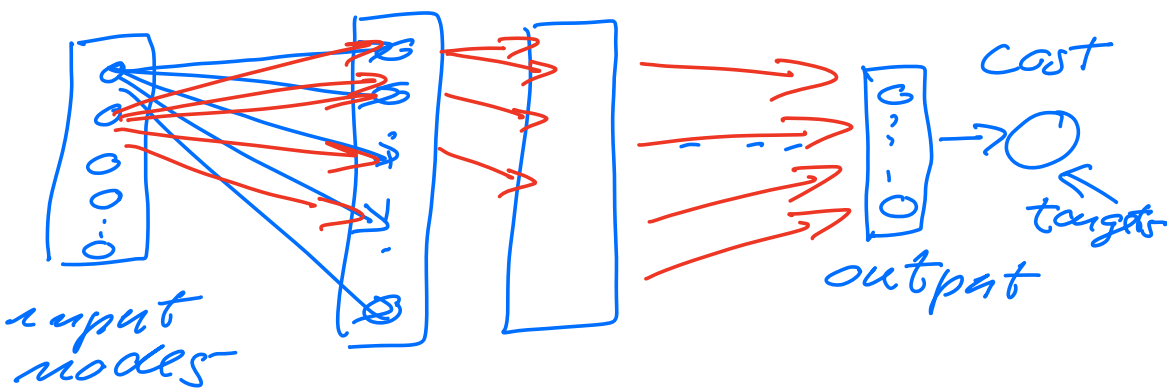
ML ERASMUS, JAN 15, 2023

Deep learning methods:

- FFNN = Feed Forward NN

  Fully connected



- Dense matrix with weights & biases, Affine operations
- Data is well structured and homogeneous

- CNN reduce dimensionality by filtering (convolution) does not work well with

data of unknown

- RNN recurrent NN can be used on data set with unknown length.
  - time series
  - Natural language studies

Time series:

$$m\frac{d^2x}{dt^2} + \eta\frac{dx}{dt} + x(t) = F(t)$$

initial conditions, $x_0 = x(t_0)$

$v_0 = v(t_0)$

$$v(t) = \frac{dx}{dt}$$

$$m\frac{dv}{dt} + \eta v + x = F$$

Discretize using Euler's method

$$x_{i+1} = x(t_i + \Delta t) = x_i + \Delta t \cdot v_i$$

$$v_{i+1} = v_i + \Delta t \, \dot{v}_i ?$$

$$\frac{dv}{dt} = -\frac{\mu}{m} v - \frac{1}{m} x + \frac{F}{m}$$

$\mu/m = \alpha$, $\frac{1}{m} = \delta$

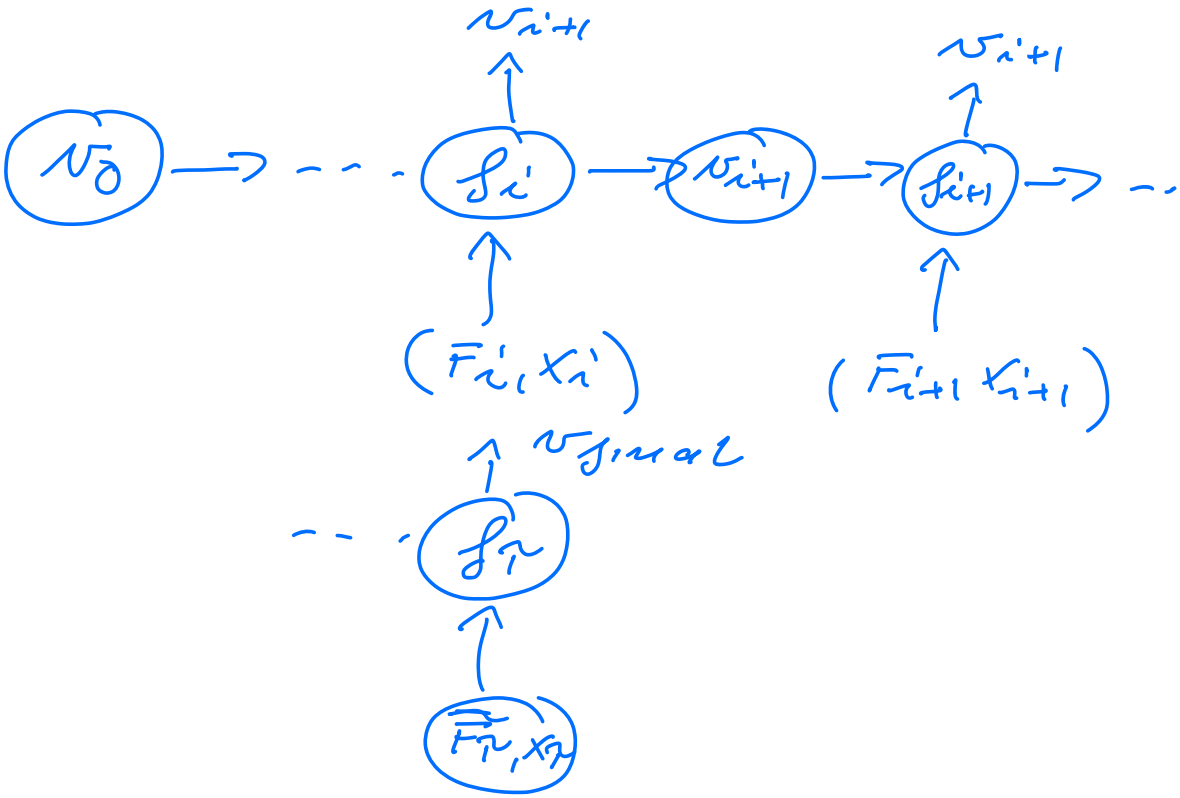
$$= \tilde{F} - \alpha v + \delta x$$

$$v_{i+1} = v_i + \Delta t (\tilde{F}_i - \alpha v_i - \delta x_i)$$

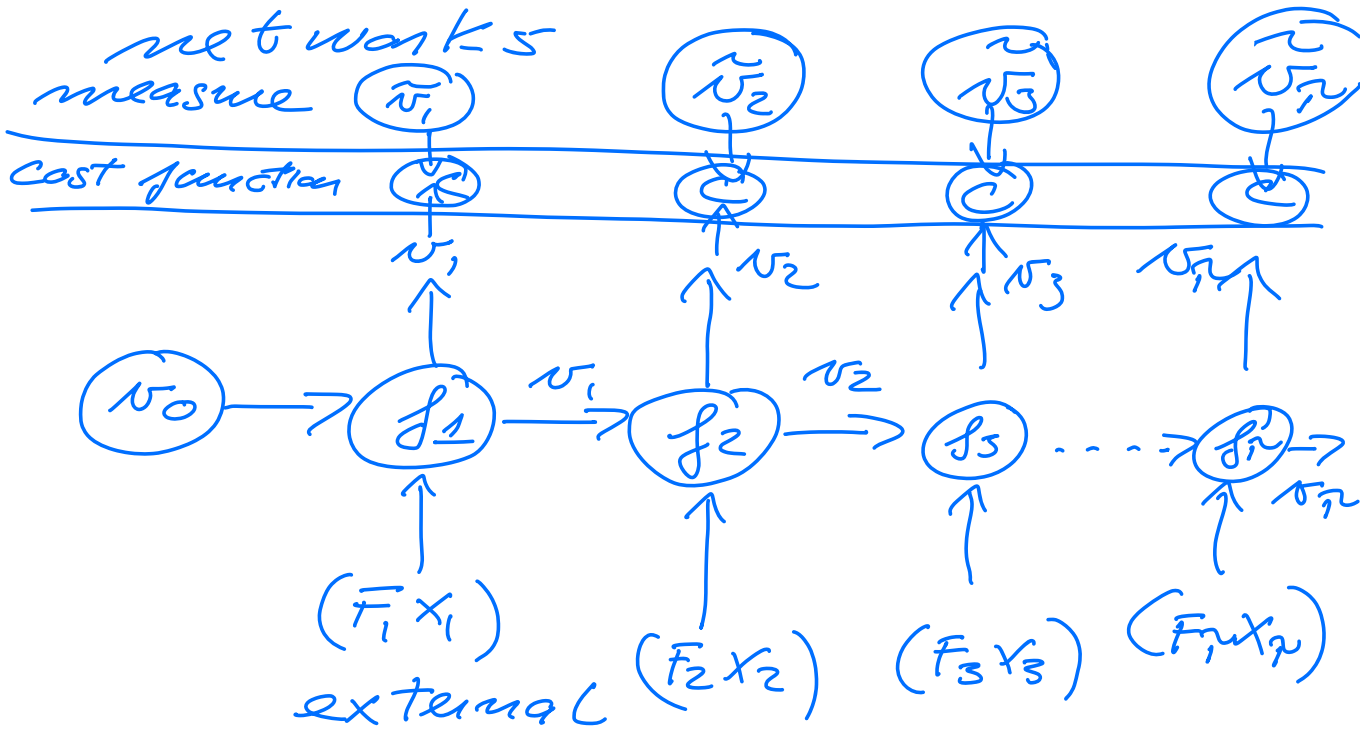
$$= v_i + f(v_i, \tilde{F}_i, x_i)$$

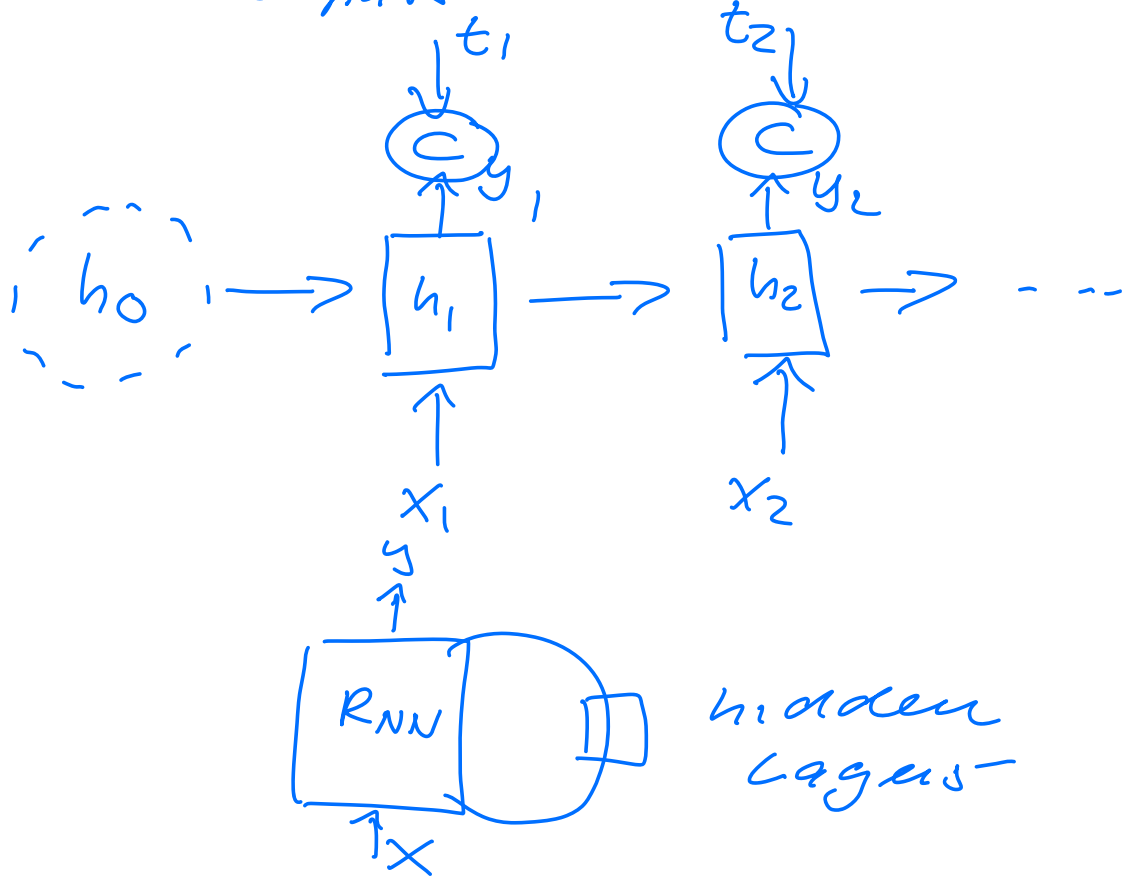
$$v_{i+1} = f(f(v_{i-1}, \tilde{F}_{i-1}, x_{i-1}), \tilde{F}_i, x_i)$$

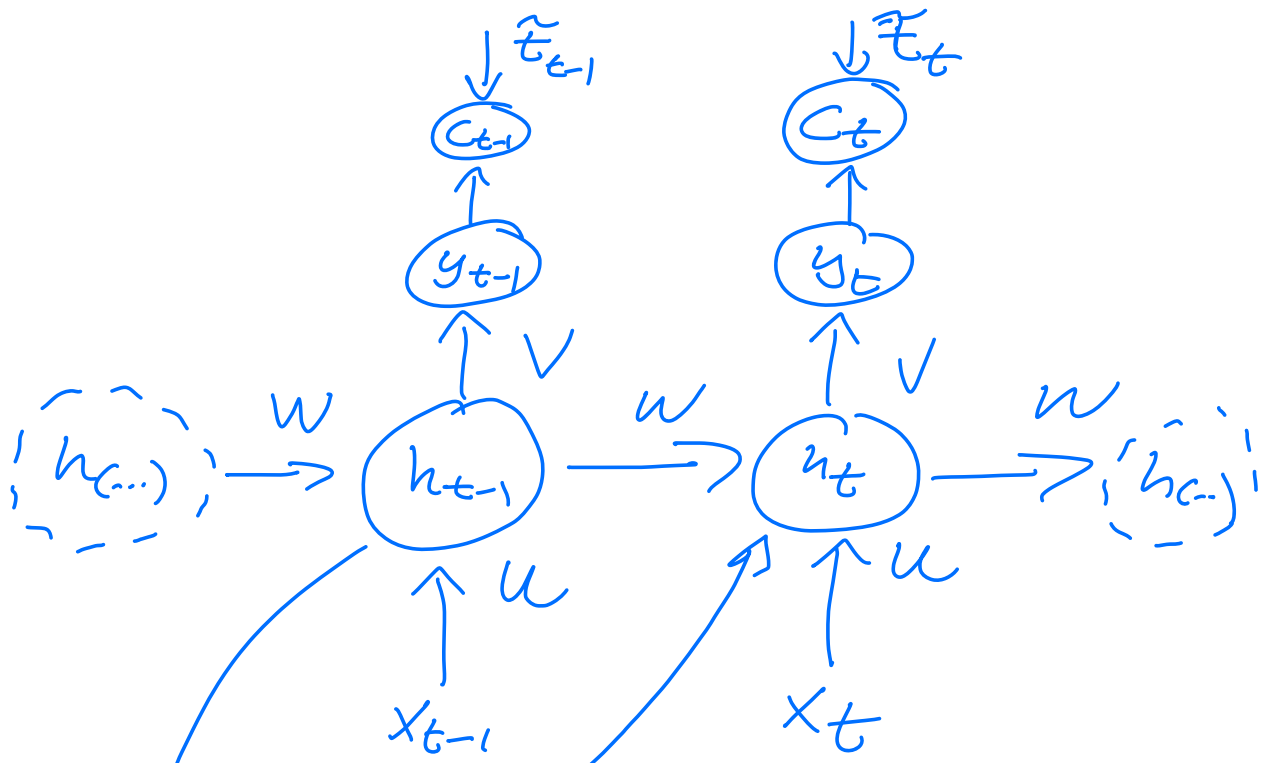
Focus on v_i



Connection with a neural networks







bias

$$z_t = b + W h_{t-1} + U x_t$$

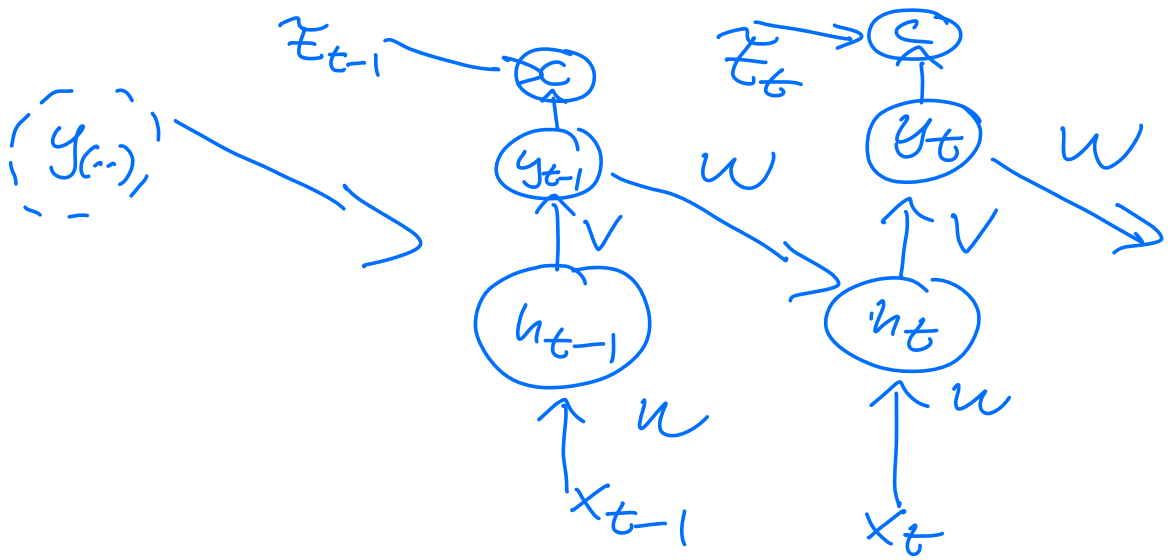
$$h_t = \sigma(z_t)$$

bias

$$y_t = g(V h_t + c)$$

BPTT = Back propagation through time,

Simplify the computational complexity



Problems with RNNs are often due to exploding gradients

$$h_t = W h_{t-1}$$

This operation is repeated for h_t $t-$ times

$$h_t = (W)^t h_0$$

$$W = S D S^T \qquad S^T S = S S^T = \mathbb{1}$$

$$D = \begin{bmatrix} \lambda_0 & 0 \\ 0 & \lambda_{d-1} \end{bmatrix} \qquad Wx = \lambda x$$

eigenvalues λ_i and eigen-
vectors w_i

$$h_0 = \sum_i \alpha_i w_i \qquad w w_i = \lambda_i w_i$$

$$W h_0 = h_1 = \sum_i \alpha_i \lambda_i w_i$$

repeat t times

$$(W)^t h_0 = h_t = \sum_i \alpha_i \lambda_i^t w_i$$

$$\lambda_0 > \lambda_1 > \lambda_2 \ldots \lambda_{d-1}$$

$$h_t \simeq \lambda_0^t w_0 \cdot \alpha_0$$

if $\lambda_0 \geq 1$ we may get
contributions to h_t which
can be large $\Rightarrow$ can
give rise to exploding
gradients

To avoid this, it is common to use gradient clipping

gradient $\vec{g}$

if $\|\vec{g}\|_2 \geq$ specific value

$$\vec{g} \leftarrow \frac{\varepsilon}{\|g\|_2} \vec{g}$$

end if