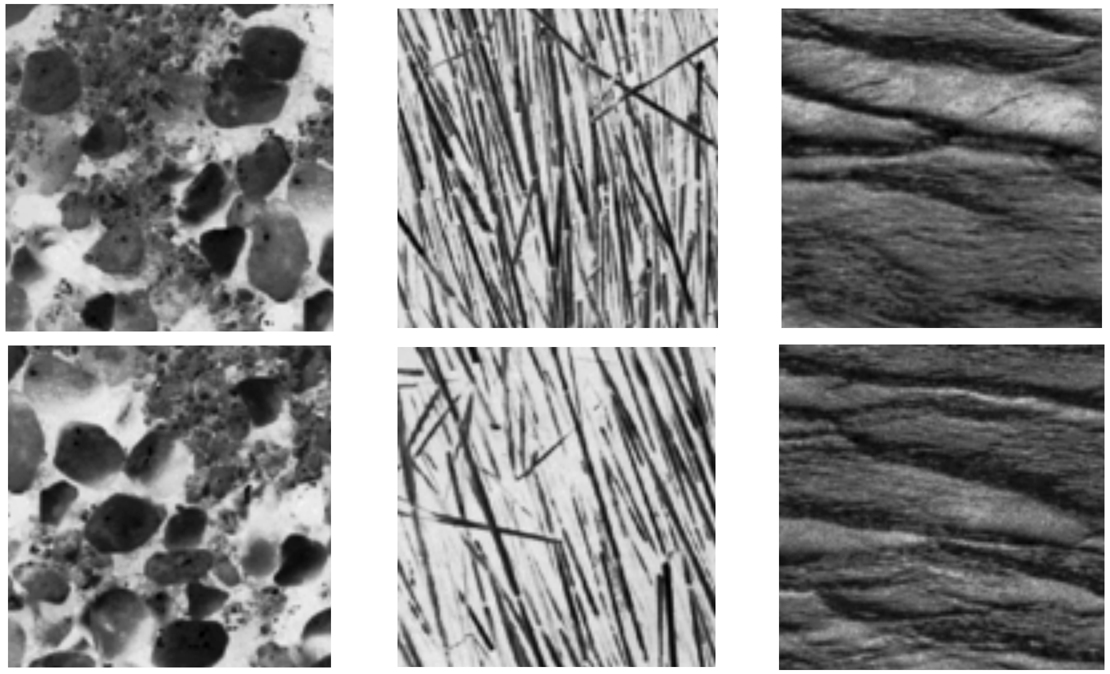
Stat 212b: Topics in Deep Learning
Lecture 13

Joan Bruna
UC Berkeley

Berkeley
UNIVERSITY OF CALIFORNIA

Review:Representation of Stationary Processes

$x(u)$: realizations of a stationary process $X(u)$ (not Gaussian)



$$\Phi(X) = \{E(f_i(X))\}_i$$

Estimation from samples $x(n)$: $\widehat{\Phi}(X) = \left\{ \frac{1}{N} \sum_n f_i(x)(n) \right\}_i$

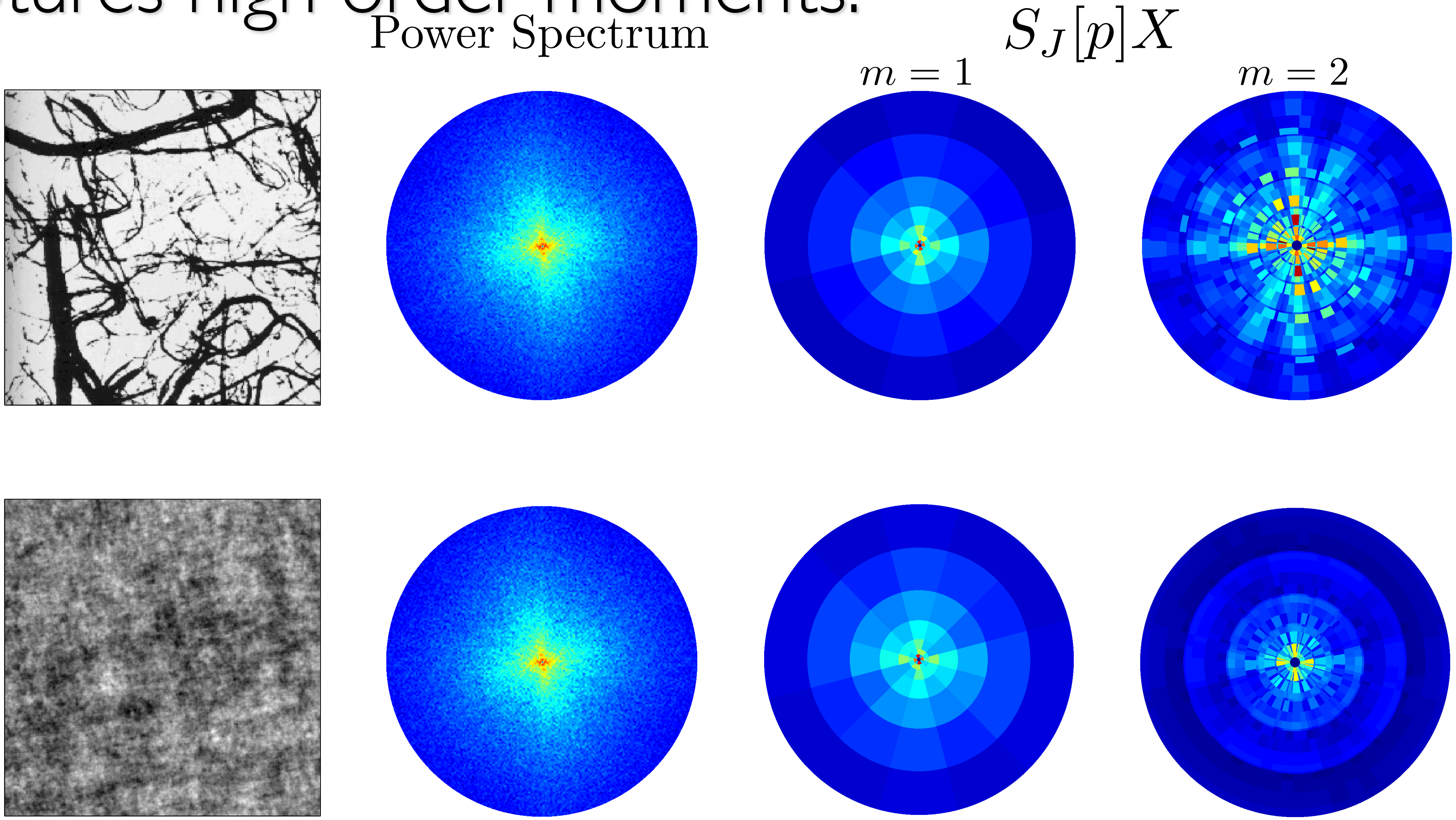
Discriminability: need to capture high-order moments

Stability: $E(\|\widehat{\Phi}(X) - \Phi(X)\|^2)$ small

Review:Properties of Scattering Moments

[Bruna, Mallat, '11,'12]

- Captures high order moments:



- Cascading non-linearities is ***necessary*** to reveal higher-order moments.

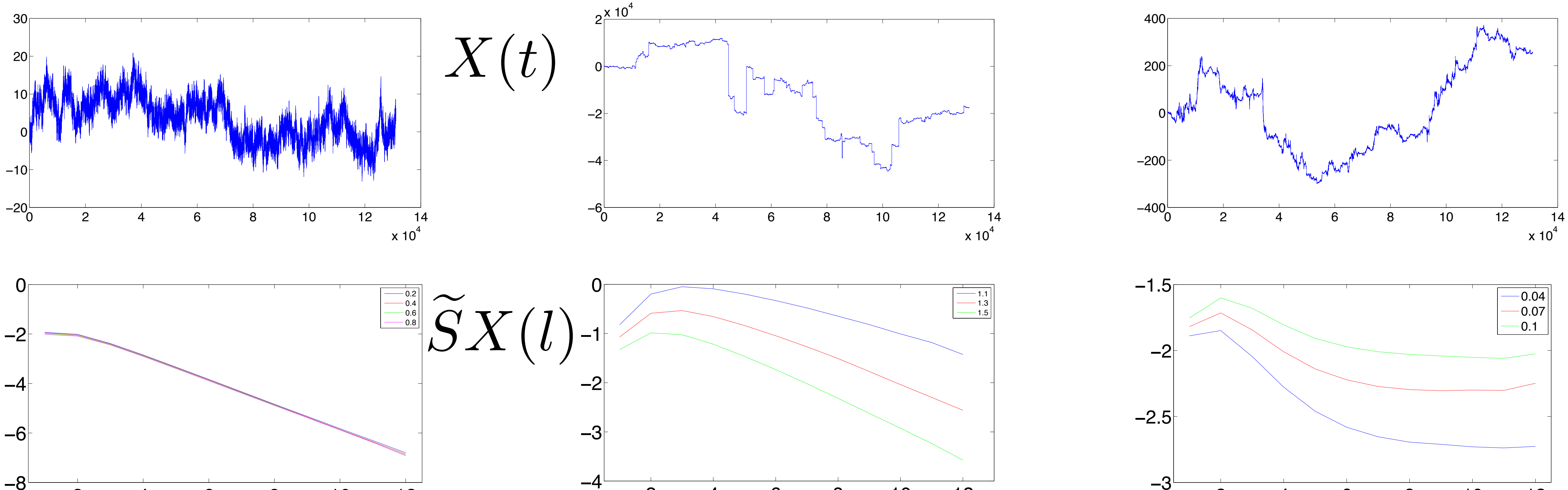
Review: Scattering and Intermittency

Theorem [BBMM'13]:

If $X(t)$ Fractional Brownian Motion, then $\tilde{S}X(l) \simeq 2^{-l/2}$,

If $X(t)$ α-stable Lévy process, then $\tilde{S}X(l) \simeq 2^{l(\alpha^{-1}-1)}$,

If $X(t)$ Multiplicative Random Cascade, then $\tilde{S}X(l) \simeq O(1)$,



$X(t) \sim$ FBM

$X(t) \sim$ Lévy

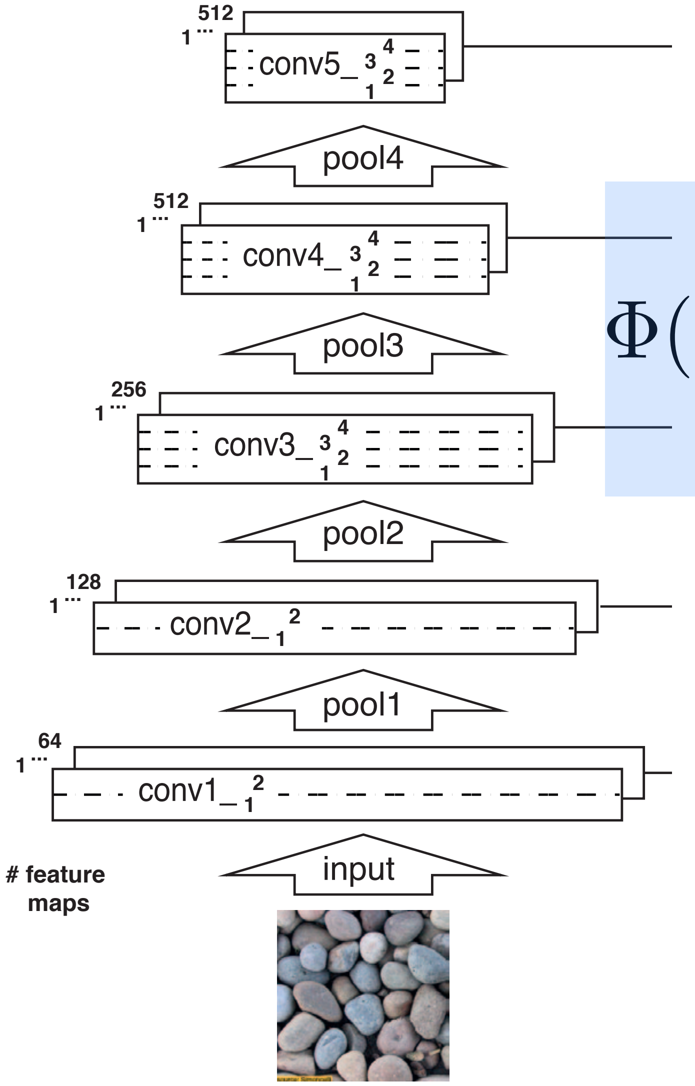
$X(t) \sim$ MRW

Second Order: Measure of Multiscale Intermittency

Review: CNNs for Texture Representation

- Q: How to obtain a texture representation from a CNN?
- Simple, yet powerful, idea [Gatys et al.'15]:

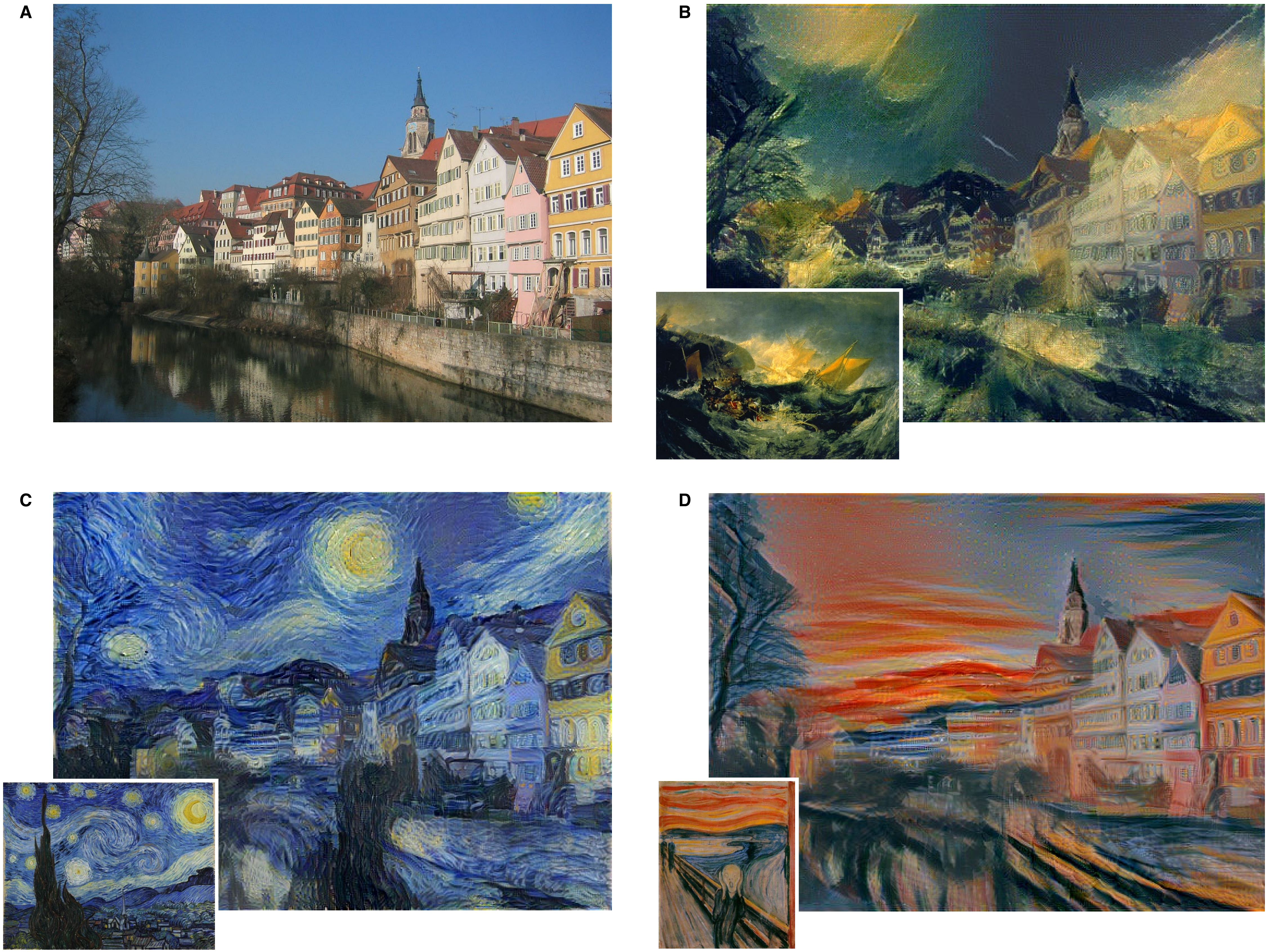
Let $(\Phi_1(x)(u_1,\lambda_1), \Phi_2(x)(u_2,\lambda_2), \dots, \Phi_K(x)(u_K,\lambda_K))$ the outputs of each layer of a pre-trained CNN



Stationary or "style" representation:

$$\Phi(x) = \left\{ \frac{1}{N_k} \sum_{u_k} \Phi_k(x)(u_k,\cdot)\Phi_k(x)(u_k,\cdot)^T \ , \ k = 1 \leq K \right\}$$

Review: Texture and Geometry



Review: State-space Models

- We can consider a hidden state Y_t with its own internal dynamics:

$$Y_{t+1} = F(Y_t, W_t)$$

W_t: Internal noise modeling uncertainty

- Hidden states influences observations X_t:

$$X_t = G(Y_t, Z_t)$$

Z_t: observational noise

- Q: How to infer the hidden states given observations?
  i.e $P(Y_t \mid X_1, \ldots, X_t)$

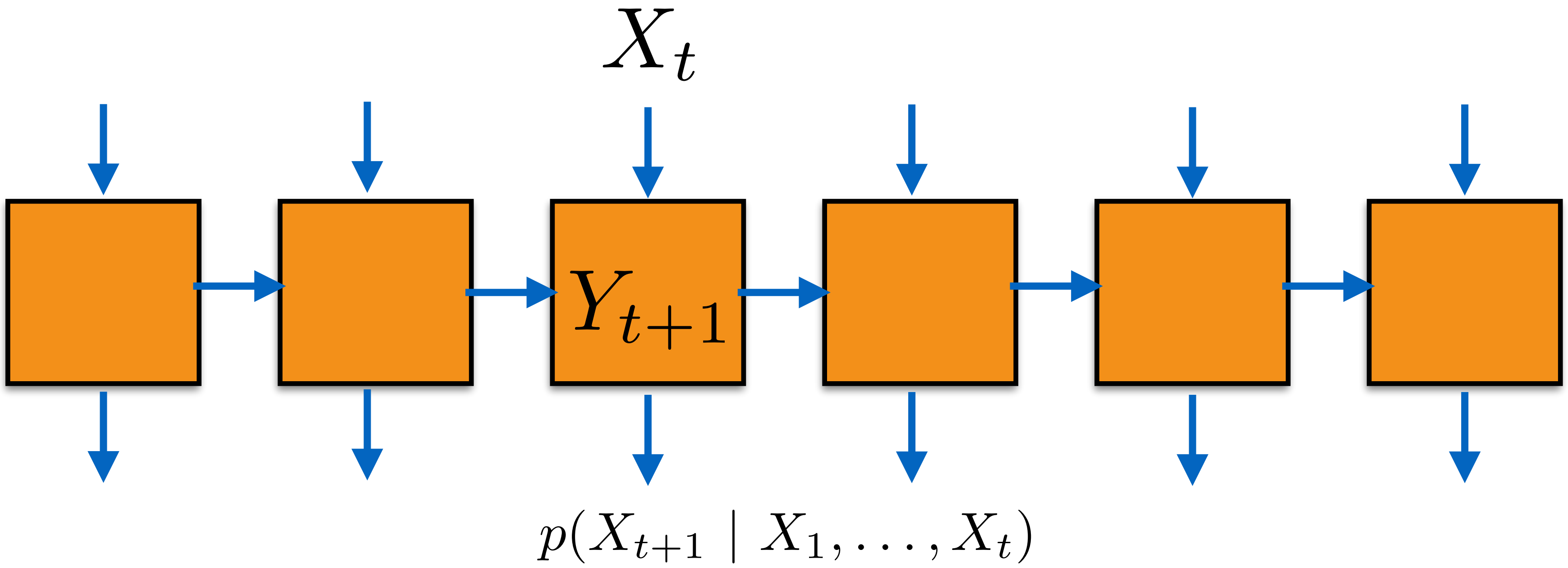
- Only tractable on particular models.

Review: Recurrent Neural Networks (RNN)

- We can combine the advantages of previous models into a non-linear continuous dynamical system:

$$p(X_1, \ldots, X_t) = \prod_{i \leq t} p(X_i \mid Y_i) \text{ with}$$

$$Y_i = F_\theta(Y_{i-1}, X_{i-1}) \qquad F_i \in \mathbb{R}^L$$



- Typically, we consider $F_\theta(Y_i, X_i) = \rho(A_{Y,Y} Y_{i-1} + A_{Y,X} X_i)$, with ρ a non-expansive point-wise nonlinearity.

Objectives

- RNNs and Memory
  - Long Short Term Memory (LSTM)
  - Explicit Memory Models

- Applications

RNN and Memory

- Q: How to efficiently store and leverage memory?

- Example: character-level prediction:

“Whenever I find myself growing grim about the mouth; whenever it is a damp, drizzly November in my soul; whenever I find myself involuntarily pausing before coffin warehouses, and bringing up the rear of every funeral I meet; and especially whenever my hypos get such an upper hand of me, that it requires a strong moral principle to prevent me from deliberately stepping into the street, and methodically knocking people's hats off - then, I account it high time to get to sea as soon as I can.”

- Can the RNN handle long memory?

“Vanishing Gradient” Problem

- The parameters of the RNN are trained by gradient descent by *unrolling* T steps of the recurrence:

$$Y_{t+1} = \rho(A_{Y,Y} Y_t + A_{Y,X} X_t)$$

“Vanishing Gradient” Problem

- The parameters of the RNN are trained by gradient descent by *unrolling* T steps of the recurrence:

$$Y_{t+1} = \rho(A_{Y,Y} Y_t + A_{Y,X} X_t)$$

- For the purpose of updating the parameters, the loss at time *t* is thus expressed in terms of *T* previous hidden states:

$$\ell(\hat{O}_t, O_t) = G(Y_t, Y_{t-1}, \dots, Y_{t-T}, X_t, \dots, X_{t-T})$$

$$\frac{\partial \ell(\hat{O}_t, O_t)}{\partial A_{Y,Y}} = \sum_{i \le T} \frac{\partial G}{\partial Y_{t-i}} \frac{\partial Y_{t-i}}{\partial A_{Y,Y}}$$

$$= \sum_{i \le T} \frac{\partial G}{\partial Y_t} \left(\prod_{j \le i} \frac{\partial Y_{t-j}}{\partial Y_{t-j-1}} \right) \frac{\partial Y_{t-i}}{\partial A_{Y,Y}}$$

“Vanishing Gradient” Problem

- The terms connecting the hidden variables are

$$\frac{\partial Y_{t-j}}{\partial Y_{t-j-1}} = \mathrm{diag}(\rho'(Y_{t-j}))A_{Y,Y}^T$$

"Vanishing Gradient" Problem

- The terms connecting the hidden variables are

$$\frac{\partial Y_{t-j}}{\partial Y_{t-j-1}} = \mathrm{diag}(\rho'(Y_{t-j}))A_{Y,Y}^T$$

- It results that communicating information for large T is unstable:
  - If $\|A_{Y,Y}\| < 1$, then $\frac{\partial Y_t}{\partial Y_{t-T}} \to 0$ exponentially fast in T. (short memory regime)
  - If $\|A_{Y,Y}\| > 1$, then the gradients might grow exponentially.

“Vanishing Gradient” Problem

- The terms connecting the hidden variables are

$$\frac{\partial Y_{t-j}}{\partial Y_{t-j-1}} = \mathrm{diag}(\rho'(Y_{t-j}))A_{Y,Y}^T$$

- It results that communicating information for large T is unstable:
  - If $\|A_{Y,Y}\| < 1$, then $\frac{\partial Y_t}{\partial Y_{t-T}} \to 0$ exponentially fast in T. (short memory regime)
  - If $\|A_{Y,Y}\| > 1$, then the gradients might grow exponentially.
- *Gradient clipping* is a popular heuristic to address.
- One can constrain/initialize the transition matrices to be unitary [Arjovsky et al’15, Henaff et al,’16] to mitigate this problem.

Unitary Recurrent Networks

- Consider a state-space evolution of the form:

$$Y_{t+1} = A_{Y,Y} Y_t + \rho(A_{Y,X} X_t) \ , \ \text{with } A^*_{Y,Y} A_{Y,Y} = \mathbf{1} \ .$$

Unitary Recurrent Networks

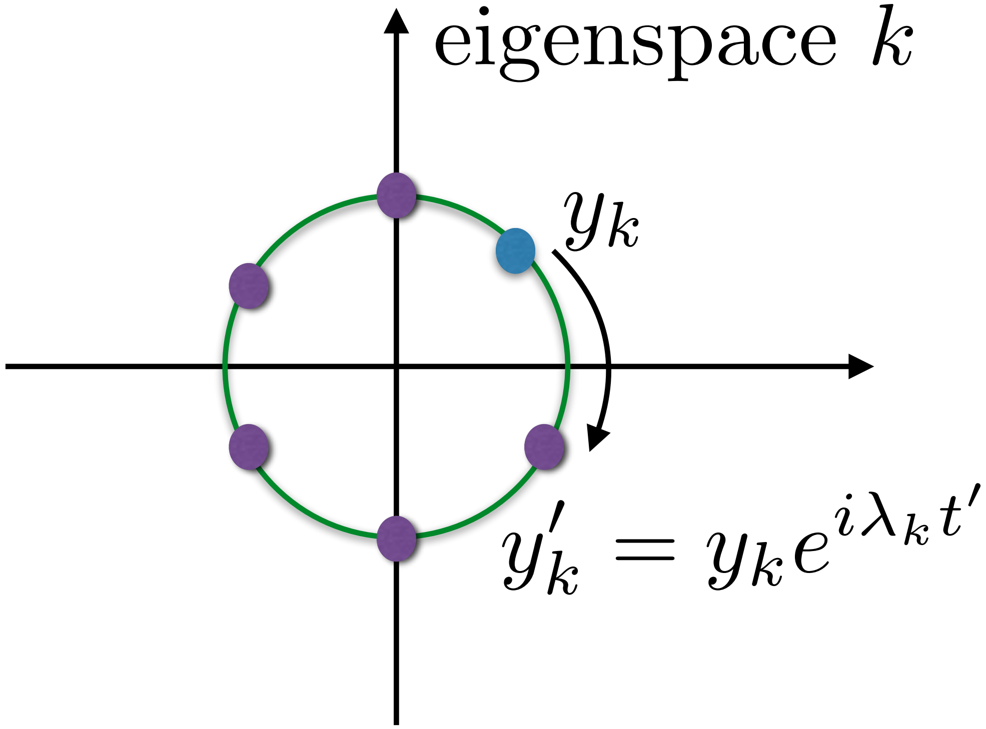
- Consider a state-space evolution of the form:

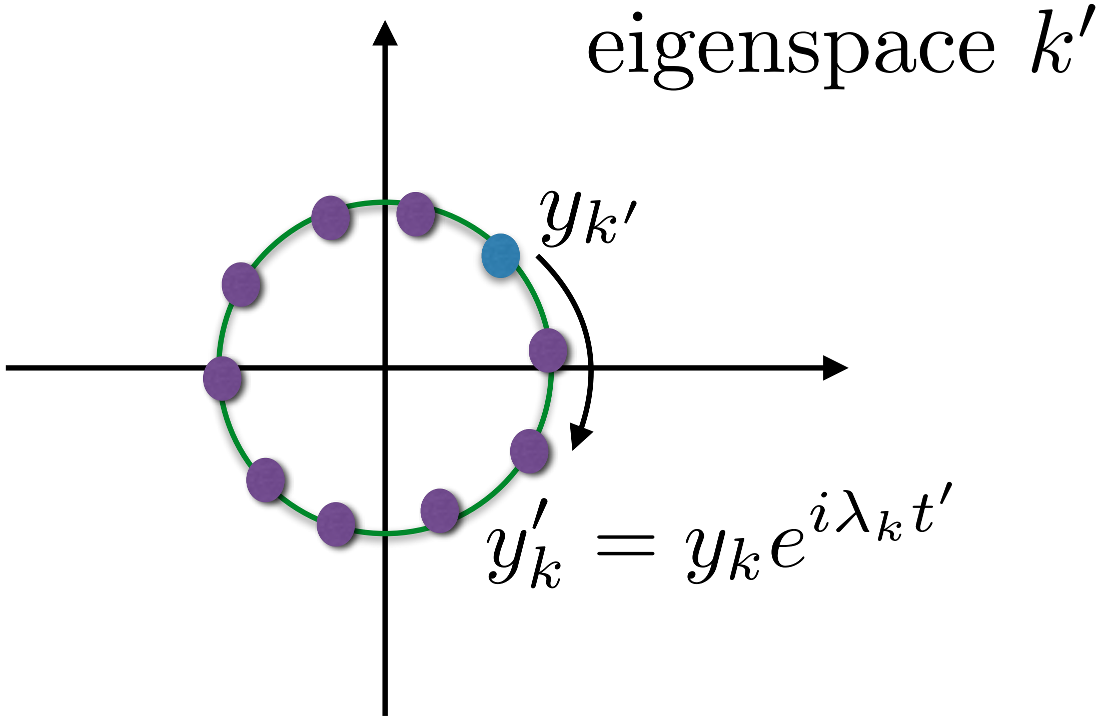
$$Y_{t+1} = A_{Y,Y} Y_t + \rho(A_{Y,X} X_t) \ , \ \text{with } A_{Y,Y}^* A_{Y,Y} = \mathbf{1} \ .$$

- The eigenvectors of $A_{Y,Y}$ are "clocks":

$$A_{Y,Y} = V \operatorname{diag}(e^{i\lambda_1}, \dots, e^{i\lambda_n}) V^*$$

Assume $\lambda_j \in \mathbb{Q}$:



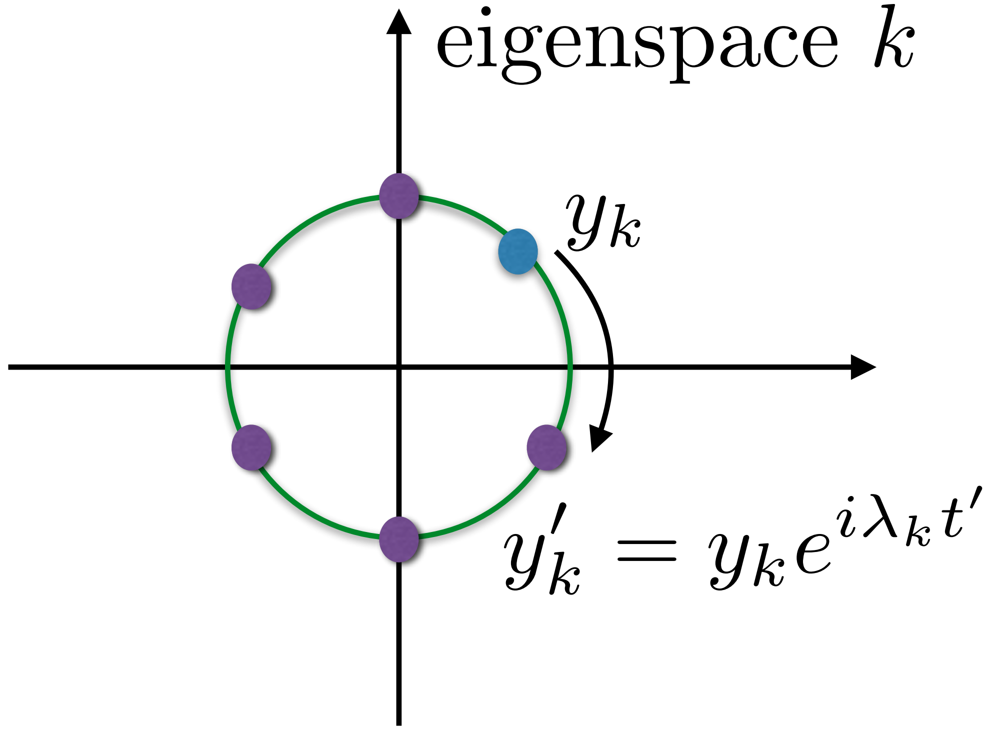


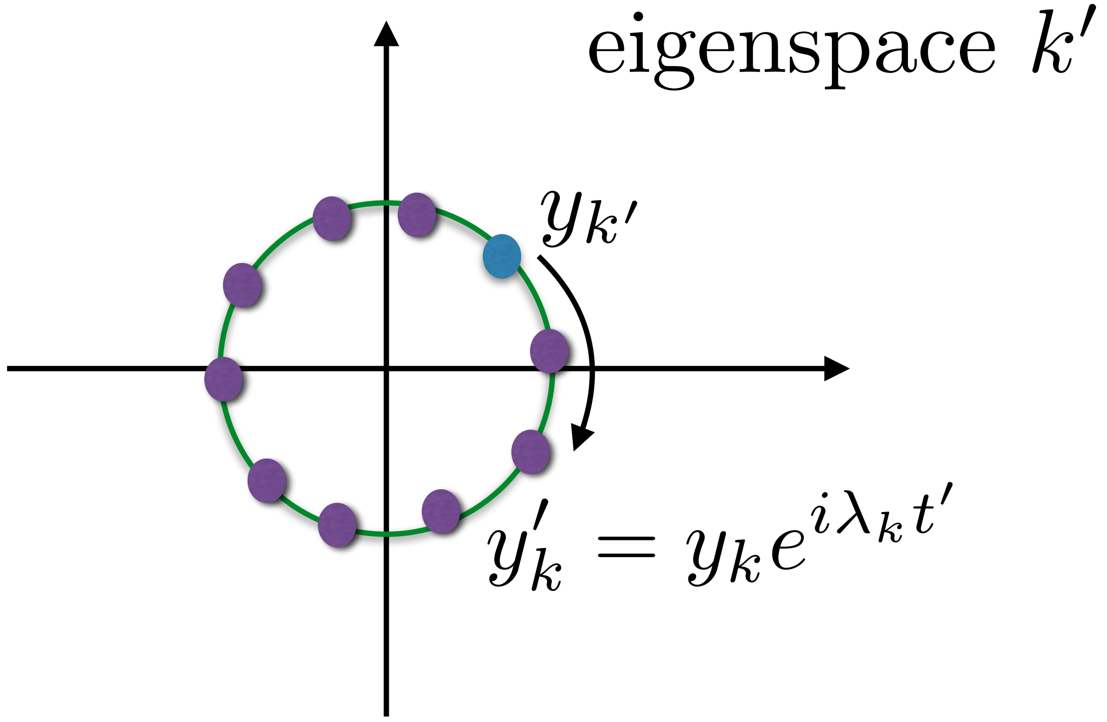
Unitary Recurrent Networks

- The eigenvectors of $A_{Y,Y}$ are "clocks":

$$A_{Y,Y} = V \mathrm{diag}(e^{i\lambda_1}, \ldots, e^{i\lambda_n}) V^*$$

Assume $\lambda_j \in \mathbb{Q}$:



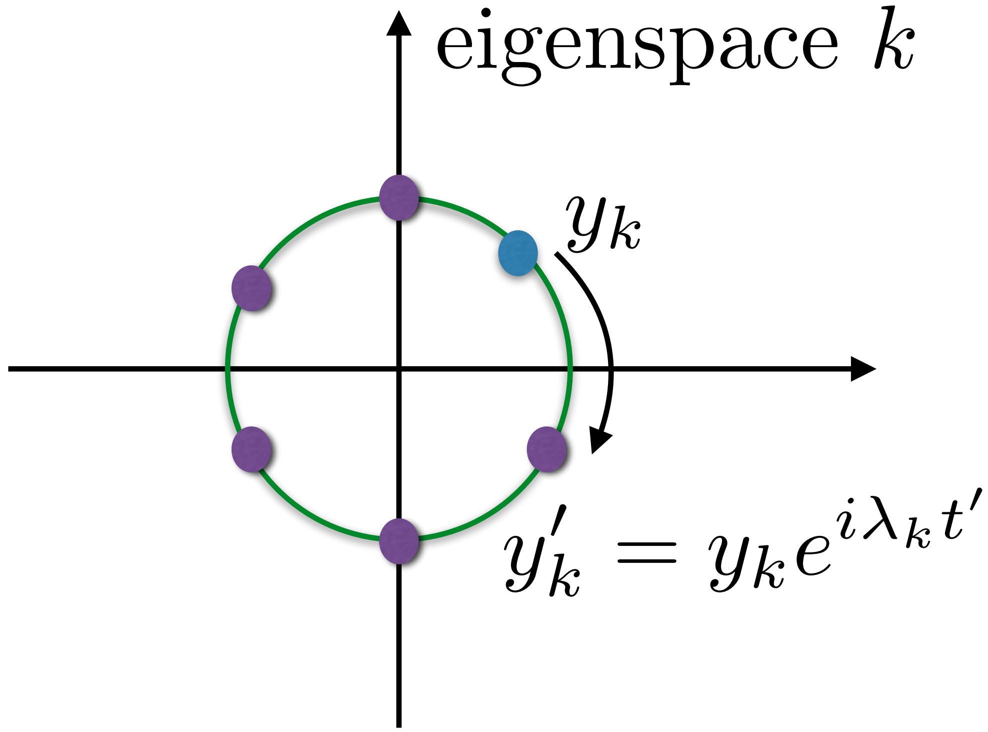


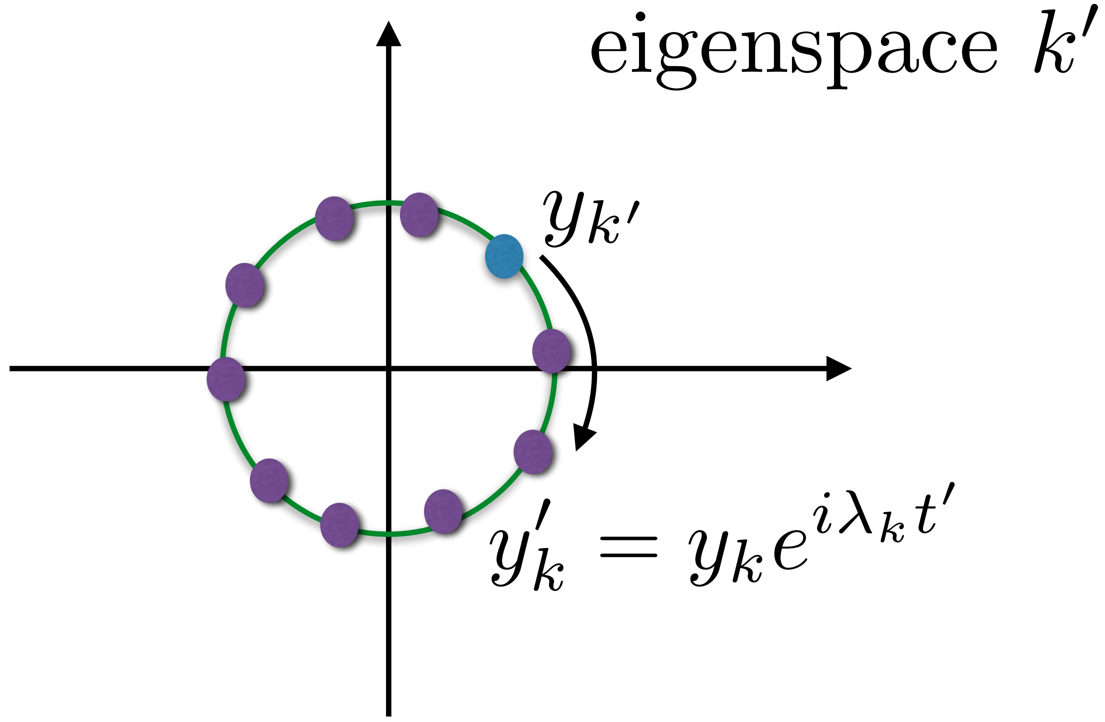
- Information is preserved and can be accessed periodically (eg for copying tasks).
- See [Henaff et al.'16] for more details.

Unitary Recurrent Networks

- The eigenvectors of $A_{Y,Y}$ are “clocks”:

$A_{Y,Y} = V \mathrm{diag}(e^{i\lambda_1}, \ldots, e^{i\lambda_n}) V^*$ Assume $\lambda_j \in \mathbb{Q}$:





- Information is preserved and can be accessed periodically (eg for copying tasks).
- See [Henaff et al.’16] for more details.
- Q: How else can we “keep” information?

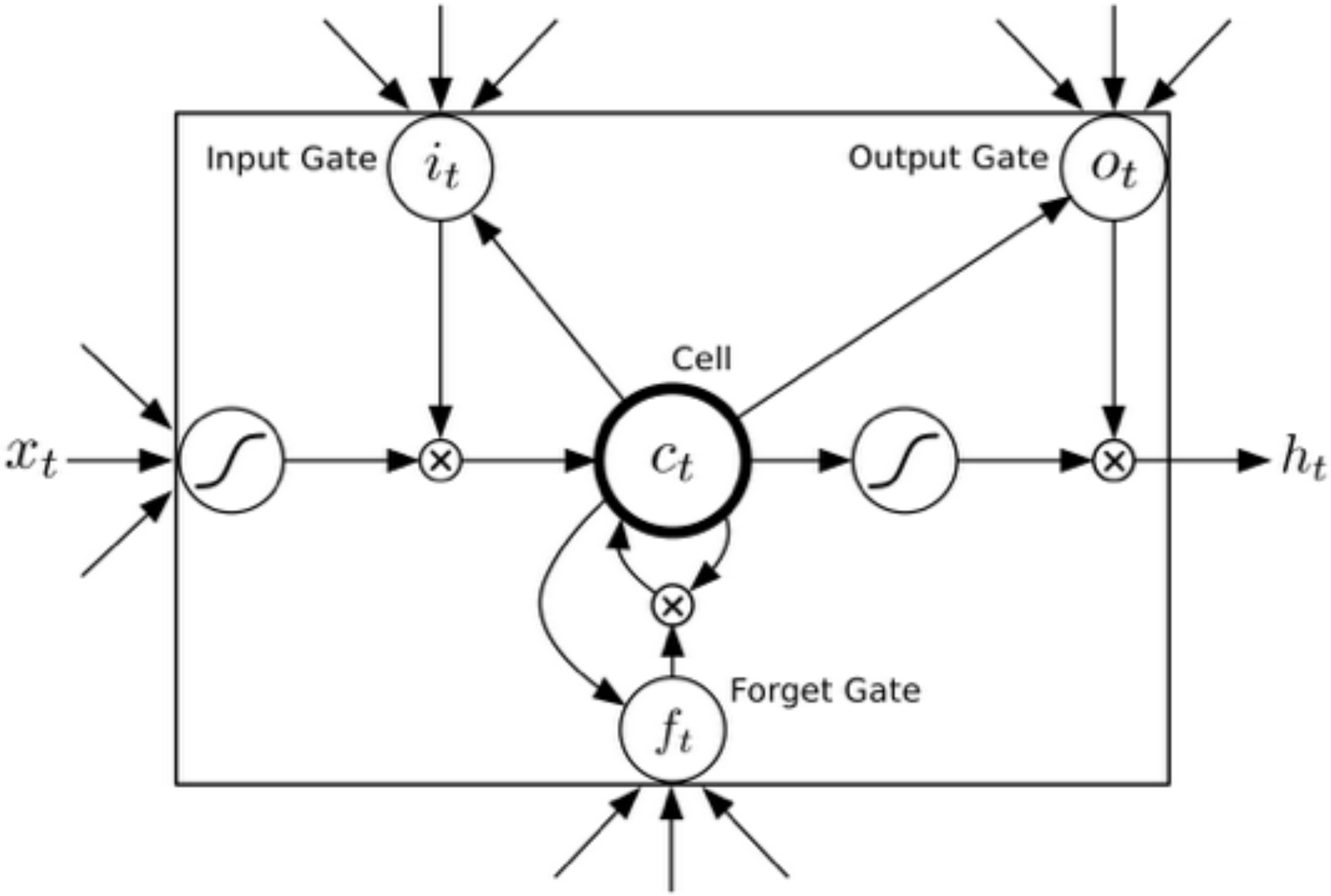
Long Short Term Memory (LSTM)

[Hochreiter & Schmidhuber'97]

Long Short Term Memory (LSTM)

[Hochreiter & Schmidhuber'97]

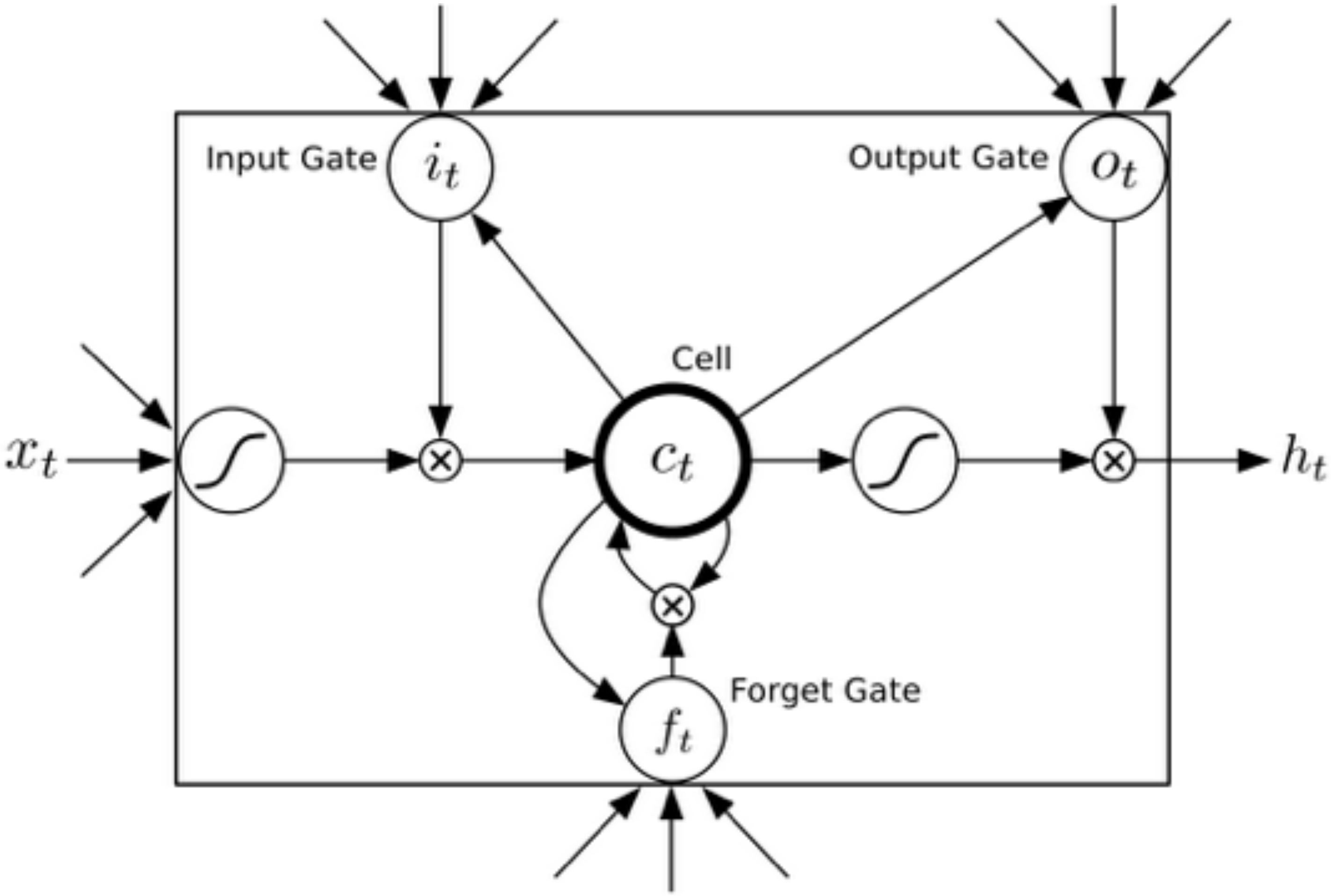
- A very popular and efficient alternative is to modify the transition operator using *gating mechanisms:*



Long Short Term Memory (LSTM)

[Hochreiter & Schmidhuber'97]

- A very popular and efficient alternative is to modify the transition operator using *gating mechanisms:*



- The *cell* is a memory that needs to be explicitly erased in order to disappear.
- What to store and when to write/erase is modeled with differentiable gates, trained with gradient descent.

RNN Perspectives and Open Questions

- Prediction Challenge: capture long-term dependencies with tractable models.

- Linear vs Non-linear state-space dynamics.
  - Can we trade-off higher dimensional linear dynamics with non-linear, lower-dimensional dynamics?
  - Role of gating and relationship with Residual Training. Optimization advantage or a more fundamental principle?

- Inference?

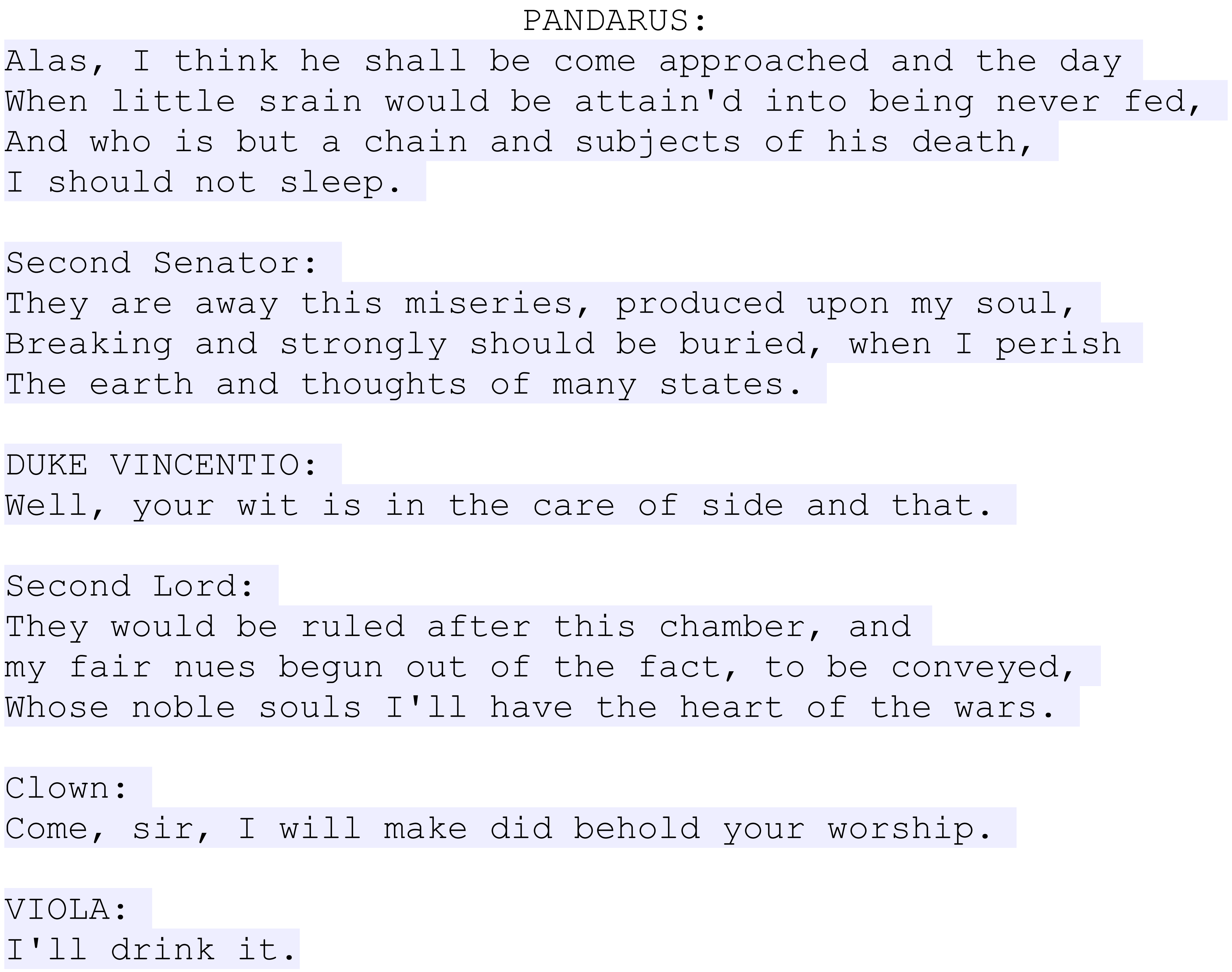
Examples: Language Modeling

- Recurrent Networks trained on character prediction. [Karpathy'15].

- The model includes several LSTM layers
- Text is generated by sampling from the multinomial output distribution, and feeding back the sample into the hidden-state evolution equation.

Language Modeling

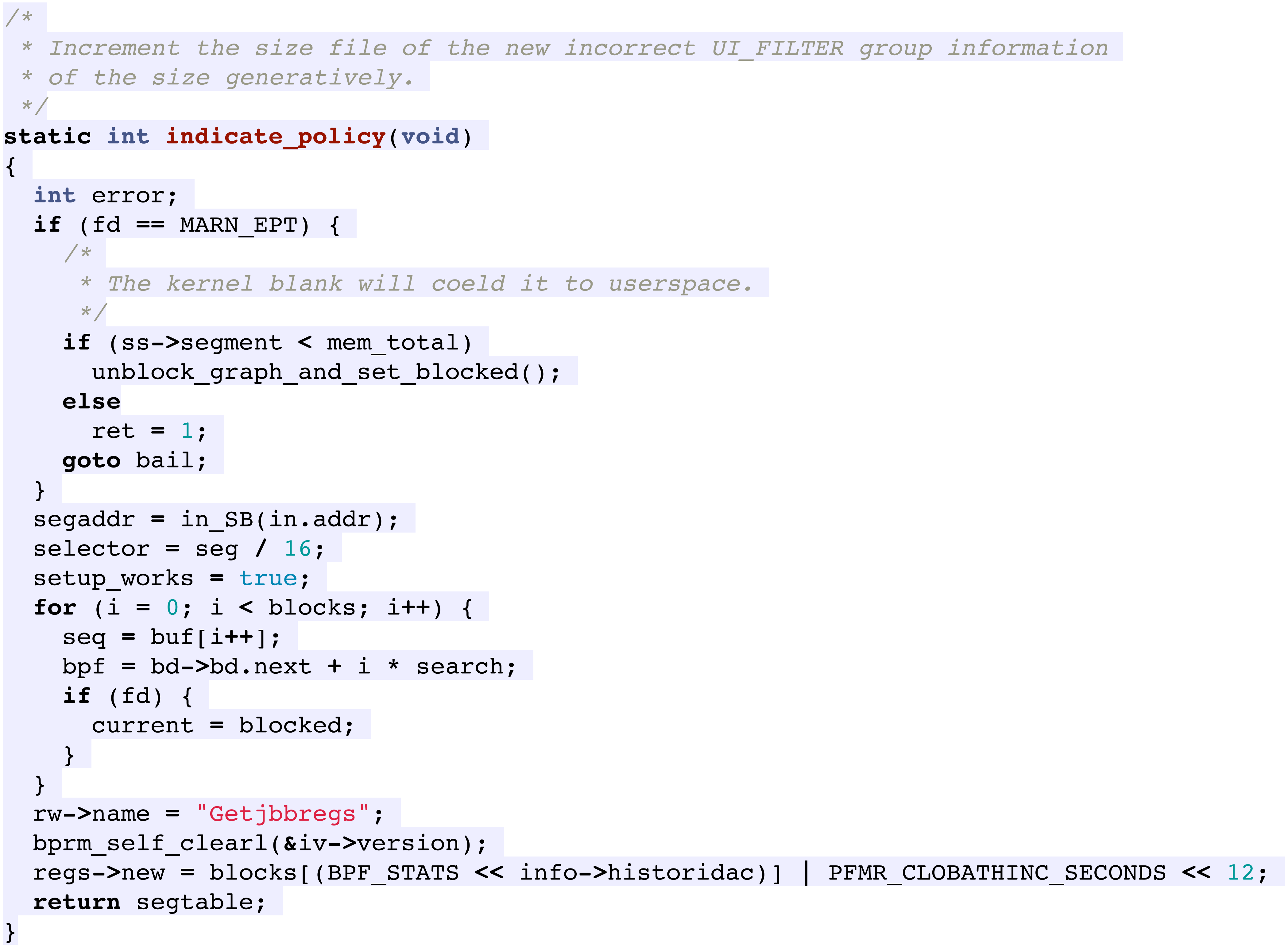
- Examples on Shakespeare training:



```
                          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.

Second Lord:
They would be ruled after this chamber, and
my fair nues begun out of the fact, to be conveyed,
Whose noble souls I'll have the heart of the wars.

Clown:
Come, sir, I will make did behold your worship.

VIOLA:
I'll drink it.
```

Language Modeling

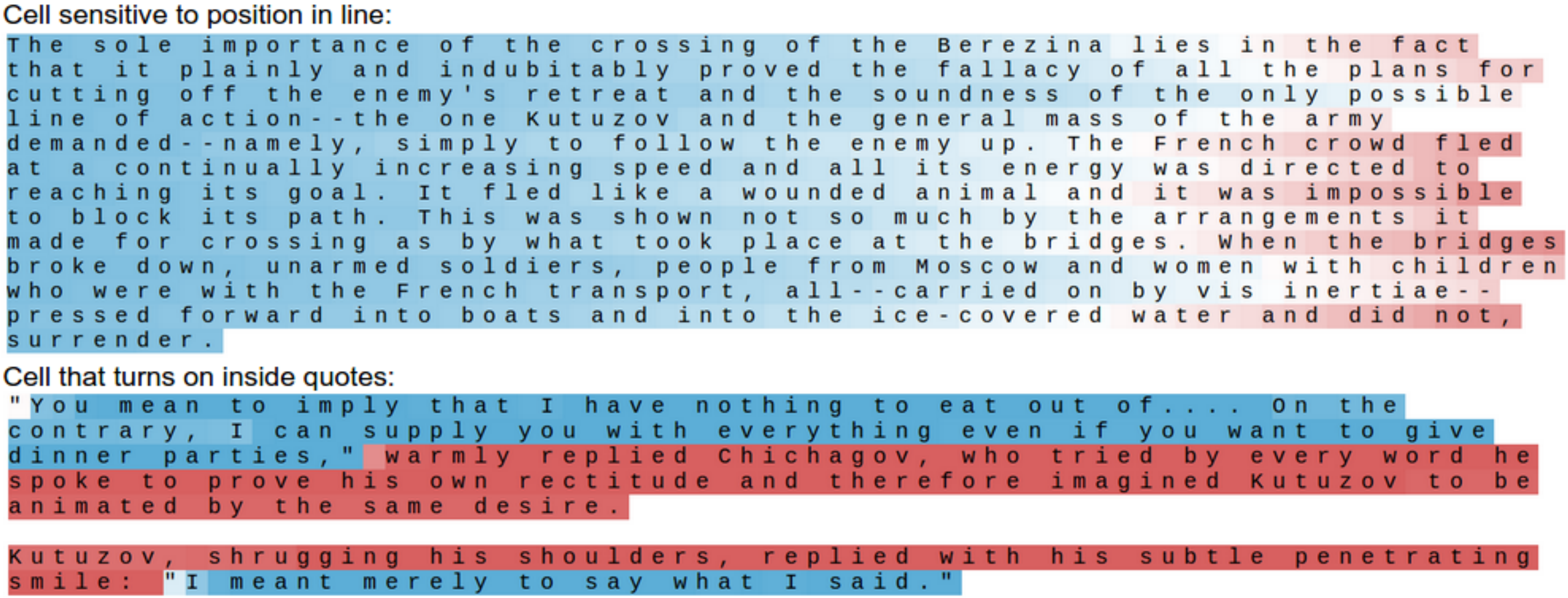
- Examples on Code generation (model has a million params):



```
/*
 * Increment the size file of the new incorrect UI_FILTER group information
 * of the size generatively.
 */
static int indicate_policy(void)
{
  int error;
  if (fd == MARN_EPT) {
    /*
     * The kernel blank will coeld it to userspace.
     */
    if (ss->segment < mem_total)
      unblock_graph_and_set_blocked();
    else
      ret = 1;
    goto bail;
  }
  segaddr = in_SB(in.addr);
  selector = seg / 16;
  setup_works = true;
  for (i = 0; i < blocks; i++) {
    seq = buf[i++];
    bpf = bd->bd.next + i * search;
    if (fd) {
      current = blocked;
    }
  }
  rw->name = "Getjbbregs";
  bprm_self_clearl(&iv->version);
  regs->new = blocks[(BPF_STATS << info->historidac)] | PFMR_CLOBATHINC_SECONDS << 12;
  return segtable;
}
```

Language Modeling

- Inspecting what the model learns:



Cell sensitive to position in line:

The sole importance of the crossing of the Berezina lies in the fact that it plainly and indubitably proved the fallacy of all the plans for cutting off the enemy's retreat and the soundness of the only possible line of action--the one Kutuzov and the general mass of the army demanded--namely, simply to follow the enemy up. The French crowd fled at a continually increasing speed and all its energy was directed to reaching its goal. It fled like a wounded animal and it was impossible to block its path. This was shown not so much by the arrangements it made for crossing as by what took place at the bridges. When the bridges broke down, unarmed soldiers, people from Moscow and women with children who were with the French transport, all--carried on by vis inertiae--pressed forward into boats and into the ice-covered water and did not, surrender.

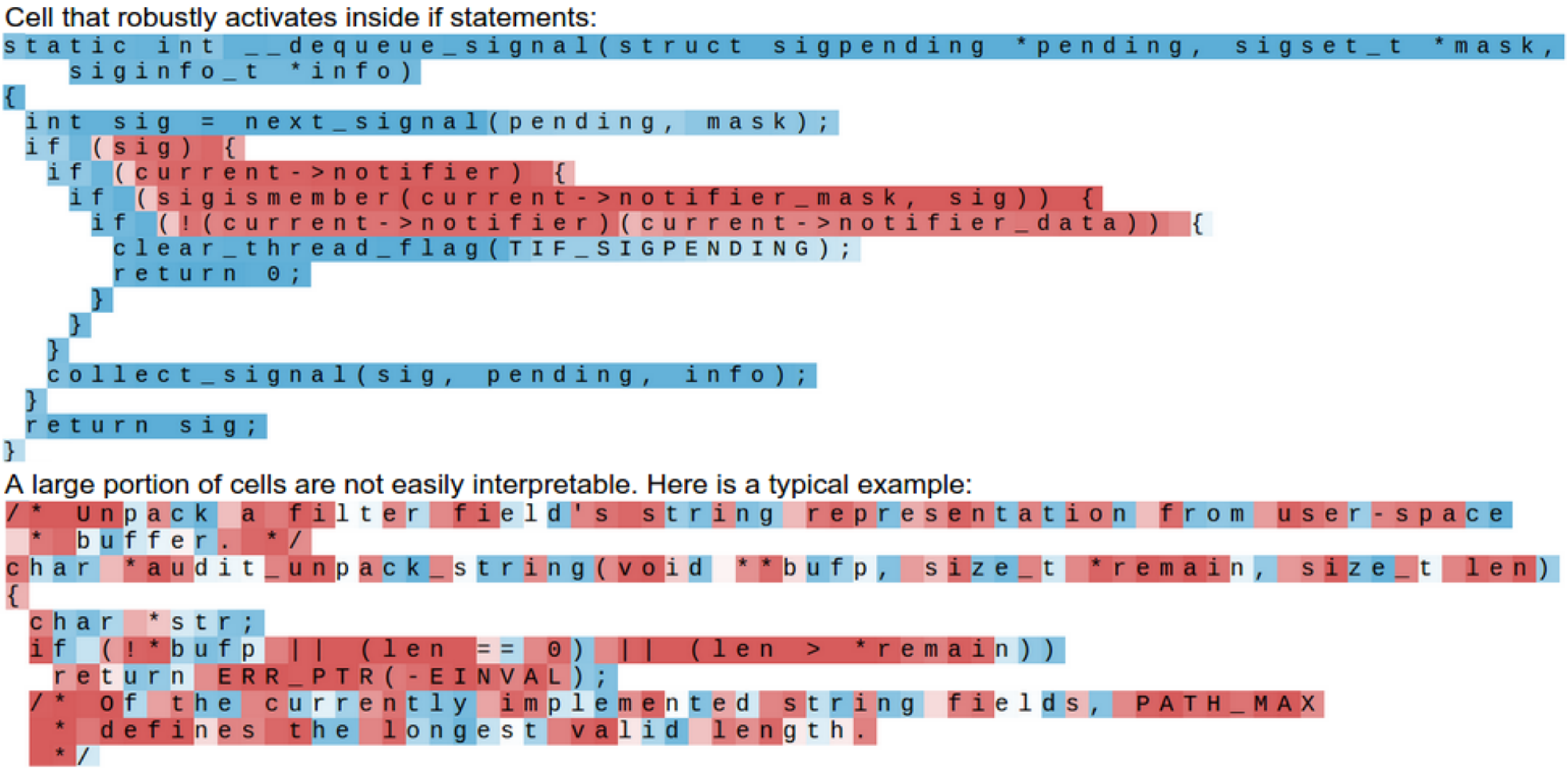
Cell that turns on inside quotes:

"You mean to imply that I have nothing to eat out of.... On the contrary, I can supply you with everything even if you want to give dinner parties," warmly replied Chichagov, who tried by every word he spoke to prove his own rectitude and therefore imagined Kutuzov to be animated by the same desire.

Kutuzov, shrugging his shoulders, replied with his subtle penetrating smile: "I meant merely to say what I said."

Language Modeling

- Inspecting what the model learns:



Cell that robustly activates inside if statements:

```
static int __dequeue_signal(struct sigpending *pending, sigset_t *mask,
   siginfo_t *info)
{
 int sig = next_signal(pending, mask);
 if (sig) {
  if (current->notifier) {
   if (sigismember(current->notifier_mask, sig)) {
    if (!(current->notifier)(current->notifier_data)) {
     clear_thread_flag(TIF_SIGPENDING);
     return 0;
    }
   }
  }
  collect_signal(sig, pending, info);
 }
 return sig;
}
```

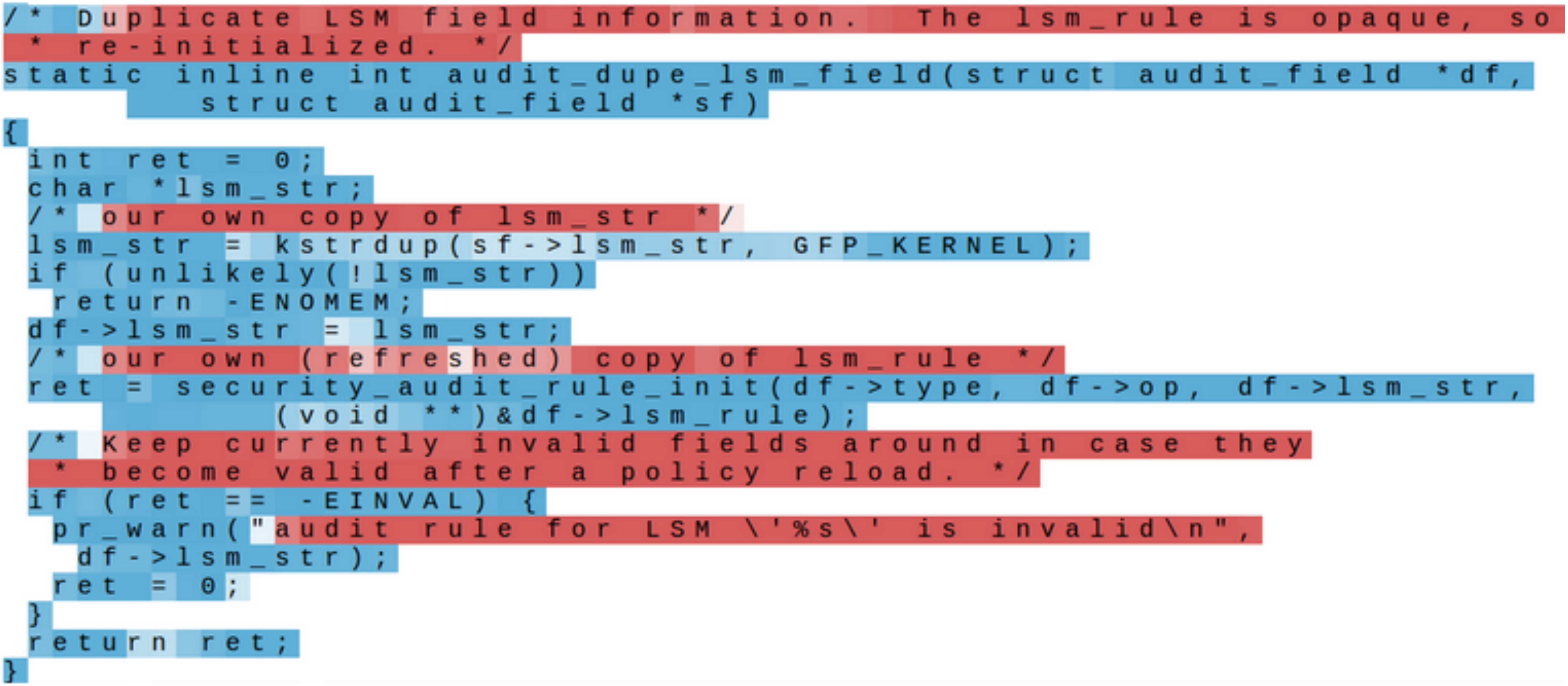
A large portion of cells are not easily interpretable. Here is a typical example:

```
/* Unpack a filter field's string representation from user-space
 * buffer. */
char *audit_unpack_string(void **bufp, size_t *remain, size_t len)
{
 char *str;
 if (!*bufp || (len == 0) || (len > *remain))
  return ERR_PTR(-EINVAL);
 /* Of the currently implemented string fields, PATH_MAX
  * defines the longest valid length.
  */
```

Language Modeling

- Inspecting what the model learns:

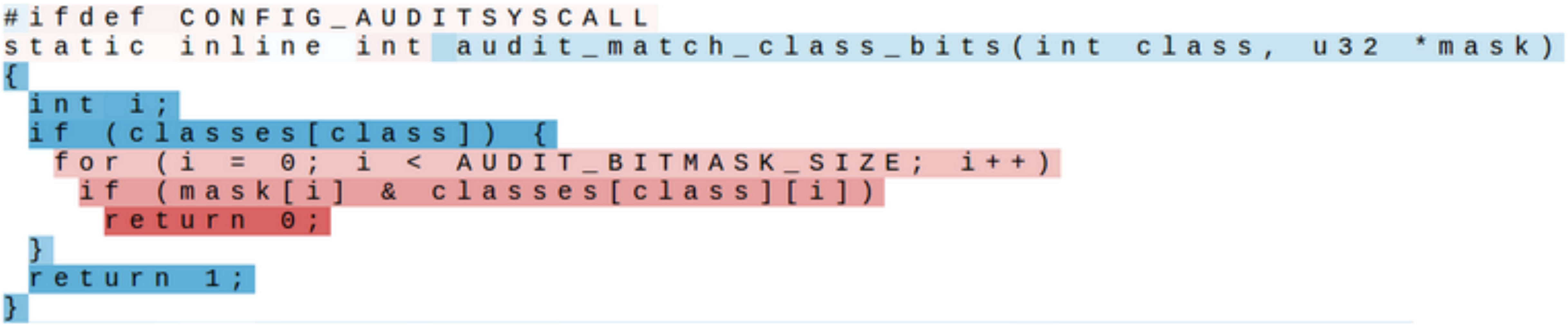
Cell that turns on inside comments and quotes:



```
/* Duplicate LSM field information.  The lsm_rule is opaque, so
 * re-initialized. */
static inline int audit_dupe_lsm_field(struct audit_field *df,
        struct audit_field *sf)
{
 int ret = 0;
 char *lsm_str;
 /* our own copy of lsm_str */
 lsm_str = kstrdup(sf->lsm_str, GFP_KERNEL);
 if (unlikely(!lsm_str))
  return -ENOMEM;
 df->lsm_str = lsm_str;
 /* our own (refreshed) copy of lsm_rule */
 ret = security_audit_rule_init(df->type, df->op, df->lsm_str,
           (void **)&df->lsm_rule);
 /* Keep currently invalid fields around in case they
  * become valid after a policy reload. */
 if (ret == -EINVAL) {
  pr_warn("audit rule for LSM \'%s\' is invalid\n",
   df->lsm_str);
  ret = 0;
 }
 return ret;
}
```

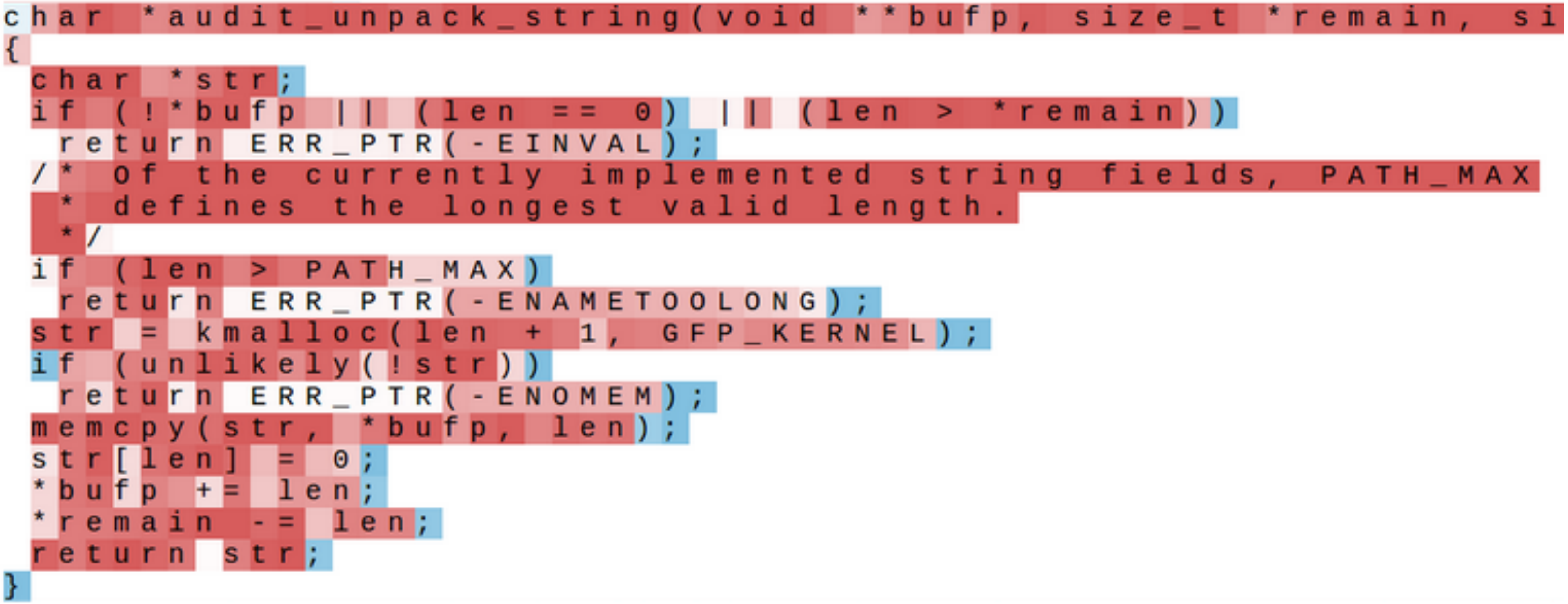
Language Modeling

Cell that is sensitive to the depth of an expression:



```
#ifdef CONFIG_AUDITSYSCALL
static inline int audit_match_class_bits(int class, u32 *mask)
{
 int i;
 if (classes[class]) {
  for (i = 0; i < AUDIT_BITMASK_SIZE; i++)
   if (mask[i] & classes[class][i])
    return 0;
 }
 return 1;
}
```

Cell that might be helpful in predicting a new line. Note that it only turns on for some ")":



```
char *audit_unpack_string(void **bufp, size_t *remain, si
{
 char *str;
 if (!*bufp || (len == 0) || (len > *remain))
  return ERR_PTR(-EINVAL);
 /* Of the currently implemented string fields, PATH_MAX
  * defines the longest valid length.
  */
 if (len > PATH_MAX)
  return ERR_PTR(-ENAMETOOLONG);
 str = kmalloc(len + 1, GFP_KERNEL);
 if (unlikely(!str))
  return ERR_PTR(-ENOMEM);
 memcpy(str, *bufp, len);
 str[len] = 0;
 *bufp += len;
 *remain -= len;
 return str;
}
```

RNNs vs n-gram statistics

- On discrete sequences such as characters, a powerful baseline is given by *n-gram models.*

$$X_i \in \mathcal{X}\ ,\ \mathcal{X} = \{c_1, \dots, c_K\}\ .$$

- multinomial distribution:

$$p(X_{t+1} = c_k \mid X_t = c_{j_n}, X_{t-1} = c_{j_{n-1}}, \dots, X_{t-n+1} = c_{j_1}) = \pi_k(j_1, \dots, j_n)$$

RNNs vs n-gram statistics

- On discrete sequences such as characters, a powerful baseline is given by *n-gram models.*

$$X_i \in \mathcal{X} \ , \ \mathcal{X} = \{c_1, \ldots, c_K\} \ .$$

- multinomial distribution:

$$p(X_{t+1} = c_k \mid X_t = c_{j_n}, X_{t-1} = c_{j_{n-1}}, \ldots, X_{t-n+1} = c_{j_1}) = \pi_k(j_1, \ldots, j_n)$$

- Maximum Likelihood estimate:

$$\widehat{\pi}_k(j_1, \ldots, j_n) = \text{empirical frequency of the sequence } (c_{j_1}, \ldots, c_{j_n}, c_k)$$

- This model explicitly remembers *all n last terms* (but nothing else).
- Efficient storing using sparsity and histogram clipping techniques.

RNNs vs n-gram statistics

[Y. Goldberg]

- Natural Text (Shakespeare) using n=2:

```
Fif thad yourty
Fare sid on Che as al my he sheace ing.

Thy your thy ove dievest sord wit whand of sold iset?

Commet laund hant.

KINCESARGANT:
Out aboy tur Pome you musicell losts, blover.

How difte quainge to sh,
And usbas ey will Chor bacterea, and mens grou:
Princeser,
'Tis a but be;
I hends ing noth much?
```

RNNs vs n-gram statistics

[Y. Goldberg]

- Natural Text (Shakespeare) using n=4:

```
THUR:
Will comfited our flight offend make thy love;
Brothere is oats at on thes:'--why, cross and so
her shouldestruck at one their hearina in all go to lives of
Costag,
To his he tyrant of you our the fill we hath trouble an over me?

KING JOHN:
Great though I gain; for talk to mine and to the Christ: a right
him out
To kiss;
And to a kindness not of loves you Gower and to the stray
Than hers of ever in this flight?
I do me,
After, wild,
Or, if I into ebbs, by fair too me knowned worship asider
thyself-skin ever is again, and eat behold speak imposed thy
hand. Give and cours not sweet you of sorrow then; for they are
gone! Then the prince, I
see your likewis, is thee; and him for is them hearts, we have a
kiss,
And it is the come, some an eanly; you that am fire: prince when
'twixt young piece, that honourish we fort
```

RNNs vs n-gram statistics

[Y. Goldberg]

- Natural Text (Shakespeare) using n=10:

```
SEBASTIAN:
Do I stand till the break off.

BIRON:
Hide thy head.

VENTIDIUS:
He purposeth to Athens: whither, with the vow
I made to handle you.

FALSTAFF:
My good knave.

MALVOLIO:
Sad, lady! I could be forgiven you, you're welcome. Give ear, sir, my
doublet and hose and leave this present death.

Second Gentleman:
Who may that she confess it is my lord enraged and forestalled ere we come
to be a man. Drown thyself?

APEMANTUS:
Ho, ho! I laugh to see your beard!

BOYET:
Madam, in great extremes of passion as she
discovers it.
```

RNNs vs n-gram statistics

[Y. Goldberg]

- Linux source code using n=10:

```
~~/*
 * linux/kernel/time.c
 * Please report this on hardware.
 */
void irq_mark_irq(unsigned long old_entries, eval);

        /*
         * Divide only 1000 for ns^2 -> us^2 conversion values don't overflow:
        seq_puts(m, "\ttramp: %pS",
                   (void *)class->contending_point]++;
    if (likely(t->flags & WQ_UNBOUND)) {
        /*
         * Update inode information. If the
         * slowpath and sleep time (abs or rel)
 * @rmtp: remaining (either due
 * to consume the state of ring buffer size. */
    header_size - size, in bytes, of the chain.
         */
        BUG_ON(!error);
        } while (cgrp) {
        if (old) {
        if (kdb_continue_catastrophic;
#endif

/*
 * for the deadlock.\n");
        return 0;
}
#endif
```

RNN vs n-gram statistics

- English language (syntactics) is reasonably well modeled with explicit short memory models.
  - The RNN just need to remember the last n characters well.

- However, other discrete time series require selective long-term memory:
  - Source code requires opening/closing brackets and indentation.

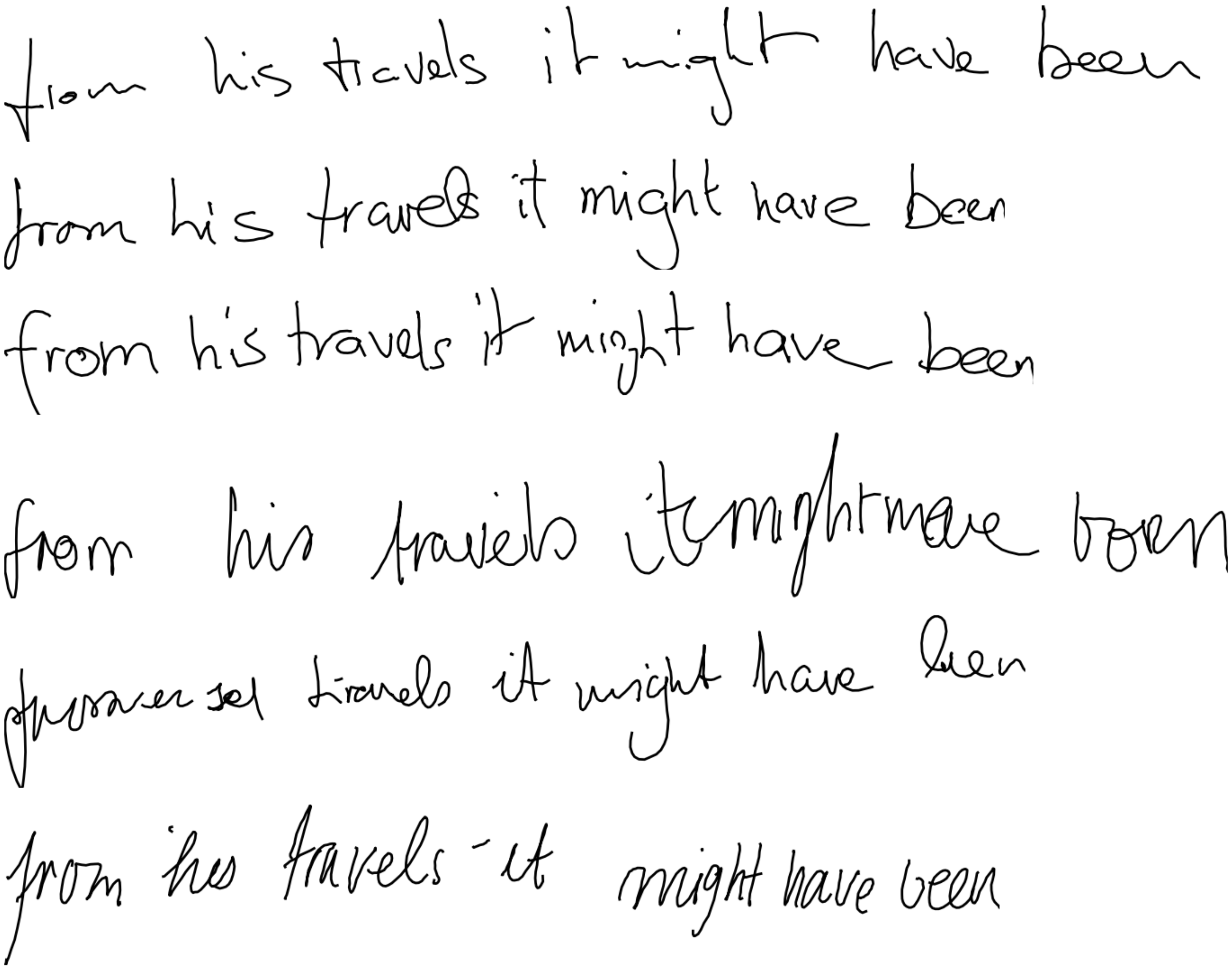
Music Generation

- Train an LSTM to model piano music using midi files

- Once the model is trained, we use it as a generator by sampling from the output distribution and feeding back the samples into the model to generate the next sample.

- Examples from [http://www.hexahedria.com/2015/08/03/composing-music-with-recurrent-neural-networks/]

Handwitten Synthesis



from his travels it might have been

from his travels it might have been

from his travels it might have been

from his travels itmightmave been

[illegible] travels it might have been

from his travels - it might have been

[A. Graves]

Sequence Structured Prediction

- Many tasks require a prediction from sequence to sequence:

Machine Translation

`There is a light that never goes out`

↕

`Il y a une lumière qui ne disparait jamais`

Question Answering

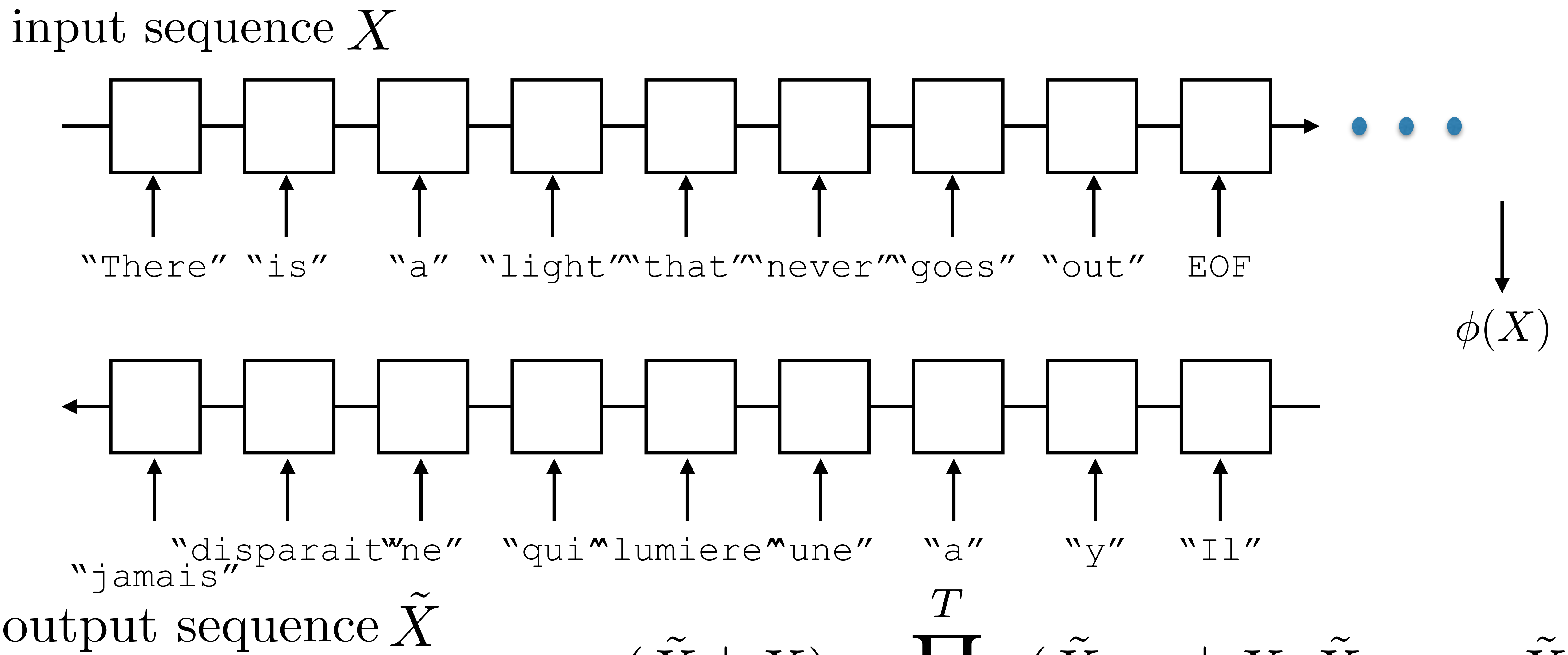
`What is the best ramen place in the Bay Area?`

↕

`Ramen Shop, in Rockridge`

Sequence Structured Prediction

- Conditional model:
  - Input sequence is used to initialize the state of the output decoder.



$$p(\tilde{X} \mid X) = \prod_{t=0}^{T} p(\tilde{X}_{t+1} \mid X, \tilde{X}_1, \ldots, \tilde{X}_t)$$

“Attention” Mechanisms

- Limits of sequence-to-sequence model.
  - All the information of the input sequence is contained in the vector $\phi(X)$
  - As the length of input increases, we require more information to perform the translation.

ВОЙНА и МИРЪ.

СОЧИНЕНІЕ

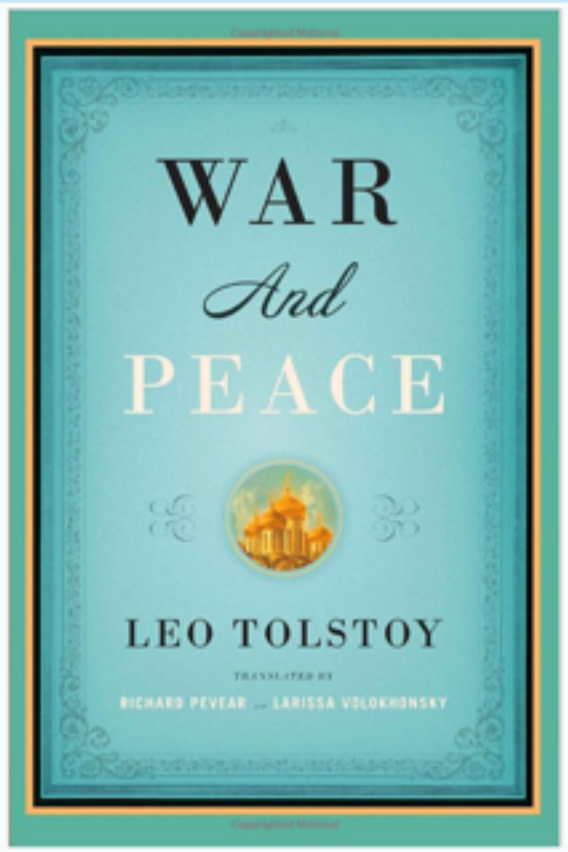
Графа Л. Н. Толстаго.

ТОМЪ ШЕСТОЙ.

МОСКВА.

[illegible]

1869.



WAR

And

PEACE

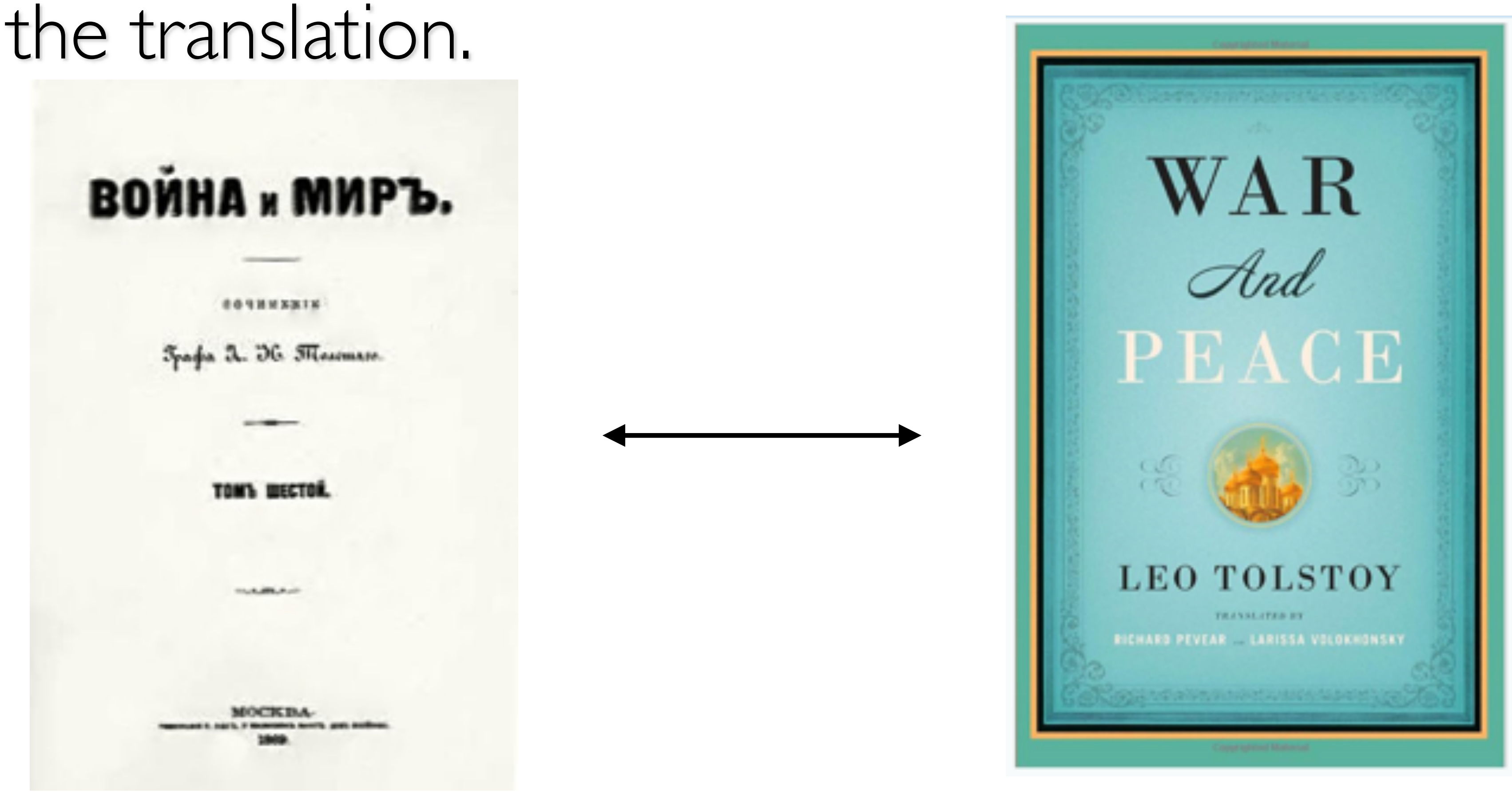
LEO TOLSTOY

TRANSLATED BY

RICHARD PEVEAR — LARISSA VOLOKHONSKY

"Attention" Mechanisms

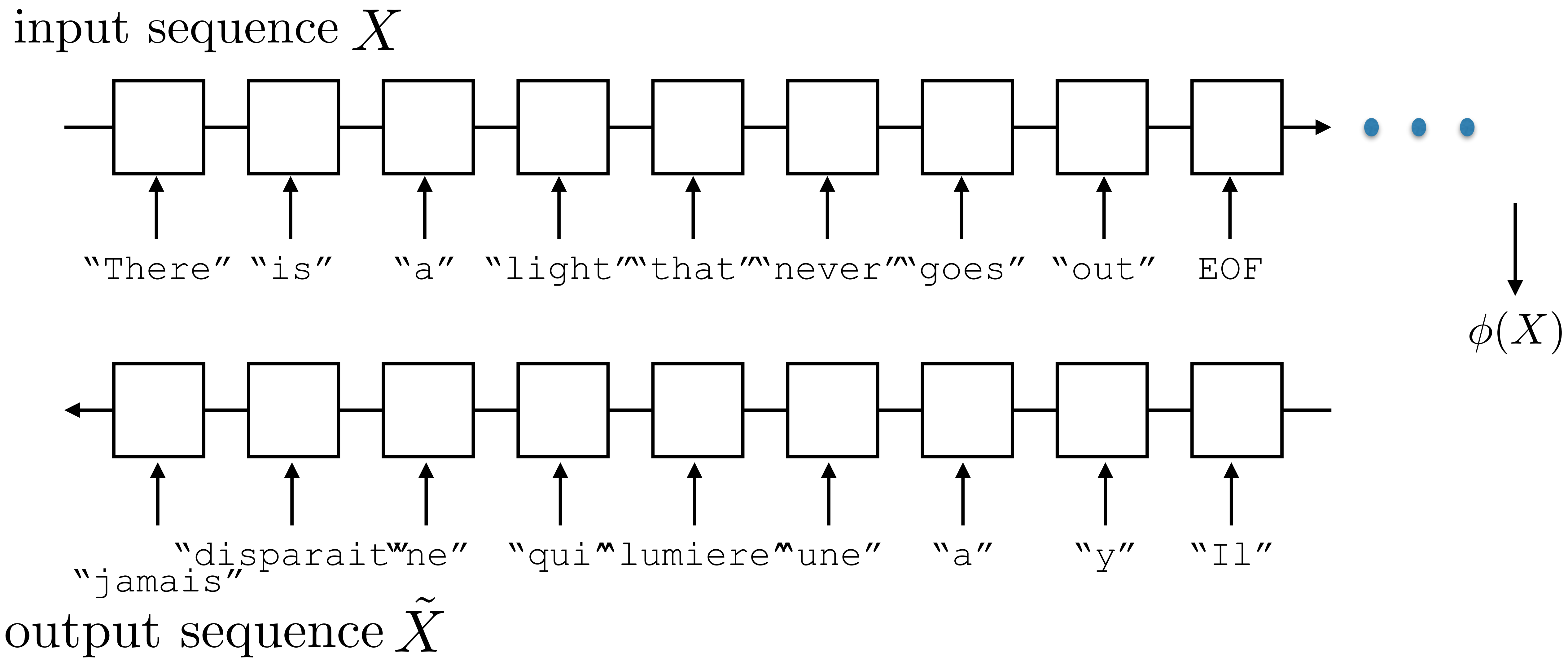
- Limits of sequence-to-sequence model.
  - All the information of the input sequence is contained in the vector $\phi(X)$
  - As the length of input increases, we require more information to perform the translation.



  - Although the ***global*** amount of information grows, the ***local*** amount of information required to translate does not. How to exploit it?

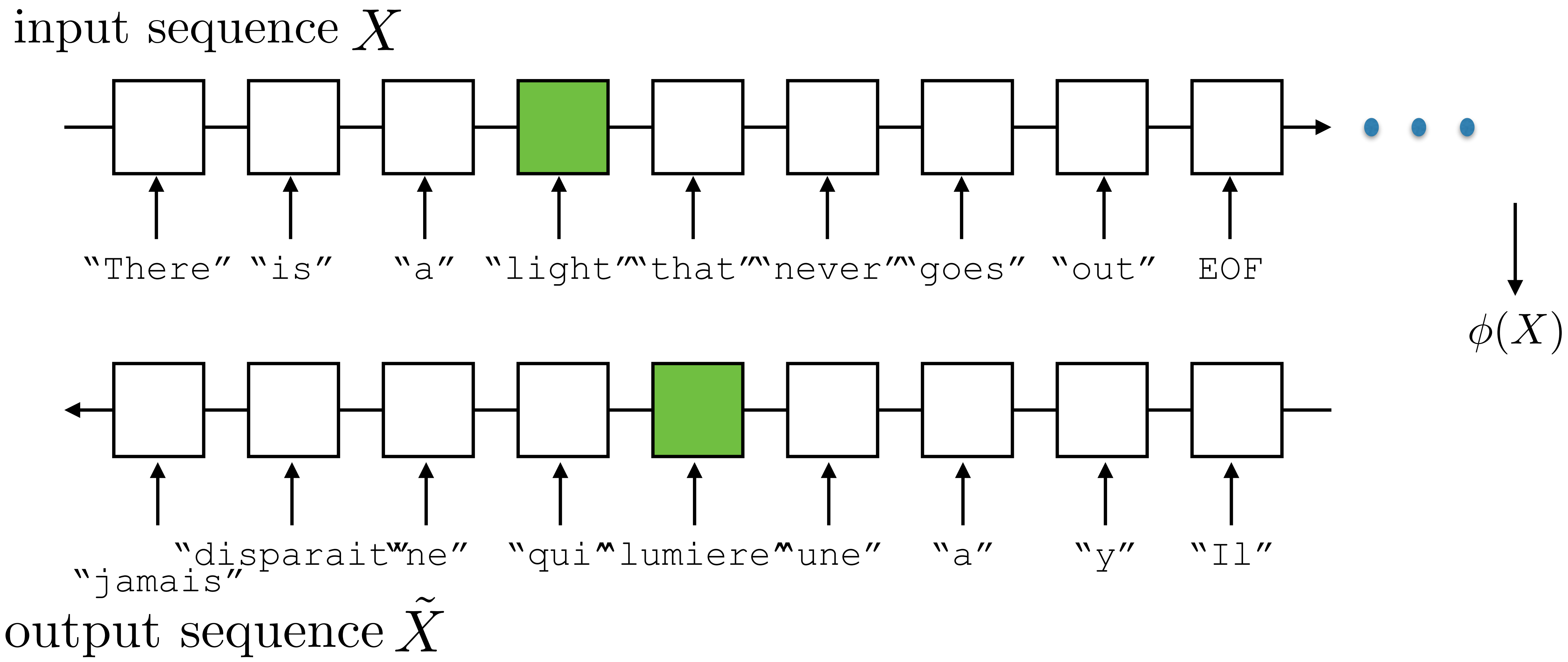
“Attention” Mechanisms

[Badhanu et al.’15]



“Attention” Mechanisms

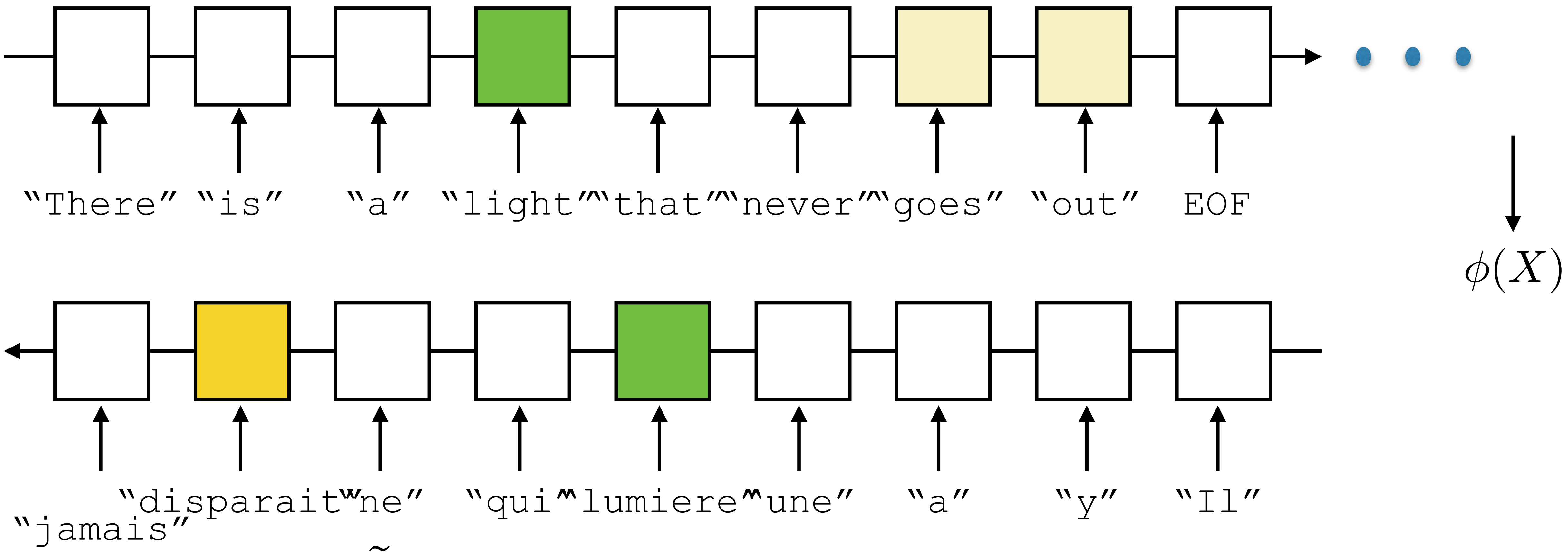
[Badhanu et al.’15]



“Attention” Mechanisms

[Badhanu et al.’15]

input sequence X



“There” “is” “a” “light”“that”“never”“goes” “out” EOF

$\phi(X)$

“jamais” “disparait”“ne” “qui”“lumiere”“une” “a” “y” “Il”

output sequence $\tilde{X}$

$$p(\tilde{X} \mid X) = \prod_{t=0}^{T} p(\tilde{X}_{t+1} \mid att(X, \tilde{X}_1, \ldots, \tilde{X}_t), \tilde{X}_1, \ldots, \tilde{X}_t)$$

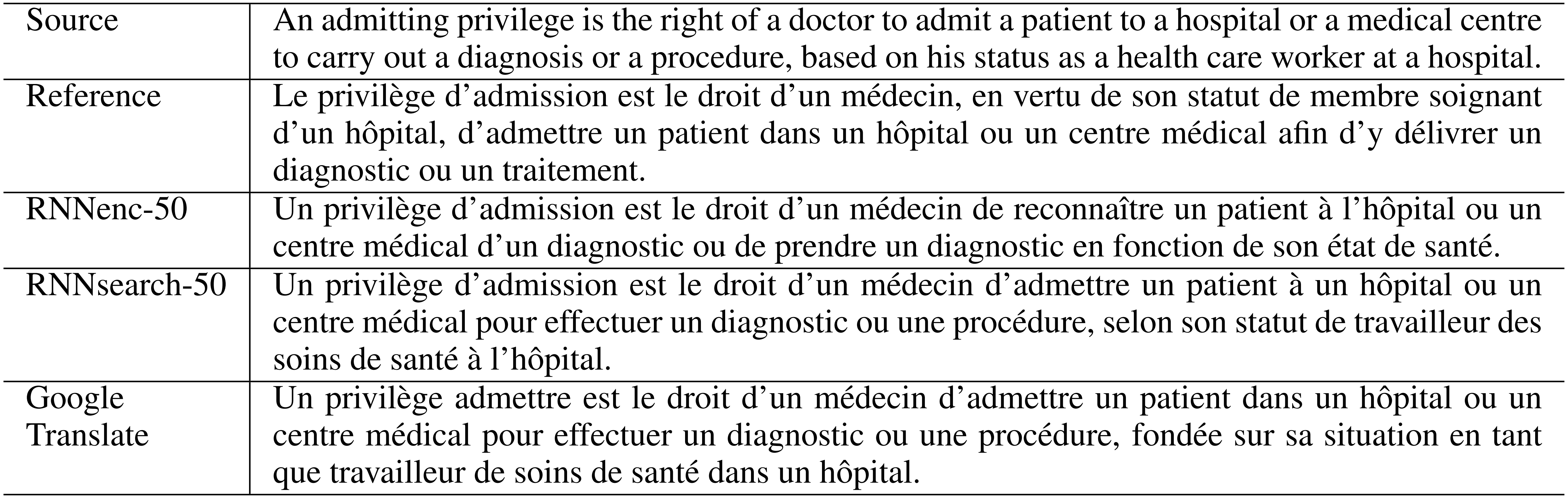
“Attention” Mechanisms

[Badhanu et al.’15]

- Pros
  - Generalizes to larger input/output sequences.

- Challenges
  - Harder to train
  - How to address larger memories efficiently?
  - Learning where to look?

Machine Translation



Source	An admitting privilege is the right of a doctor to admit a patient to a hospital or a medical centre to carry out a diagnosis or a procedure, based on his status as a health care worker at a hospital.
Reference	Le privilège d'admission est le droit d'un médecin, en vertu de son statut de membre soignant d'un hôpital, d'admettre un patient dans un hôpital ou un centre médical afin d'y délivrer un diagnostic ou un traitement.
RNNenc-50	Un privilège d'admission est le droit d'un médecin de reconnaître un patient à l'hôpital ou un centre médical d'un diagnostic ou de prendre un diagnostic en fonction de son état de santé.
RNNsearch-50	Un privilège d'admission est le droit d'un médecin d'admettre un patient à un hôpital ou un centre médical pour effectuer un diagnostic ou une procédure, selon son statut de travailleur des soins de santé à l'hôpital.
Google Translate	Un privilège admettre est le droit d'un médecin d'admettre un patient dans un hôpital ou un centre médical pour effectuer un diagnostic ou une procédure, fondée sur sa situation en tant que travailleur de soins de santé dans un hôpital.

[Badhanu et al., '15]