NYU CDS, SPRING 2018

INTRODUCTION TO DATA SCIENCE

TEXT MINING

- Natural language models
  - Bag of words
  - N-gram
  - Feature vector
  - Markov model
  - Recurrent neural network

BAG OF WORDS

- Multiset
- Does not preserve order information

"Alice sent a message to Bob"

TERM FREQUENCY

- How prevalent a term is in a single document.

d1 jazz music has a swing rhythm

d2 swing is hard to explain

d3 swing rhythm is a natural rhythm

	a	explain	hard	has	is	jazz	music	natural	rhythm	swing	to
d1	1	0	0	1	0	1	1	0	1	1	0
d2	0	1	1	0	1	0	0	0	0	1	1
d3	1	0	0	0	1	0	0	1	2	1	0

TERM FREQUENCY

- Normalization (to lowercase), stemming (remove suffix), stopwords (removed).

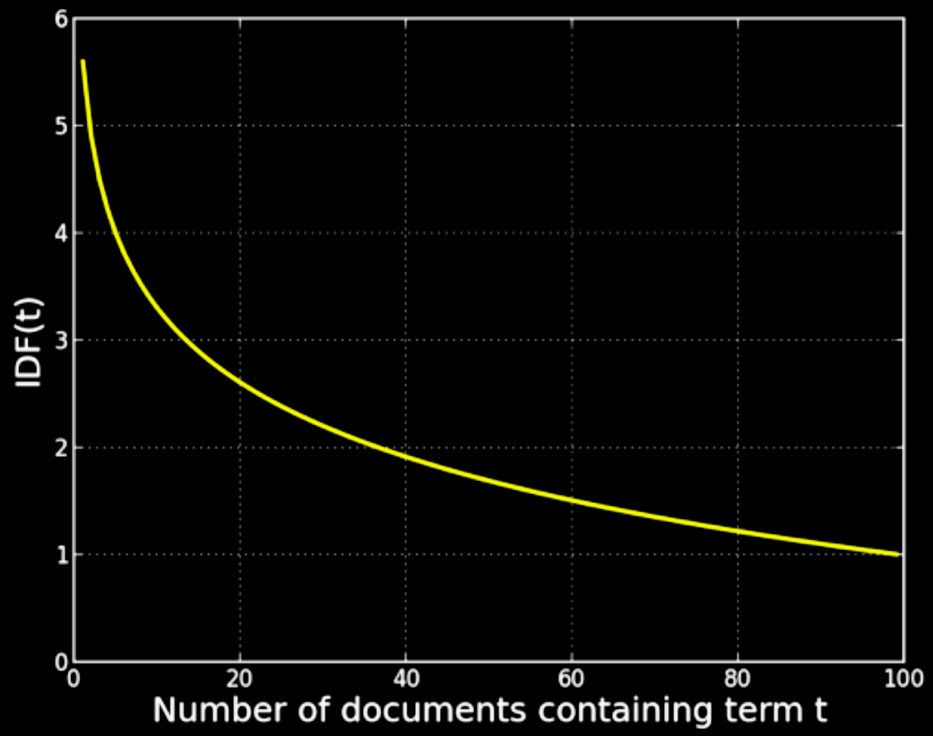
> Microsoft Corp and Skype Global today announced that they have entered into a definitive agreement under which Microsoft will acquire Skype, the leading Internet communications company, for $8.5 billion in cash from the investor group led by Silver Lake. The agreement has been approved by the boards of directors of both Microsoft and Skype.

Term	Count	Term	Count	Term	Count	Term	Count
skype	3	microsoft	3	agreement	2	global	1
approv	1	announc	1	acquir	1	lead	1
definit	1	lake	1	communic	1	internet	1
board	1	led	1	director	1	corp	1
compani	1	investor	1	silver	1	billion	1

INVERSE DOCUMENT FREQUENCY

- Boost of term for being rare in entire corpus of documents

$$\mathrm{IDF}(t) = 1 + \log\left(\frac{\text{Total number of documents}}{\text{Number of documents containing } t}\right)$$



TFIDF

- Product of term frequency & inverse document frequency

$$\mathrm{TFIDF}(t, d) = \mathrm{TF}(t, d) \times \mathrm{IDF}(t)$$

TFIDF EXAMPLE

- Jazz musician biographies from Wikipedia

Charlie Parker

Charles "Charlie" Parker, Jr., was an American jazz saxophonist and composer. Miles Davis once said, "You can tell the history of jazz in four words: Louis Armstrong. Charlie Parker." Parker acquired the nickname "Yardbird" early in his career and the shortened form, "Bird," which continued to be used for the rest of his life, inspired the titles of a number of Parker compositions, […]

Duke Ellington

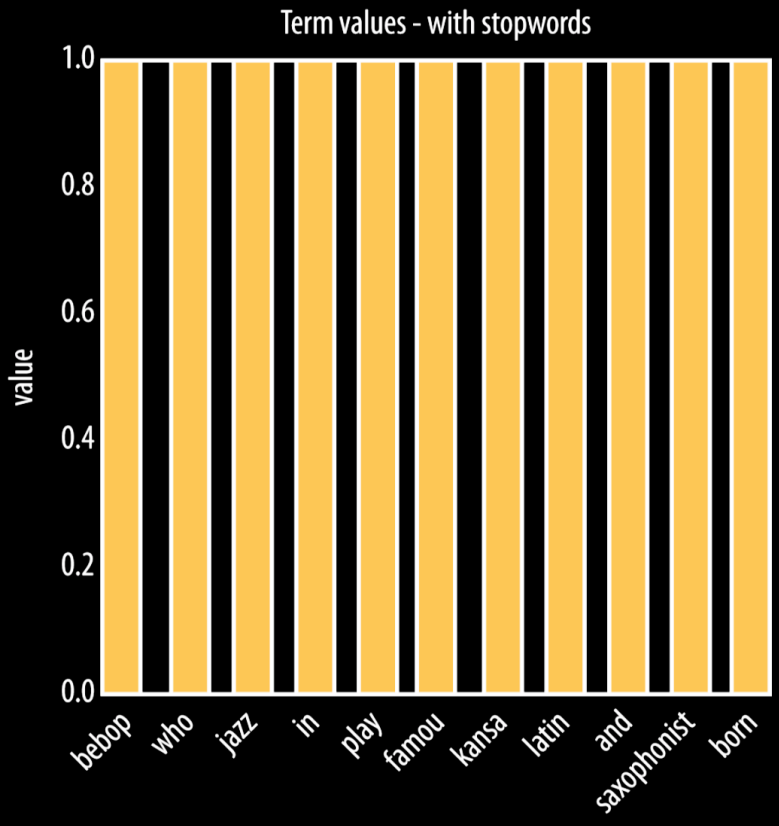
Edward Kennedy "Duke" Ellington was an American composer, pianist, and big-band leader. Ellington wrote over 1,000 compositions. In the opinion of Bob Blumenthal of *The Boston Globe*, "in the century since his birth, there has been no greater composer, American or otherwise, than Edward Kennedy Ellington." A major figure in the history of jazz, Ellington's music stretched into various other genres, including blues, gospel, film scores, popular, and classical.[…]

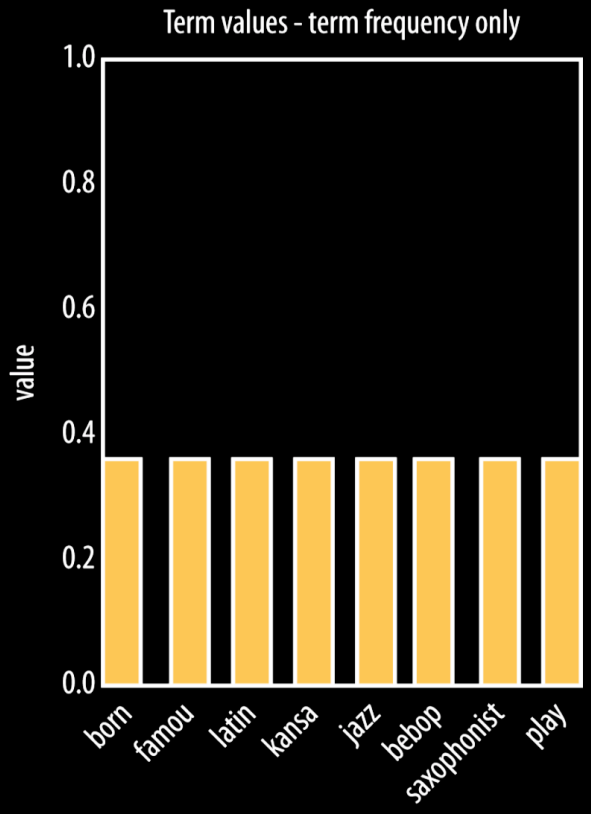
Miles Davis

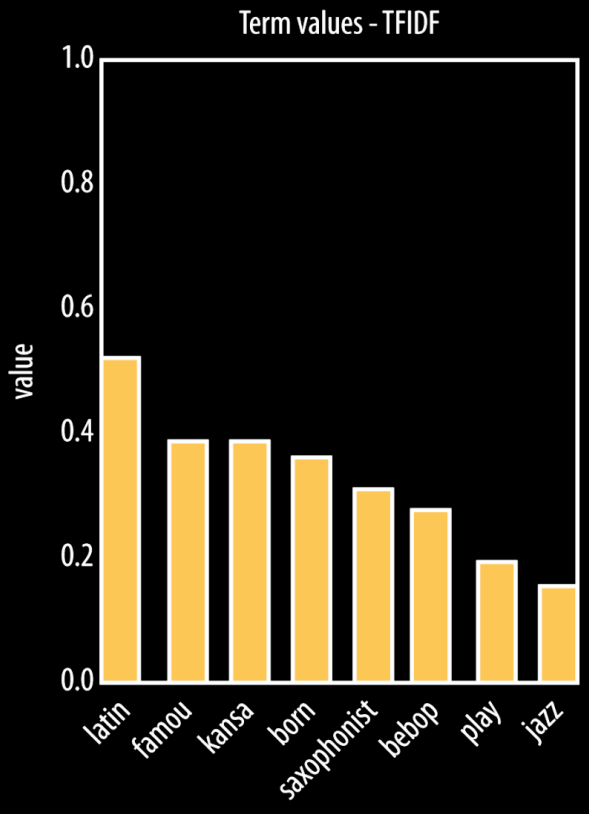
Miles Dewey Davis III was an American jazz musician, trumpeter, bandleader, and composer. Widely considered one of the most influential musicians of the 20th century, Miles Davis was, with his musical groups, at the forefront of several major developments in jazz music, including bebop, cool jazz, hard bop, modal jazz, and jazz fusion.[…]

TFIDF EXAMPLE

- "Famous jazz saxophonist born in Kansas who played bebop and latin." Query is treated as a document:







TFIDF EXAMPLE

- Cosine similarity between TFIDF of each musician's biography and query.

Musician	Similarity	Musician	Similarity
Charlie Parker	0.135	Count Basie	0.119
Dizzie Gillespie	0.086	John Coltrane	0.079
Art Tatum	0.050	Miles Davis	0.050
Clark Terry	0.047	Sun Ra	0.030
Dave Brubeck	0.027	Nina Simone	0.026
Thelonius Monk	0.025	Fats Waller	0.020
Charles Mingus	0.019	Duke Ellington	0.017
Benny Goodman	0.016	Louis Armstrong	0.012

ENTROPY AS EXPECTED VALUE OF IDF(T) AND IDF(NOT T)

$$p(t) = \frac{\text{Number of documents containing } t}{\text{Total number of documents}}$$

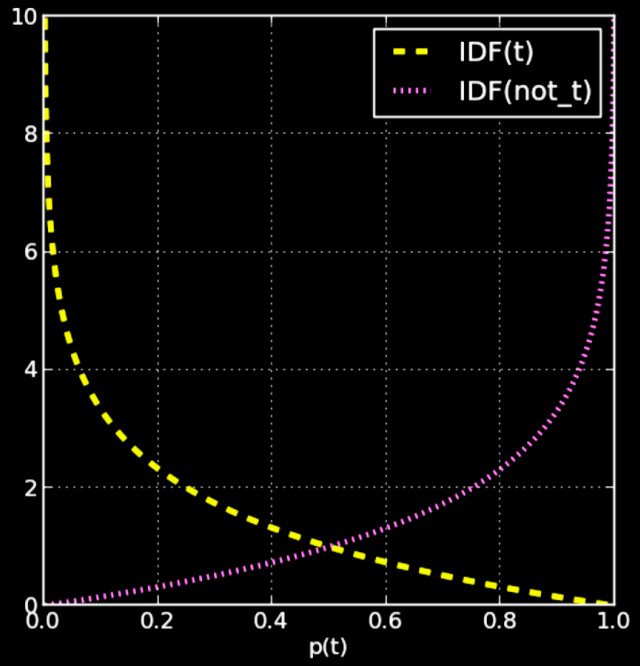
$$\text{IDF}(t) = 1 + \log\left(\frac{\text{Total number of documents}}{\text{Number of documents containing } t}\right)$$

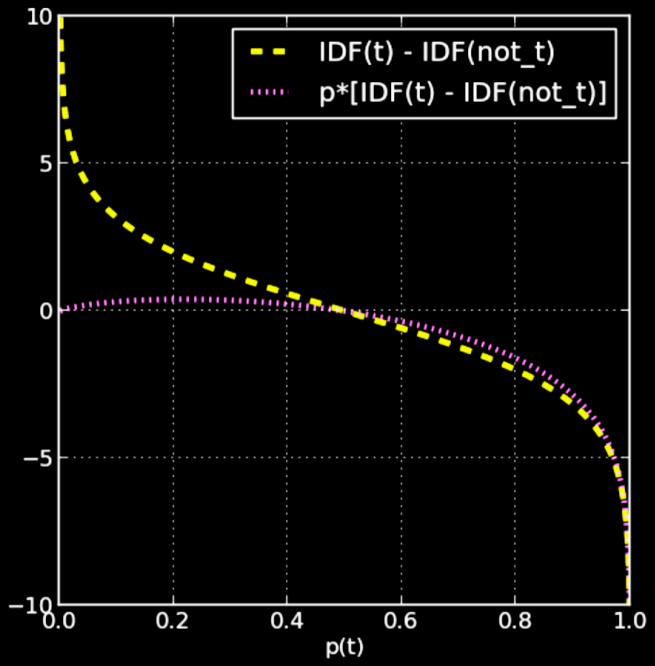
IDF(t) is basically $log(1/p)$

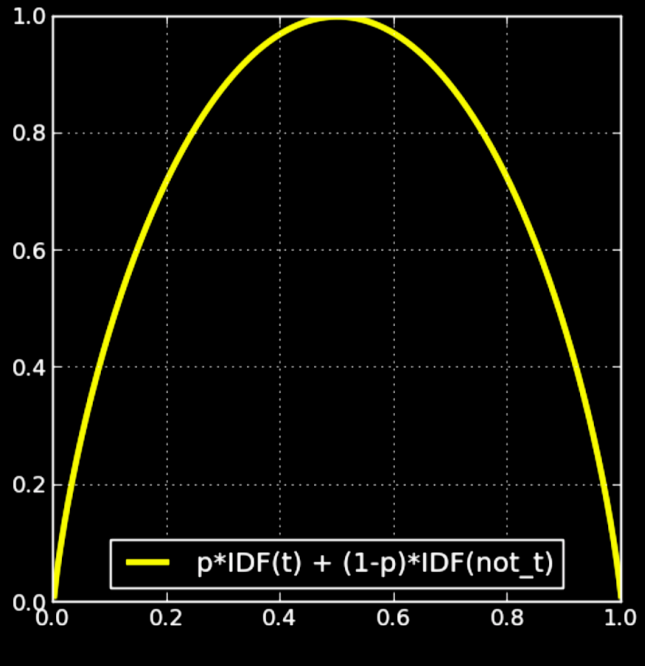
$$\text{IDF}(not\_T) = \log 1 / (1 - p) = - \log (1 - p)$$

$$\begin{aligned} \text{entropy}(t) &= -p \log (p) - (1 - p) \log (1 - p) \\ &= p \cdot \text{IDF}(t) - (1 - p)[- \text{IDF}(not\_t)] \\ &= p \cdot \text{IDF}(t) + (1 - p)[\text{IDF}(not\_t)] \end{aligned}$$

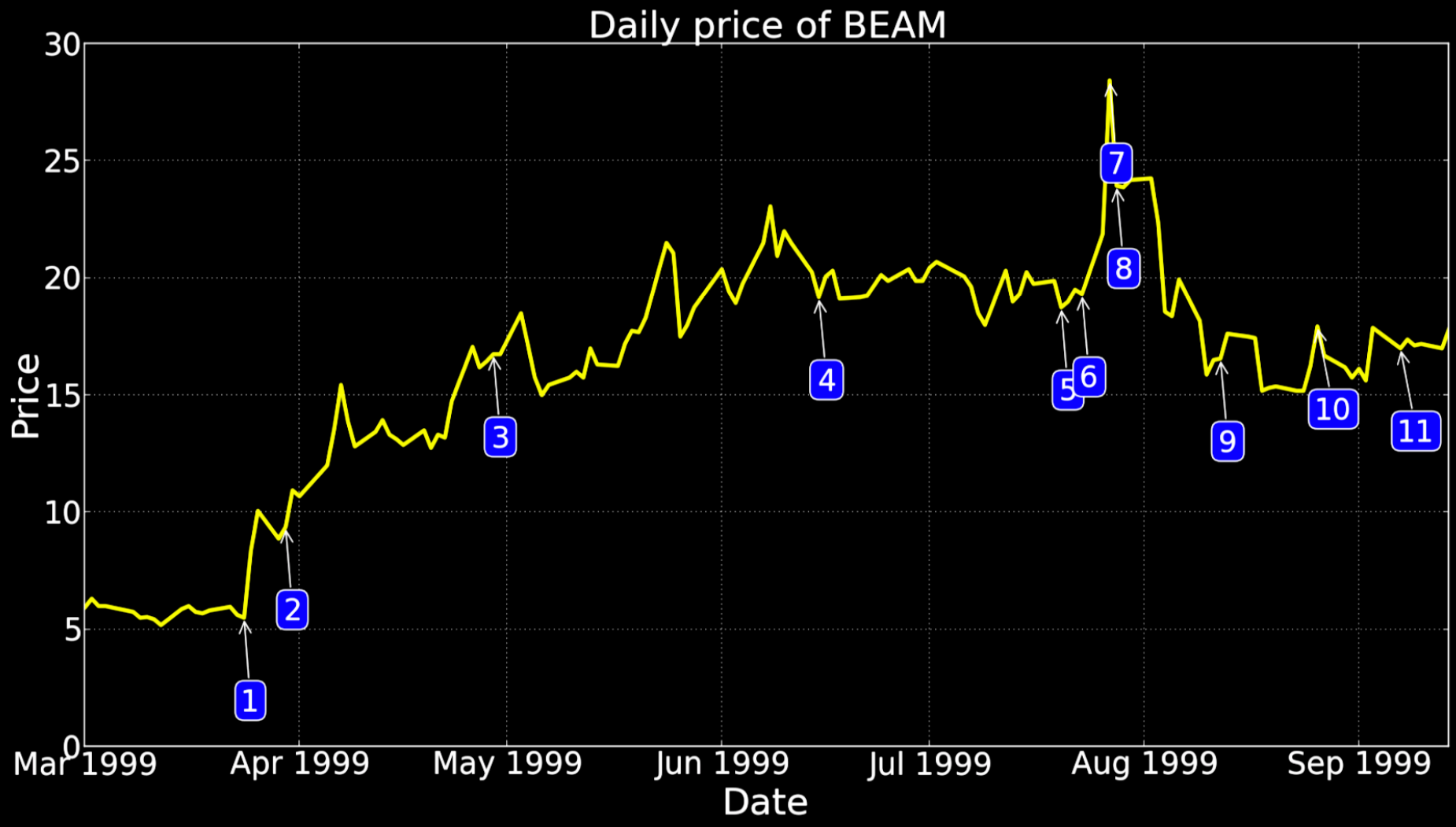
ENTROPY AS EXPECTED VALUE OF IDF(T) AND IDF(NOT T)







EXAMPLE: PREDICTING STOCK CHANGE FROM NEWS STORIES

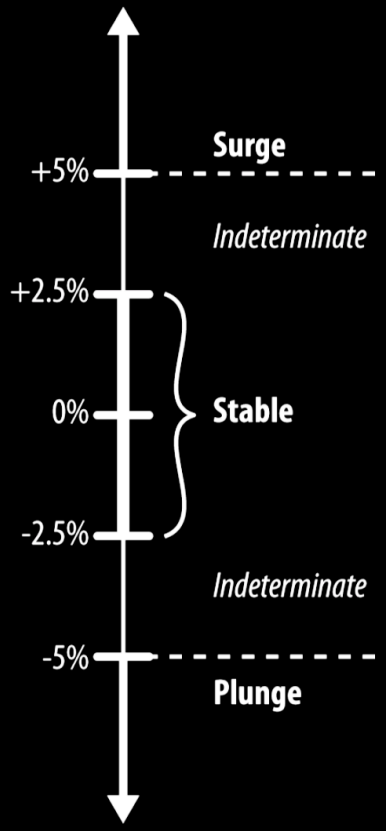


EXAMPLE: PREDICTING STOCK CHANGE FROM NEWS STORIES

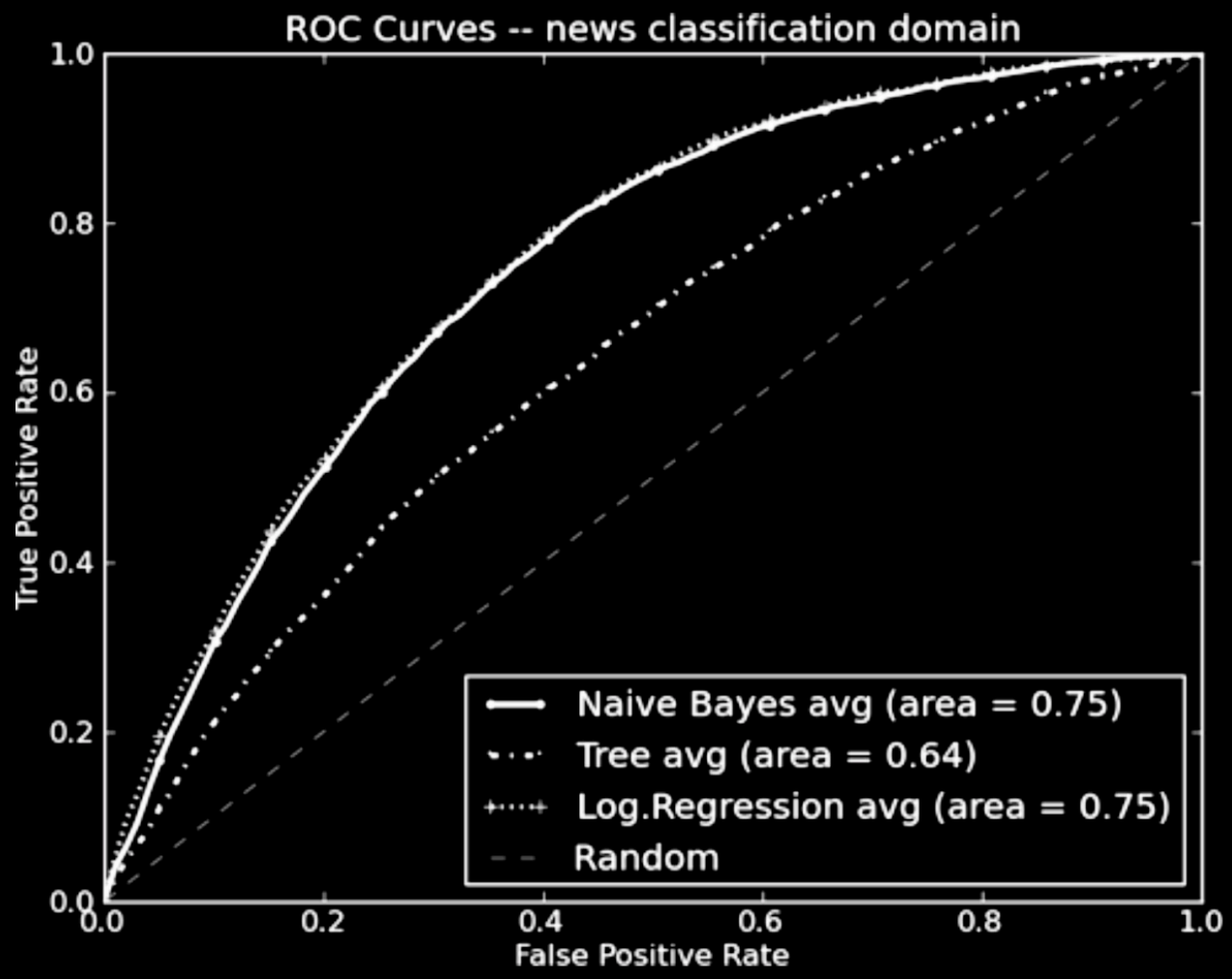
1. Summit Tech announces revenues for the three months ended Dec 31, 1998 were $22.4 million, an increase of 13%.
2. Summit Tech and Autonomous Technologies Corporation announce that the Joint Proxy/Prospectus for Summit's acquisition of Autonomous has been declared effective by the SEC.
3. Summit Tech said that its procedure volume reached new levels in the first quarter and that it had concluded its acquisition of Autonomous Technologies Corporation.
4. Announcement of annual shareholders meeting.
5. Summit Tech announces it has filed a registration statement with the SEC to sell 4,000,000 shares of its common stock.
6. A US FDA panel backs the use of a Summit Tech laser in LASIK procedures to correct nearsightedness with or without astigmatism.
7. Summit up 1-1/8 at 27-3/8.
8. Summit Tech said today that its revenues for the three months ended June 30, 1999 increased 14%…
9. Summit Tech announces the public offering of 3,500,000 shares of its common stock priced at $16/share.
10. Summit announces an agreement with Sterling Vision, Inc. for the purchase of up to six of Summit's state of the art, Apex Plus Laser Systems.
11. Preferred Capital Markets, Inc. initiates coverage of Summit Technology Inc. with a Strong Buy rating and a 12-16 month price target of $22.50.

EXAMPLE: BINARY CLASSIFICATION

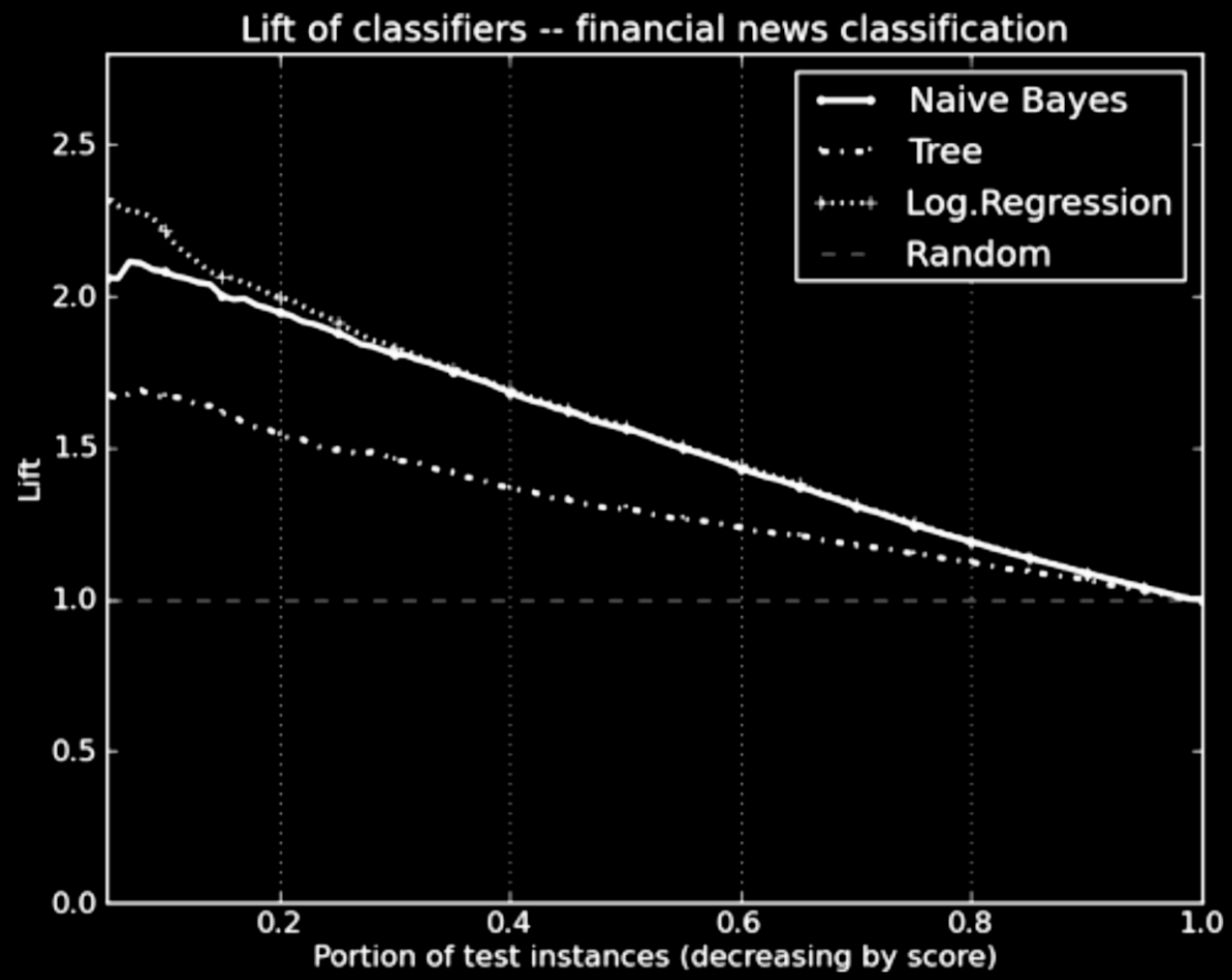
- Predicting change (surge or plunge) or no change (stable)



EXAMPLE: CLASSIFICATION RESULTS



EXAMPLE: LIFT RESULTS



EXAMPLE: WORDS WITH HIGH INFORMATION GAIN

alert(s,ed), architecture, auction(s,ed,ing,eers), average(s,d), award(s,ed), bond(s), brokerage, climb(ed,s,ing), close(d,s), comment(ator,ed,ing,s), commerce(s), corporate, crack(s,ed,ing), cumulative, deal(s), dealing(s), deflect(ed,ing), delays, depart(s,ed), department(s), design(ers,ing), economy, econtent, edesign, eoperate, esource, event(s), exchange(s), extens(ion,ive), facilit(y,ies), gain(ed,s,ing), higher, hit(s), imbalance(s), index, issue(s,d), late(ly), law(s,ful), lead(s,ing), legal(ity,ly), lose, majority, merg(ing,ed,es), move(s,d), online, outperform(s,ance,ed), partner(s), payments, percent, pharmaceutical(s), price(d), primary, recover(ed,s), redirect(ed,ion), stakeholder(s), stock(s), violat(ing,ion,ors)

PROBABILISTIC LANGUAGE MODEL

- Spelling correction

p(about ten minutes) > p(about ten minuets)

- Machine translation

p(he briefed reporters on the main contents of the statement) >
p(he introduced reporters to the main contents of the statement)

- Speech recognition
- Handwritten recognition

PROBABILISTIC LANGUAGE MODEL

$$p(w_1 w_2 \ldots w_n)$$

$$p(w_n \mid w_1 w_2 \ldots w_{n-1})$$

N-GRAM

- Sequences of n adjacent words

"The quick brown fox jumps"

2-grams (bigrams)

{quick, brown, fox, jumps, quick_brown, brown_fox, fox_jumps}

Markov: p(jumps | the quick brown fox) ~ p(jumps | fox)

1-GRAM (UNIGRAM) MODEL

$$p(w_1 w_2 \ldots w_n) = \prod p(w_i)$$

2-GRAM (BIGRAM) MODEL

$$p\left(w_n \mid w_1 w_2 \ldots w_{n-1}\right) \approx p\left(w_n \mid w_{n-1}\right)$$

$$p\left(w_n \mid w_{n-1}\right) = \frac{c\left(w_{n-1} w_n\right)}{c\left(w_{n-1}\right)}$$

3,4,5-GRAM MODELS

- Cannot model long term dependencies

$$p(w_n \mid w_1 w_2 \ldots w_{n-1}) \approx p(w_n \mid w_{n-1} w_{n-2} w_{n-3})$$
$$p(w_n \mid w_1 w_2 \ldots w_{n-1}) \approx p(w_n \mid w_{n-1} w_{n-2} w_{n-3} w_{n-4})$$
$$p(w_n \mid w_1 w_2 \ldots w_{n-1}) \approx p(w_n \mid w_{n-1} w_{n-2} w_{n-3} w_{n-4} w_{n-5})$$

N-GRAMS: SHAKESPEARE

- 1-gram: To him swallowed confess hear both. Which. Of save on trail for are ay device and rote life have. Hill he late speaks; or! a more to leg less first you enter.
- 2-gram: Why dost stand forth thy canopy, forsooth; he is this palpable hit the King Henry. Live king. Follow. What means, sir. I confess she? then all sorts, he is trim, captain.
- 3-gram: Fly, and will rid me these news of price. Therefore the sadness of parting, as they say, 'tis done. This shall forbid it should be branded, if renown made it empty.
- 4-gram: King Henry. What! I will go seek the traitor Gloucester. Exeunt some of the watch. A great banquet serv'd in. It cannot be but so.

FEATURE VECTOR

- Requires learning each word order

"Alice sent a message on Sunday"

MARKOV MODEL

- Does not model long term dependencies

"Alice and Bob Communicate. Alice sent Bob a message"

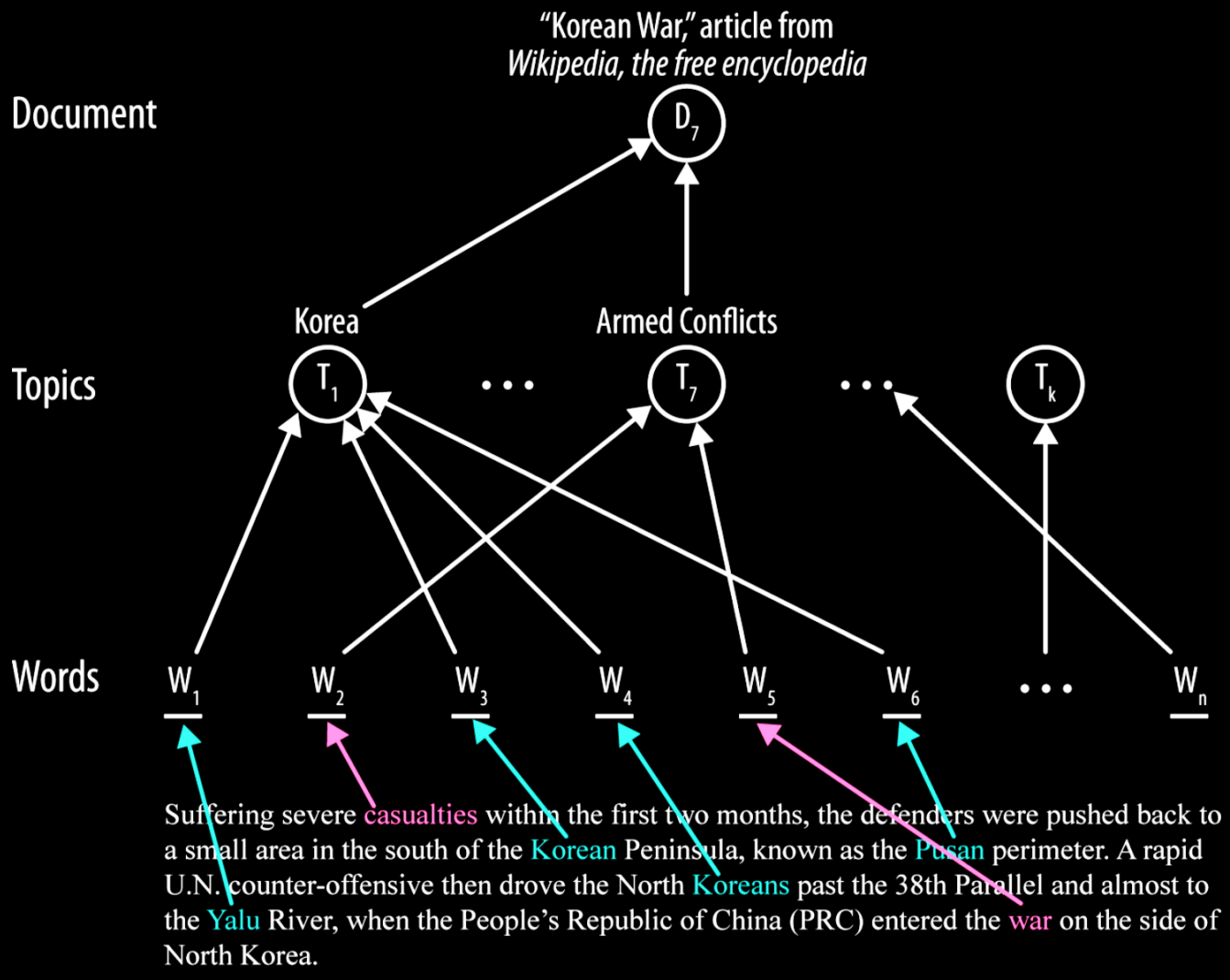
RECURRENT NEURAL NETWORK

- Maintain word order
- Model long term dependencies by sharing parameters across time
- Allow for examples (and labels) of different lengths

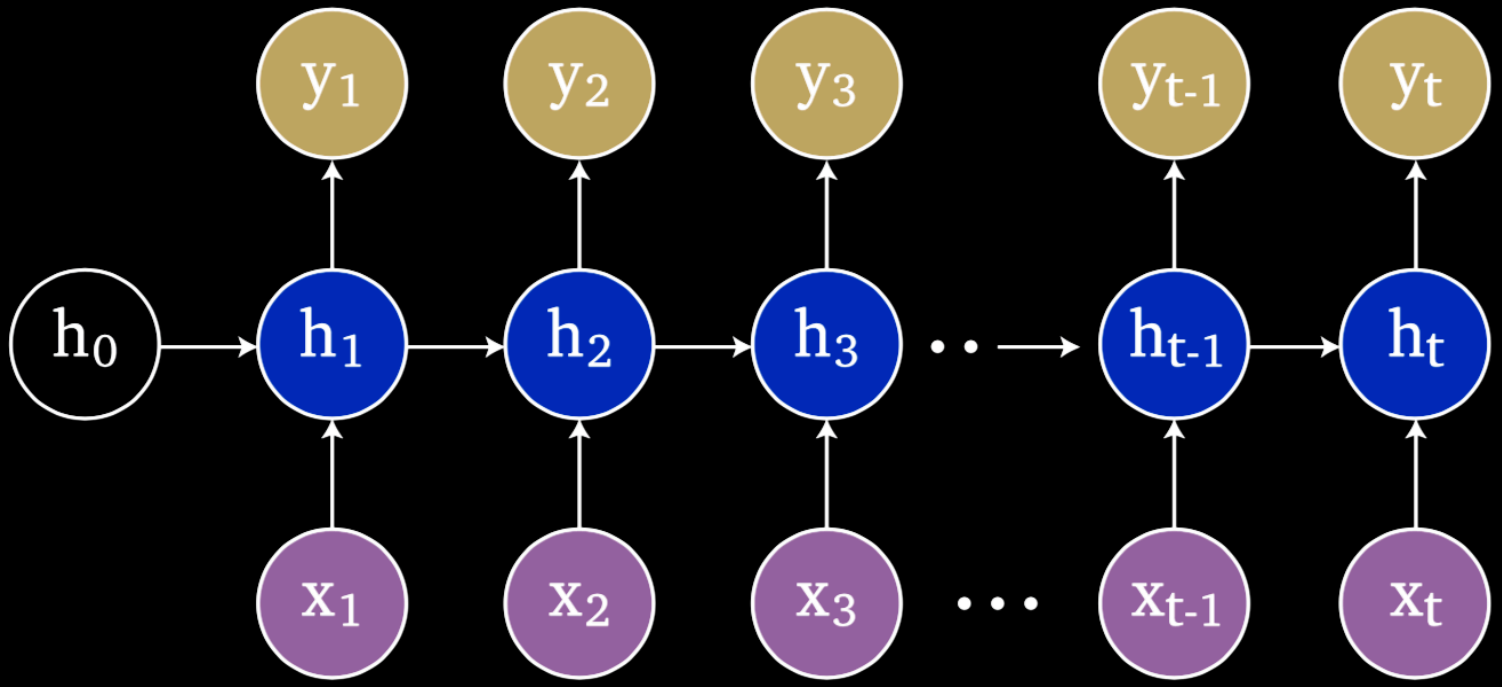
NAMED ENTITY EXTRACTION

- Recognize common named entities in documents.

TOPIC MODELS

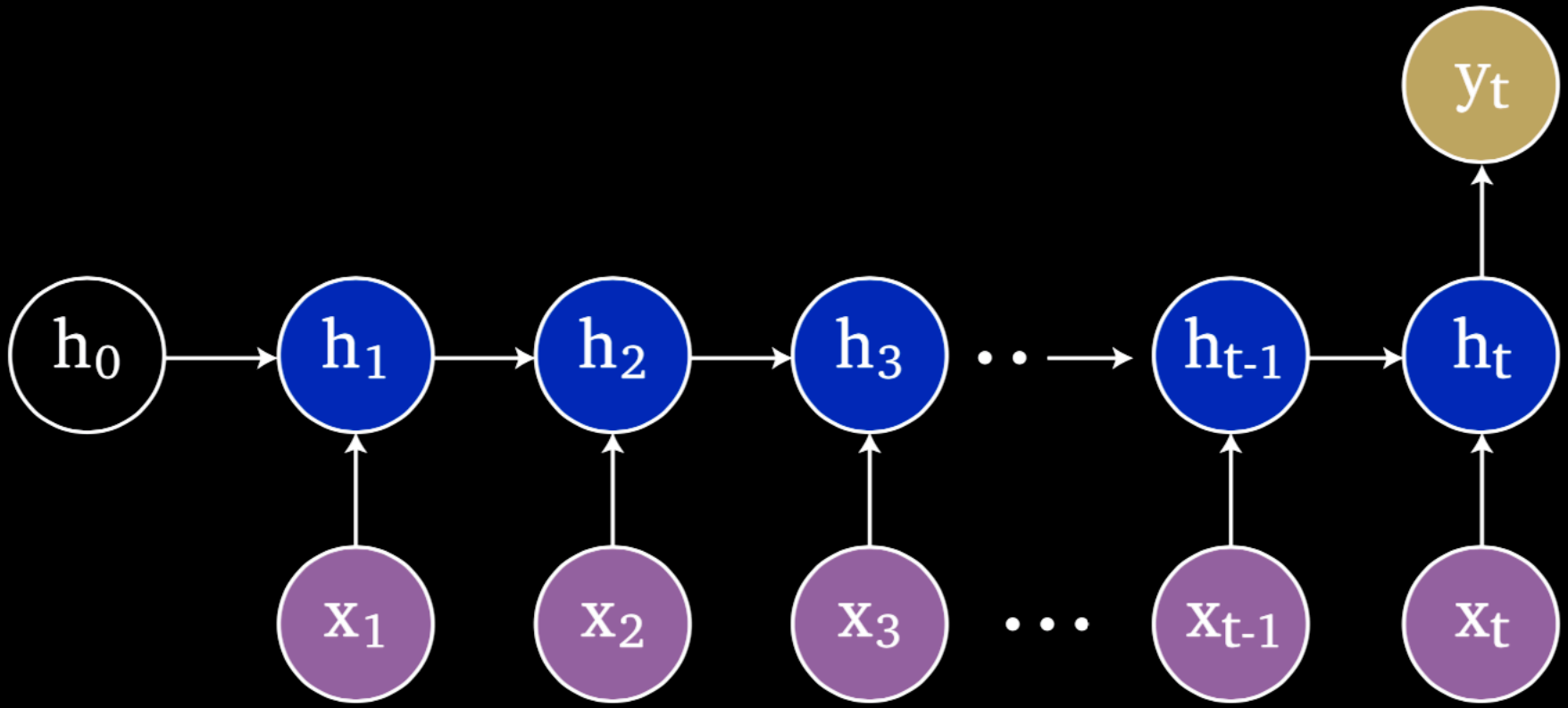


RECURRENT NEURAL NETWORK



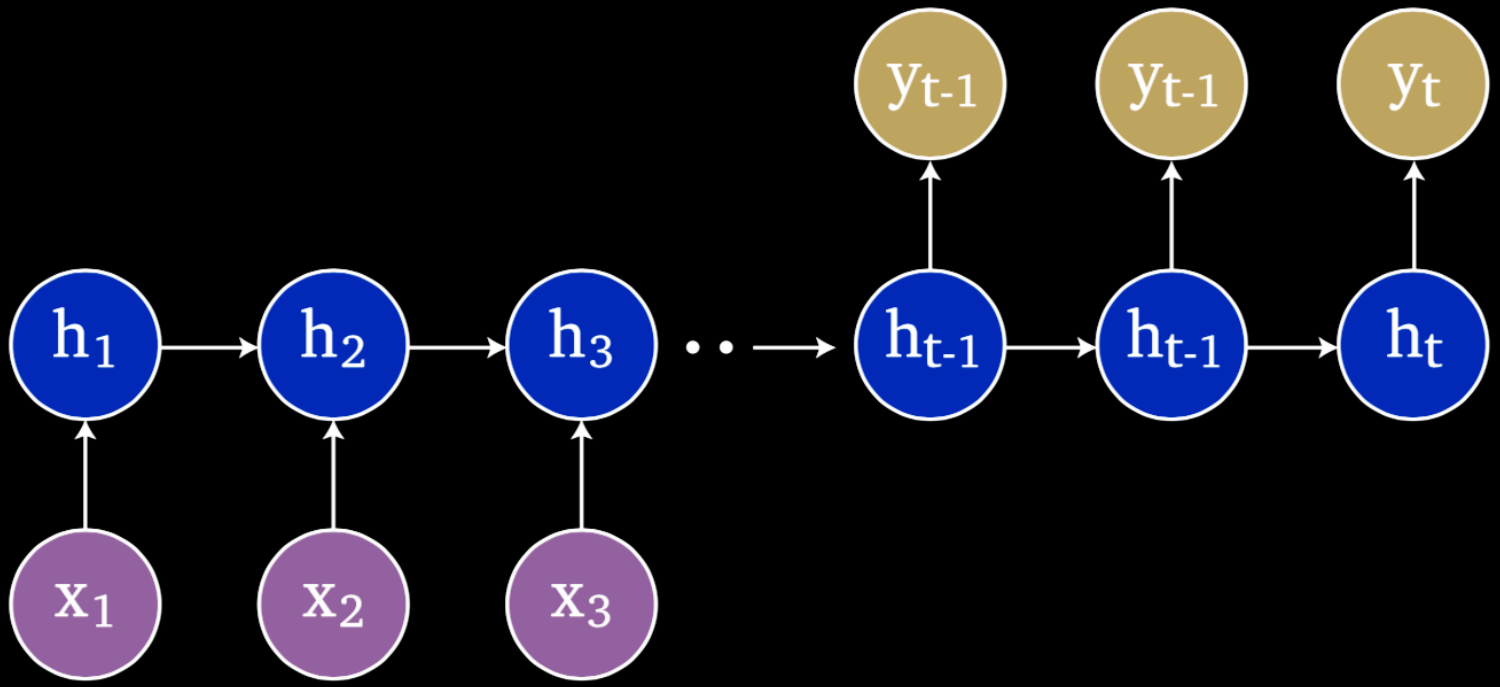
RECURRENT NEURAL NETWORK

- Sentiment classification: sequence of words to sentiment



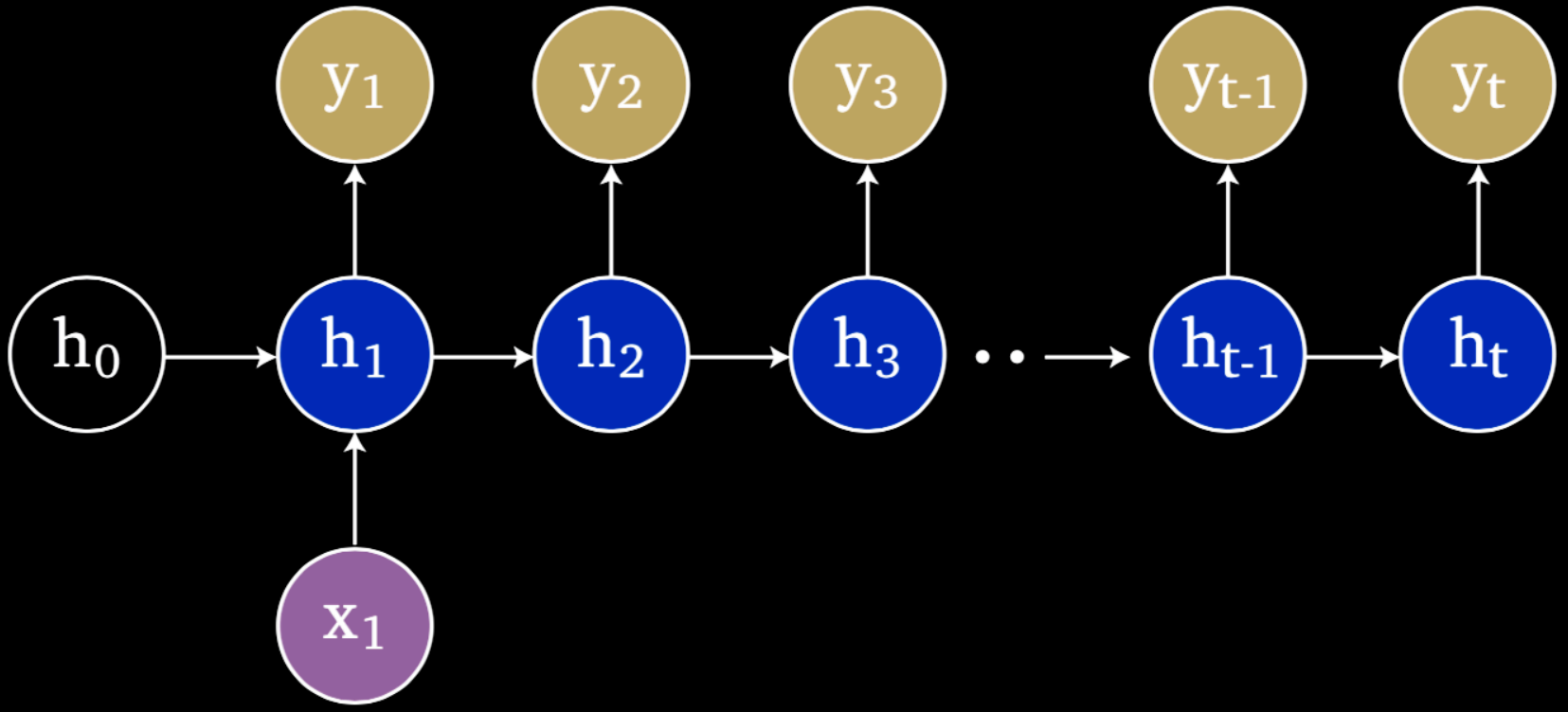
RECURRENT NEURAL NETWORK

- Machine translation: sequence of words to sequence of words



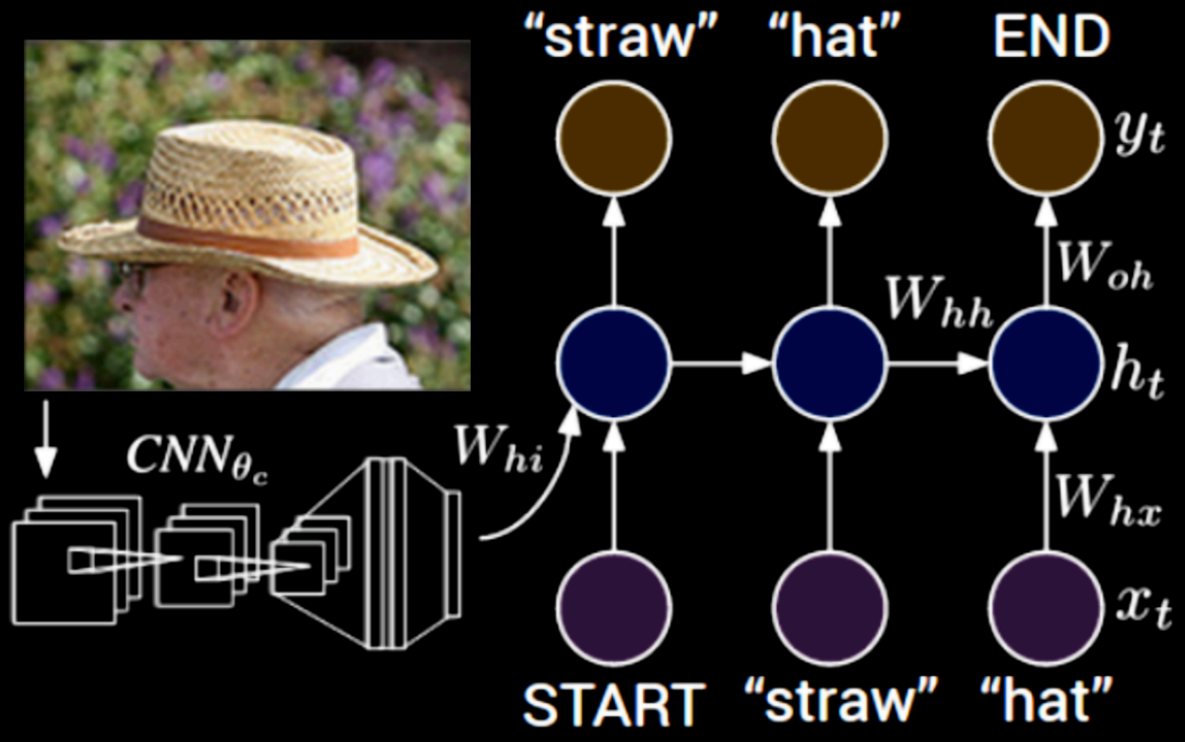
RECURRENT NEURAL NETWORK

- Image captioning: image to sequence of words



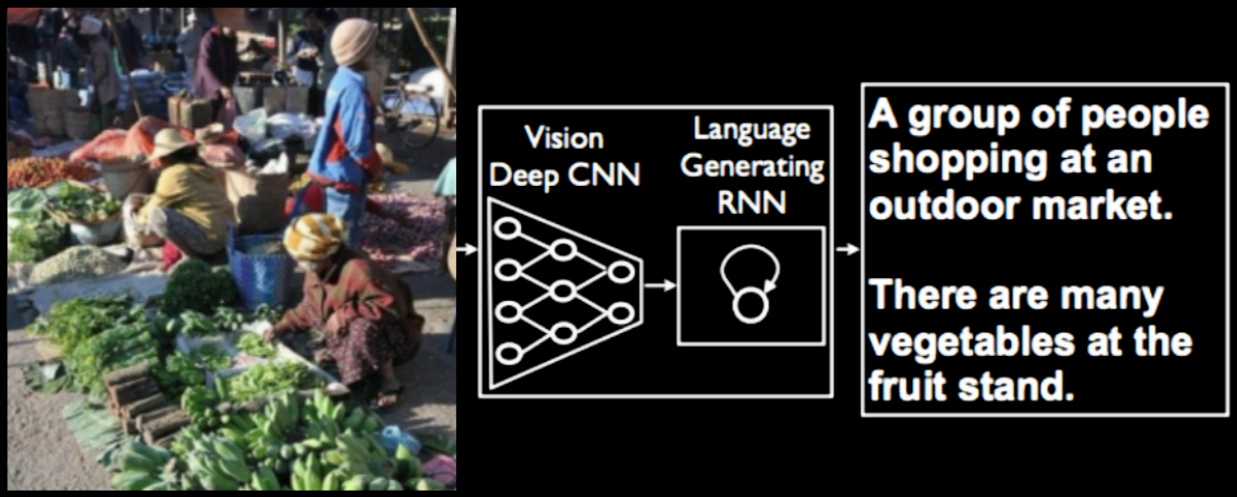
RECURRENT NEURAL NETWORK

- Image captioning: image to sequence of words



RECURRENT NEURAL NETWORK

- Image captioning: image to sequence of words



VINYALS ET AL

A person riding a motorcycle on a dirt road.

Two dogs play in the grass.

A skateboarder does a trick on a ramp.

A dog is jumping to catch a frisbee.

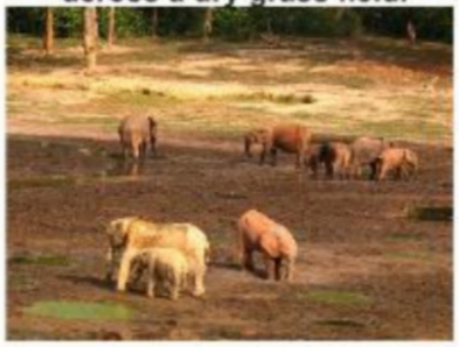
A group of young people playing a game of frisbee.

Two hockey players are fighting over the puck.

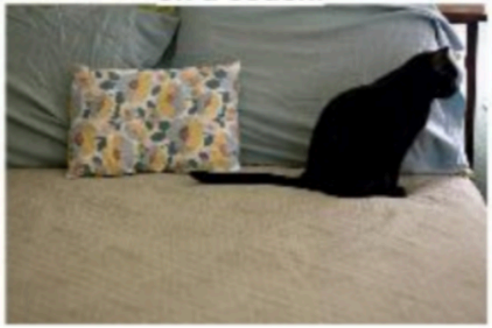
A little girl in a pink hat is blowing bubbles.

A refrigerator filled with lots of food and drinks.

A herd of elephants walking across a dry grass field.



A close up of a cat laying on a couch.



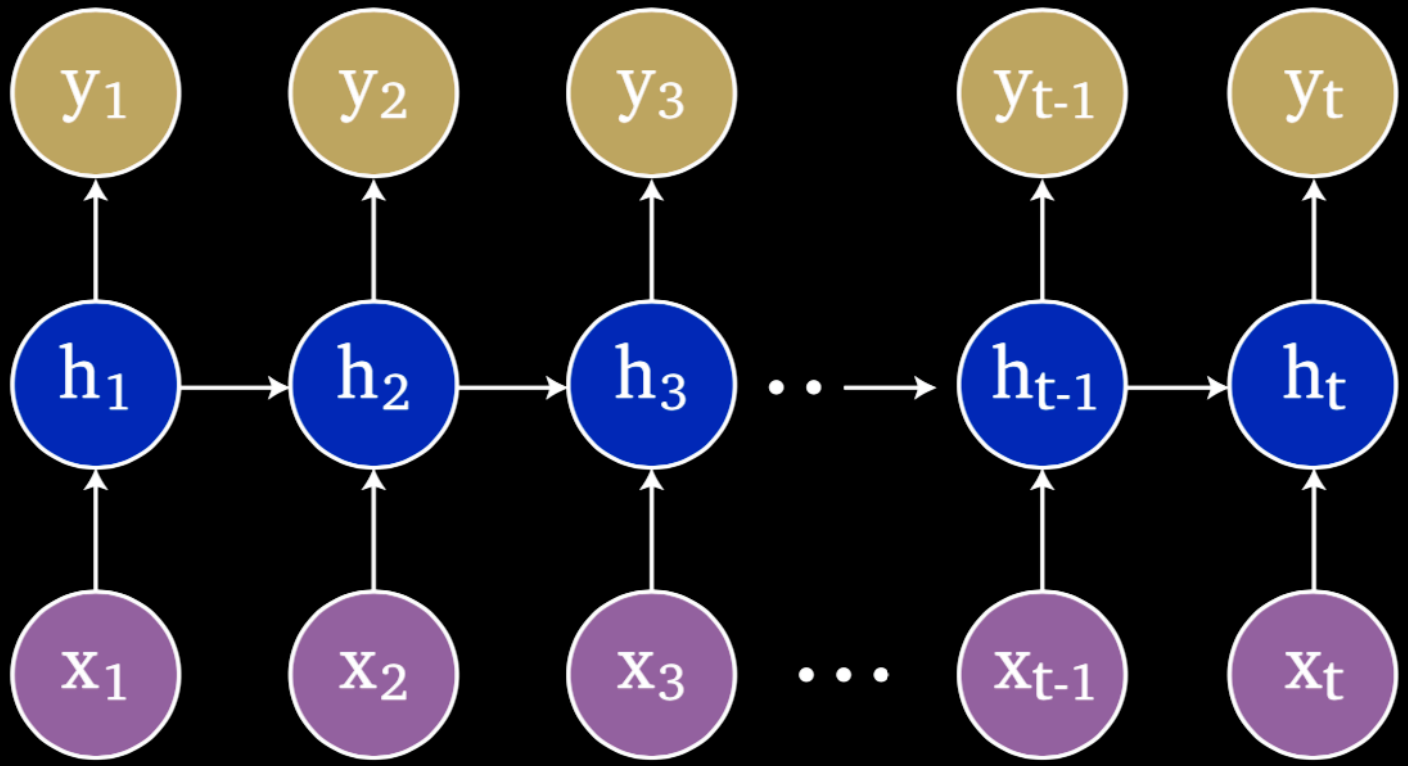
A red motorcycle parked on the side of the road.

A yellow school bus parked in a parking lot.

Describes without errors	Describes with minor errors	Somewhat related to the image	Unrelated to the image

RECURRENT NEURAL NETWORK

- Video classification: sequence of video frames to labels

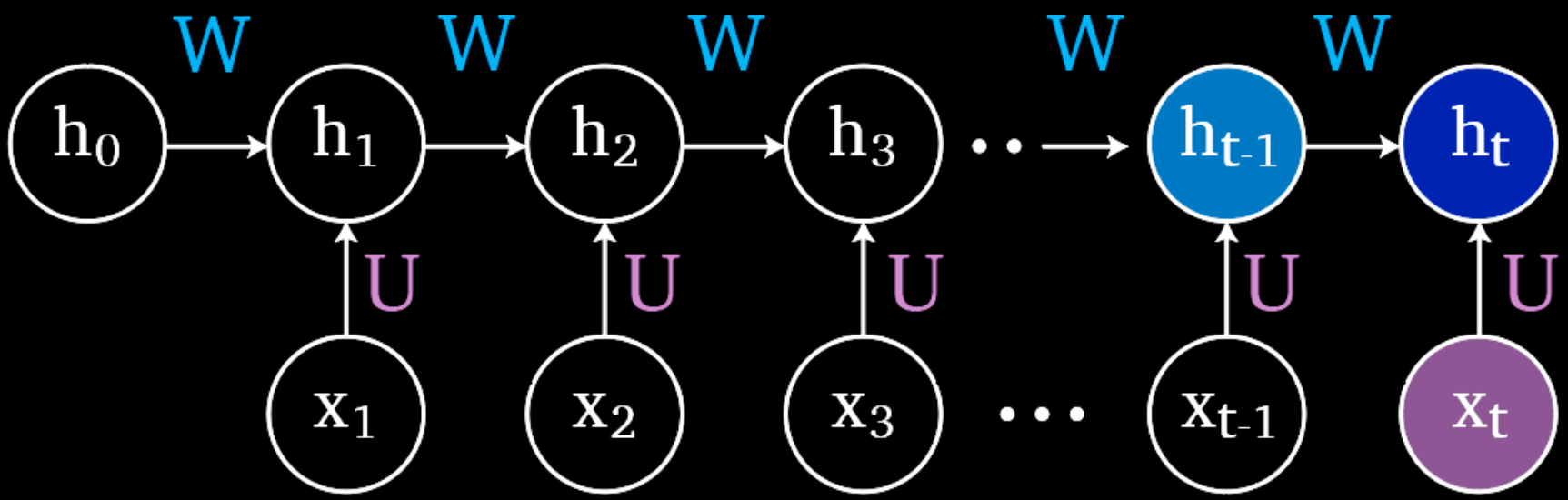


RECURRENT NEURAL NETWORK

- Share weight matrices U, W across all time steps t

$$h_t = f(h_{t-1}, x_t)$$

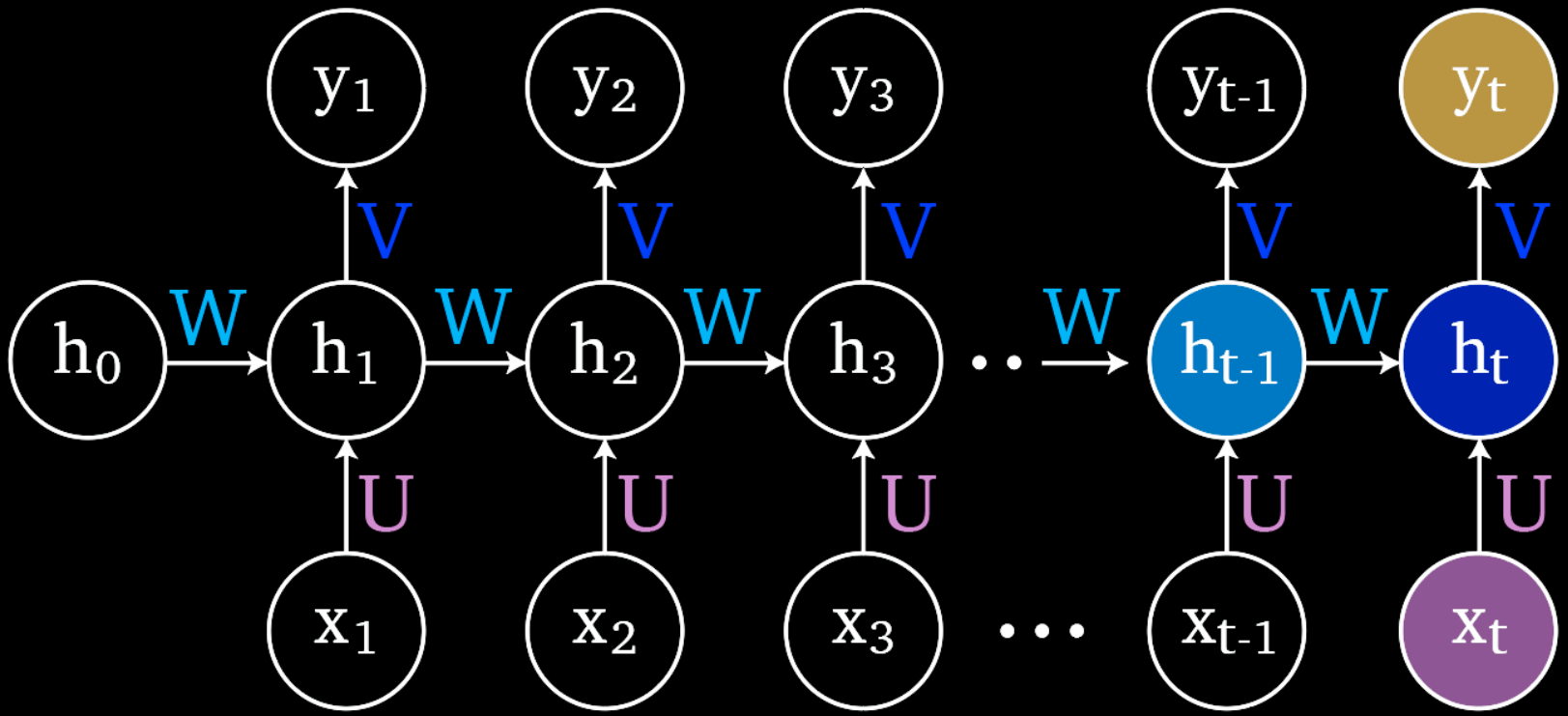
$$h_t = g(W h_{t-1} + U x_t)$$



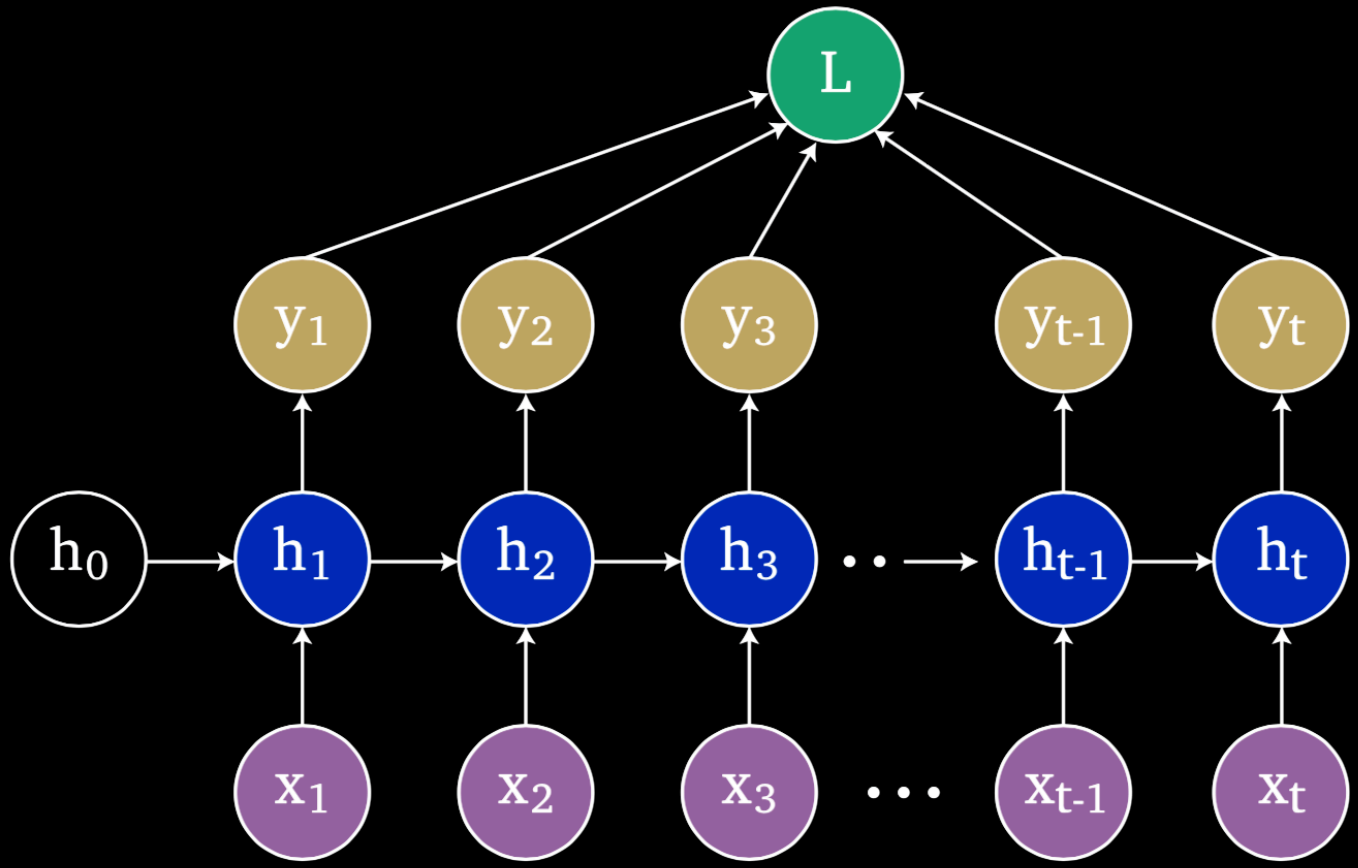
RECURRENT NEURAL NETWORK

- Share weight matrices U, W, V across all time steps t

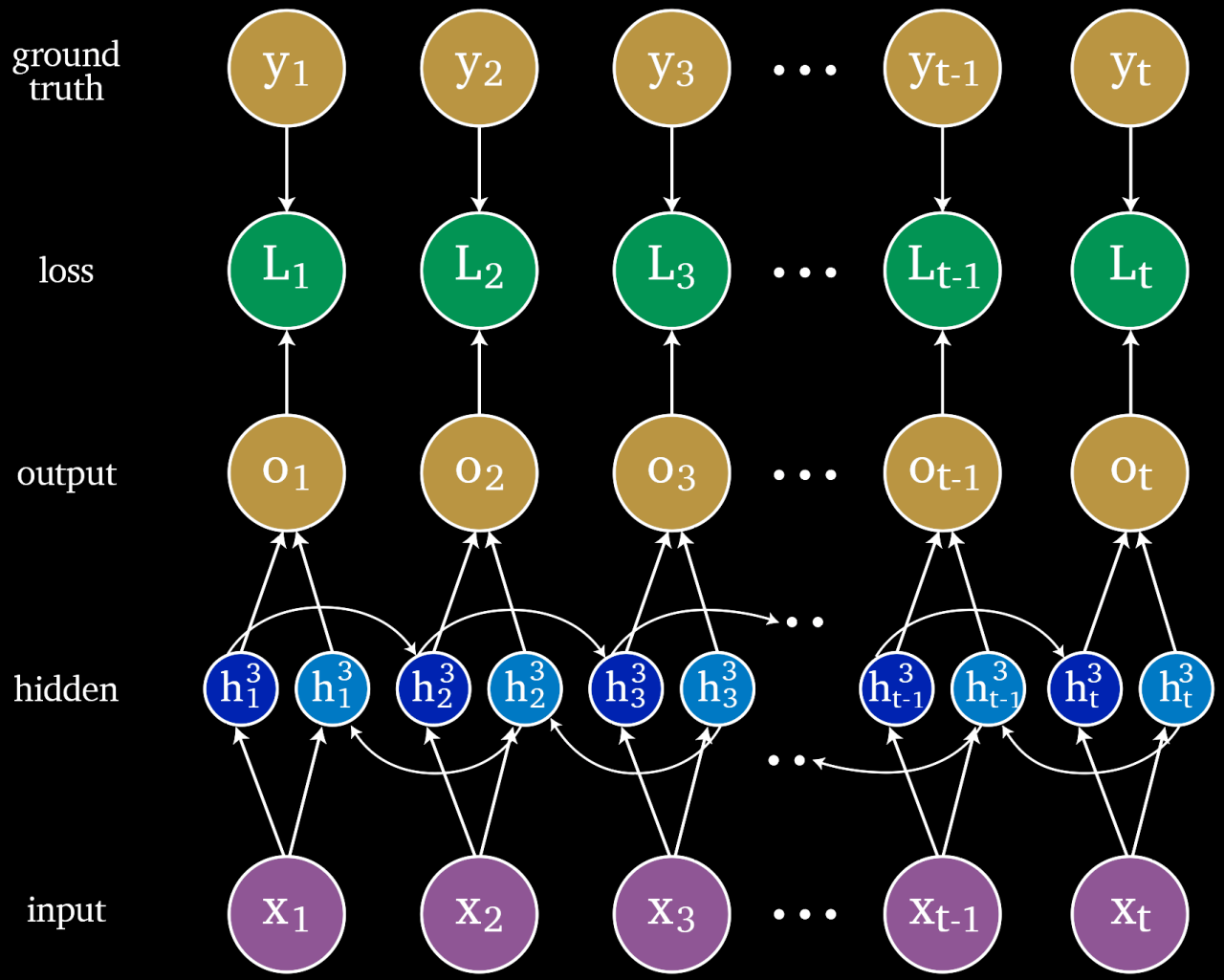
$$y_t = V h_t$$



RECURRENT NEURAL NETWORK



RECURRENT NEURAL NETWORK



RNN CHARACTER LEVEL SHAKESPEARE GENERATION

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.

RNN CHARACTER LEVEL LINUX CODE GENERATION

```
/*
 * 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;
```

SPELLING CORRECTION

- Non-word errors: graffe -> giraffe
  - Detect a word not in dictionary and correct by most similar real word
- Real word errors: three->there, piece->peace, too->two
  - Find best candidate among words with similar pronunciation and spelling

SPELLING CORRECTION: NOISY CHANNEL

- Original word w -> noisy channel -> misspelled word x

$$\arg\max_{w} p(w \mid x) = \arg\max_{w} p(x \mid w)\, p(w)$$

EDIT DISTANCE

- Minimal edit distance between two strings:

Insertion: acress : cress, acres

Deletion: acress : actress

Substitution: acress : access, across

Transposition: acress : caress

- 80% of spelling errors are within edit distance 1.
- Almost all spelling errors are within edit distance 2.

REFERENCES

- Reading: DSB Ch 10