CSE 446
Dimensionality Reduction, Sequences

Administrative

- Final review this week
  - Practice exam questions will come out Wed
- Final exam next week Wed 8:30 am
- Today
  - Dimensionality reduction examples
  - Sequence models

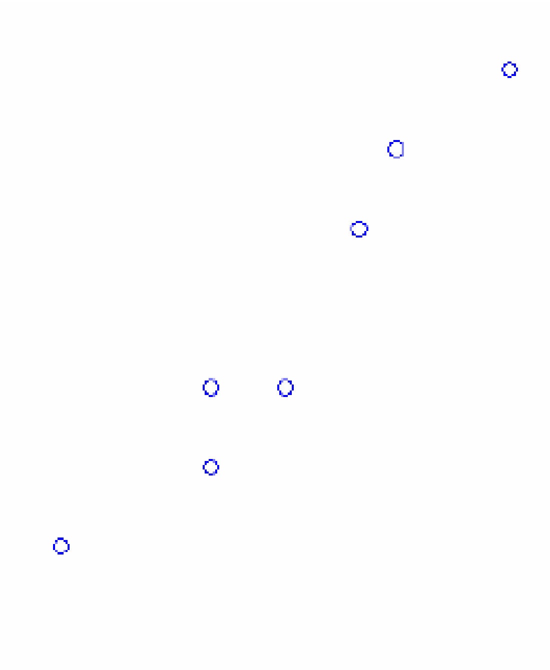
Dimensionality Reduction

- Principal Component Analysis (PCA)
  - Perform eigenvalue decomposition $\Sigma = \mathbf{U}\Lambda\mathbf{U}^T$
  - Use columns of **U** that correspond to K largest eigenvalues
  - You should know why this works…
- Singular value decomposition (SVD)
  - Faster than eigenvalue decomposition, especially for high-dimensional data: $\mathbf{X} = \mathbf{W}\mathbf{S}\mathbf{V}^T$
  - Take columns of **V** corresponding to largest singular values
  - You should know why this works…

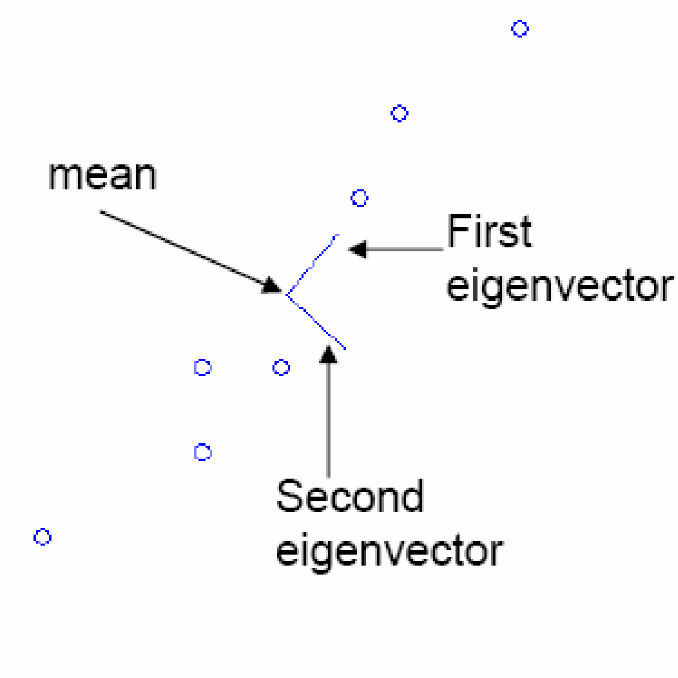
PCA example

$$\hat{\mathbf{x}}^i = \bar{\mathbf{x}} + \sum_{j=1}^{k} z_j^i \mathbf{u}_j$$

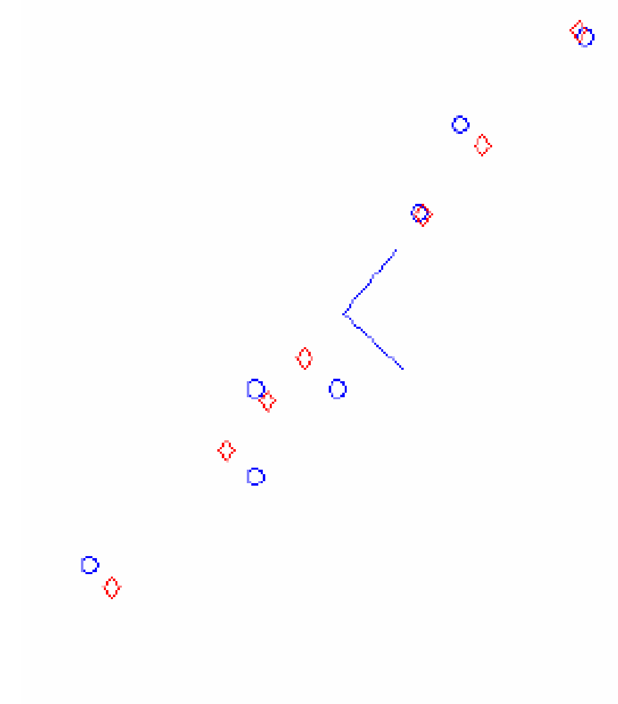
Data:



Projection:

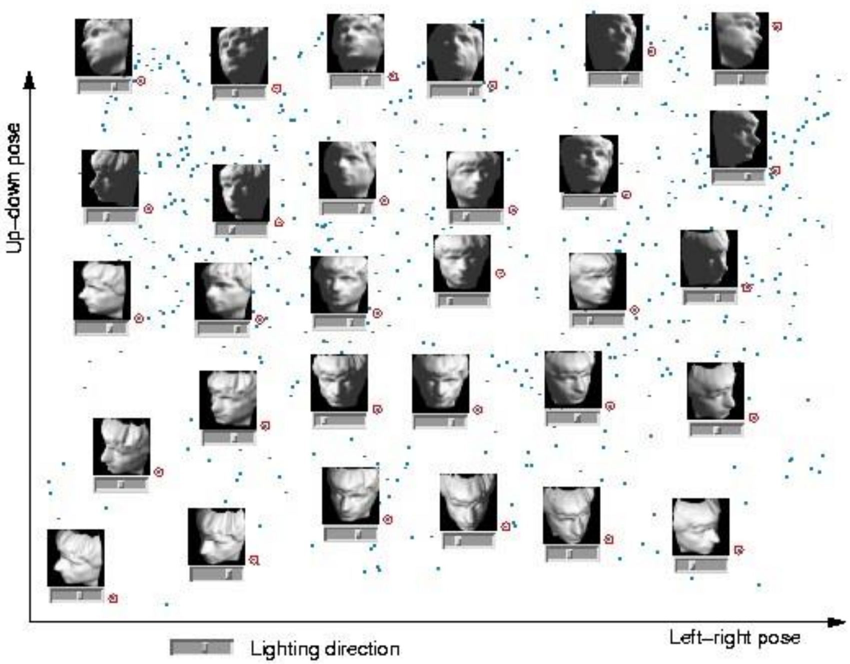


Reconstruction:



Dimensionality Reduction

raw data = pixels
(# dimensions = # pixels)



reduced dimensionality: only dimensions that matter

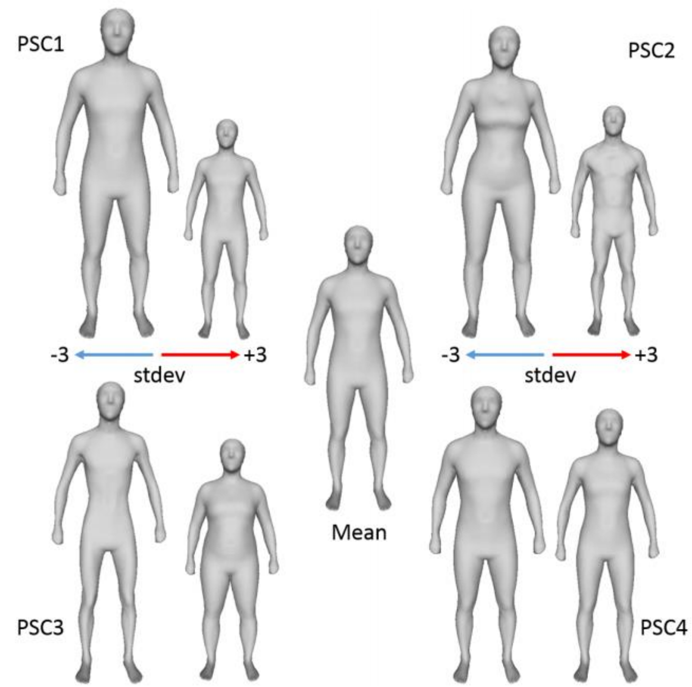
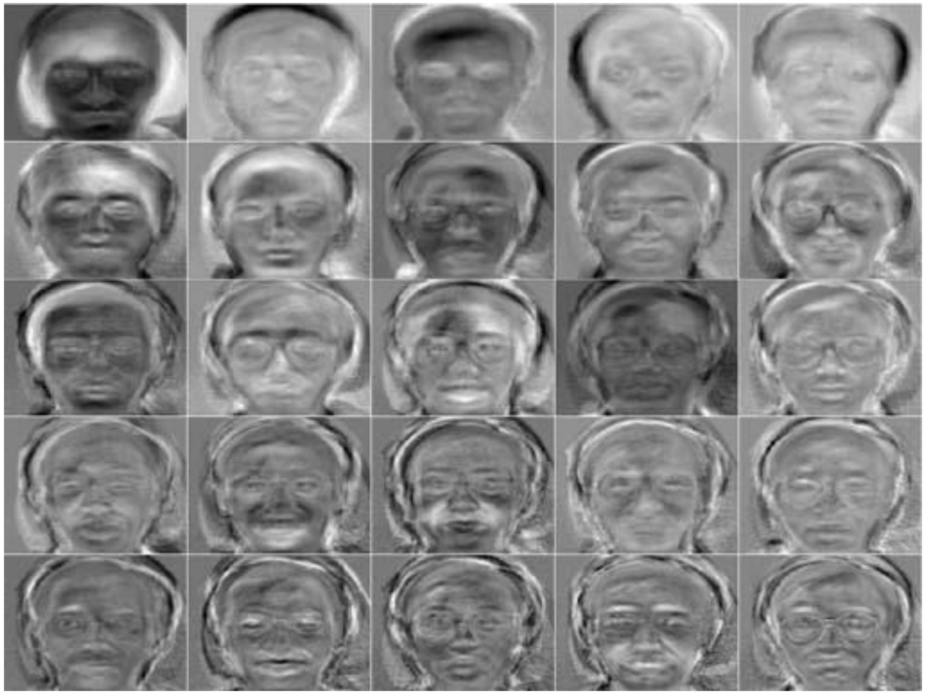


Image: Chen et al. '13

Eigenfaces [Turk, Pentland '91]

- Input images:

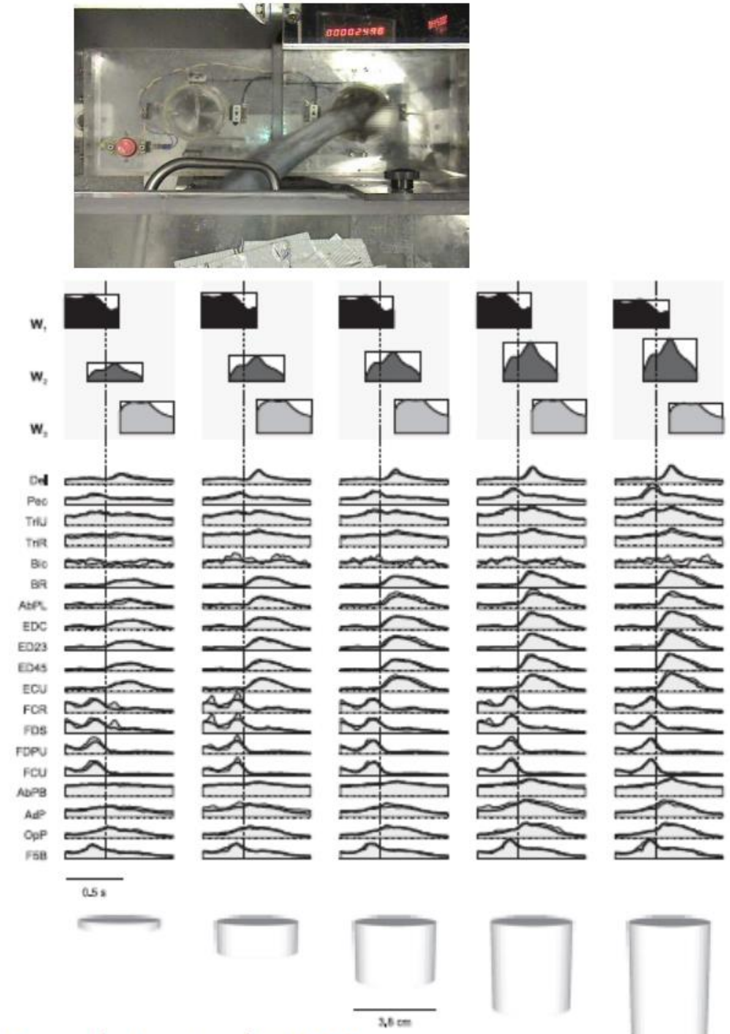
- Principal components:



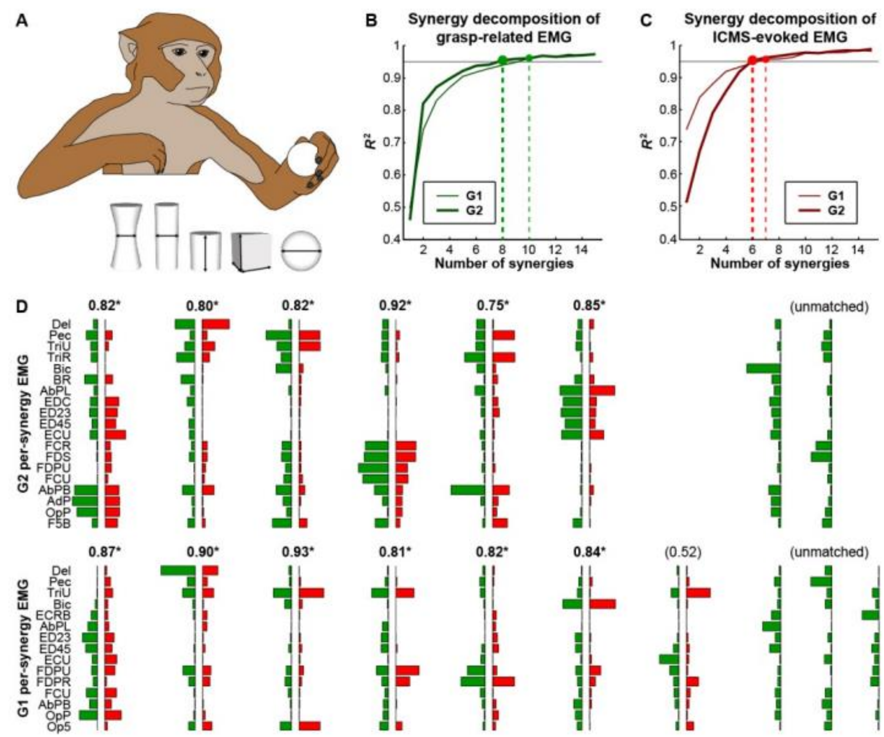
Eigenfaces reconstruction

- Each image corresponds to adding together the principal components:

Discovering muscle synergies

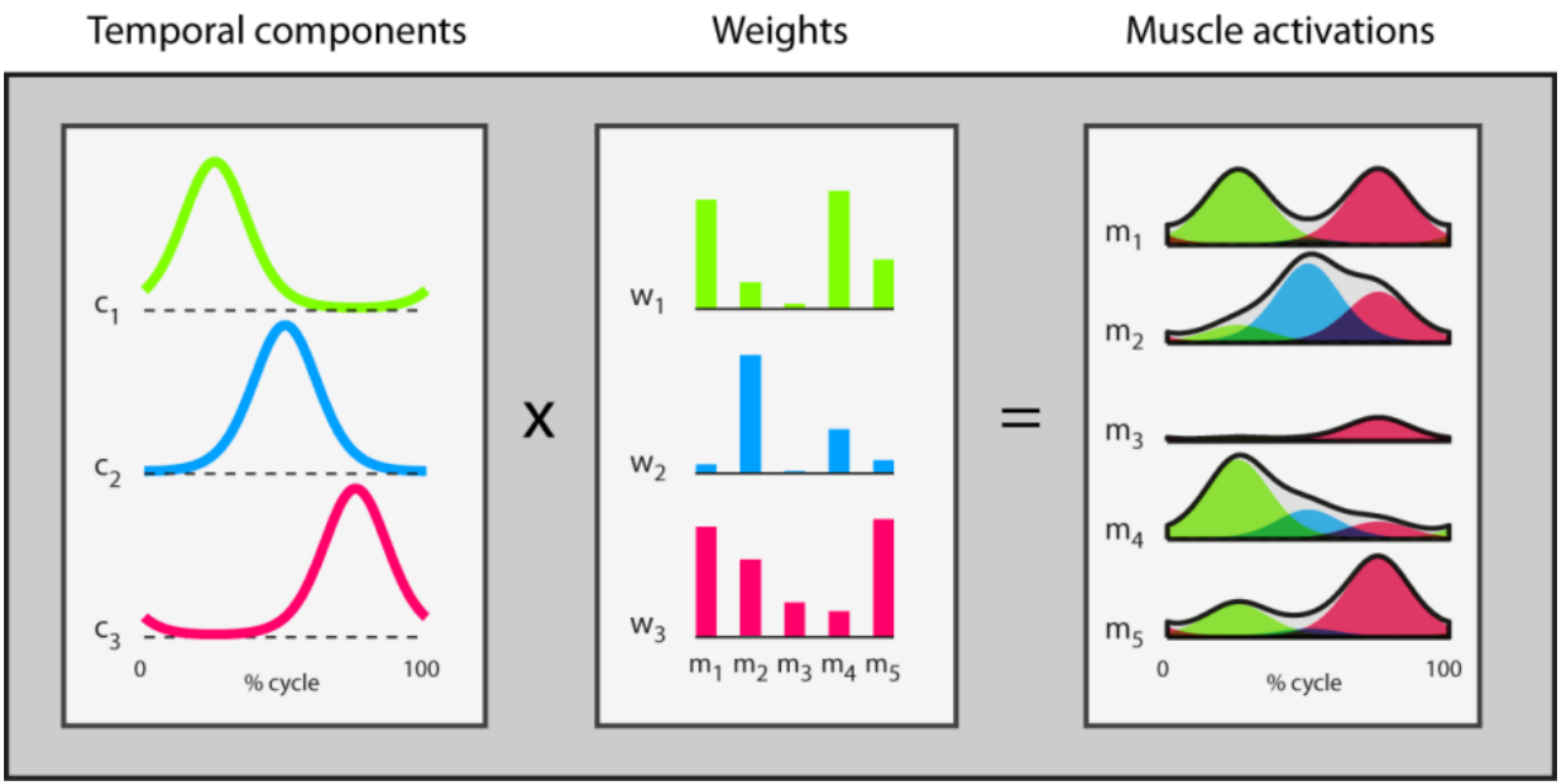


Overduin et al. 2008



Overduin et al. 2012

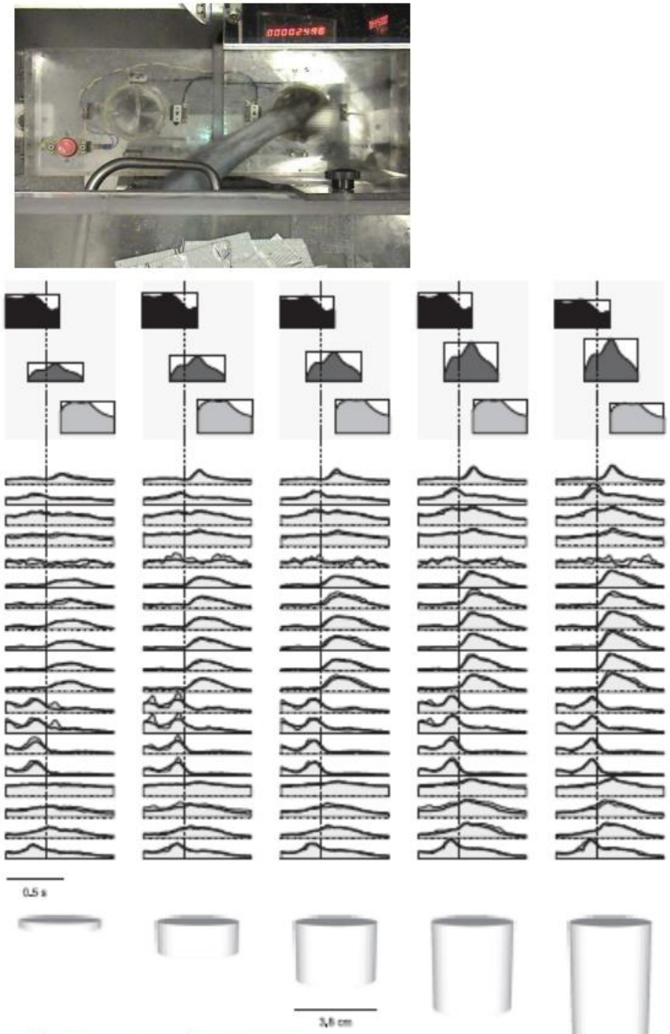
Discovering muscle synergies



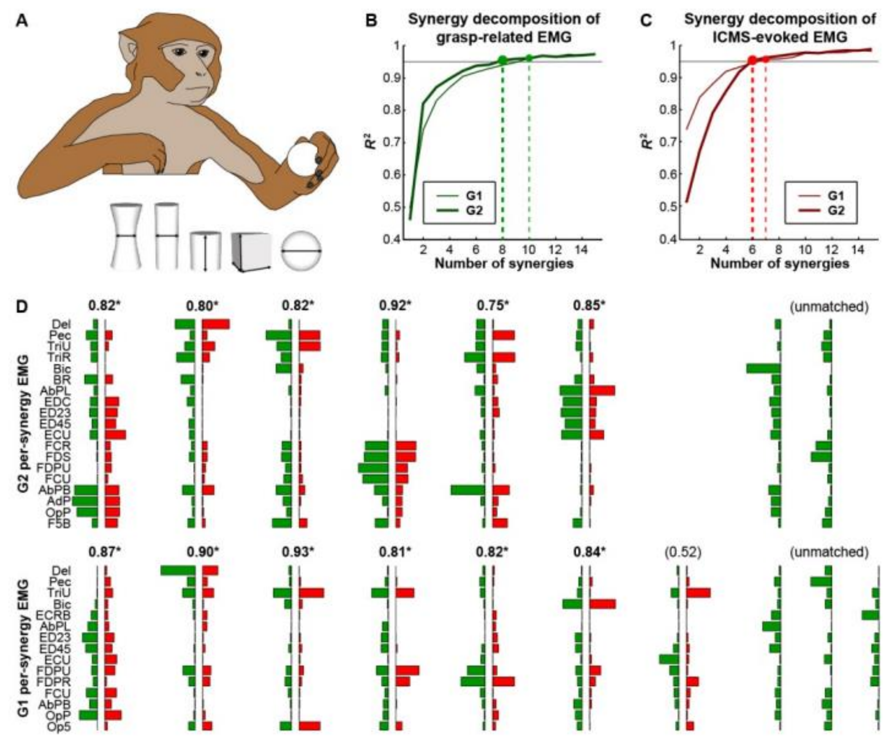
$$\mathbf{m}(t) = \sum_{i=1}^{N} c_i(t)\,\mathbf{w}_i$$

Temporal components capture temporal regularities in the motor output

Discovering muscle synergies



Overduin et al. 2008



Overduin et al. 2012

Dimensionality Reduction

- Design spaces:

  http://jerrytalton.net/research/tgyhk-emcds-09/video.mov

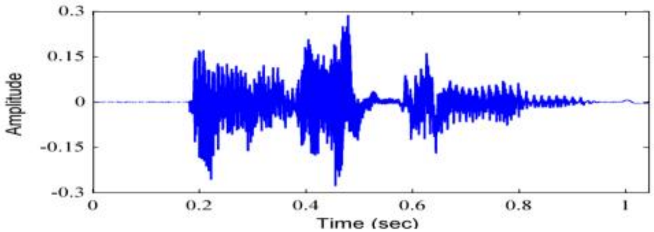
Sequence Models

- High level overview of *structured data*
- What kind of structure? Temporal structure:

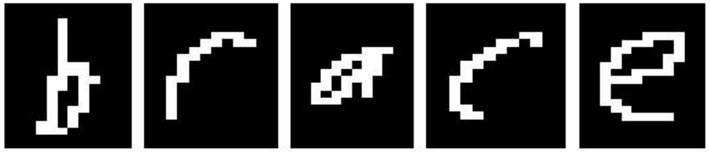
$$\mathbf{x}_i = \begin{bmatrix} \mathbf{x}_{1,i} \\ \mathbf{x}_{2,i} \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{x}_{T,i} \end{bmatrix}$$

- Sequential data
  - Time-series data

    E.g. Speech



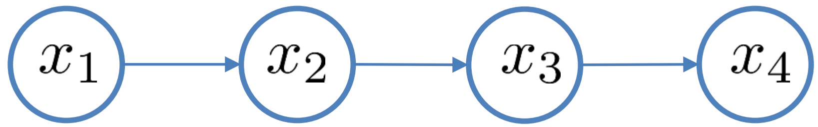
  - Characters in a sentence



  - Base pairs along a DNA strand

Markov Model

$$\mathbf{x}_i = \begin{bmatrix} x_{1,i} \\ x_{2,i} \\ \cdot \\ \cdot \\ \cdot \\ x_{T,i} \end{bmatrix}$$



$$x_{i,t} \in \{1, 2, \ldots, K\}$$

$$p(\mathbf{x}) = p(x_1)p(x_2|x_1)p(x_3|x_2)\ldots p(x_T|x_{T-1})$$

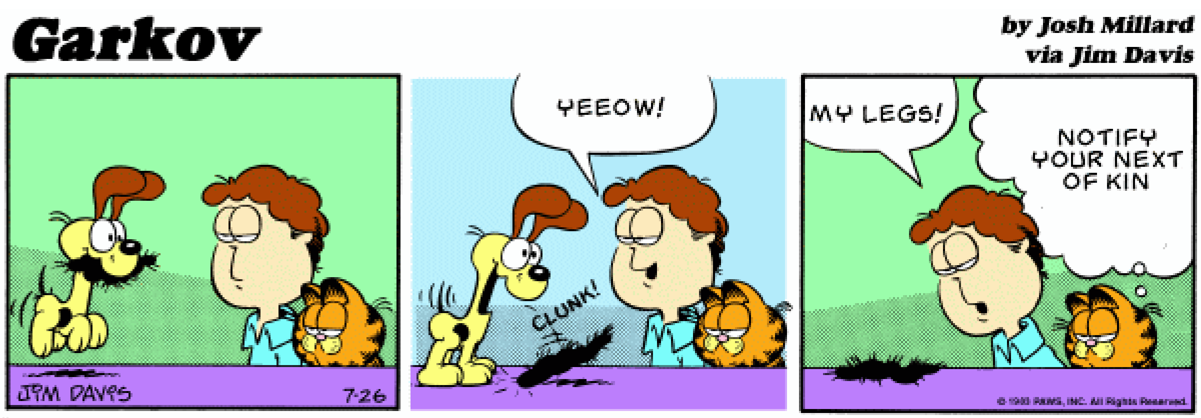
Markov Model: Learning

- How to learn? First, assume all time steps are the same: $p(x_2|x_1) = p(x_3|x_2) = p(x_t|x_{t-1})$
- Next, do exactly the same thing as in naïve Bayes $$p(x_t = i|x_{t-1} = j) = \frac{\text{Count}(x_t = i \wedge x_{t-1} = j)}{\text{Count}(x_{t-1} = j)}$$

Markov Model Applications

- What can we model?

His heard." "Exactly he very glad trouble, and by Hopkins! That it on of the who difficentralia. He rushed likely?" "Blood night that.



from Jeff Atwood

and Josh Millard

Markov Model for Classification

- Just like naïve Bayes, can use Markov model for classification (e.g. of text)
- Condition transitions on label – different transition model for each label

$$p(x_t = i | x_{t-1} = j, y = \ell) = \frac{\mathrm{Count}(x_t = i \wedge x_{t-1} = j \wedge y = \ell)}{\mathrm{Count}(x_{t-1} = j \wedge y = \ell)}$$

- Essentially trains different model for each label (often written as such…)

Markov Model for Classification

- Use just like naïve Bayes: evaluate probability of a test sequence given every possible label

$$p(\mathbf{x}|y=\ell) \propto p(x_1|y=\ell)p(x_2|x_1,y=\ell)p(x_3|x_2,y=\ell)\ldots p(x_T|x_{T-1},y=\ell)$$

- Classify text (type of article, author, sentiment)
- Classify DNA sequence as intron or exon

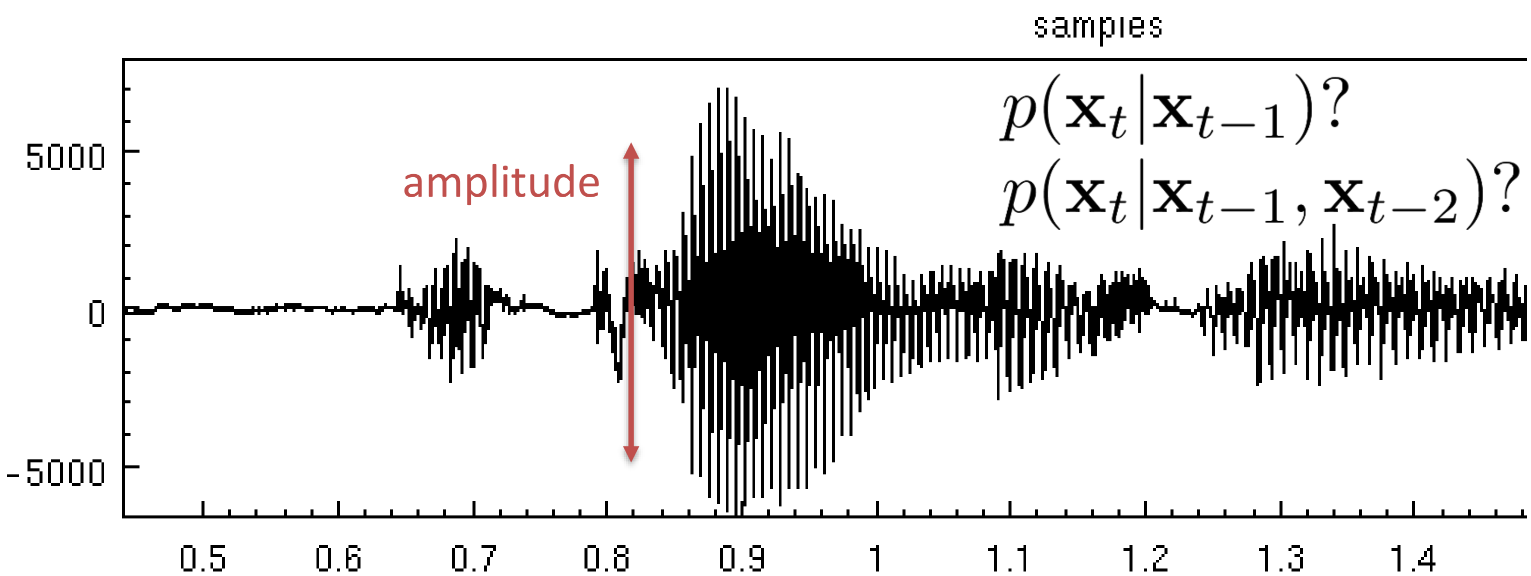
Markov Model Problems

- We often want longer temporal relationship: a word doesn't depend *only* on the preceding word! $p(x_t|x_{t-1}, x_{t-2}, x_{t-3})$
- How many entries in table if each state x has K values, and we condition on T past states?
- Can we do better?

Markov Model for Continuous Data

- What about continuous data?
- Linear-Gaussian model: $p(\mathbf{x}_t|\mathbf{x}_{t-1}) \sim \mathcal{N}(\mathbf{A}\mathbf{x}_{t-1}, \Sigma)$
  - Only good if transitions are linear
- Make it discrete?
  - K-means clustering to get discrete state
  - Gaussian mixture model (EM) cluster to get discrete state?
- Can we do better?

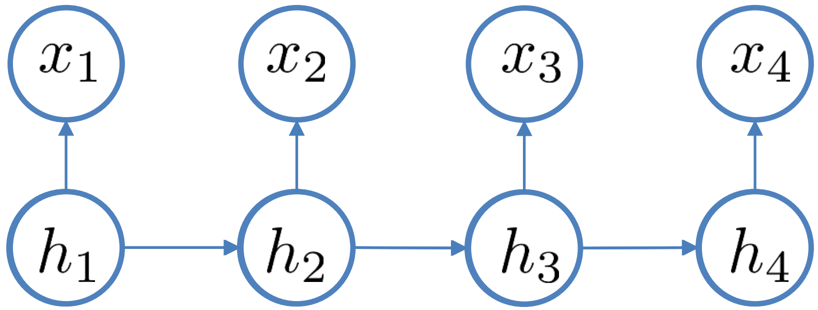
Hidden Markov Model



	nothing	loud	soft	LOUD	loud	soft	loud
h_t	A	B	C	D	E	F	G

$$p(\mathbf{x}_t|h_t) \qquad p(h_t|h_{t-1})$$

Hidden Markov Model



$p(\mathbf{x}_t | h_t)$ $\qquad$ $p(h_t | h_{t-1})$

Hidden Markov Model

- Observations (continuous or discrete): $\mathbf{x}_t$
- Hidden state (discrete): h_t
- Hidden state has dynamics: $p(h_t|h_{t-1})$
- Hidden state gives rise observations: $p(\mathbf{x}_t|h_t)$
- Just like clustering, but with dynamics!
- Why?
  - Continuous observations
  - Simple discrete state
  - No long temporal dependence! Tractable form
  - Learn the state that makes 1-step temporal dependence work

Hidden Markov Model: EM

- How to learn?
- Unobserved hidden state: h_t
- Just like clustering: use EM
  - E-step: $q(h_t) \leftarrow p(h_t|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)$

    $q(h_t, h_{t-1}) \leftarrow p(h_t, h_{t-1}|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)$
  - M-step:

$$\mu_k \leftarrow \frac{\sum_i \sum_t q(h_{i,t} = k)\mathbf{x}_{i,t}}{\sum_i \sum_t q(h_{i,t} = k)} \qquad p(h_1 = k) = \frac{\sum_i q(h_{i,1} = k)}{N}$$

$$\Sigma_k \leftarrow \frac{\sum_i \sum_t q(h_{i,t} = k)(\mathbf{x}_{i,t} - \mu_k)(\mathbf{x}_{i,t} - \mu_k)^T}{\sum_i \sum_t q(h_{i,t} = k)}$$

$$p(h_t = k|h_{t-1} = m) = \frac{\sum_i \sum_t q(h_{i,t} = k, h_{i,t-1} = m)}{\sum_i \sum_t q(h_{i,t-1} = m)}$$

$$p(\mathbf{x}_t|h_t)$$

$$p(h_t|h_{t-1})$$

$$p(h_1)$$

Hidden Markov Model: EM

- E-step requires inference
  - Somewhat more complex
  - Requires Viterbi algorithm
- We will not cover in detail
- Important to know:
  - Hidden state
  - Use EM
  - Similar to clustering (but with temporal model)

Hidden Markov Model for Classification

- Condition transitions on label – different transition model for each label
- Use just like naïve Bayes: evaluate probability of a test sequence given every possible label
- Often label is left out of the math, but it's there…

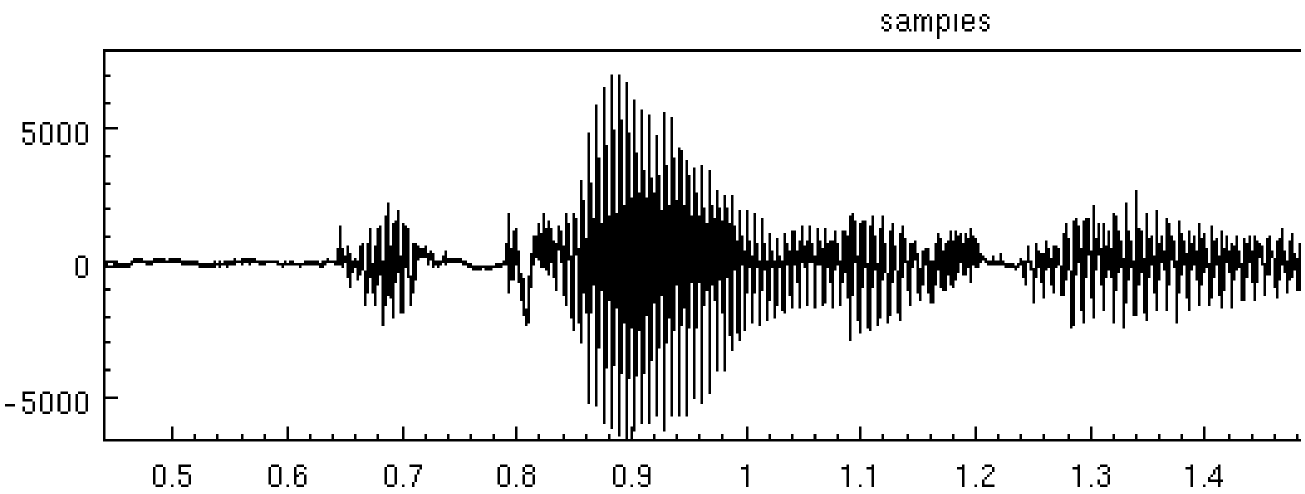
$$p(\mathbf{x}_{1:T}|y=\ell) \propto \sum_{h_1,h_2,\ldots,h_T} p(h_{1:T},\mathbf{x}_{1:T}|y=\ell)$$

different model for each label
same thing

$$p(\mathbf{x}_{1:T}|y=\ell) \propto \sum_{h_1,h_2,\ldots,h_T} p_\ell(h_{1:T},\mathbf{x}_{1:T})$$

Hidden Markov Model Applications

- Extremely popular for speech recognition
- 1 HMM = 1 phoneme
- Given a segment of audio, figure out which HMM gives it highest probability



Continuous *and* Nonlinear?

- HMM: continuous observations, but state is still discrete!
- Doesn’t scale well:
  - What if we want to track N different facts, each can be true or false?
    - Example: modeling structured text with different syntax, like parens “(“ (remember to close them…), quotes, etc.
    - Need pow(2,N) states!

Continuous *and* Nonlinear?

- Nonlinear continuous sequence model:

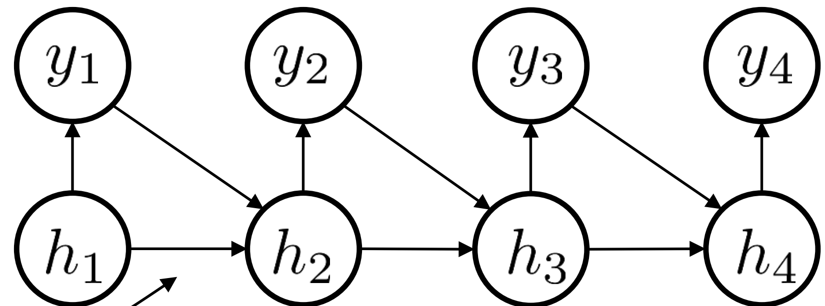
recurrent neural network

$$p(y_t = k|\mathbf{h}_t) = \frac{\exp(-\mathbf{W}_k \mathbf{h}_t)}{\sum_{k'=1}^{K} \exp(-\mathbf{W}_{k'} \mathbf{h}_t)}$$

$$\mathbf{h}_{t+1} = \sigma(\mathbf{W}_h \mathbf{h}_t + \mathbf{b}_h)$$
$$\mathbf{h}_{t+1} = \sigma(\mathbf{W}_h \mathbf{h}_t + \mathbf{W}_x \mathbf{x}_t + \mathbf{b}_h)$$
$$\mathbf{h}_{t+1} = \sigma(\mathbf{W}_h \mathbf{h}_t + \mathbf{W}_x \mathbf{x}_t + \mathbf{W}_y \mathbf{y}_t + \mathbf{b}_h)$$
$$\mathbf{h}_{t+1} = \sigma(\mathbf{W}_h \mathbf{h}_t + \mathbf{W}_y \mathbf{y}_t + \mathbf{b}_h)$$

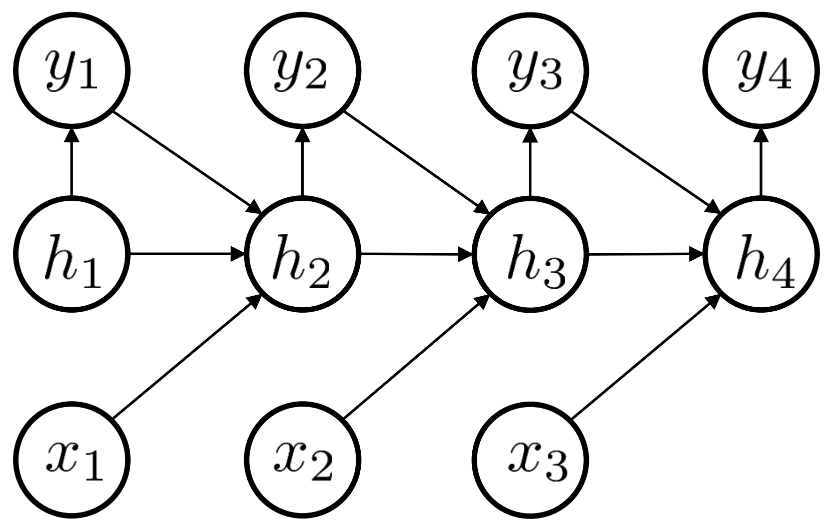
RNN in Pictures

$$\mathbf{h}_{t+1} = \sigma(\mathbf{W}_h \mathbf{h}_t + \mathbf{W}_y \mathbf{y}_t + \mathbf{b}_h)$$



(these are not probabilities!)

$$\mathbf{h}_{t+1} = \sigma(\mathbf{W}_h \mathbf{h}_t + \mathbf{W}_x \mathbf{x}_t + \mathbf{b}_h)$$

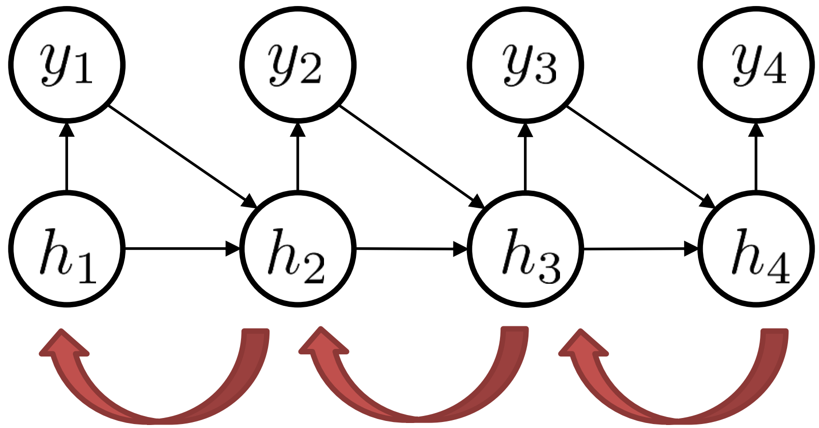


many many many other designs…

RNN Training

- Almost always use backpropagation + stochastic gradient descent/gradient ascent
  - No different than any other neural network
  - Just have many outputs
  - Compute gradients and use chain rule
    - Per time step instead of per layer
    - Math is exactly the same
- But it’s very hard to optimize…

Why RNN Training is Hard



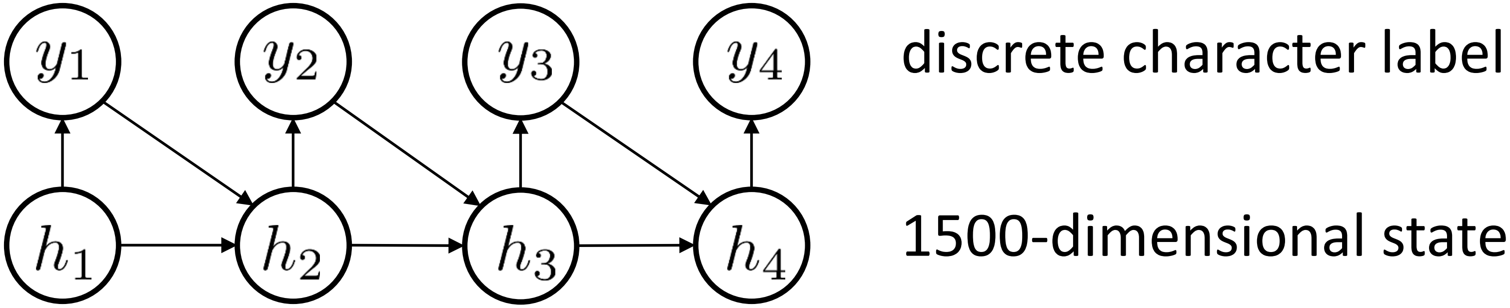
$$\frac{d\mathcal{L}(y_T)}{dh_2} = \underbrace{\frac{d\mathcal{L}(y_T)}{dh_T} \frac{dh_T}{dh_{T-1}} \cdots \frac{dh_3}{dh_2}}_{\text{lots of multiplication}}$$

lots of multiplication
very unstable numerically

- Backpropagation = chain rule
- Derivative multiplied by new matrix at each time step (time step in RNN = layer in NN)
- Lots of multiplication by values less than 1 = gradients become tiny
- Lots of multiplication by values greater than 1 = gradients explode
- Many tricks for effective training
  - Clever nonlinearity (e.g. LSTM)
  - Better optimization algorithms (more advanced than gradient descent)

RNN Application: Text Generation

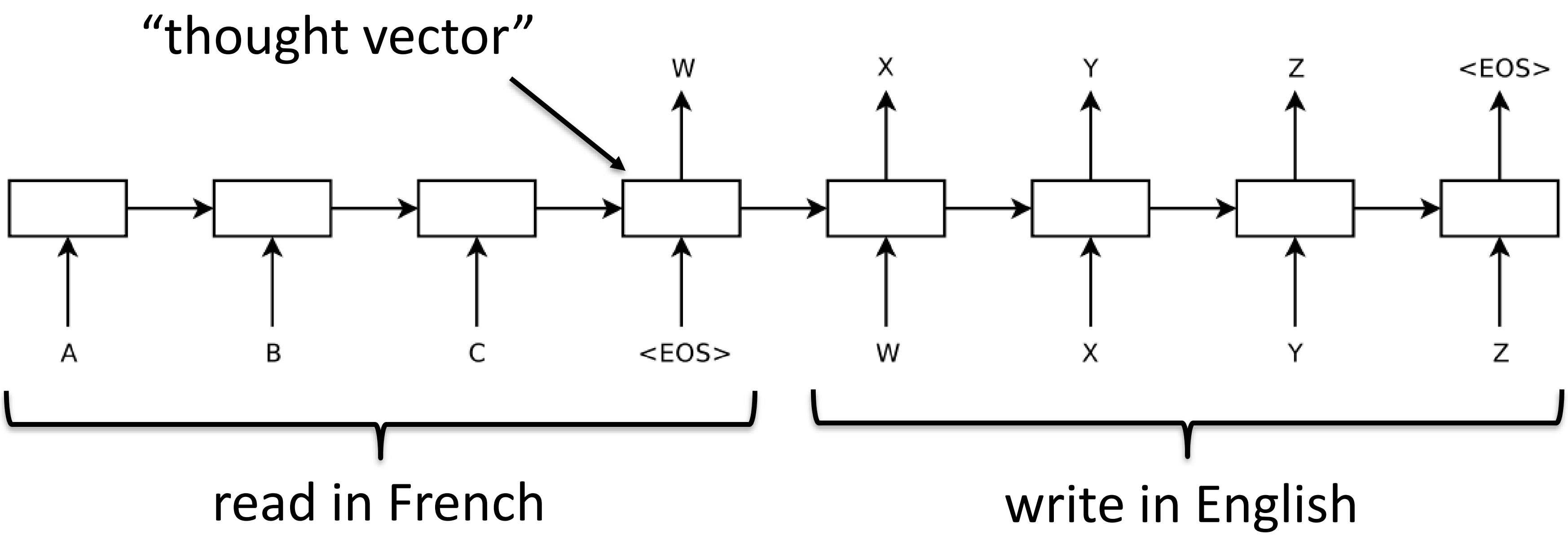
- http://www.cs.toronto.edu/~ilya/fourth.cgi



The meaning of life is any older bird. Get into an hour performance, in the first time period in

RNN Application: Machine Translation

- Sequence to sequence model



Sutskever et al. 2014

RNN does Shakespeare

PANDARUS:
Alas, I think he shall be come approached and the day
When little srain would be attain'd into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:
They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.

DUKE VINCENTIO:
Well, your wit is in the care of side and that.

From Andrej Karpathy

RNN does algebraic geometry (maybe it can write my lecture notes?)

For $\bigoplus_{n=1,\ldots,m}$ where $\mathcal{L}_{m_\bullet} = 0$, hence we can find a closed subset $\mathcal{H}$ in $\mathcal{H}$ and any sets $\mathcal{F}$ on X, U is a closed immersion of S, then $U \to T$ is a separated algebraic space.

Proof. Proof of (1). It also start we get

$$S = \mathrm{Spec}(R) = U \times_X U \times_X U$$

and the comparicoly in the fibre product covering we have to prove the lemma generated by $\coprod Z \times_U U \to V$. Consider the maps M along the set of points Sch_{fppf} and $U \to U$ is the fibre category of S in U in Section, **??** and the fact that any U affine, see Morphisms, Lemma **??**. Hence we obtain a scheme S and any open subset $W \subset U$ in $Sh(G)$ such that $\mathrm{Spec}(R') \to S$ is smooth or an

$$U = \bigcup U_i \times_{S_i} U_i$$

which has a nonzero morphism we may assume that f_i is of finite presentation over S. We claim that $\mathcal{O}_{X,x}$ is a scheme where $x, x', s'' \in S'$ such that $\mathcal{O}_{X,x'} \to \mathcal{O}'_{X',x'}$ is separated. By Algebra, Lemma **??** we can define a map of complexes $\mathrm{GL}_{S'}(x'/S'')$ and we win. $\square$

From Andrej Karpathy

RNN does operating system code

```
/*
 * If this error is set, we will need anything right after that BSD.
 */
static void action_new_function(struct s_stat_info *wb)
{
  unsigned long flags;
  int lel_idx_bit = e->edd, *sys & ~((unsigned long) *FIRST_COMPAT);
  buf[0] = 0xFFFFFFFF & (bit << 4);
  min(inc, slist->bytes);
  printk(KERN_WARNING "Memory allocated %02x/%02x, "
    "original MLL instead\n"),
    min(min(multi_run - s->len, max) * num_data_in),
    frame_pos, sz + first_seg);
  div_u64_w(val, inb_p);
  spin_unlock(&disk->queue_lock);
  mutex_unlock(&s->sock->mutex);
  mutex_unlock(&func->mutex);
  return disassemble(info->pending_bh);
}
```

From Andrej Karpathy

RNN does clickbait…

Romney Camp : ‘ I Think You Are A Bad President ‘
Here ‘ s What A Boy Is Really Doing To Women In Prison Is Amazing
L . A . ‘ S First Ever Man Review
Why Health Care System Is Still A Winner
Why Are The Kids On The Golf Team Changing The World ?
2 1 Of The Most Life – Changing Food Magazine Moments Of 2 0 1 3
More Problems For ‘ Breaking Bad ‘ And ‘ Real Truth ‘ Before Death
Raw : DC Helps In Storm Victims ‘ Homes
U . S . Students ‘ Latest Aid Problem
Beyonce Is A Major Woman To Right – To – Buy At The Same Time
Taylor Swift Becomes New Face Of Victim Of Peace Talks
Star Wars : The Old Force : Gameplay From A Picture With Dark Past (Part 2)
Sarah Palin : ‘ If I Don ‘ t Have To Stop Using ‘ Law , Doesn ‘ t Like His Brother ‘ s Talk On His ‘ Big Media ‘
Israeli Forces : Muslim – American Wife ‘ s Murder To Be Shot In The U . S .
And It ‘ s A ‘ Celebrity ‘
Mary J . Williams On Coming Out As A Woman
Wall Street Makes $ 1 Billion For America : Of Who ‘ s The Most Important Republican Girl ?
How To Get Your Kids To See The Light
Kate Middleton Looks Into Marriage Plans At Charity Event
Adorable High – Tech Phone Is Billion – Dollar Media

From Lars Eidnes