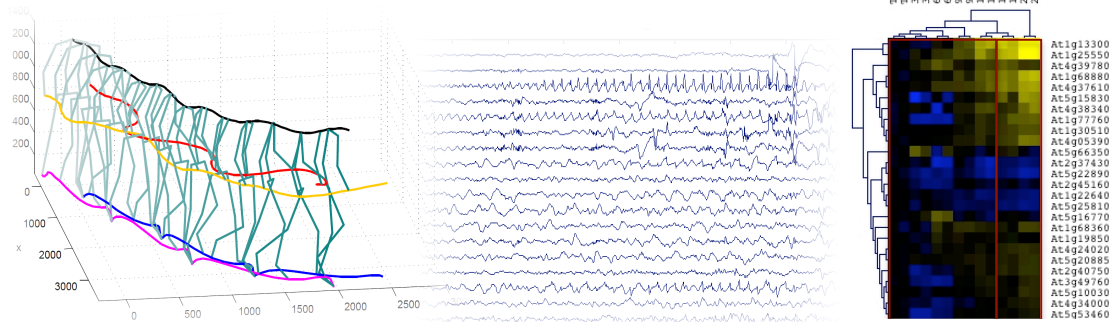
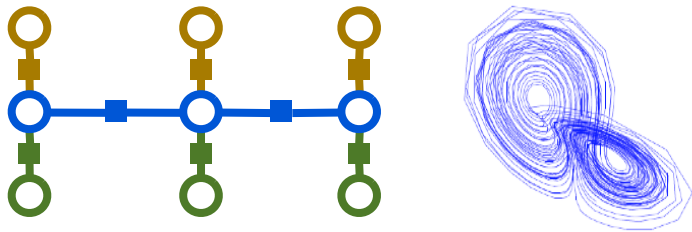


# Dynamical Factor Graphs (DFG) for Time Series Modeling



Piotr Mirowski

Advisor: Prof. Yann LeCun

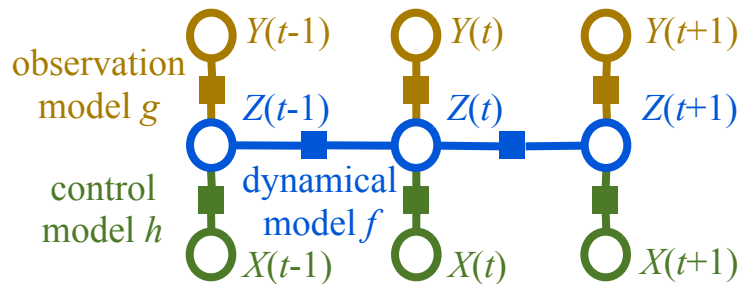
Courant Institute of Mathematical Sciences,  
New York University

[mirowski@cs.nyu.edu](mailto:mirowski@cs.nyu.edu)

<http://cs.nyu.edu/~mirowski>

# Motivation for DFG

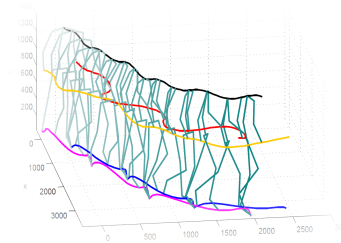
- State-space model



- Unknown **latent states**
- Potentially high-dimensional continuous **latent states**
- Highly nonlinear **dynamics** or **observation/control** models (convolutional net)
- Handle long sequences in linear time

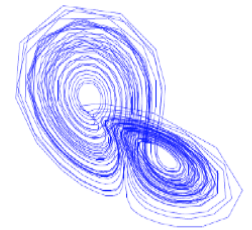
- Human MoCap

- Few visible markers
- Many (hidden) joint angles



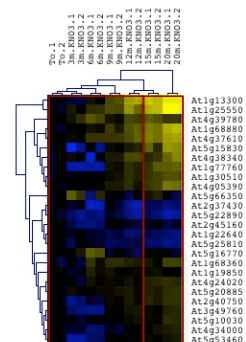
- Chaotic time series

- Unobserved data
- Complex, deterministic dynamics



- Gene regulation networks

- Missing micro-array data



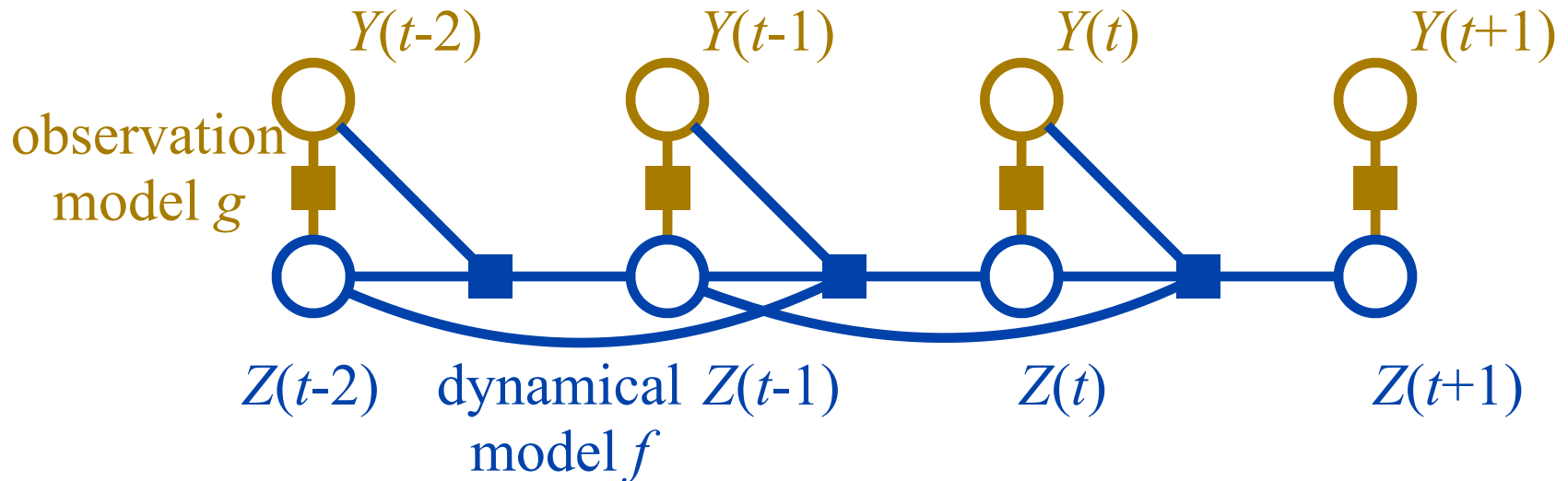
# Dynamical Factor Graph (DFG)

$n$ -dimensional **observed variable**

$$Y(t)$$

**observation model**

$$Y(t) \equiv g(Z(t)) + \omega(t) \quad \text{Gaussian noise}$$



$m$ -dimensional **latent variable**

$$Z(t)$$

time-embedded sequence  
of  $p$  **latent variables**

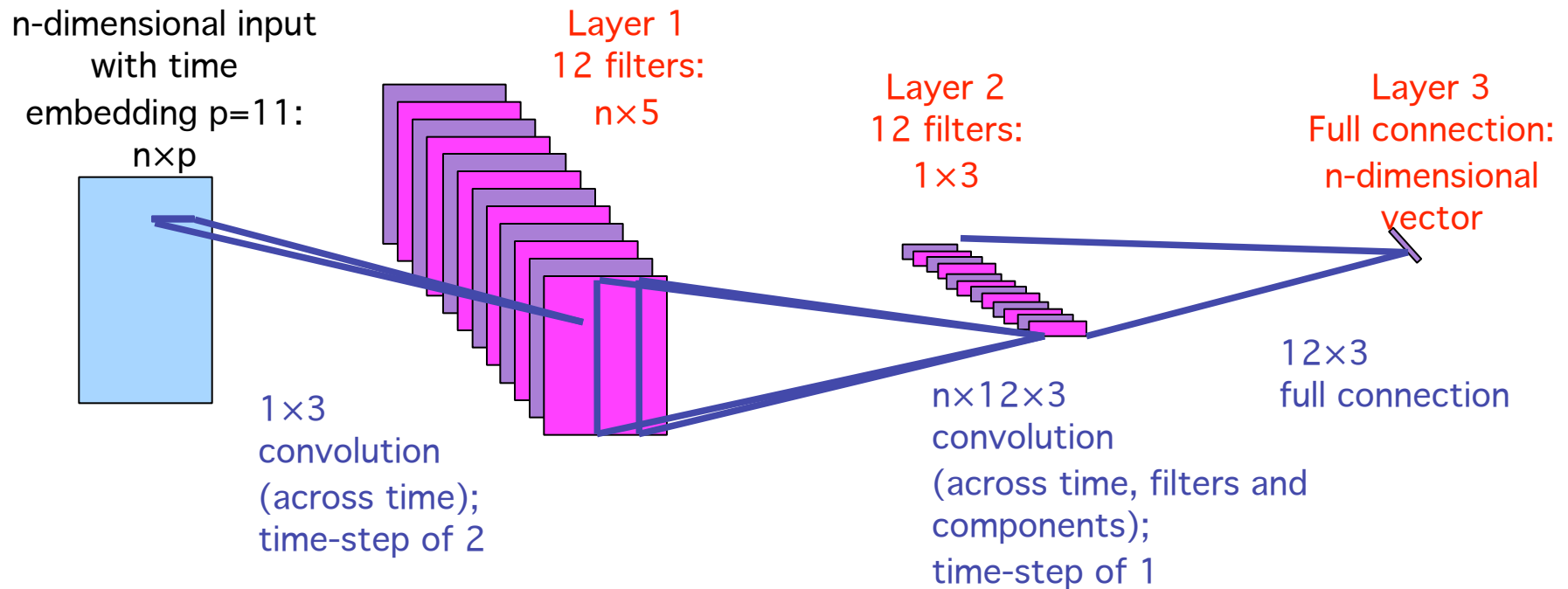
$$\mathbf{Z}_{t-p}^{t-1}$$

**dynamical model** ( $p^{\text{th}}$  order Markov)

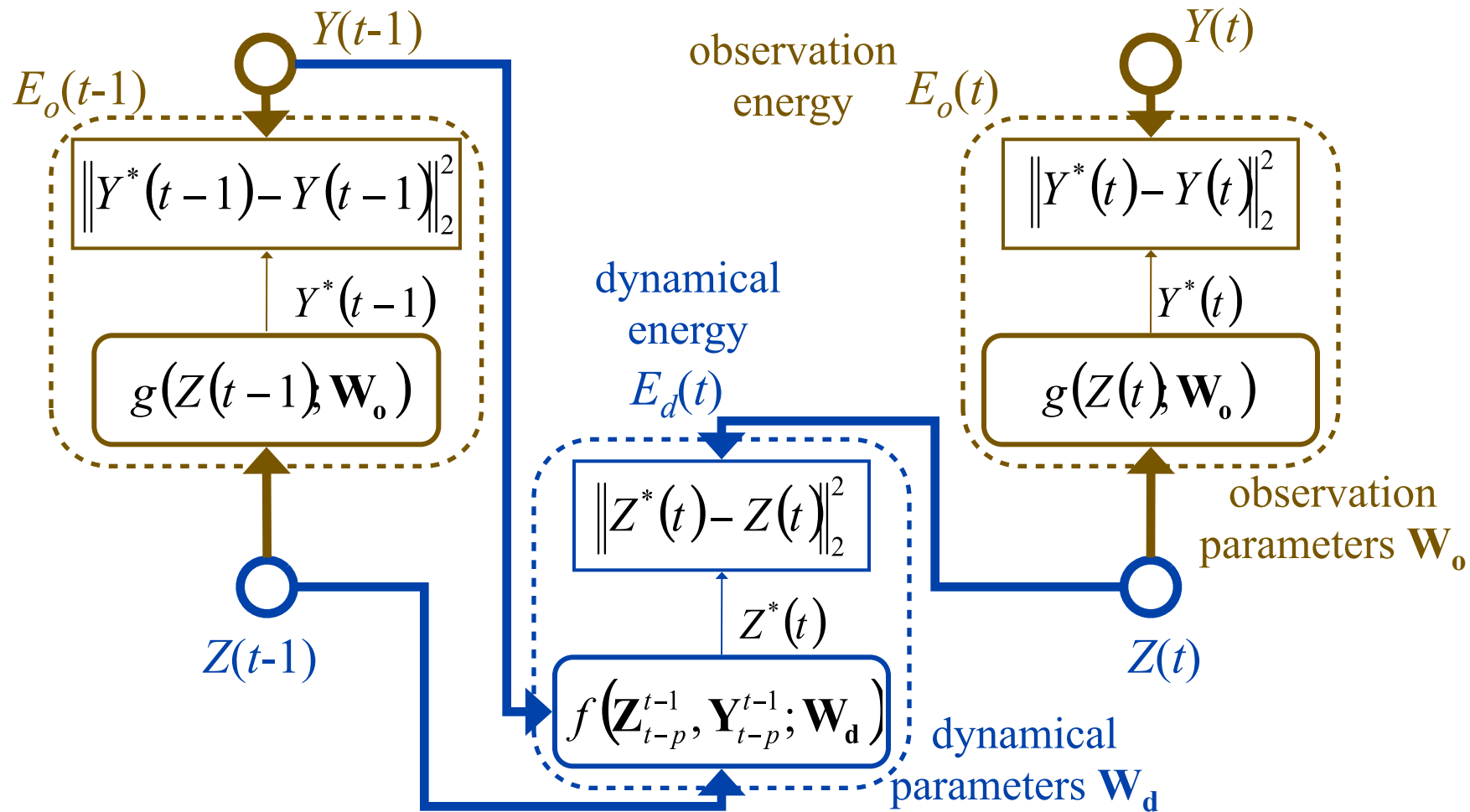
$$Z(t) \equiv f(\mathbf{Z}_{t-p}^{t-1}, Y(t-1)) + \varepsilon(t) \quad \text{Gaussian noise}$$

# Highly nonlinear factors: convolutional networks

- Higher-order nonlinearity than:
  - radial basis functions
  - single hidden-layer Perceptrons
- No closed-form optimization, but gradient-based



# Energy-based graph of a DFG



Learning and inference: deterministic gradient-based EM

# Smoothness penalty on latent variables

- Underconstrained **latent variable** inference
  - $L_2$ -norm smoothness penalty  
to reduce high-frequency noise  
(or Brownian Motion random walk assumption)

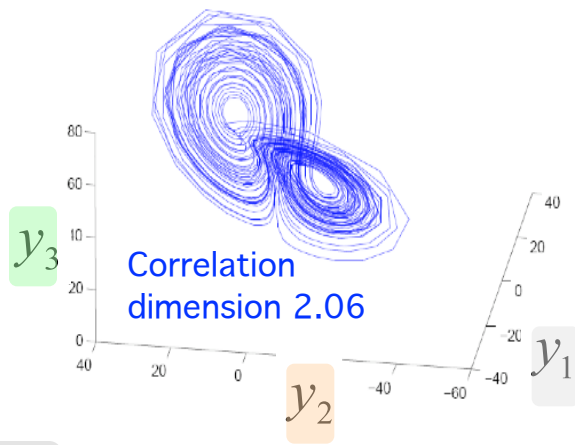
$$R_Z(\mathbf{Z}) = \sum_t \|Z(t) - Z(t+1)\|_2^2 = \sum_t \sum_{k=1}^m (Z_k(t) - Z_k(t+1))^2$$

- $L_1$ -norm smoothness penalty to tolerate “news”  
(or random walk assumption with likely “shocks”)
- Smoothness penalty can add to  
or replace the dynamical model

# Results<sup>1</sup>: Inferring the Lorenz chaotic attractor

## Data

Lorenz dynamical model



$$\frac{\partial y_1}{\partial t} = -16y_1 + 16y_2$$

$$\frac{\partial y_2}{\partial t} = 45.92y_1 - y_2 - y_1 \times y_3$$

$$\frac{\partial y_3}{\partial t} = y_1 \times y_2 - 2y_3$$

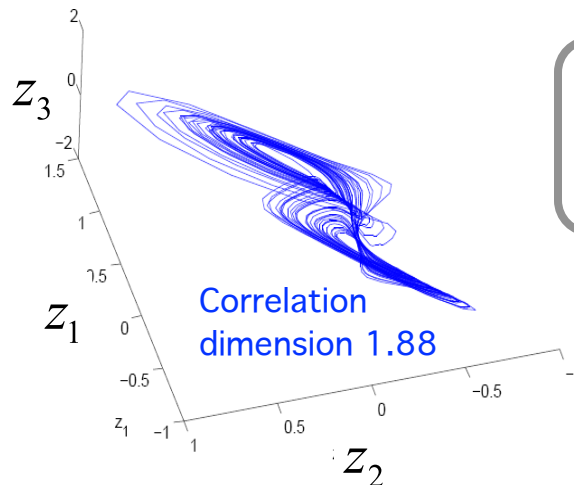
## Problem

**Partial observation**  $y(t) = y_1(t) + y_2(t) + y_3(t)$

Learn the DFG on train data  
with **latent variables** of dimension  $m=3$

## Results

**Latent state attractor** inferred on test data  
is similar to Lorenz attractor



**Lorenz attractor  
reconstructed  
on latent variables**

**1-step prediction  
error of -46.2dB  
smaller than in  
SVR (-41.6dB)**

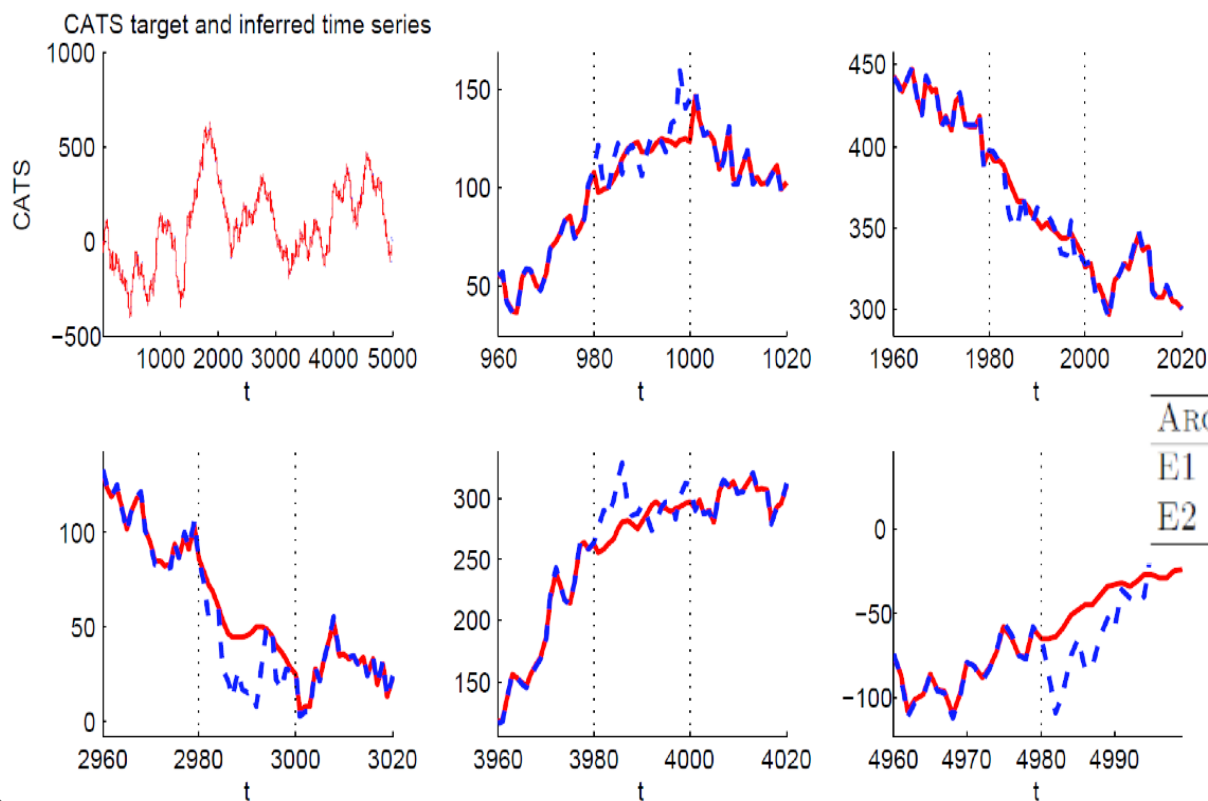
# Results<sup>2</sup>: CATS time series

## Data and problem

CATS time series prediction competition

Noisy **chaotic time series** (5000 points) with missing data (100 points)

## Results



**Predictions of 5  
segments  
of missing data  
beat the CATS  
benchmark**

ARCHITECTURE	KALMAN SMOOTHER	DFG
E1 (5 SEGMENTS)	408	390
E2 (4 SEGMENTS)	346	288



# Results<sup>3</sup>: Missing MoCap markers

## Data

- **Observations  $Y$ :**  
49-dimensional Motion Capture markers

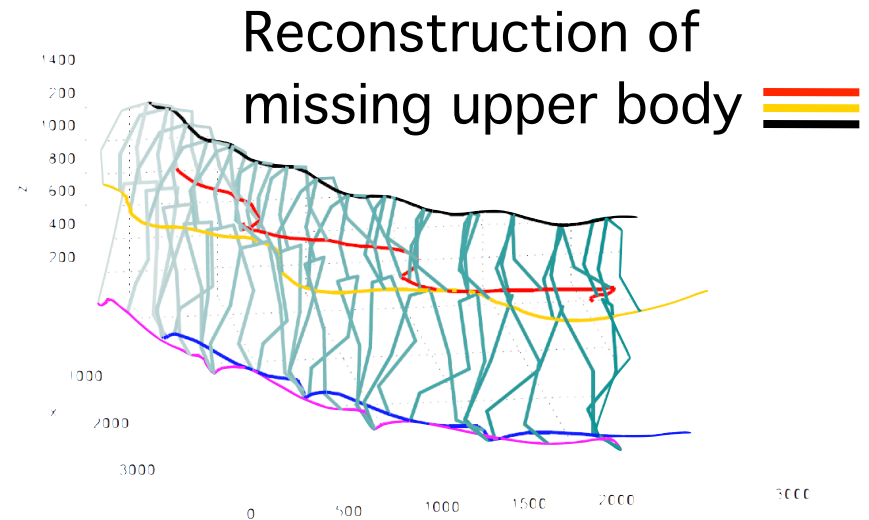
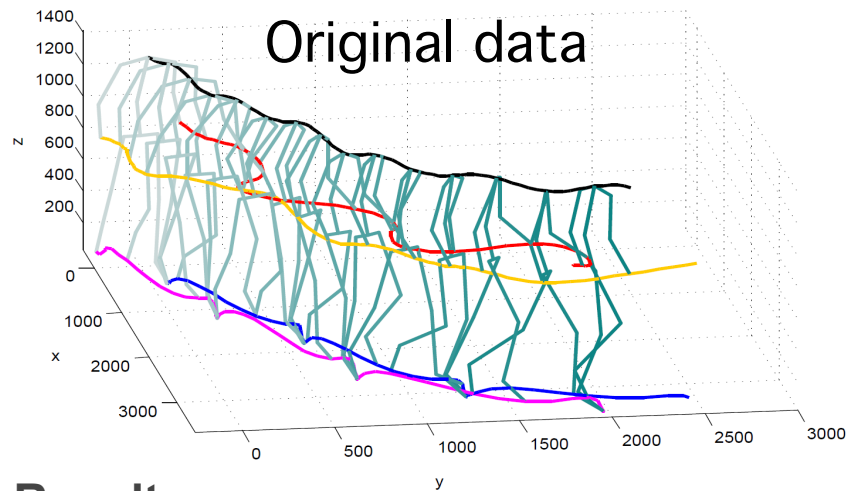
## Problem

- Model missing data (e.g. occlusions...)
  - Test sequence: 260 frames
  - 2 subsequences of 65 frames with missing data:
    - Left leg
    - Entire upper body

## Approach

- Infer **latent variables** (E-step) on test sequence,  
(without gradient from **missing  $Y_i(t)$** ), **generate  $Y$  from  $Z$**

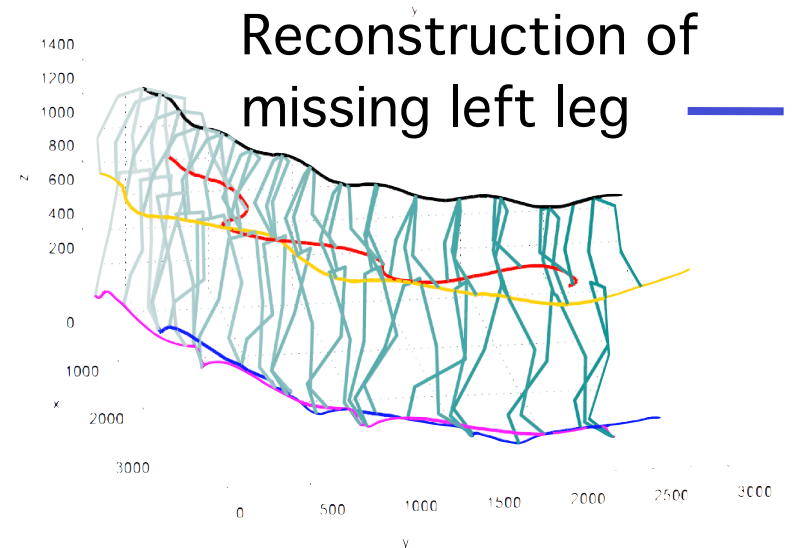
# Results<sup>3</sup>: Missing MoCap markers



## Results

**Lower NMSE than nearest neighbors;  
Inferred smooth, realistic motion**

METHOD	NEAREST NEIGHB. DFG	
MISSING LEG 1	0.77	0.59
MISSING LEG 2	0.47	0.39
MISSING UPPER BODY 1	1.24	0.9
MISSING UPPER BODY 2	0.8	0.48



# Results<sup>4</sup>: Learning Genetic Regulatory Networks (GRN)

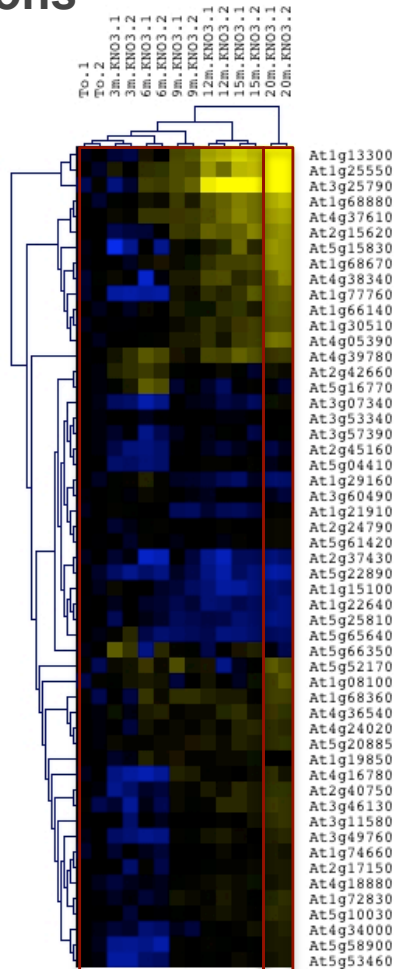
## Data and assumptions

Micro-arrays of protein expression levels for *Arabidopsis Thaliana* (plant), in reaction to  $\text{NO}_3^-$  sampled in time

$$Y(t) = f(Y(t-1)) + e(t)?$$

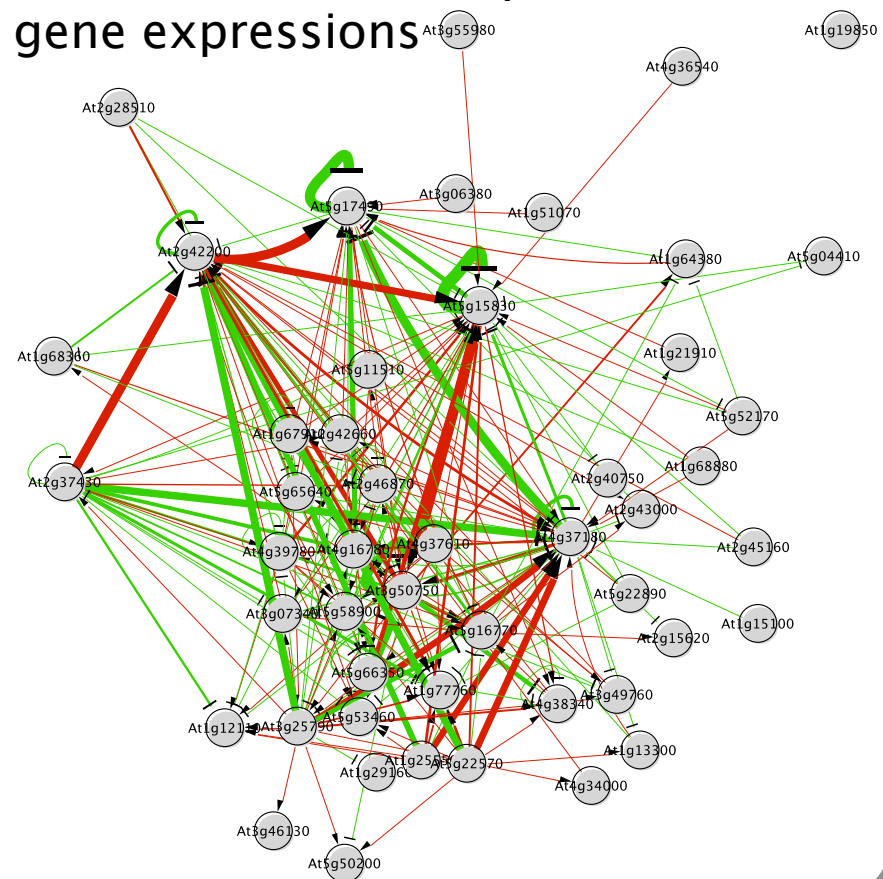
>22k genes,  
only 7 time points  
and 2 “replicates”

Clustering restricts  
to subset of  
76 genes



## Problem (and results)

Infer the GRN from dynamics on gene expressions



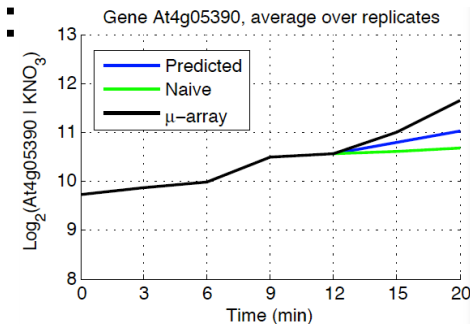
# Results<sup>4</sup>: Learning Genetic Regulatory Networks (GRN)

## Approach

- Linear dynamics  $f$  + Gaussian noise
- GRN sparsity through  $L_1$ -norm regularization of  $f$
- Latent variables  $Z$  = “smoothed” gene expression  $Y$   
i.e. identity observation model  $h$  + Gaussian noise
- Cross-validate on last or 2 last time points

## Results

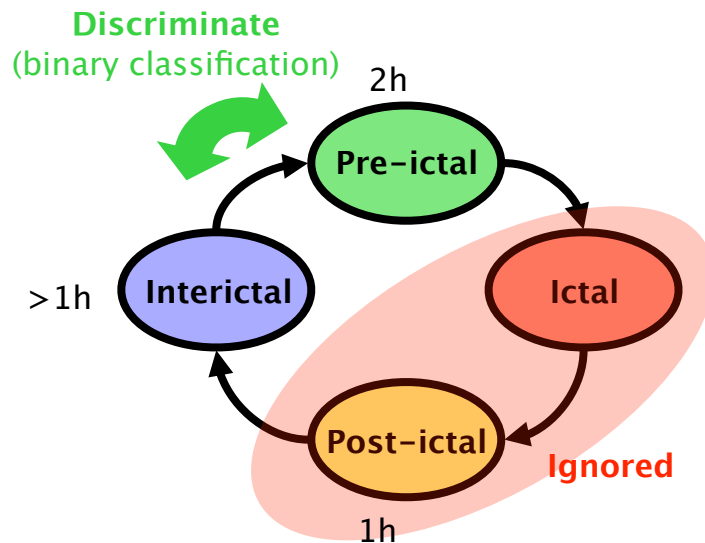
- Leave-out-last and leave-out-2-last predictions:  
70% +/- 3% correct direction  
(vs. 51% or 64% naive extrapolation)
- Incorporate prior knowledge about impossible gene interactions: set sparse connections in  $f$
- Impute micro-array data (infer missing values)



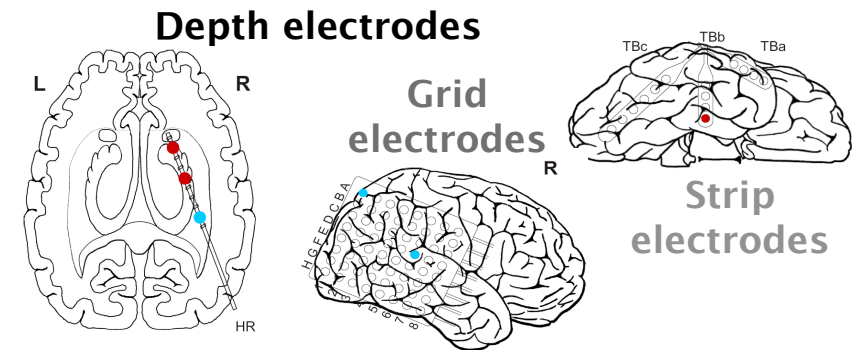
# Results<sup>5</sup>: Epileptic seizure prediction from EEG

## Problem

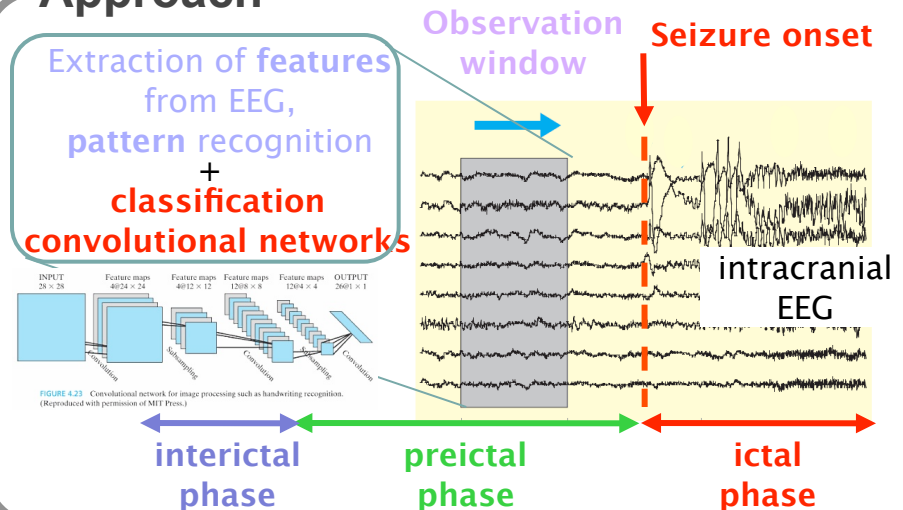
**Epilepsy**: a chronic illness  
Affects **1% of world population**  
Seizures are **harrowing**  
40% of patients medication **refractory**  
Avoid **resective surgery** treatment



## Data



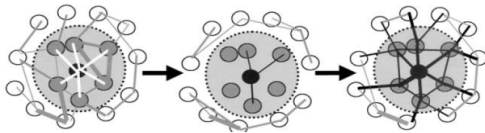
## Approach



# Results<sup>5</sup>: Epileptic seizure prediction from EEG

## Approach

Classify patterns of synchronization of EEG

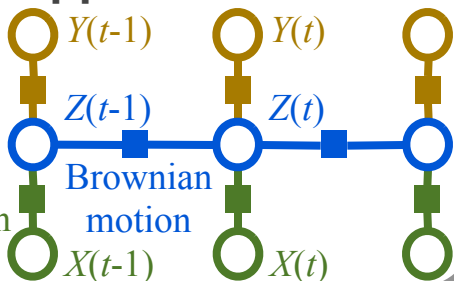


## Current results

21-patient public dataset:  
predicted **71% seizures**,  
**no false positives**,  
**>30min ahead** of seizure  
**Best results ever achieved**

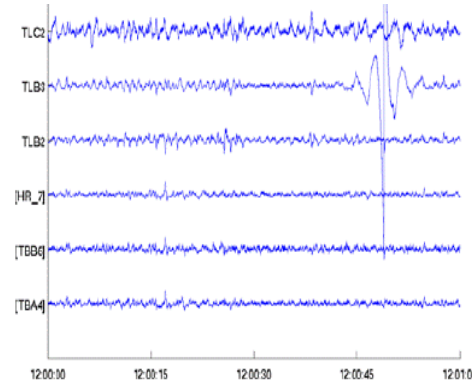
## Revised approach

time-to-seizure  $Y$

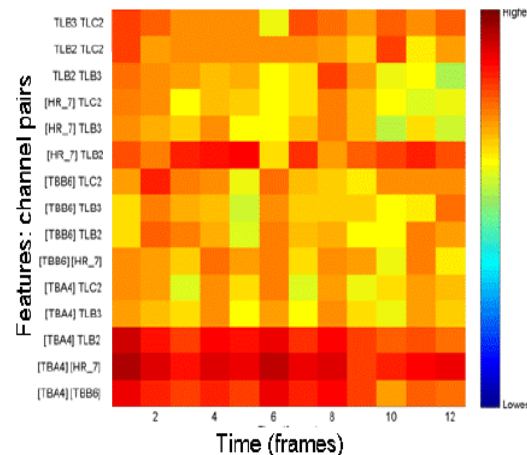


latent  
patient state  $Z$   
synchronization  
pattern  $X$

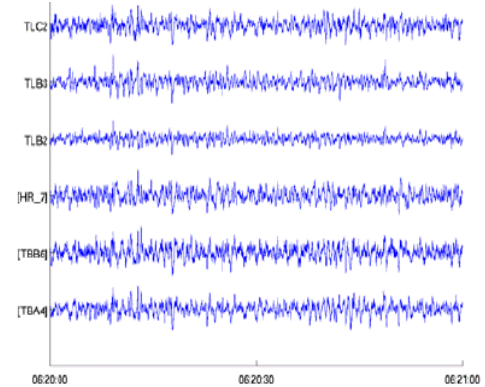
1min of **interictal** EEG



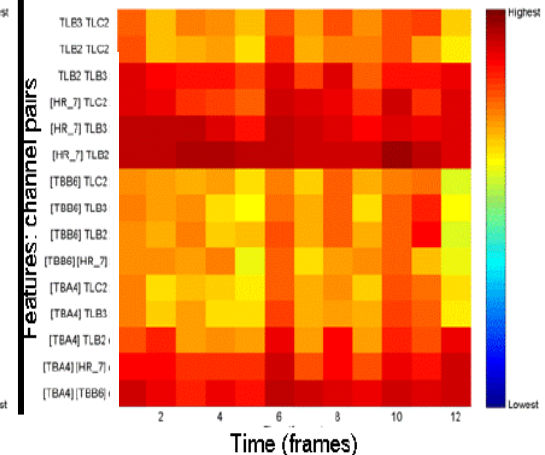
1min **interictal** pattern



1min of **preictal** EEG



1min **preictal** pattern



Examples of **patterns** of cross-correlation



# Thank you

- Further references:

- Mirowski P., LeCun Y., “Dynamic Factor Graphs for Time Series Modeling”, [Lecture Notes in Artificial Intelligence](#), 2009.
- Mirowski P., Madhavan D., LeCun Y., Kuzniecky R.I., “Classification of Patterns of EEG Synchronization for Seizure Prediction”, [Clinical Neurophysiology](#), 2009.
- Mirowski P., LeCun Y., Madhavan D., Kuzniecky R., “A Method for Classifying Ictal States of a Subject”, [US Patent Publication](#), submitted in June 2009.
- Mirowski P., LeCun Y., Madhavan D., Kuzniecky R., “Comparing SVM and Convolutional Networks for Epileptic Seizure Prediction from Intracranial EEG”, [IEEE MLSP](#), 2008.
- Mirowski P., Madhavan D., LeCun Y., “Time-Delay Neural Networks and Independent Component Analysis for EEG-Based Prediction of Epileptic Seizures Propagation”, [Conference of the AAAI](#), 2007.
- Krouk G., Mirowski P., LeCun Y., Shasha D., Coruzzi G., “High resolution dynamic transcriptome of Arabidopsis roots in response to NO<sub>3</sub><sup>-</sup>: Molecular physiology and predictive modeling”, submitted to [Proceedings of the National Academy of Sciences](#).

# Additional material



# Inference of latent variables

dynamical energy  $E_d(\mathbf{Z}_{t-p}^{t-1}, Z(t), \mathbf{W}_d) = \frac{1}{2} \|f(\mathbf{Z}_{t-p}^{t-1}; \mathbf{W}_d) - Z(t)\|_2^2$

observation energy  $E_o(Z(t), Y(t), \mathbf{W}_o) = \frac{1}{2} \|g(Z(t), \mathbf{W}_o) - Y(t)\|_2^2$

Inference of latent variables  $Z$

=

minimization w.r.t.  $Z$  of

dynamical energy  $E_d(\mathbf{Y}_{t-p}^t; \mathbf{W}_d) = \min_Z E_d(\mathbf{Z}_{t-p}^{t-1}, Z(t), \mathbf{W}_d)$

observation energy  $E_o(Y(t), \mathbf{W}_o) = \min_Z E_o(Z(t), Y(t), \mathbf{W}_o)$

Total energy of the DFG given a sequence  $Y$  and model parameters  $\mathbf{W}_o$ ,  $\mathbf{W}_d$ :

$$E(\mathbf{Y}_{t_a}^{t_b}; \mathbf{W}_d, \mathbf{W}_o) = \alpha \sum_{t=t_a+p}^{t_b} E_d(\mathbf{Y}_{t-p}^t; \mathbf{W}_d) + \beta \sum_{t=t_a}^{t_b} E_o(Y(t), \mathbf{W}_o)$$

# Learning of DFG model parameters

Loss function to minimize

e.g.  $L_1$  regularization  
of model parameters

$$L(\mathbf{Y}, \mathbf{Z}; \mathbf{W}) = \sum_t (\alpha E_d(\mathbf{Z}_{t-p}^{t-1}, \mathbf{Z}(t), \mathbf{W}_d) + \beta E_o(\mathbf{Z}(t), \mathbf{Y}(t), \mathbf{W}_o)) + \boxed{R(\mathbf{W})} + \boxed{R_Z(\mathbf{Z})})$$

Sparsity or smoothness  
penalties  
during state inference

Deterministic gradient-based version of Expectation-Maximization

**E-step** (latent variable inference)

annealed gradient descent

on minibatch of  $\mathbf{Z}$  until convergence

$$\tilde{\mathbf{Z}} = \arg \min_{\mathbf{Z}} L(\tilde{\mathbf{W}}, \mathbf{Y}, \mathbf{Z})$$

**M-step** (parameter learning)

1 step of stochastic gradient descent  
(diagonal Levenberg-Marquard)

$$\tilde{\mathbf{W}} = \arg \min_{\mathbf{W}} L(\mathbf{W}, \mathbf{Y}, \tilde{\mathbf{Z}})$$