# Reservoir Riddles:
# New Directions in ESN Research

Herbert Jaeger

International University Bremen (IUB)

# Overview

1. The ESN model
2. Current R&D at IUB/Fraunhofer
    1. Learning to work with ESNs
    2. Optimization of global control parameters
    3. Continuous-time ESNs
    4. Coping with time-warped data
    5. Bi-directional ESNs
3. The Reservoir Riddle
    1. The Problem(s)
    2. Non-solutions
    3. A partial solution
    4. Some speculation

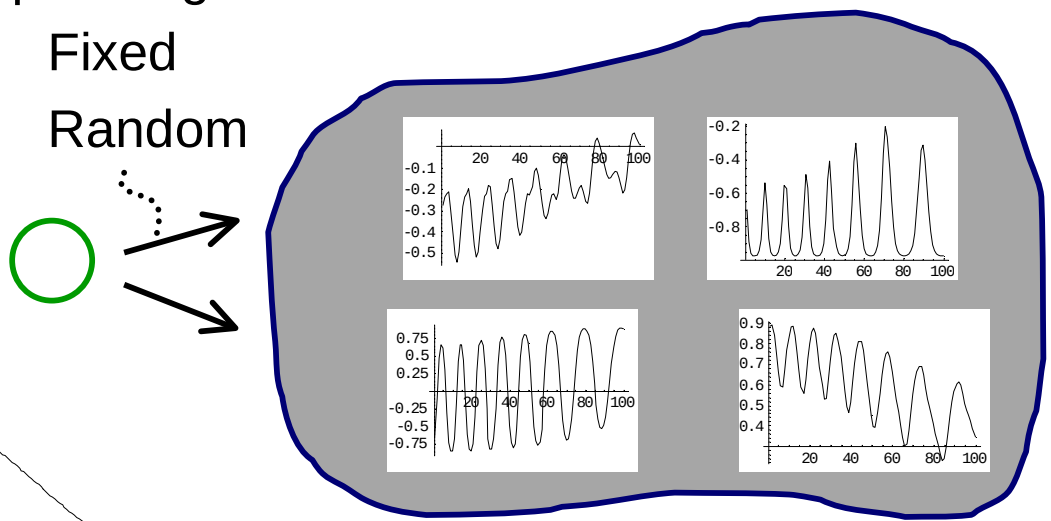
# 1. The ESN/LSM principle

Input weights:

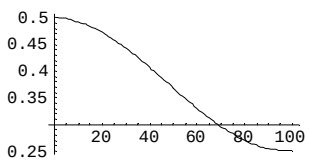
- Fixed
- Random

Output feedback weights (optional):

- Fixed
- Random



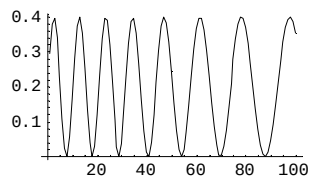
Output weights learnt by linear regression



Input signal

"Dynamical Reservoir":

- Large
- Fixed
- Random
- Recurrent



Output (teacher) signal

# 2 Current R&D

# 2.1 Learning to train ESNs

**Problem:** novices take weeks to months to achieve good ESN training results, or quit prematurely

**Reason:** it's apparently not so easy to get a "gut feeling" for excited nonlinear dynamics

- judge degree of nonlinearity
- judge short-term memory requirements
- negotiate on the bias-variance scale, master noise regularization

**Non-solution:** refer novice to techreport[1)] with "tricks of the trade"

**Solution:** develop Matlab tutorial with didactic exercise cases

1) http://www.faculty.iu-bremen.de/hjaeger/pubs/ESNTutorialRev.pdf

# 2.2 Optimizing global control parameters

**Problem:** network size, spectral radius, input scaling, input shift, noise level are important global learning controls

**How ESN experts "solve" it:** tuning by hand

**Novices**: can't do it that way!

**Better solution:** automated search for locally optimal global parameters by empirical gradient descent (student project)

**Desirable solution:** thorough theoretical understanding (open research issue)

# 2.3 Continuous-time ESNs

**Problem:** discrete-time OOMs don't work well for slow dynamics

**Approach:** use (Euler approximation of) leaky integrator neuron ESN, where $\Delta = \Delta t/c$, $a$ leaking rate, $c$ time constant:

$$\mathbf{x}(n+1) = (1 - a\Delta)\mathbf{x}(n) + \Delta\ f(\mathbf{W}_{in}\mathbf{u}(n+1); \mathbf{W}\mathbf{x}(n))$$

**Difficulty 1:** guarantee echo state property

**Difficulty 2:** prevent sigmoid $f$ from saturation

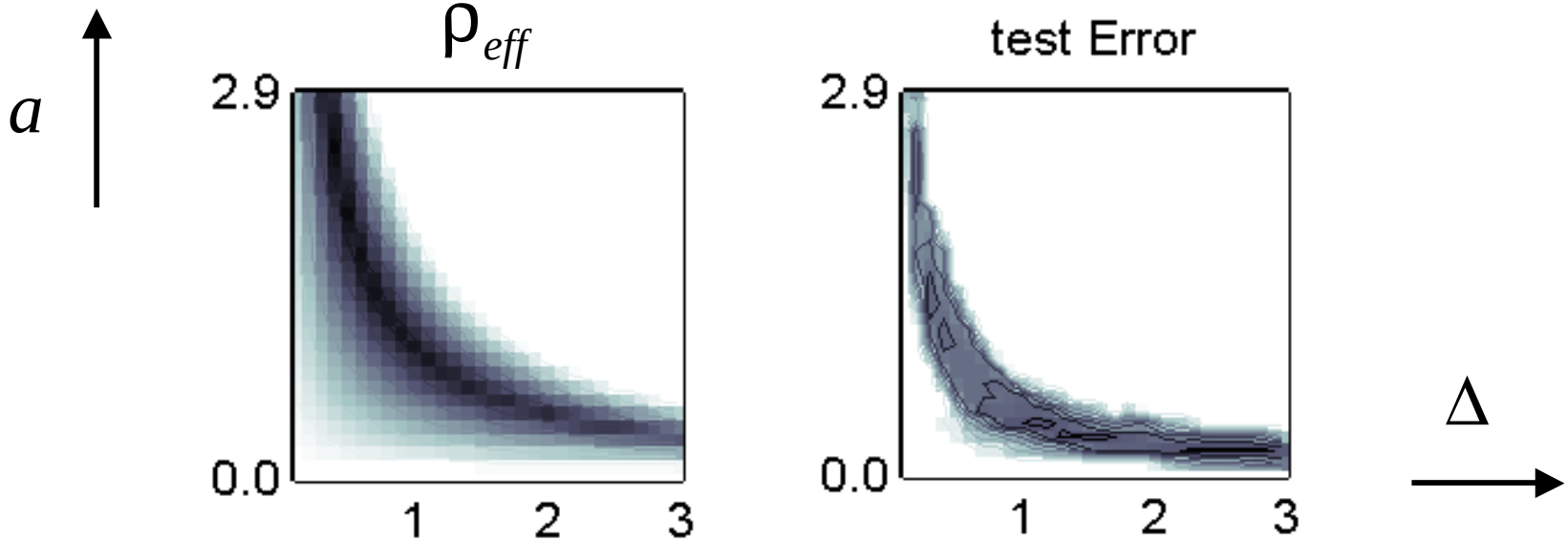
**Solution:** observe algebraic constraints detailed out in techreport[1)]

1) http://www.faculty.iu-bremen.de/hjaeger/pubs/ESNTutorialRev.pdf

**Difficulty 3:** we must tune two more global control parameters (leaking rate $a$, ratio $\Delta = \Delta t/c$)

**Facilitation:** effects of $a$ and $\Delta$ can in practice be combined into single "effective spectral radius"

$$\rho_{eff} = specrad(\Delta \mathbf{W} + (1 - a\Delta)\mathbf{I})$$



# 2.4 Coping with time-warped data

**Problem:** in handwriting and speech recognition, target patterns occur in time-warped variants

**Approach**[1)]**:** use (Euler-discretized) continuous-time ESNs, adapt old idea[2)] : adjust $\Delta$ dynamically according to metric input distance:

$$\Delta_n = K \left\| \mathbf{x}(n) - \mathbf{x}(n-1) \right\|$$

**Difficulty 1:** how to choose the scaling constant $K$?

**Heuristic solution:** first optimize for fixed $\Delta$, then set $K$ such that $E[\Delta_n] = \Delta$ .

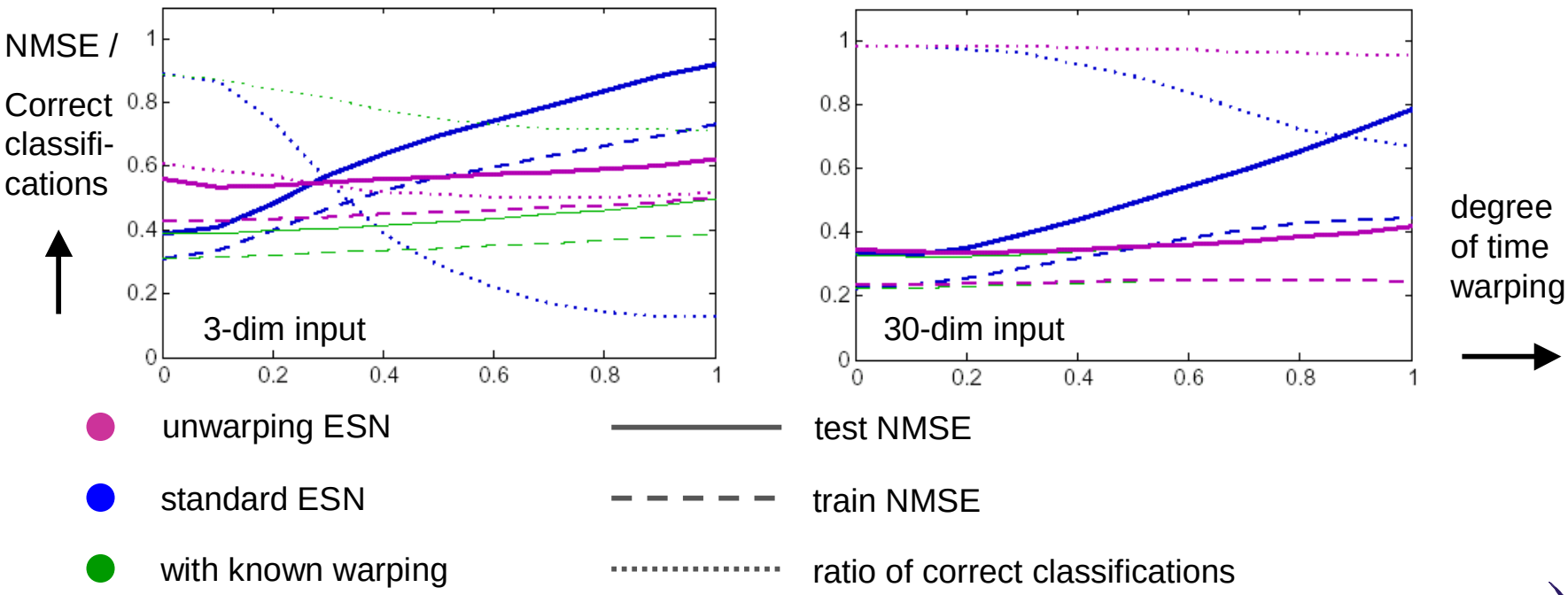
1) Joint project with Planet AG, Raben-Steinfeld, Germany

2) G-Z Sun, H-H Chen, Y-C Lee. Time warping invariant neural networks. NIPS 5 (1993)

**Difficulty 2:** stiff input dynamics → very large $\Delta_n$ → unstable (Euler discretized) network dynamics

**Heuristic solutions:** soft-bound $\Delta_n$ by wrapping with sigmoid, *or* interpolate high-slope input segments

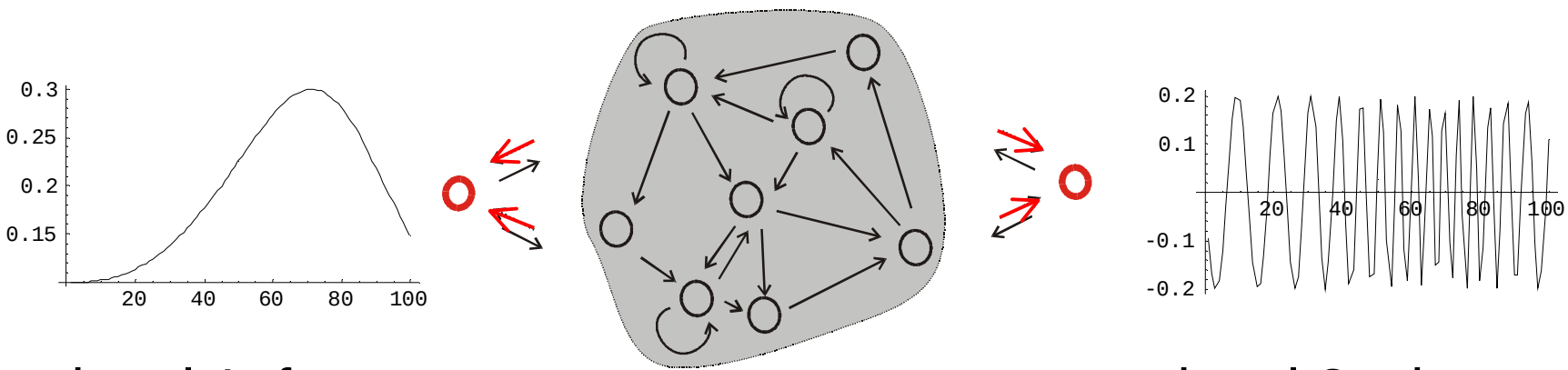
**Result**[1)]**:** (on synthetic dynamic pattern detection task )




1) Lukosevicius / Popovici / Jaeger / Siewert, Time Warping Invariant Echo State Networks (subm.)

# 2.5 Bidirectional dynamics: frequency generating/measuring device

Teaching signals: sine wave and its frequency (N = 1000, here section with N = 100)

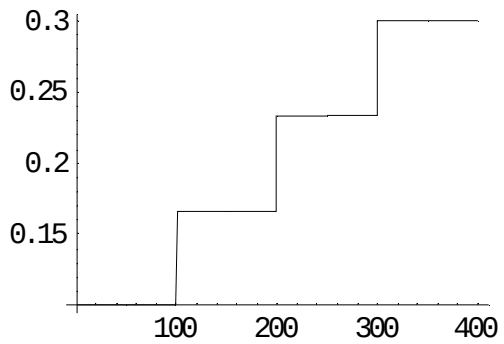



signal 1: frequency

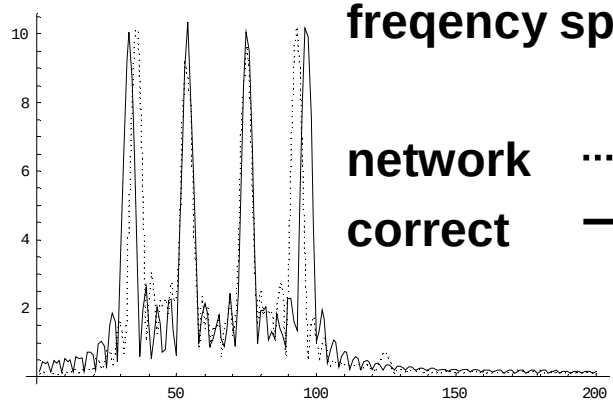
signal 2: sines

# Trained ESN used as **frequency generator**:

**input**



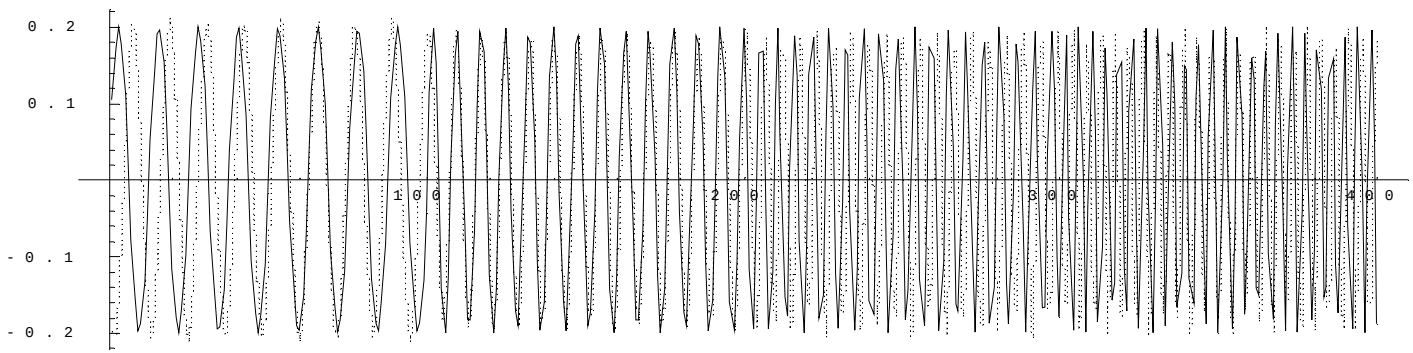

**freqency spectrum**




**network** .....

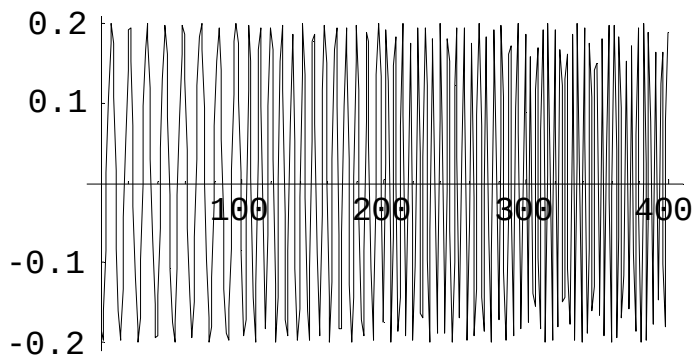
**correct** ——

**network output** ..... **correct sine signal** ——



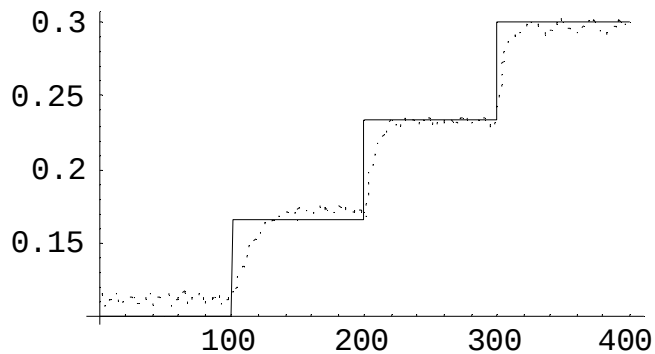
Trained ESN used for **frequency measurement**:

**input**




**network output** .....

**correct frequency** ——

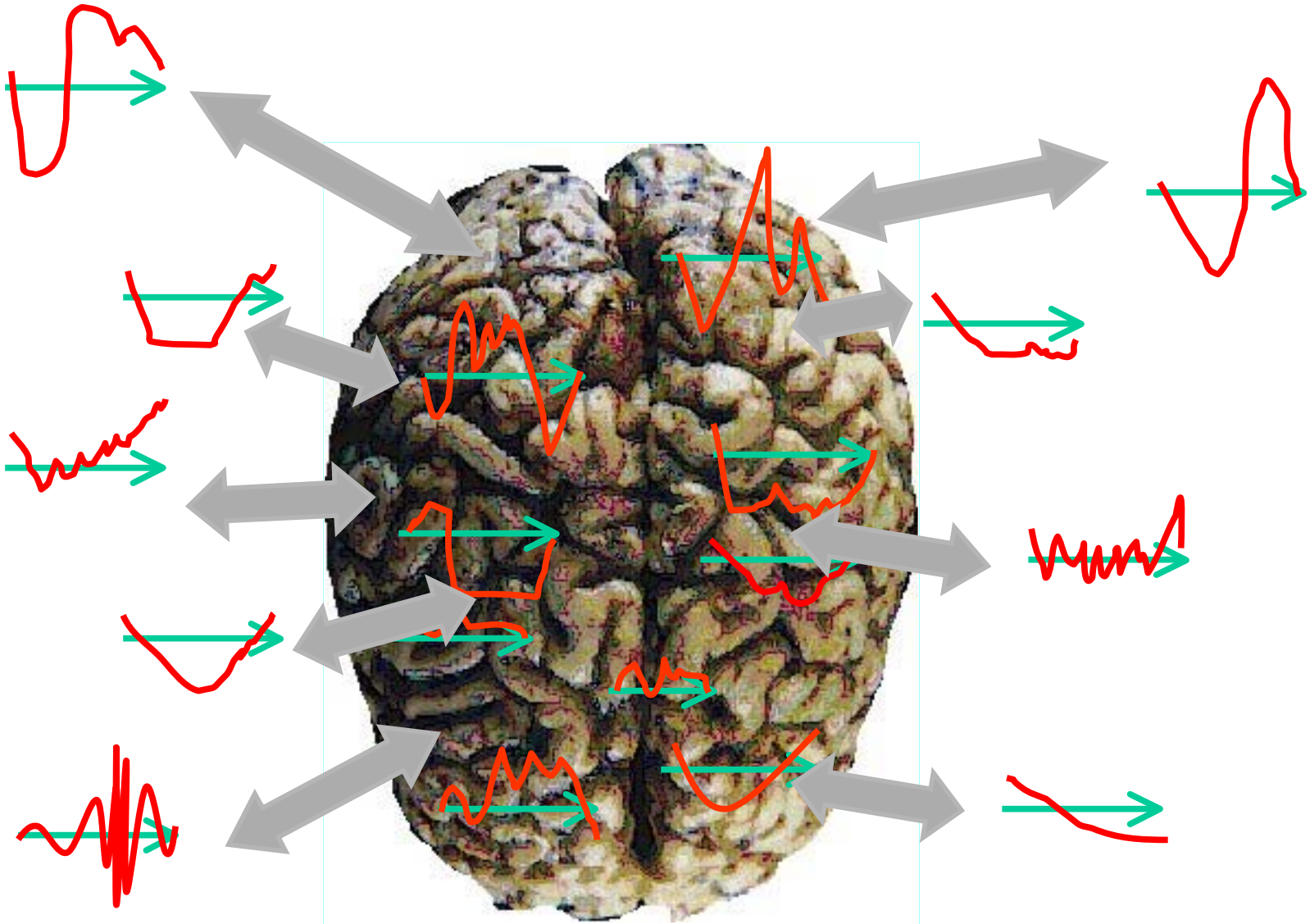


**Also tried:** bidirectional controller / predictor device

**In our brains:** multi-directional, cross-modal association of dynamic patterns, e.g.

- visual display of dance ↔ music score
- limb motion (proprioceptive) ↔ limb motion (motor command) ↔ music score
- listening to speech ↔ watching lip motion
- motor command generation ↔ expectation of proprioceptive feedback
- coordinating actuators
- etc.

# 3 The Riddle



# 3.1 The problem(s)

**Problem 1:** On some problems (ask Danil), ESN models inferior to BPTT or EKF models.

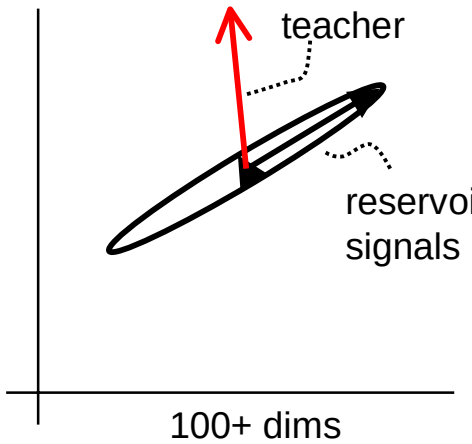
Reservoir doesn't offer the requisite signal components.

**Problem 2:** On some (deterministic) problems, trained output weights grow very large.

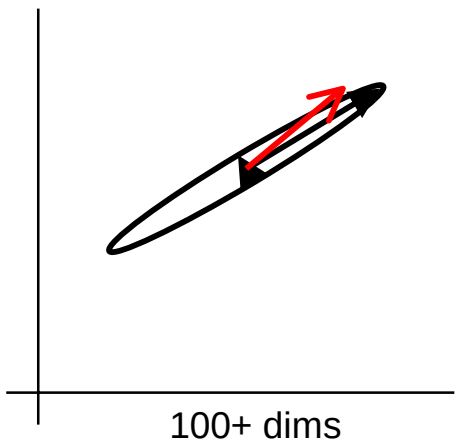
Relevant reservoir signal components are in "thin" directions of the linear reservoir signal space.

**Problem 3:** The reservoir signal correlation matrix $\left(E[x_i x_j]\right)$ typically has an extreme eigenvalue spread (order of 1E14 and higher) $\rightarrow$ impossible to use cheap online learning by LMS, impossible to use analog VLSI or integer arithmetics for reservoir.

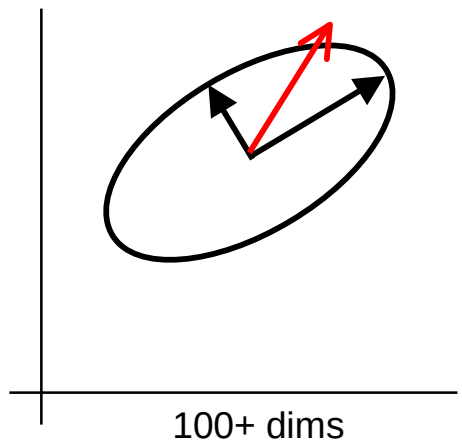
Reservoirs tend to have very similar unit signals due to mutual entrainment of unit's dynamics.



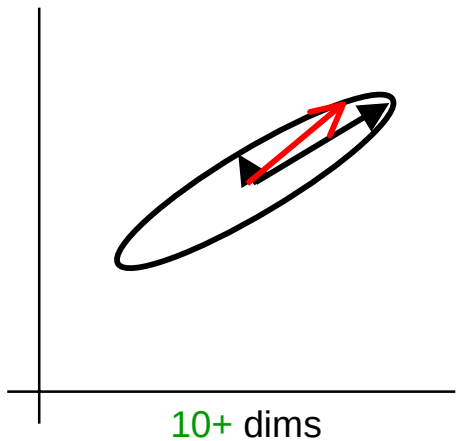

What we get




What we want...




What we want even more...




What BPTT or EKF gives us

**Wish, the first:** ESN reservoir signal space with flat spectrum (many orthogonal reservoir signal components of similar power),

and / or

**wish, the second:** ESN reservoir signal space which neatly embraces teacher signal (small teacher component orthogonal to reservoir signal span).

# 3.2 Non-solutions for flat spectrum

- Unsupervised "preshaping" of reservoir by anti-Hebbian learning

  Spectrum stays same or even spreads further!

- Unsupervised "preshaping" of reservoir by Hebbian learning
  Spectrum spreads further!

- Use non-random reservoir connection graphs (scale-free, small-world (= clustered), hierarchical, grid topology... )

  No significant effects on spectrum found![1)]

- Train reservoir neurons (with ESN rule) individually to orthogonal signals
  Doesn't work -- training one neuron disrupts
  training results of previously trained ones

**1) Benjamin Liebald: Exploration of effects of different network topologies on the ESN signal crosscorrelation matrix spectrum. Guided Research Thesis, IUB 2004**

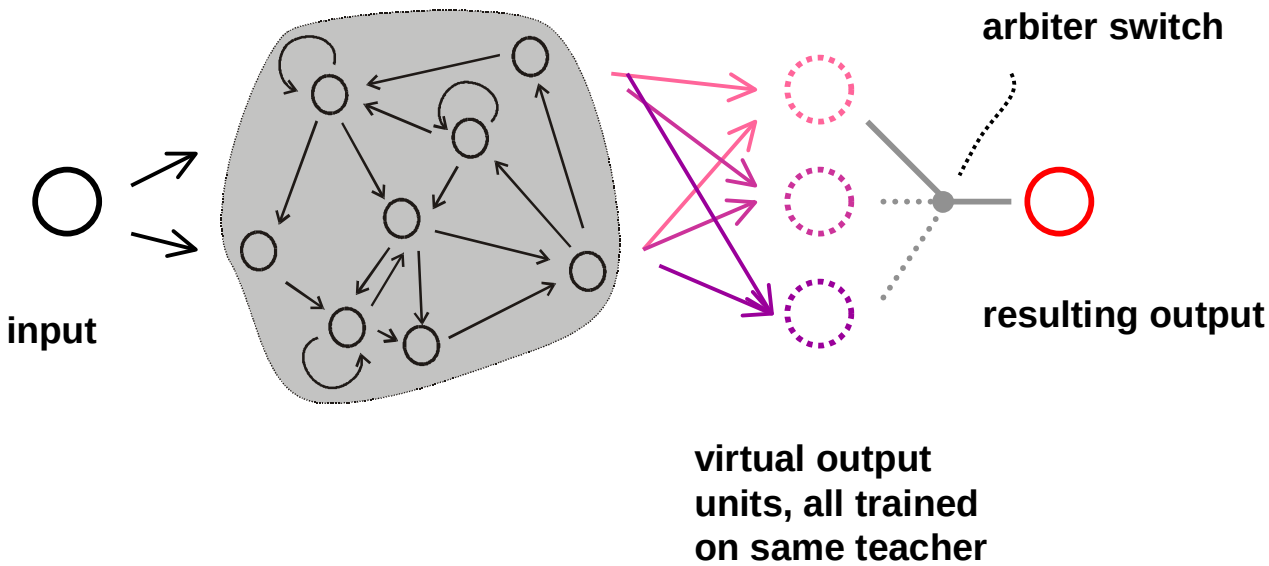
# Non-solutions for embracing teacher signal in reservoir dynamics

- Evolutionary algorithms, "traditional" learning algorithms like BPTT or EKF

  Defies raison d'être for ESNs!

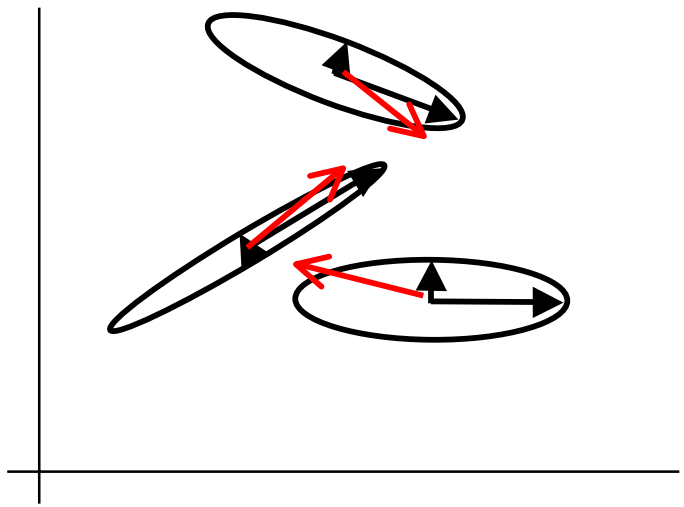
# 3.3 A partial solution

**Idea:** train several sets of output weights, switch between them by arbiter




**Switching criteria:** reservoir state partitioning, input partitioning, history-dependent criteria... Or train virtual output units competitively (did not work well so far)

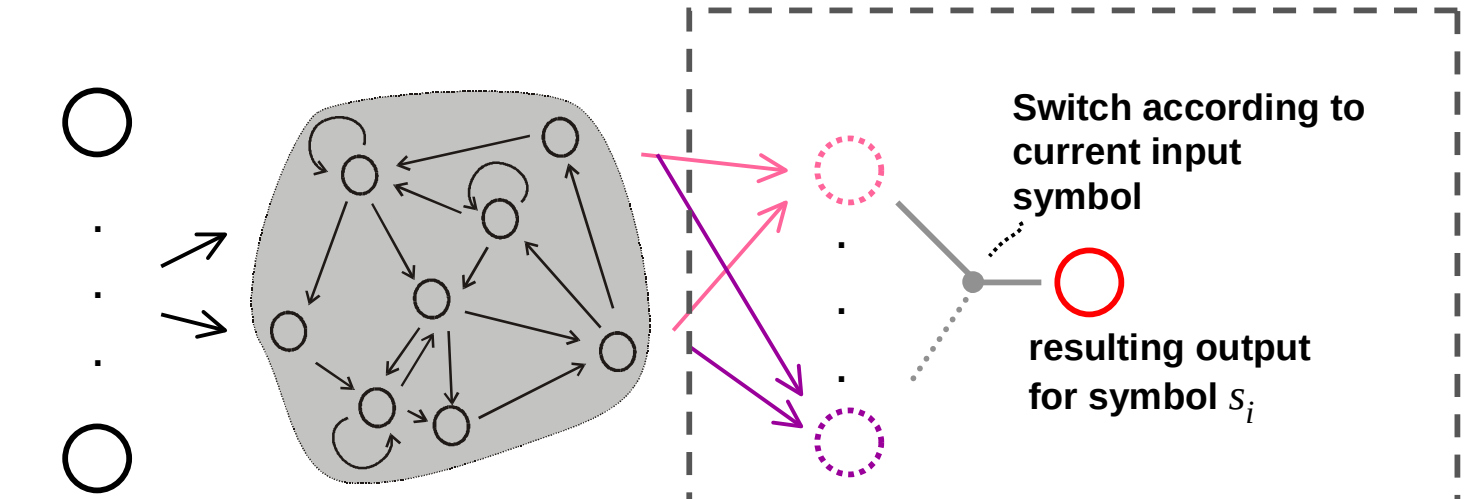
This decomposes reservoir and teacher signals in time -- has aspects of both flattening spectrum and embracing teacher signal.



# Case study: learning probability distribution of a belletristic text

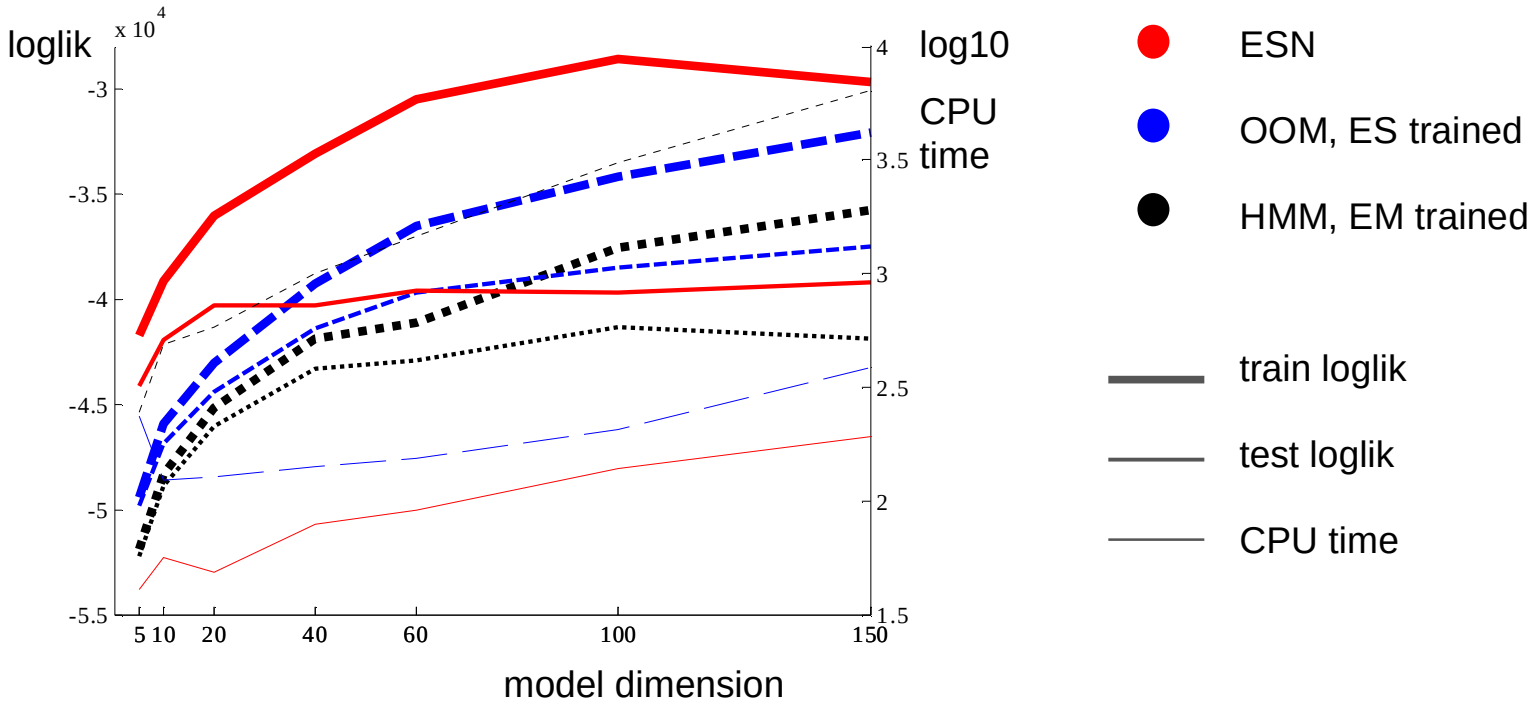
**Data:** $2 \times 21{,}000$ step 27-symbol sequence from Mark Twain short story

**Task:** Prediction of next symbol (yields distribution for next symbols)




This box replicated 27 times (for each symbol $s_i$), makes $N \times 27 \times 27$ trained weights

# Results




- ESN test performance better than HMM
- Best ESN test loglik reached with 20 dim reservoir, $20 \times 27 \times 27 =$ 14,580 trained weights (best HMM dim 100 = 12,700 parameters)
- 2 orders of magnitude shorter learning time than HMM

# 3.4 Some Speculation

**Assumption:** biological (vertebrate?) brains use ESN principle at some places

**Riddle:** how is flat spectrum / teacher embracement accomplished?

**Speculation:**

- component signals come from specialized and spatially separated regions
- global brain network structure (small-world, hierarchic, clustered, spatially organized) correlates with hierarchy of ± orthogonal feature detectors
- evolutionary pressure toward formation of ± orthogonal feature detectors
- no reservoir -- output unit distinction: brain = reservoir = trainable units

**Suggestion for artificial networks:**

- complex architectures needed: complex network topology, local learning
- no simple trick with ± homogeneous reservoir

# Thank you.

**(Interested in test-using an ESN Matlab toolbox?**

**Then tell me your contact coordinates right here.)**