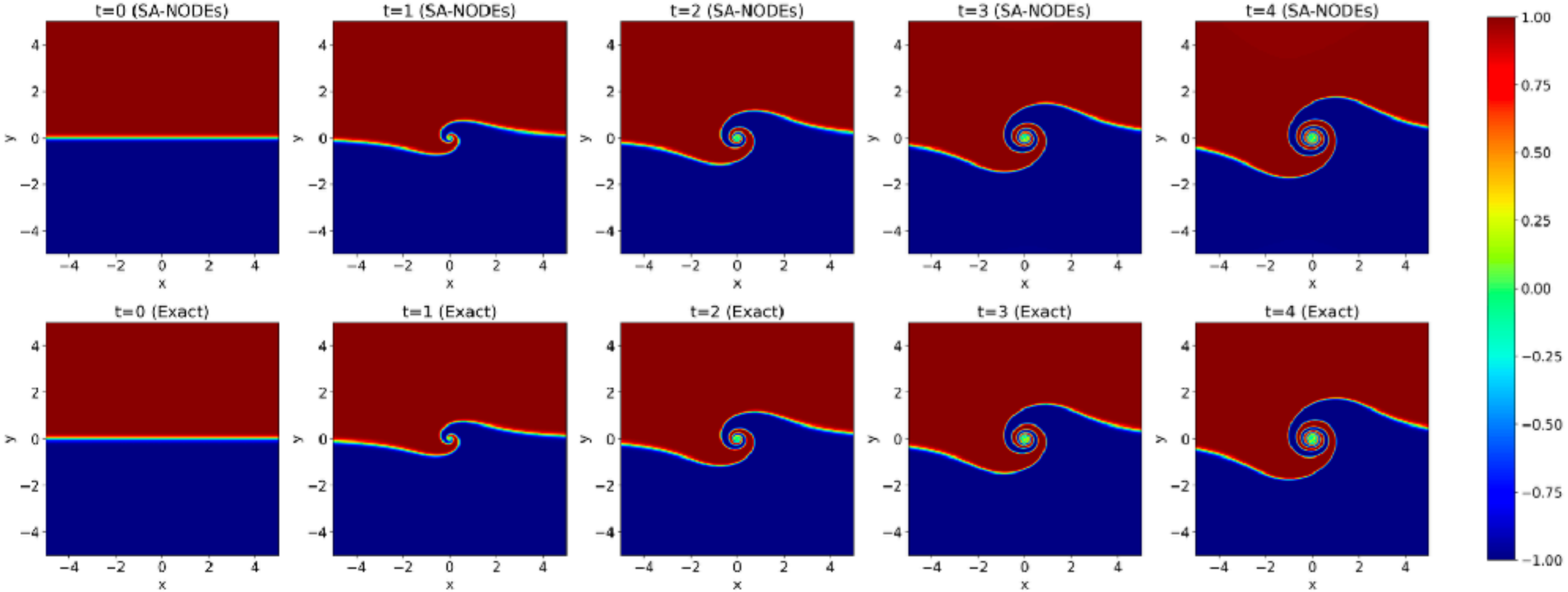
- The structure is motivated by the Universal Approximation property of ReLU activation functions (Pinkus, 1999)

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), t) \;\rightarrow\; \mathbf{f}(\mathbf{x}, t) \sim \sum_{i=1}^{p} w_i\, \sigma(a_i^1 \cdot \mathbf{x} + a_i^2 t + b_i)$$

- The coefficients are now time-independent, greatly reducing the complexity of the model
- The obtained model can be employed to anticipate future evolution of trajectories.

Numerical Results: Transport Equations

Ongoing work with Weiwei Hu on optimal fluid mixing



SA-NODEs and exact solution of the transport equation modeling Doswell frontogenesis

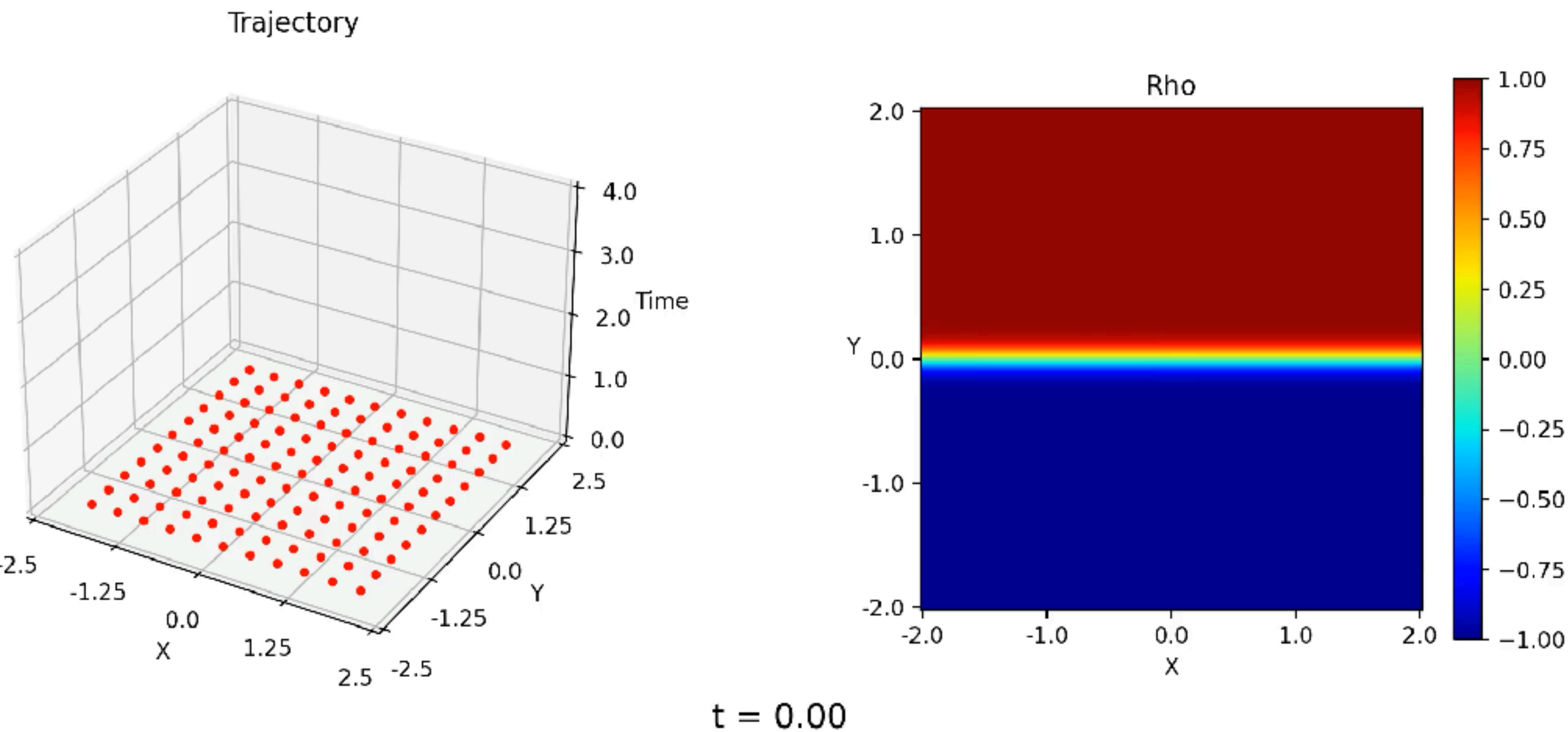
$$\partial_t \rho(x,y,t) + \operatorname{div}\left(\rho(x,y,t)\left(-yg(r), xg(r)\right)\right) = 0,$$

where $(x,y,t) \in \mathbb{R}^2 \times [0, T]$ and,

$$g(r) = c\, r^{-1}\, \operatorname{sech}^2 r \tanh r, \quad \rho_0(x,y) = \tanh\left(y/\delta\right).$$

The exact solution:

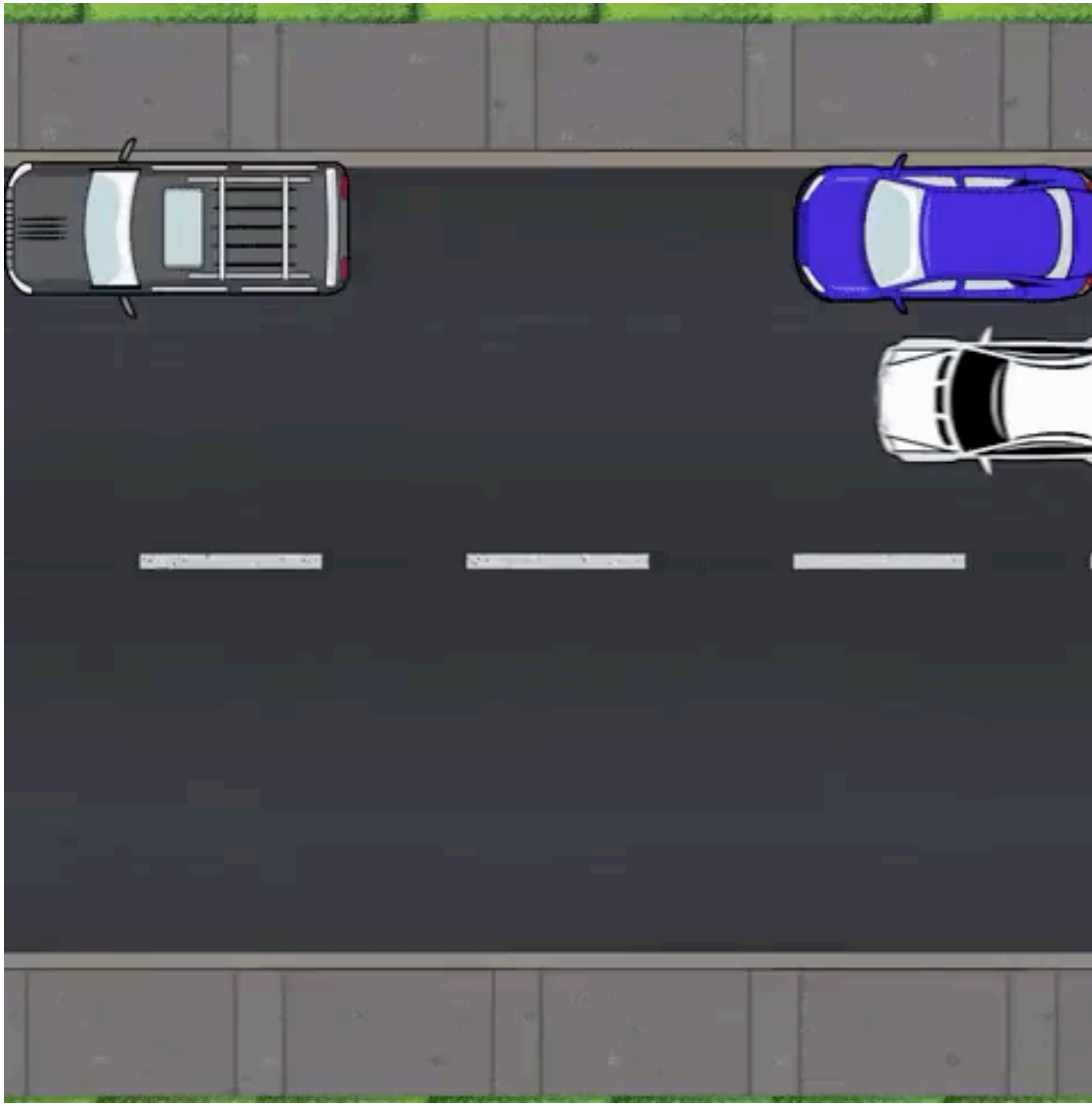
$$\rho(x,y,t) = \tanh\left(\frac{y\cos(gt) - x\sin(gt)}{\delta}\right).$$

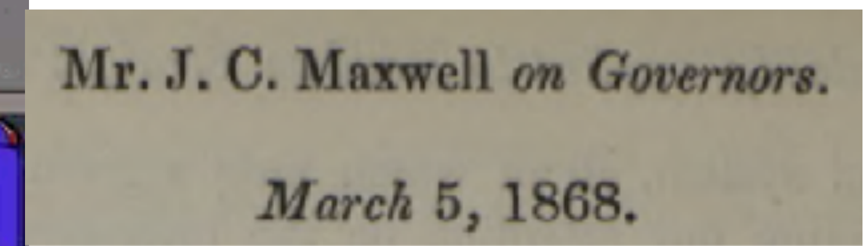

Trajectory
Time
4.0
3.0
2.0
1.0
0.0
X
-2.5
-1.25
0.0
1.25
2.5
Y
-2.5
-1.25
0.0
1.25
2.5
Rho
Y
X
2.0
1.0
0.0
-1.0
-2.0
-2.0
-1.0
0.0
1.0
2.0
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
t = 0.00

?2

PDE+D(ata)

An example: Nelson's car.




Mr. J. C. Maxwell *on Governors.*

March 5, 1868.

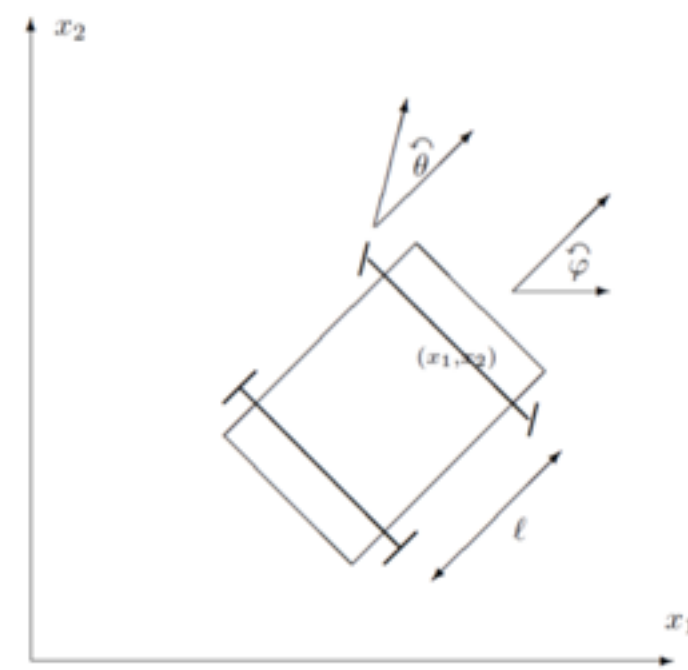
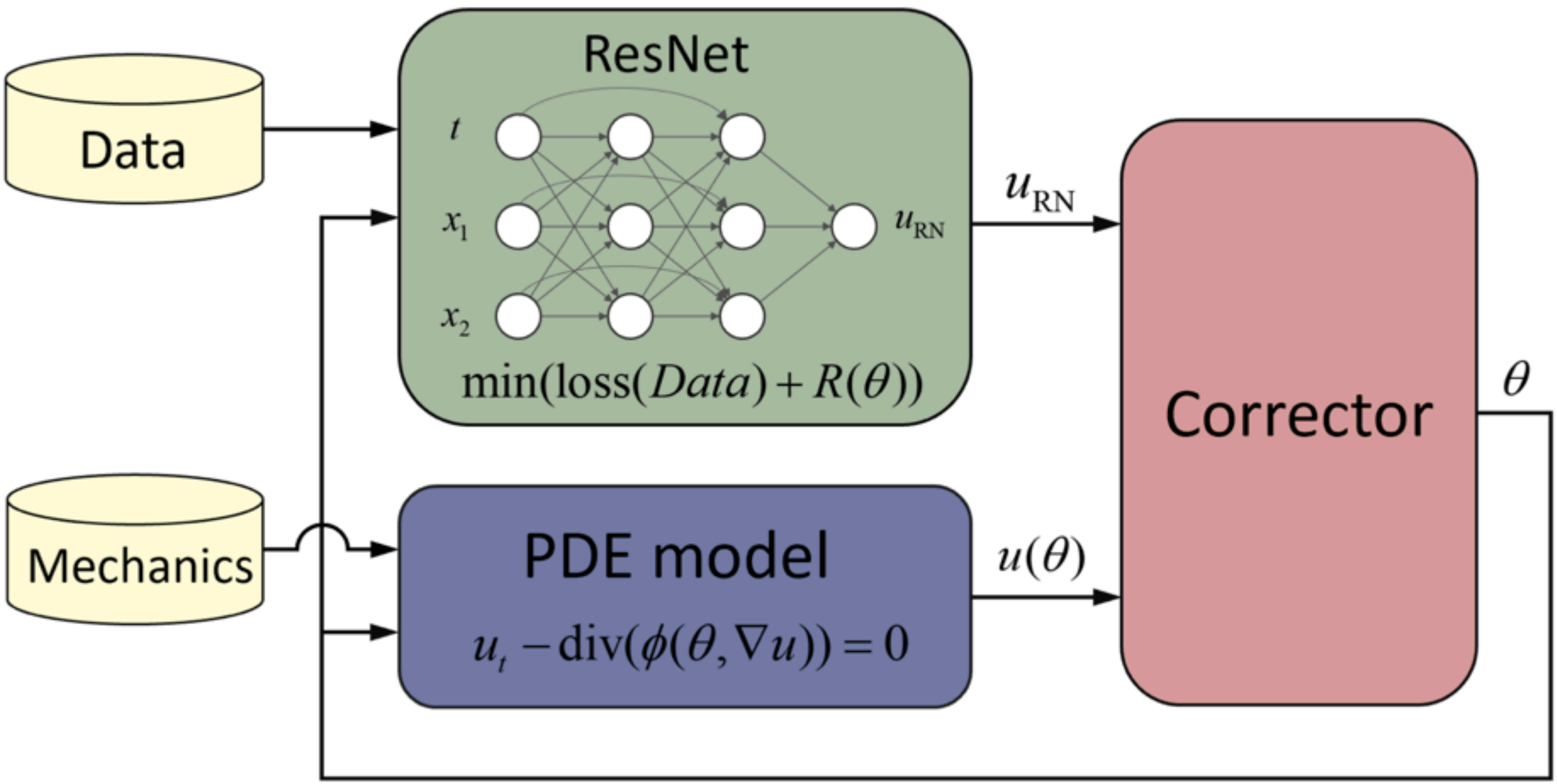


Figure 4.1: *4-dimensional car model.*

Two controls suffice to control a four-dimensional dynamical system.[1]

[1] E. Sontag, *Mathematical control theory*, 2nd ed.,Springer-Verlag, NewYork,1998.

Hybrid Data driven + PDE modelling + Collapse



Training:

Curse of Dimensionality

+ Evil of Non-Convexity

Training & Generalization

joint work with Kang Liu

General architecture of NNs

$$f: \mathbb{R}^d \times \prod_{i=1}^{P} \mathbb{R}^{d_i} \to \mathbb{R}^m, \ (x, \Theta) \mapsto f(x, \Theta),$$

where

- x is the feature (input),
- Θ is the parameter (control),
- $f(x, \Theta)$ is the prediction (output).

Three training scenarios

Consider a dataset: $\{(x_i, y_i)\}_{i=1}^{N}$.

1. **Exact representation**:

$$f(x_i, \Theta) = y_i, \quad \text{for } i = 1, \ldots, N.$$

2. **Approximate representation**:

$$\|f(x_i, \Theta) - y_i\| \leq \epsilon, \quad \text{for } i = 1, \ldots, N.$$

3. **Regression**:

$$\inf_{\Theta} \frac{1}{N} \sum_{i=1}^{N} \ell(f(x_i, \Theta) - y_i).$$

Problems: Existence, regularization, generalization property, numerical algorithms, etc.

Primal sparsified problems

Let Ω be a compact subset of $\mathbb{R}^{d+1}$. Consider the following three optimization problems: Let $\Theta = (\omega_j, a_j, b_j)_{j=1}^P$.

- The sparse **exact representation** problem:

$$\inf_{\Theta\in(\mathbb{R}\times\Omega)^P} \|\omega\|_{\ell^1}, \quad \text{s.t. } \sum_{j=1}^P \omega_j \sigma(\langle a_j, x_i\rangle + b_j) = y_i, \quad \text{for } i = 1, \dots, N. \qquad (\mathrm{P}_0)$$

- The sparse **approximate representation** problem:

$$\inf_{\Theta\in(\mathbb{R}\times\Omega)^P} \|\omega\|_{\ell^1}, \quad \text{s.t. } \left|\sum_{j=1}^P \omega_j \sigma(\langle a_j, x_i\rangle + b_j) - y_i\right| \leq \epsilon, \quad \text{for } i = 1, \dots, N, \qquad (\mathrm{P}_\epsilon)$$

  where $\epsilon > 0$ is a hyperparameter.

- The sparse **regression** problem:

$$\inf_{\Theta\in(\mathbb{R}\times\Omega)^P} \|\omega\|_{\ell^1} + \frac{\lambda}{N}\sum_{i=1}^N \ell\left(\sum_{j=1}^P \omega_j \sigma(\langle a_j, x_i\rangle + b_j) - y_i\right), \qquad (\mathrm{P}_\lambda^{\mathrm{reg}})$$

  where $\lambda > 0$ is a hyperparameter.

Mean-field relaxation

Primal problems (P_0), (P_ϵ), and (P_λ^{reg}) are non-convex optimization problems, where the non-convexity is from the non-linearity of shallow NNs, e.g.,

$$\left\{ \Theta \,\middle|\, \sum_{j=1}^{P} \omega_j \sigma(\langle a_j, x_i \rangle + b_j) = y_i, \ \forall i = 1, \ldots, N \right\} \text{ is a non-convex set.}$$

The mean-field relaxation technique is commonly employed in shallow NNs, see [Mei-Montanari-Nguyen, 2018] and [Chizat-Bach, 2018].

Shallow NN

The original Shallow NN writes:

$$\sum_{j=1}^{P} \omega_j \sigma(\langle a_j, x \rangle + b_j),$$

where $(\omega_j, a_j, b_j) \in \mathbb{R} \times \Omega$ for all j.

Cost function: $\|\omega\|_{\ell^1}$.

Mean-field shallow NN

The mean-field shallow NN writes:

$$\int_\Omega \sigma(\langle a, x \rangle + b) d\mu(a, b),$$

where $\mu \in \mathcal{M}(\Omega)$. The outcome is linear with respect to μ.

Cost function: $\|\mu\|_{\text{TV}}$.

Lack of relaxation gap

Theorem

Assume that $P \geq N$. Then, there is no gap between the original primal problems and the relaxed ones.
Moreover, the extreme points of the solution sets of relaxed problems have the following form:

$$\mu^* = \sum_{j=1}^{N} \omega_j^* \delta_{(a_j^*, b_j^*)}.$$

Based on the "Representer Theorem" from [S. D. Fisher and J. W. Jerome. Spline solutions to L^1 extremal problems in one and several variables. Journal of Approximation Theory, 13.1 (1975), pp. 73 – 83.]

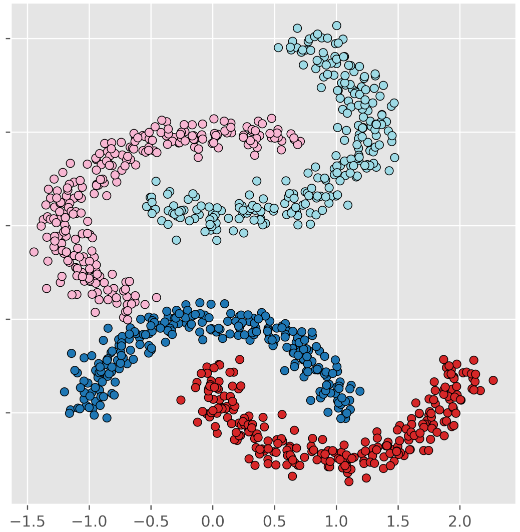
Numerical algorithms and generalization

Two numerical schemes

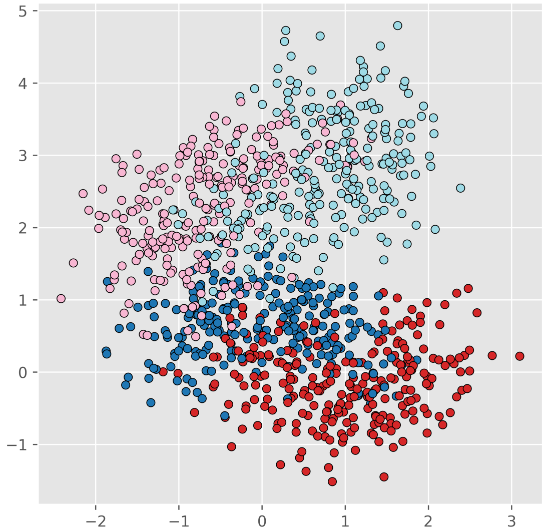
1. Discretization of Ω combined with the simplex method:
   - **Advantage**: Guarantees a global minimizer.
   - **Limitation**: Suffers from the curse of dimensionality.
2. Stochastic Gradient Descent (for overparameterized shallow NNs [Chizat-Bach, 2018]) combined with a sparsification method:
   - **Advantage**: Free from the curse of dimensionality.
   - **Limitation**: Lacks global convergence guarantees.

Conclusion: We apply Scheme 1 for low-dimensional data, while Scheme 2 is more suitable for high-dimensional data.

Generalization



If the datasets have clear separable boundaries, consider (P_0), (P_ϵ) with $\epsilon \to 0$, or $(\mathrm{P}_\lambda^{\mathrm{reg}})$ with $\lambda \to \infty$.



If the datasets have heavily overlapping areas, consider the regression problem $(\mathrm{P}_\lambda^{\mathrm{reg}})$ with $\lambda \sim \mathcal{O}(N^{1/d})$.

Our recent contributions

E. Zuazua, *Control and Machine Learning*, SIAM News, October 2022

D. Ruiz-Balet, E. Zuazua, *Neural ODE control for classification, approximation and transport*, SIAM Review, 65 (3)3 (2023), 735-773.

B. Geshkovski, E. Zuazua, *Turnpike in optimal control of PDEs, ResNets, and beyond*, Acta Numer., 31 (2022), 135–263

D. Ruiz-Balet, E. Zuazua, *Control of neural transport for normalizing flows, Journal de mathématiques pures et appliquées*, JMPA, 181 (2024), 58-90.

Z. Li, K. Liu, L. Liverani, E. Zuazua,Universal Approximation of Dynamical Systems by Semi-Autonomous Neural ODEs and Applications, arXiv:2407.17092v2, (2024).

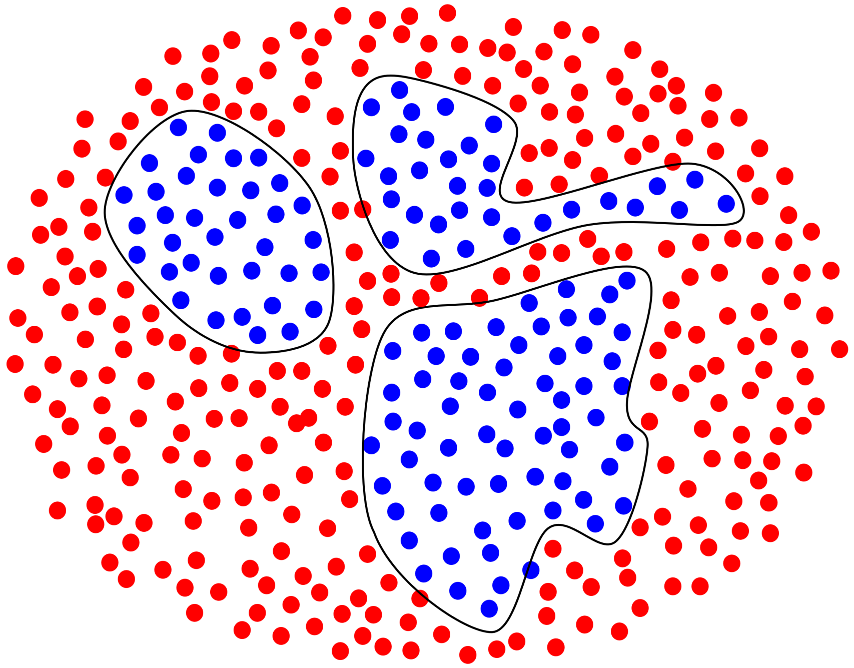
A. Alcalde, G. Fantuzzi, E. Zuazua, Clustering in pure-attention hardmax transformers, arXiv:2407.01602, (2024).

K. Liu, E. Zuazua, Representation and regression problems in neural networks: relaxation, generalization and numerics, in progress.

Clustering versus Complexity

A. Álvarez, R. Orive, & E. Z., 2023

- Clustering of data allows to diminish the number of switches

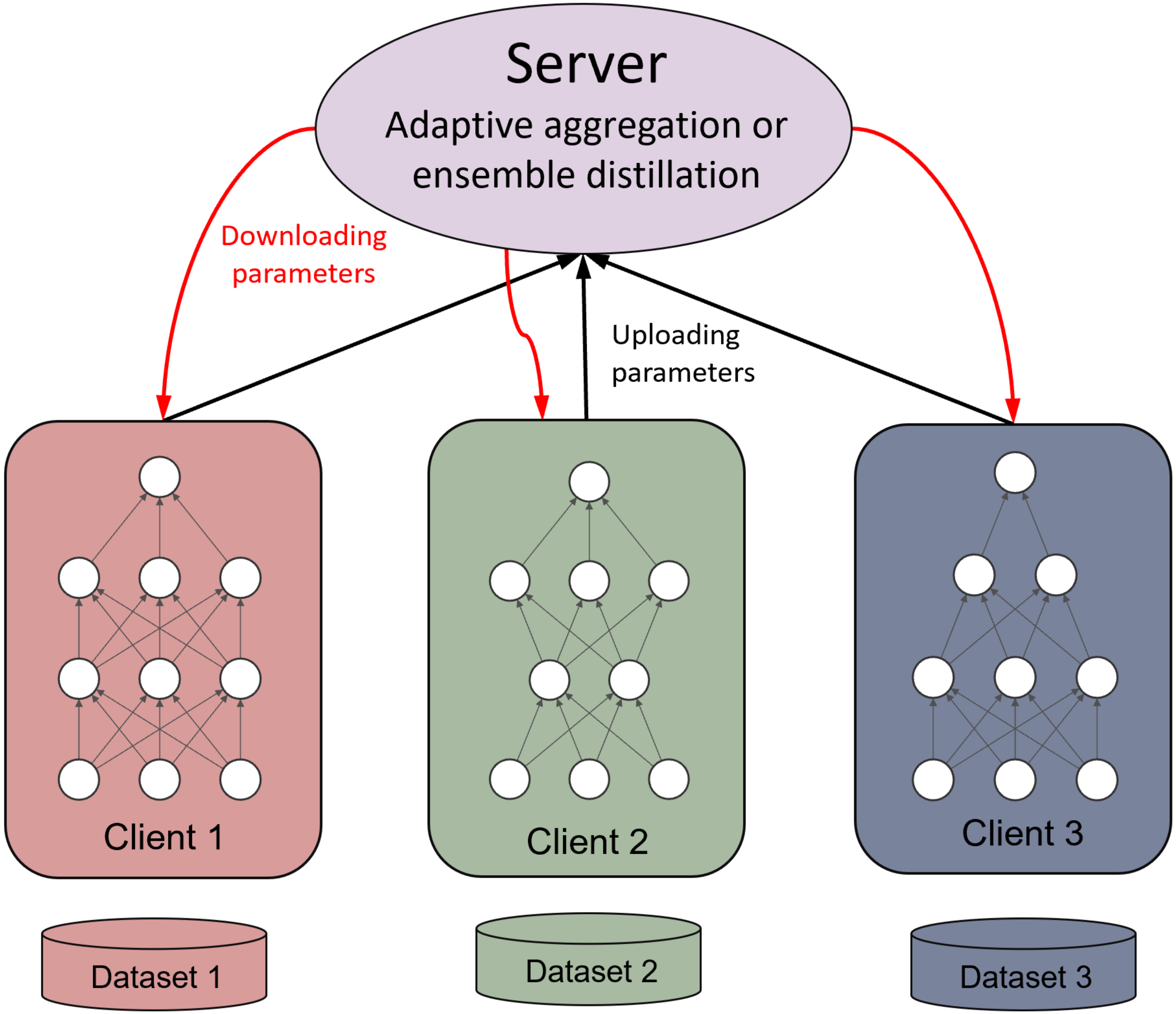


- Increasing the space dimension d diminishes the number of switches:

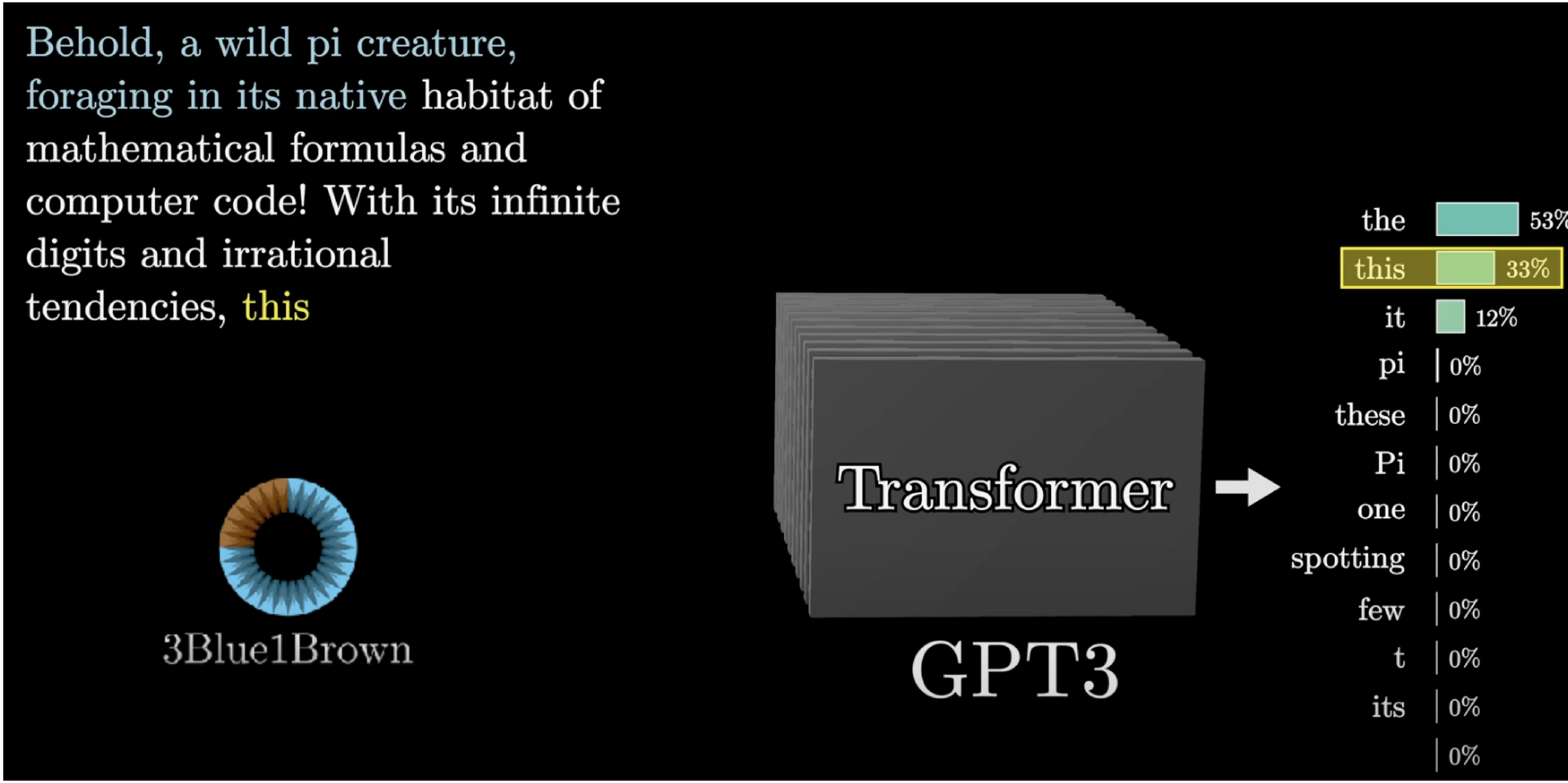
$$N \to N/d.$$

Federated Learning

joint work with K. Liu, Y. Song & Z. Wang



Transformers



The **pure-attention hardmax transformer** is given by

$$z_i^{k+1} = z_i^k + \frac{\alpha}{1+\alpha} \frac{1}{|\mathcal{C}_i(Z^k)|} \sum_{j \in \mathcal{C}_i(Z^k)} \left(z_j^k - z_i^k \right)$$

where

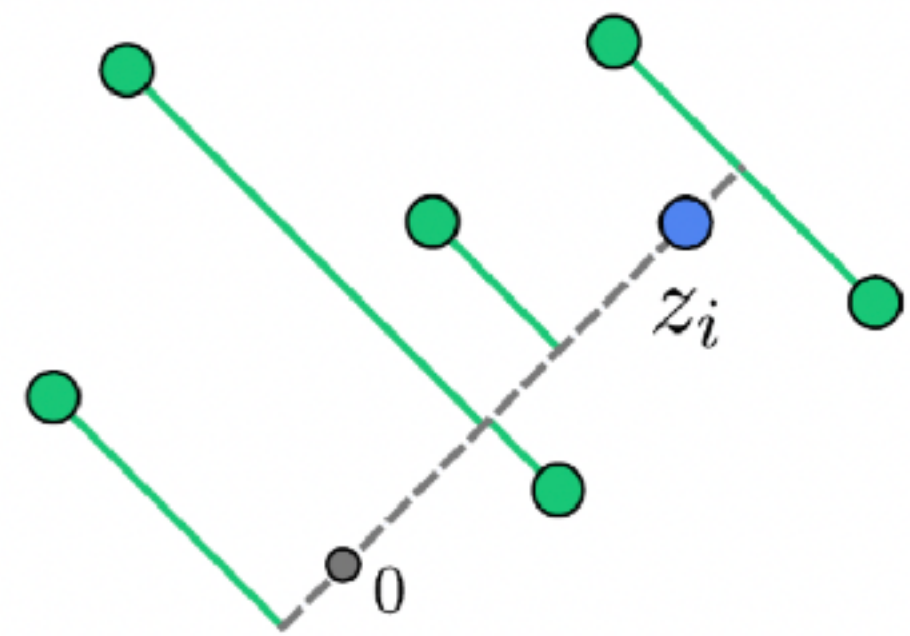
$$\mathcal{C}_i(Z) = \left\{ j \in [n] : \langle \; z_i, z_j \rangle = \max_{\ell \in [n]} \langle \; z_i, z_\ell \rangle \right\}.$$

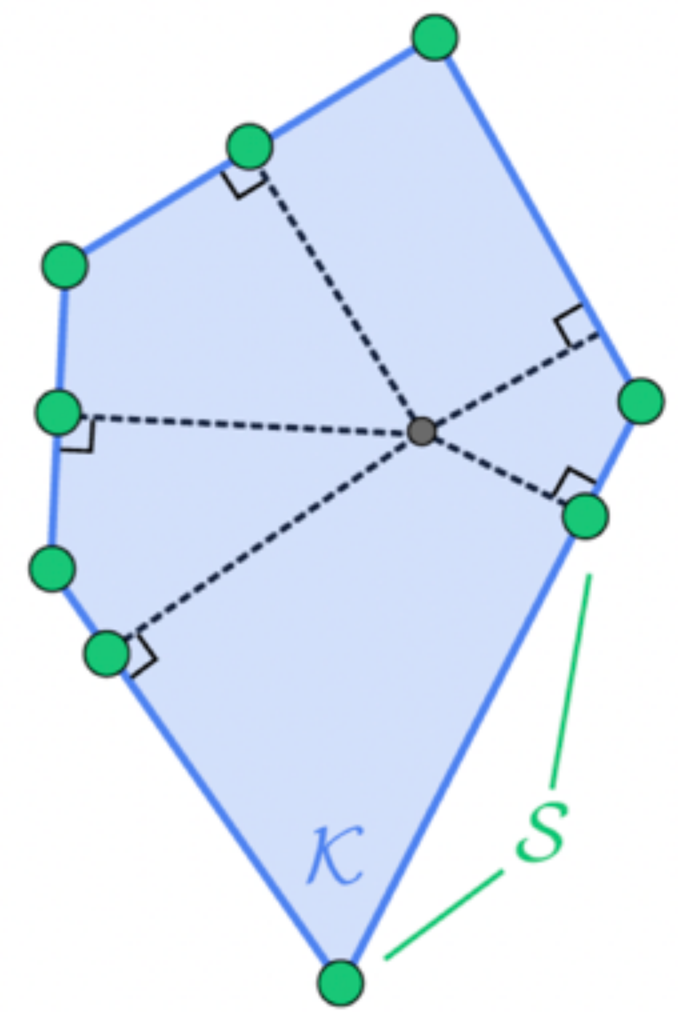
Theorem (Emergence and identification of clusters)

Let $z_1^0, \ldots, z_n^0 \in \mathbb{R}^d$ be nonzero. There exists a finite set $\mathcal{S} = \{s_1, \ldots, s_p\} \subset \mathbb{R}^d$, $p \leq n$, such that

$$z_i^K \to s_j \quad \text{as} \quad K \to \infty.$$

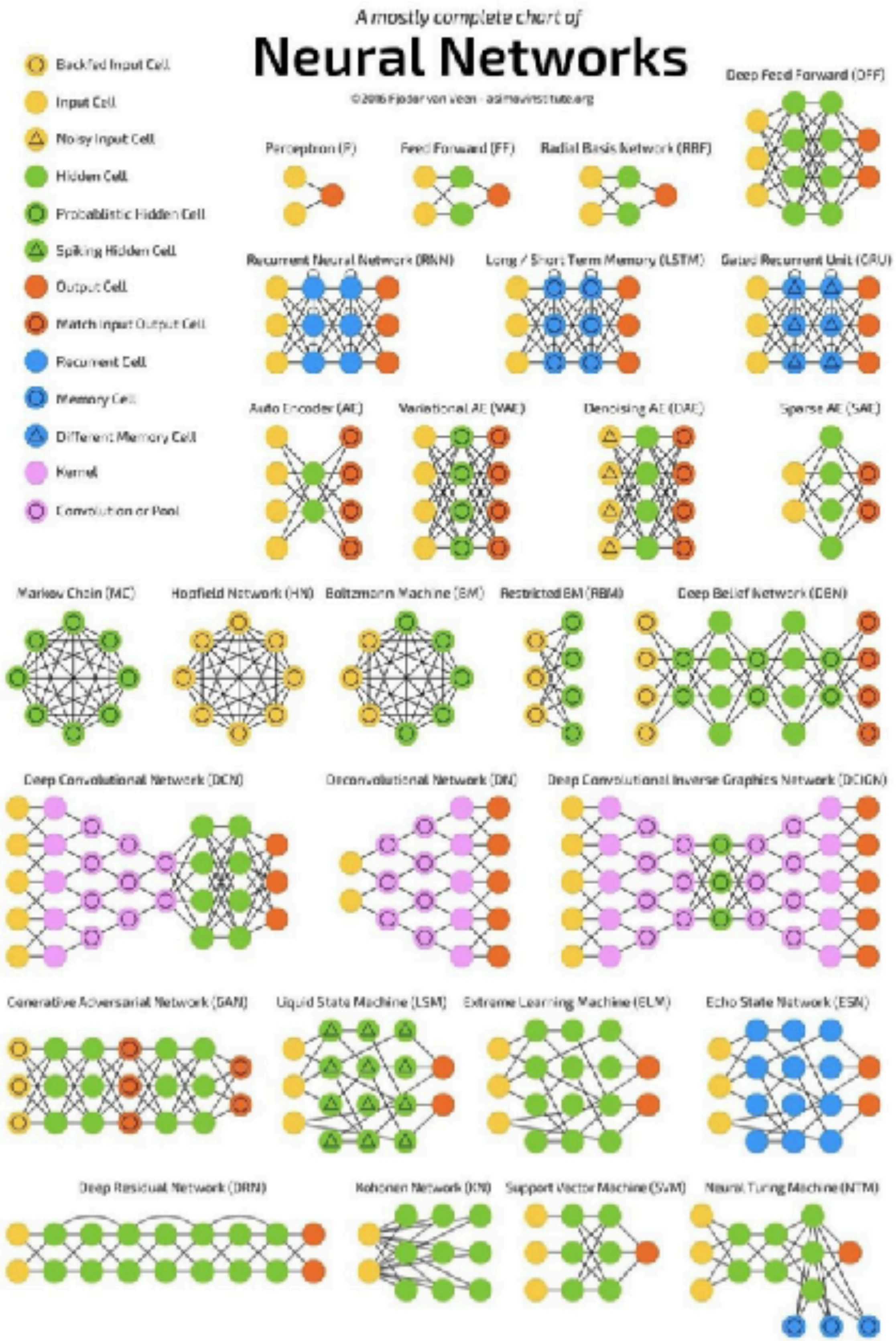
Moreover, the first $m \leq p$ elements of $\mathcal{S}$ are the vertices of a convex polytope which are the leaders, and the remaining elements are the projections of the origin onto the faces of such polytope.





Promising Field

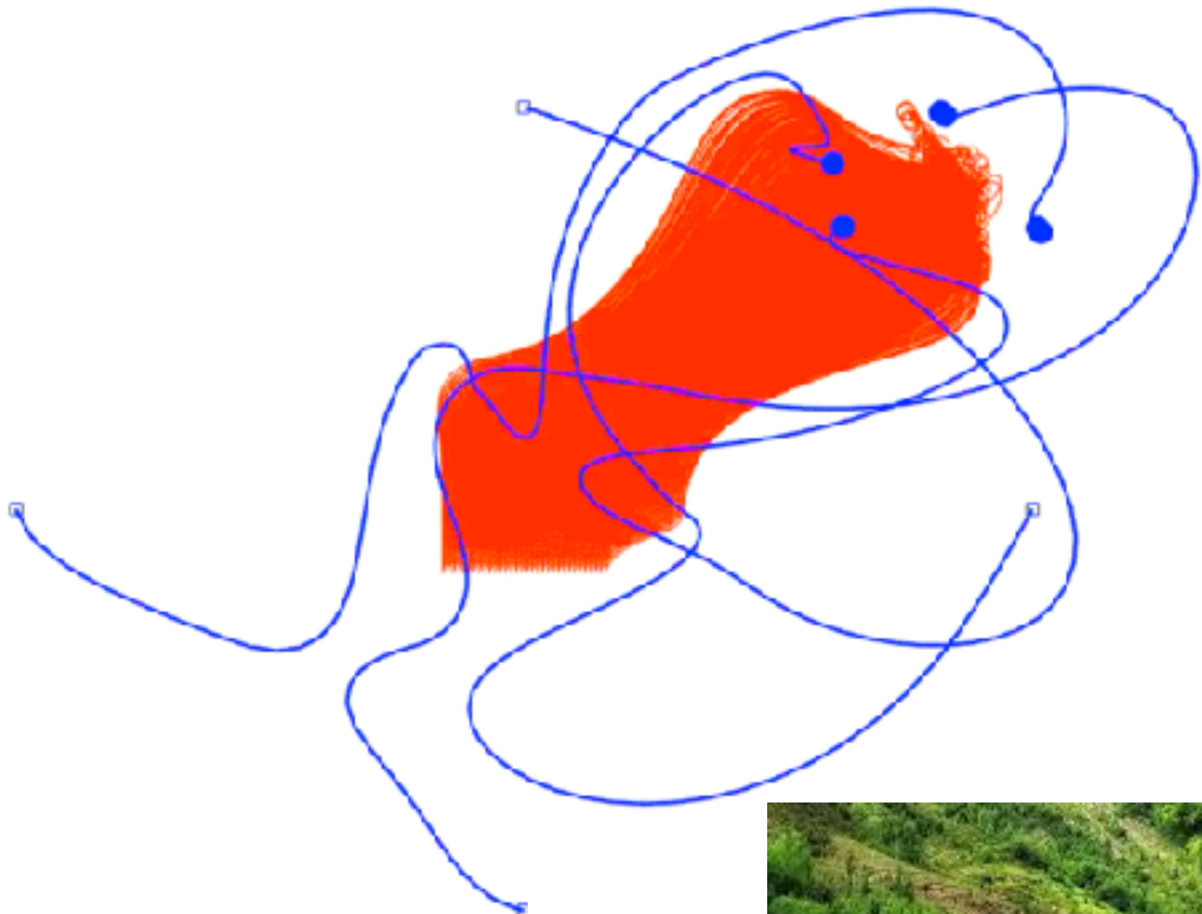
Lots to do

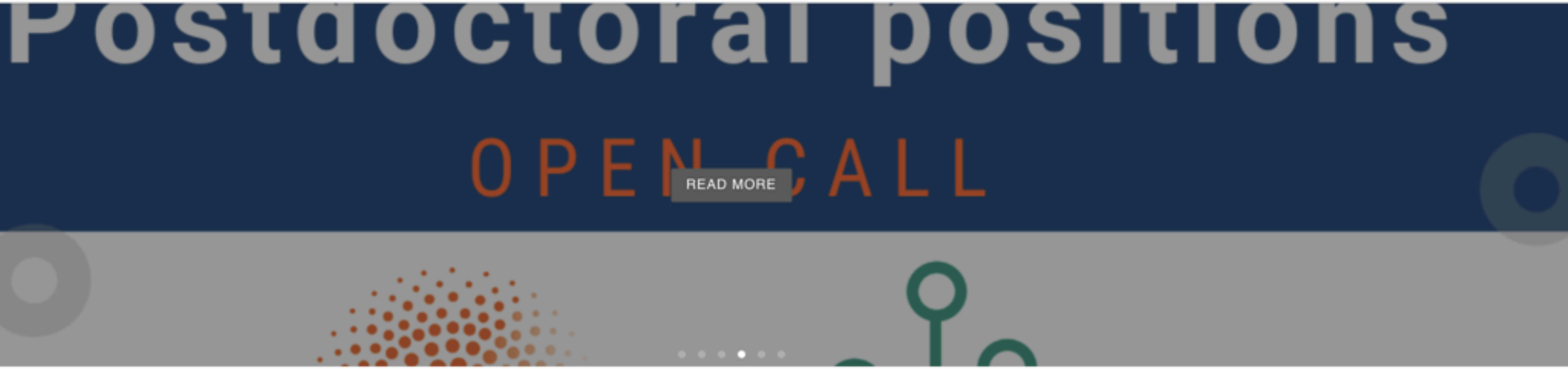

A mostly complete chart of
Neural Networks
Backfed Input Cell
Input Cell
Noisy Input Cell
Hidden Cell
Probablistic Hidden Cell
Spiking Hidden Cell
Output Cell
Match Input Output Cell
Recurrent Cell
Memory Cell
Different Memory Cell
Kernel
Convolution or Pool
Perceptron (P)
Feed Forward (FF)
Radial Basis Network (RBF)
Deep Feed Forward (DFF)
Recurrent Neural Network (RNN)
Long / Short Term Memory (LSTM)
Gated Recurrent Unit (GRU)
Auto Encoder (AE)
Variational AE (VAE)
Denoising AE (DAE)
Sparse AE (SAE)
Markov Chain (MC)
Hopfield Network (HN)
Boltzmann Machine (BM)
Restricted BM (RBM)
Deep Belief Network (DBN)
Deep Convolutional Network (DCN)
Deconvolutional Network (DN)
Deep Convolutional Inverse Graphics Network (DCIGN)
Generative Adversarial Network (GAN)
Liquid State Machine (LSM)
Extreme Learning Machine (ELM)
Echo State Network (ESN)
Deep Residual Network (DRN)
Kohonen Network (KN)
Support Vector Machine (SVM)
Neural Turing Machine (NTM)

CoDeFeL

CONTROL FOR DEEP AND FEDERATED LEARNING

Bridging two neighbouring fields

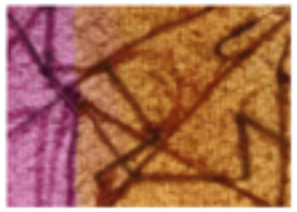




Meet us!
About us, our team

Dynamics, Control, Machine Learning and Numerics
Mathematics is everywhere; we show you how!

Upcoming events
Join our next FAU DCN-AvH seminar, workshop, ...

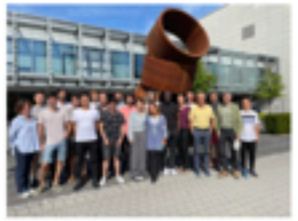


Math to go!
Looking to re-play or missed an event? Find it here!

We are the **Chair for Dynamics, Control, Machine Learning and Numerics – Alexander von Humboldt Professorship**.
Located at Friedrich-Alexander-Universität, Erlangen-Nürnberg, a beautiful bavarian region in Germany, FAU DCN-AvH is co-funded by Alexander von Humboldt Foundation and led by Prof. Dr. DhC. Enrique Zuazua
Read more about us

Working actively in the broad area of Applied Mathematics and Machine Learning, we are passionate people developing and applying methods of Mathematical and Computational Mathematics to model, understand, design and control the dynamics of various phenomena arising in the interface of Mathematics with Engineering, Physics, Biology and Social Sciences.

Our Head
Prof. Enrique Zuazua is the Head of the Chair for Dynamics, Control, Machine



Our Team
We believe in people and the unlimitless power of a multicultural, open and

Maths to the World!
We do research to make a better world. Our passion led us here to give Society

Conclusions and Perspectives

- Fantastic horizon for mathematical research and in particular for the fields of Control and Optimization:
  - Training
  - Generalization
  - Generation
  - Complexity: Width/Depth
  - Dimensionality and probabilities and statistics.
  - Federated Learning
  - ....

- Digital Twins Methodologies pose specific challenges:
  - Scalability / Adaptivity / Personalized / Goal oriented (Model Predictive Control?)
  - Control of control for DT modelling
  - Reliability / generalization / synthetic data
  - Merging with Physics and Mechanics

Thank you for the invitation and attention