Deep Learning and its Recent Development

Professor Francis Chin 錢玉麟教授
Department of Computing 電子計算系
Hang Seng Management College 恒生管理學院
Emeritus Professor, University of Hong Kong
香港大學榮休教授

April 2018

恒生管理學院
HANG SENG
MANAGEMENT COLLEGE

nature International weekly journal of science

NATURE | NEWS

عربي

Google AI algorithm masters ancient game of Go

Deep-learning software defeats human professional for first time.

Elizabeth Gibney

27 January 2016

27 January 2016

PDF | Rights & Permissions

Deep-learning software (AlphaGo) defeated human, Fan Hui 樊麾, European Professional Go Championship, for the first time in 2015.

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Au

News & Comment > News > 2016 > July > Article

NATURE | BLOG

The Go Files: AI computer wraps up 4-1 victory against human champion

***Nature* reports from AlphaGo's victory in Seoul.**

Tanguy Chouard

15 March 2016

SEOUL, SOUTH KOREA

Rights & Permissions

15 March 2016

Seoul, South Korea

AlphaGo wraps up 4-1 victory against human champion (Lee Sedol) in Seoul, for a $1 million prize

中文网

學者：AlphaGo表明人腦已斗不過AI

自從2016年12月底起，就有一位網路代號"Master"的圍棋棋士，在網路上不斷以連勝之姿擊敗多位圍棋好手。

最後AlphaGo在擊敗中國高手古力後，達成60勝的壯舉。此前，AlphaGo連續擊敗了中國"棋聖"聶衛平以及"棋王"柯潔、陳耀燁等、還有多位韓國與日本的高手接連也被他擊敗，宛若圍棋界"獨孤求敗"。

早在2016年3月時，黃士傑所研發的AlphaGo就擊敗了韓國的九段棋士李世乭。當時聶衛平曾批評李世乭棋藝"下得太差"，柯潔也認為李世乭棋藝糟到"看不下去"。而如今，2位高手也成了AI手下敗將。

林順喜表示，根據中國圍棋等級分來看，AlphaGo目前估算已經超過4500分，遠遠超過許多中國圍棋高手，如古力(2773分)與柯潔(2764分)。他也直接說："人類超越不了電腦，這是必然結果"。

AlphaGo beat Ke Jie 3-0 in May 2017

AlphaGo Zero

- Published in Nature in October 2017
- Created without using data from human games
- Stronger than any previous AlphaGo version by playing games against itself
  - beat AlphaGo Lee in three days
  - reached the level of AlphaGo Master in 21 days
  - exceeded all the old versions in 40 days

Why Google's Go Win Such a Big Deal?

Deep Blue 深藍 (IBM) beat chess grandmaster Garry Kasparov in 1997

- Go vs Chess
  - Bigger board, 19x19 vs 8x8
  - Simpler rules, more move possibilities, 361 vs 28
  - Longer game, 150 vs 80
- AlphaGo vs Deep Blue
  - Deep Blue can only play Chess.
  - AlphaGo is general-purpose, can win 49 different arcade games

Why AlphaGo? What is next?

Silver: ” … can learn from data and self-play to figure out knowledge ...”

- no built-in expert knowledge
  – reinforcement learning

Silver: “… more exciting was the manner in which AlphaGo did it…”

- creativity and intuitive insights

Silver: “…many many different domains. We're by no means done with AlphaGo.”

- Deep-learning can revolutionize everything

Google CEO Eric Schmidt:

This technology will be using in every one of the Alphabet companies.

Google新策略「人工智能優先」

Google周二（4日）在三藩市推出一系列電子硬件，包括旗下首款智能手機、智能家居、虛擬現實設備等，這些硬件都融入了Google的人工智能技術，突出該公司向「人工智能優先」轉型的新發展策略，與蘋果公司和三星一較高下。

Google總裁皮柴（Sundar Pichai）稱，過去10年，Google一直在建立「流動優先」（mobile-first）的世界，在接下來10年，Google將轉向建立「人工智能優先」（AI-first）的世界，人機互動將變得更加自然。Pixel系列智能手機成爲Google推出新硬件中的重頭作。

Google推出自家高檔品牌智能手機 Pixel(左)，以及「白日夢視野」虛擬現實頭戴設備(右)。（路透社）

Pixel手機售價5000 賣點快速充電

Pixel由HTC代工，但設計由Google全力包辦，機身將會刻上「Google製造」字樣，被視爲Google挑戰蘋果和三星高級手機的主力，它有5吋和5.5吋屏幕兩種尺寸，擁有4GB RAM，內置儲存量分別是32GB和128GB。它其中大賣點是充電15分鐘就可以使用7小時，另一賣點是拍照質素。它配備1200萬像素鏡頭，感光度較一般手機的手機。「Google助理」類似於蘋果的Siri，可根據用户語音指令執行一系列任務。Google强調，「Google助理」將是Google下一個活躍而開放的生態系統。

VR頭戴裝置發「白日夢」

Google今年初才成立硬件部門。分析指出，

明報

Oct 2017

Machine Learning Daily Life Examples

Filtering Spam emails

CBS NEWS NEWS SHOWS VIDEO CBSN MORE

By **JOEL ARAK** / **AP** / *January 24, 2004, 5:36 PM*

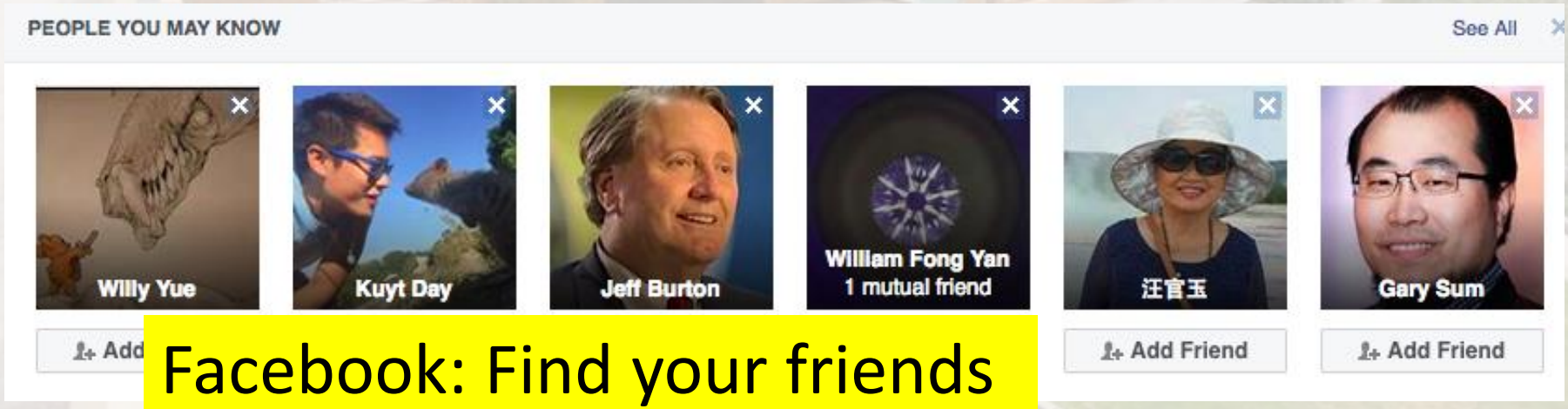
Gates: Spam To Be Canned By 2006

by James Denison / March 11, 2004 9:23 AM PST

http://www.cnn.com/2004/TECH/internet/03/05/spam.charge.ap/index.html

Machine Learning Daily Life Examples

Filtering Spam emails



Facebook: Find your friends

Recent Quantum Leaps in Technology

- Machine translation
  (google translate between over 100 languages)
- Speech recognition
- Image recognition

Google search - “dog”









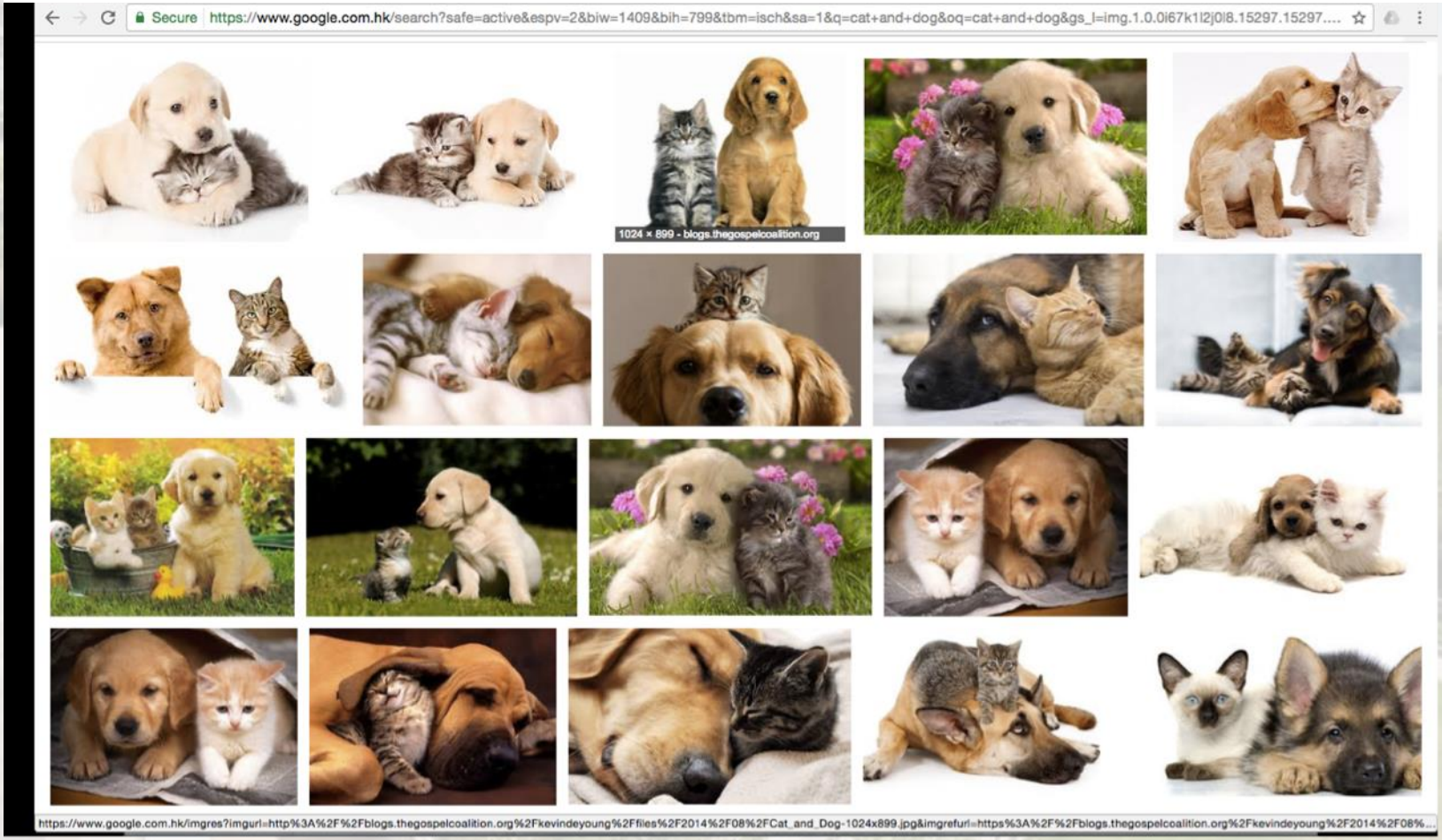









Google search - “cat and dog”



Recent Quantum Leaps in Technology

- Machine translation
  (google translate between over 100 languages)
- Speech recognition
- Image recognition
- Robotics
- Self-driving cars

Self-driving cars (自駕車)

Nvidia taught a self-driving car to drive like a human without telling it ...
Quartz - 2016年10月1日
The self-driving Lincoln was able to navigate around real New Jersey ... "With deep learning, the vehicle can be trained to have superhuman ...

ChinaTopix

Baidu, nVidia To Build Independent Self-Driving Car Platform
Forbes - 2016年9月29日
Integrating nVidia's deep learning expertise in processing the data and ... It should be noted that for a safe experience in a self driving vehicle, ...

百度和 Nvidia 合作建自駕車

人工智慧和深度學習
無限可能

深度學習的強大威力

深度學習是人工智慧中成長最為快速的領域，可協助電腦理解影像、聲音和文字等資料。現在透過多層級的神經網路，電腦可以和人類一樣針對複雜的情況進行觀察、學習和反應，甚至表現得更好。藉由這個方式，以完全不同的角度深入然後重新思考如何應用您的資料、您的技術，及創新您所提供的產品和服務。

瞭解深度學習的熱門趨勢

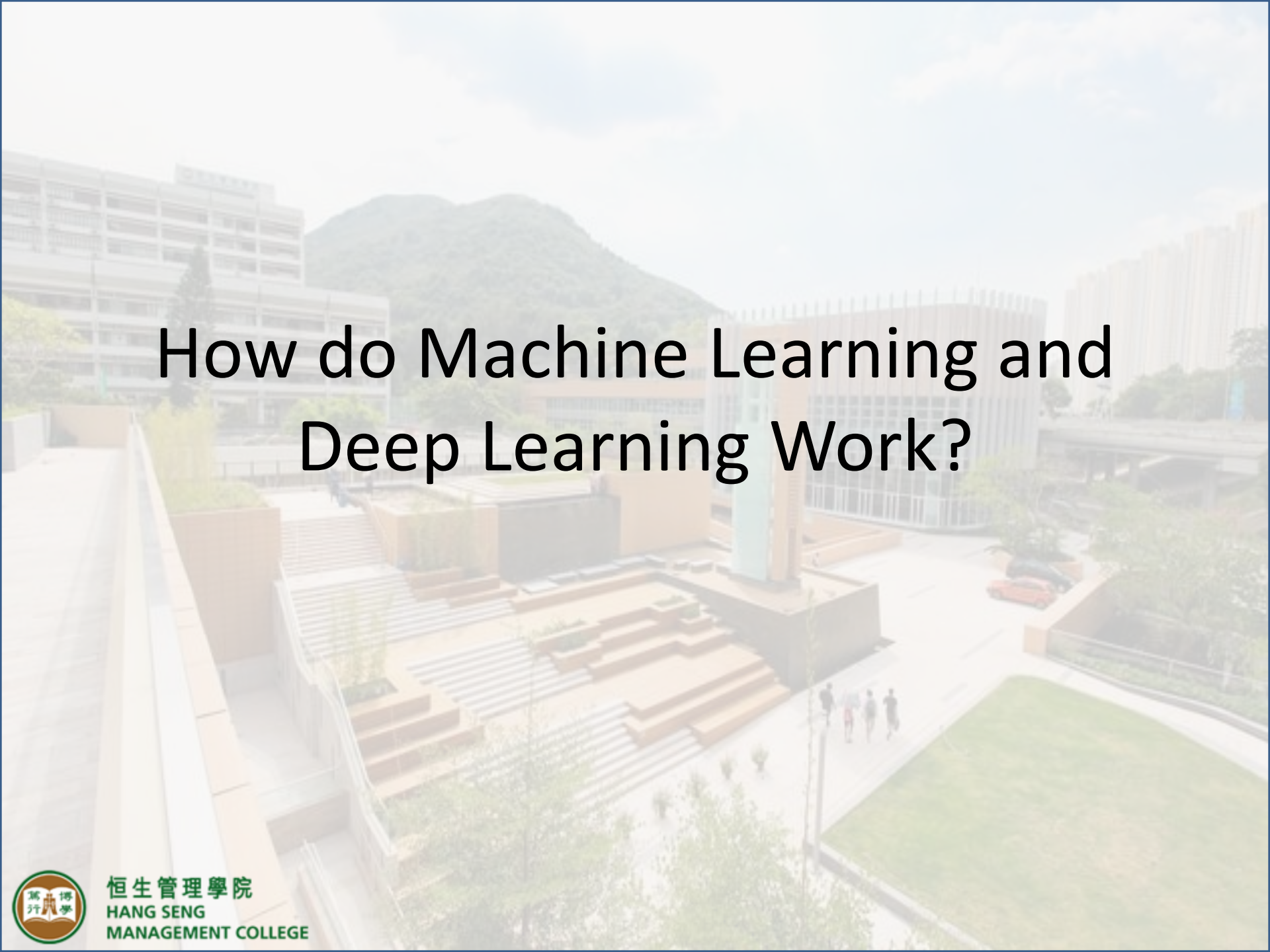
閱讀 NVIDIA CEO 黃仁勳所撰寫的「透過 GPU 加快人工智慧運算速度」部落格文章

Recent Quantum Leaps in Technology

- Machine translation
  (google translate between over 100 languages)
- Speech recognition
- Image recognition
- Robotics
- Self-driving cars
- …

All these technology breakthroughs derive from the same breakthrough in Artificial Intelligence machine learning – Deep Learning

How do Machine Learning and Deep Learning Work?



April 21, 2017

ARTIFICIAL INTELLIGENCE

A Computer Is Deciding What Movies You're Going to Watch Next (电脑决定你要看什么电影)

AI Predicts This Summer's Opening Weekend Box-Office Hauls

人工智能预测这个夏天的电影票房

- **BAYWATCH**
  $21.58m on 3,500 screens
- **DIARY OF A WIMPY KID: THE LONG HAUL**
  $17.3m on 3,400 screens
- **ROUGH NIGHT**
  $15.6m on 3,000 screens
- **SNATCHED**
  $13.3m on 3,000 screens

Algorithms can use inputs like Amy Schumer's popularity to project ticket sales.

A Machine Learning Example

Will this movie be popular? 這部電影會受歡迎嗎？

x_1 = Type (romance, violence, detective, fiction, …類型（浪漫，暴力，偵探 …

x_2 = Language (English, Chinese, Korean, … 語言（英語，中國，韓國, …

x_3 = Main actor (Tom Cruise, Tom Hanks, 周潤發, 成龍, …

x_4 = Main actress (Angelina Jolie, Meryl Streep, 章子怡, 张曼玉

x_5 = Background (modern, ancient, animation, 现代, 古装, 动画电影…

x_6 = Length of the movie 片長

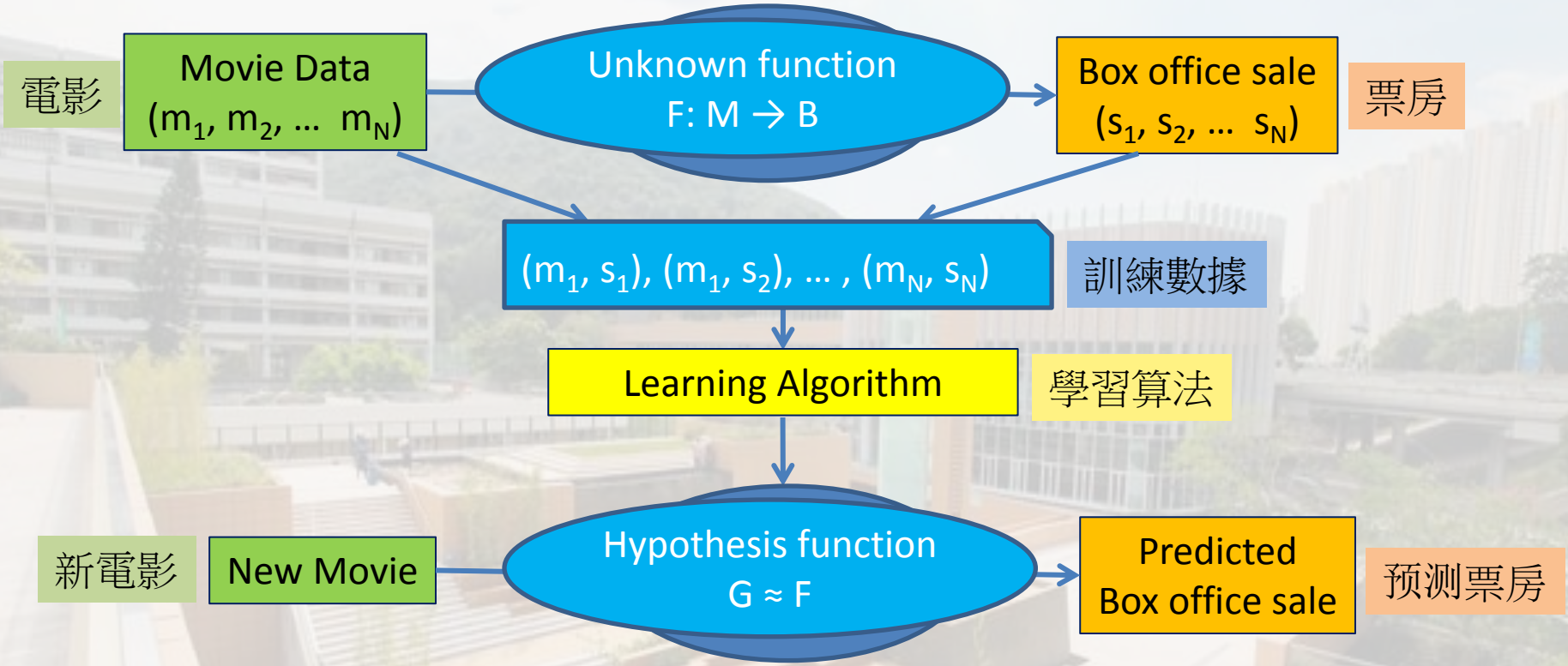
x_7 = Director (Steven Spielberg, 張藝謀, …

x_8 = Show time (summer, Christmas, New year, … 放映时段

…

- Study previous movies 電影 and box office sales 票房
- Data : $(m_1, s_1), (m_1, s_2), \ldots, (m_N, s_N)$
  - movie m described by $<x_1, x_2, x_3, \ldots, x_p>$

Machine Learning Model

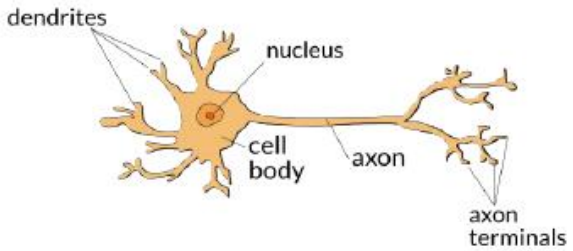


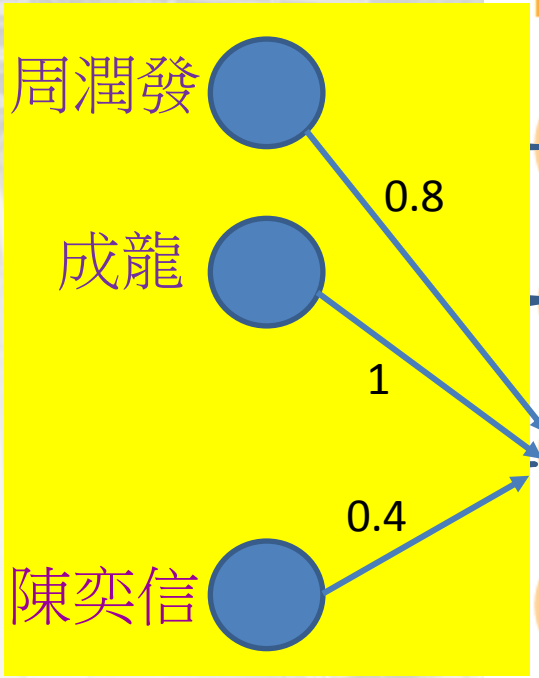
- What is G ?
- Example: 2.5xActor(成龍) + 3xDirector(張藝謀) + …
  $s = w_0 + w_1x_1 + w_2x_2 + \ldots + w_px_p$ (linear dependence)
- Learn values of w_i's (also x_i's, e.g., weights of 成龍,周潤發…)

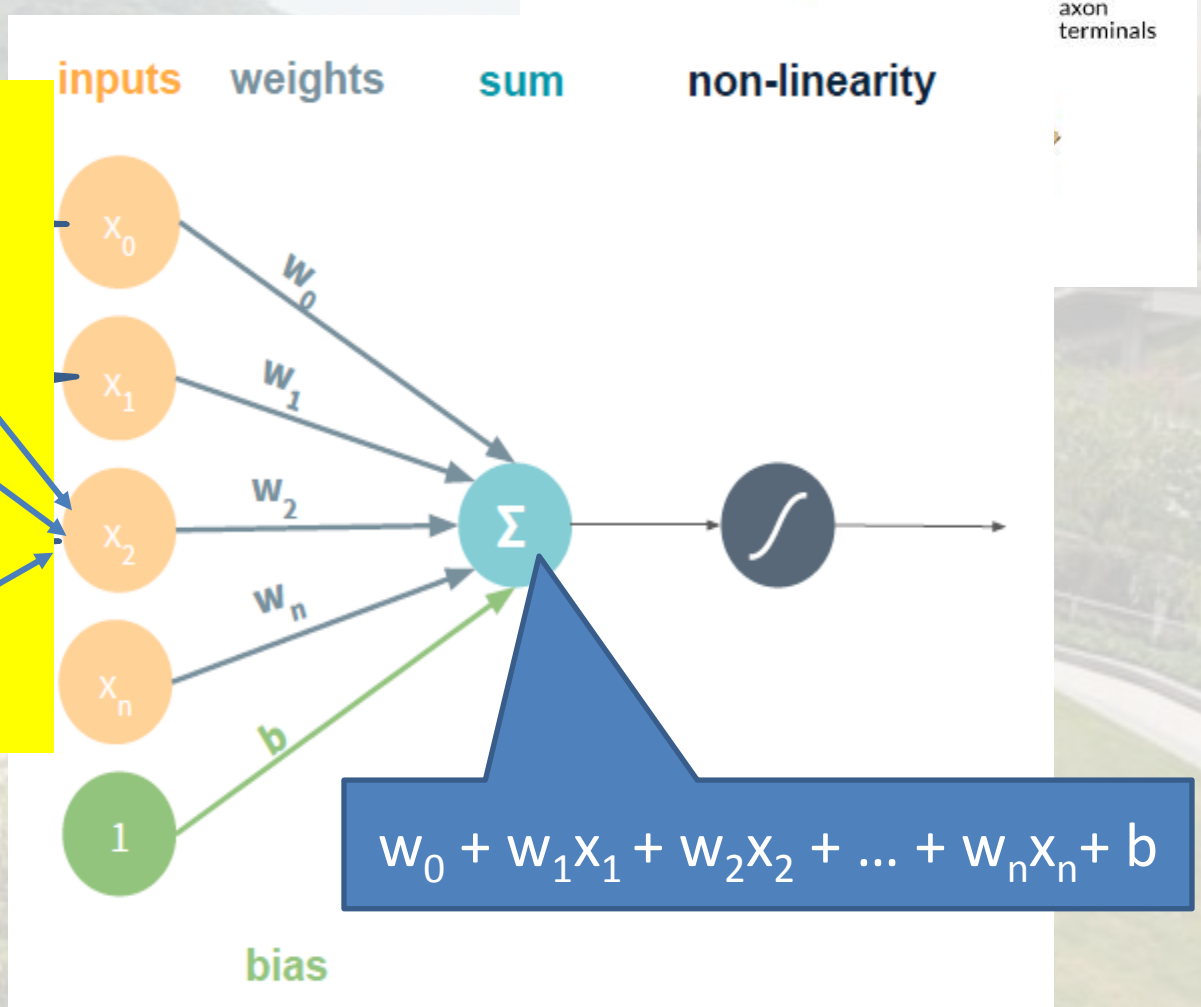
Perceptron

Inspired by neurobiology (1954)

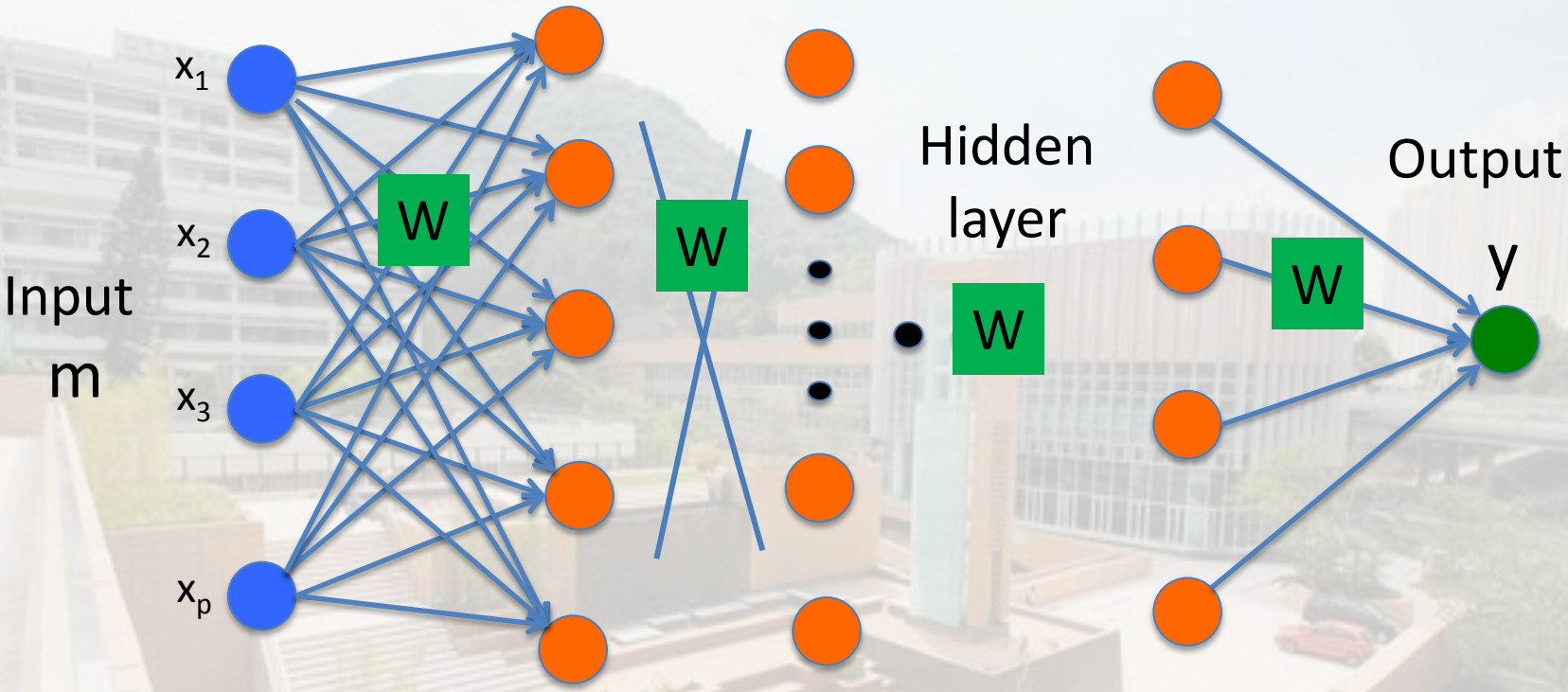
灵感来自神经生物学







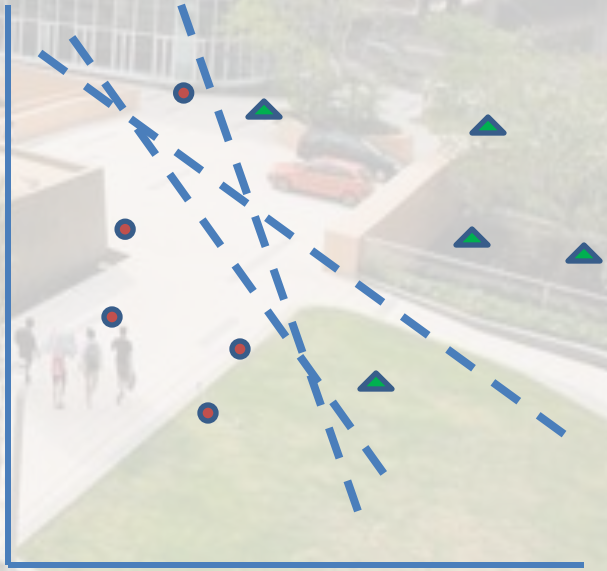
Deep Neural Network (深层学习网络)



- Data $\{(m_1, s_1), (m_1, s_2), \ldots, (m_N, s_N)\}$
- To train model (adjust W)
  to minimize error (difference between s_i and y_i)

Training Perceptron

- Calculate output for each training sample
  計算每個訓練樣本的輸出
- For each sample, weights w_i are adjusted so as to minimize output error
  在每個樣本之後，調整 b_i 以使輸出誤差最小化



It is linear regression if mean square error is applied.

Historic review of Perceptron 歷史回顧

- Artificial neural network developed in 1950-60
- Perceptron (a book by M.Minsky) gives a very pessimistic view on neural network
  - cannot do XOR, connectedness, parity.
- Funding for neural network was cut and direction on AI research was changed.
  削減神經網絡的研究資金

Limitation of Perceptron

Perceptron can only learn linearly separable functions.

只能學習線性可分離函數

Weather : Raining		
outside	umbrella	result
Y	Y	✔

XOR cannot be classified by linear separator

XOR不能線性分類

(01) and (10) → 0 ✘

(00) and (11) → 1 ✔

Historic review of Perceptron 歷史回顧 continue….

Deep Learning

- Nonlinear activation function
- The neuron should connect to every input (instead of some local inputs)
- Multi-levels with millions neurons
- Any function can be computed.

Neural network is now basic for machine learning

Why Deep Learning (DL)?

- Starting 2006, DL starts to outperform other ML techniques
- What make DL work?
  - Faster machines and multicore CPU/GPU
  - Big data – large dataset
  - New models and algorithms
  - Major companies, Google, Microsoft, SAS, Amazon, eBay, Facebook, Alibaba, Baidu, Huawei and Tencent, invest heavily on Deep Learning and Data Analytics

HSMC Awarded HK$7million

Establishment of DLC

Deep Learning Research & Applications Centre

Three-year grant starting 2016

One of the ongoing projects:
Machine Translation on IPO documents

Phrase based Machine Translation

Translate :

- I want to go to the market today.

Break sentence into chunks

- I | want | to go | to the market | today.

Translate model

- 我| 想 |去 | 市場 | 今天

Language model

Rule base : adding rule to reorder the words

- 我今天想去市場 or 今天我想去市場

Google translate 我想今天去市場

More and more rules are added.

Machine Translation system becomes very complicated

Translate using statistics

1. Break sentence into parts
   I | want | to go | to the market | today.
2. Find all possible translates

我	想	去	市場	今天
	會	走	交易	當今
	要求	達到	上市	現在
	需要	消失	推銷	
	希望	結束		
	必須			

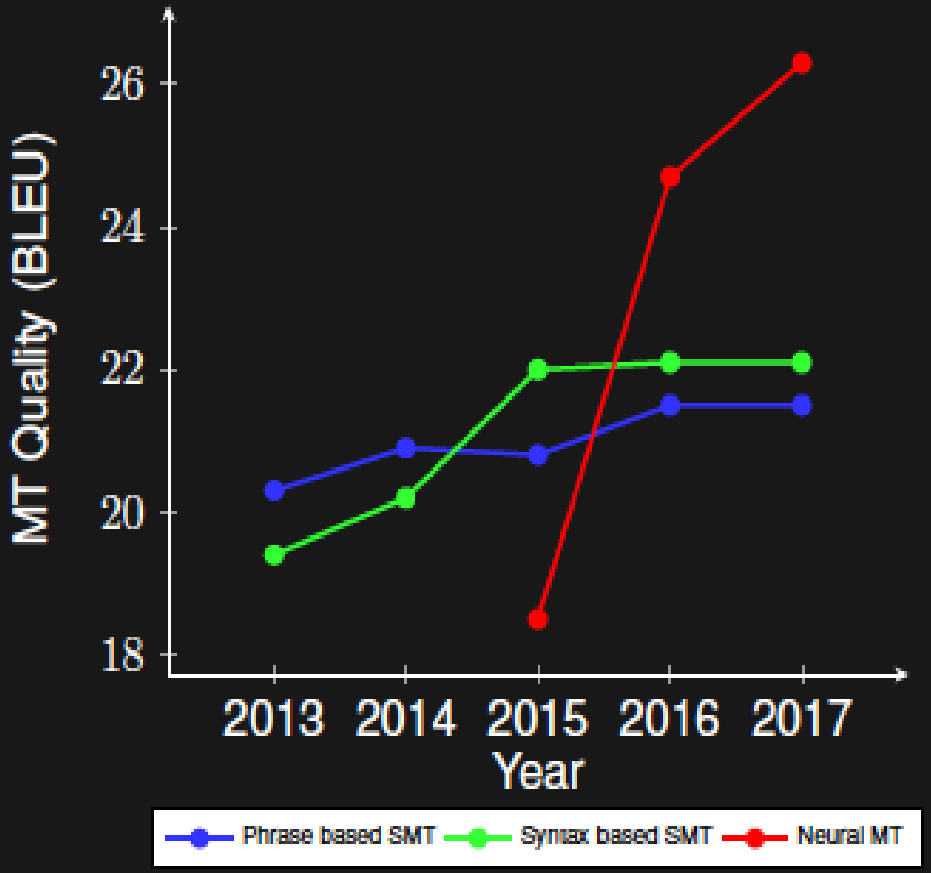
3. Generate all possible sentences (combinations)
   - 我 要求 走 交易 今天
   - 我 會 消失 市場 現在
   - 消失 我 會 今天 市場
4. Find the most common and similar one
   - 今天 我 想 去 市場

Machine Translation and Natural Language Processing (NLP)

- 1970-80's, early NLP is rule-based (by hand)
- Using statistical inference to learn rules through large collections of documents
  - Spelling 拼写(dtwhe, rtfgu, …)
  - Grammar 语法 (I go, he goes, one book, two books …)
  - Structure 结构 (subject – verb – object,…)
  - Style 样式
  - Meaning 意义
- Hypothesis: spelling, grammar, structure… are all hidden patterns in documents (拼写

What about word meaning?

Progress in Machine Translation



Data from [Sennrich 2016, http://www.meta-net.eu/events/meta-forum-2016/slides/09_sennrich.pdf]

Representing Words in NLP

One-hot representation:

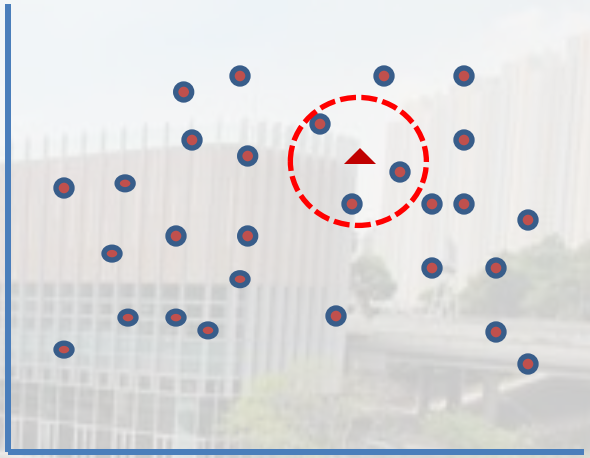
One token for each word as an N-dim vector

man = (0, 0, …, 0, 1, 0, 0, …, 0, 0)
woman = (0, 0, …, 0, …0, 1, 0, 0, …0)
human = (0, 0, …, 1, …0, 0, 0, 0, …0)
manly = (0, 0, …, 0, 0, 0, 0, 1, 0,.. 0)

- shortcomings:
  - Size of space, R^N
    (size of vocabulary, N=13 million English words),
  - semantic independent (each word is independent)

Word Vectors encode similarity and difference between words in the vector.

History of the development of Vector Space Models (VSM)



- SMART information retrieval system (Salton, 1971)
  - Mappi… documents to a vector
  - M… same vector space
  - A… close to query in vector space

How are library books represented in the system?

- Words are mapped to vectors, similar words mapped to nearby vectors (embedding)

Word-context (similarity)

- Distributional hypothesis (in linguistics)
  - words in similar contexts have similar meanings (Harris 1954)
  - "You shall know a word by the company it keeps." 物以類聚 (Firth 1957)
- Hypothesis: Related words often appear in same documents

{"economic", "money", "finance", "banks",…},

{"dogs", "cat", "pets", ..} likely appear together

"money", "shrimp", "liver", "north" seldom appear together

Word2Vec by Deep Learning

Fake task (continuous bag of words CBOW)

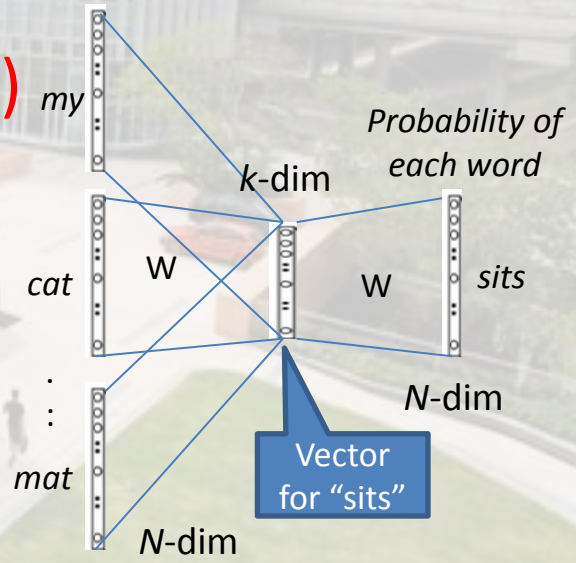
- To predict word by surrounding context
- i.e., guess ? from {“My”, “cat”, “on”, “the”, “mat”}
  (supervised learning)

Possible words (“sits”, “sat”, “sleeps”, “slept”, “crawls”, “pees”…)

Definitely not (“iron”, “whale”, “north”,…)

Train DL network with many documents

The hidden layer represents word “sits”



Observations:

Similar words, “smart” and “intelligent” likely to have similar context, and similar word vectors.

Word Vector Analogies

Given incomplete analogy a : b = c : x

For example: boy : girl = man : x

Identify the word x, s.t., girl - boy = x - man,

i.e., x = girl – boy + man

Download English Wikipedia Dump

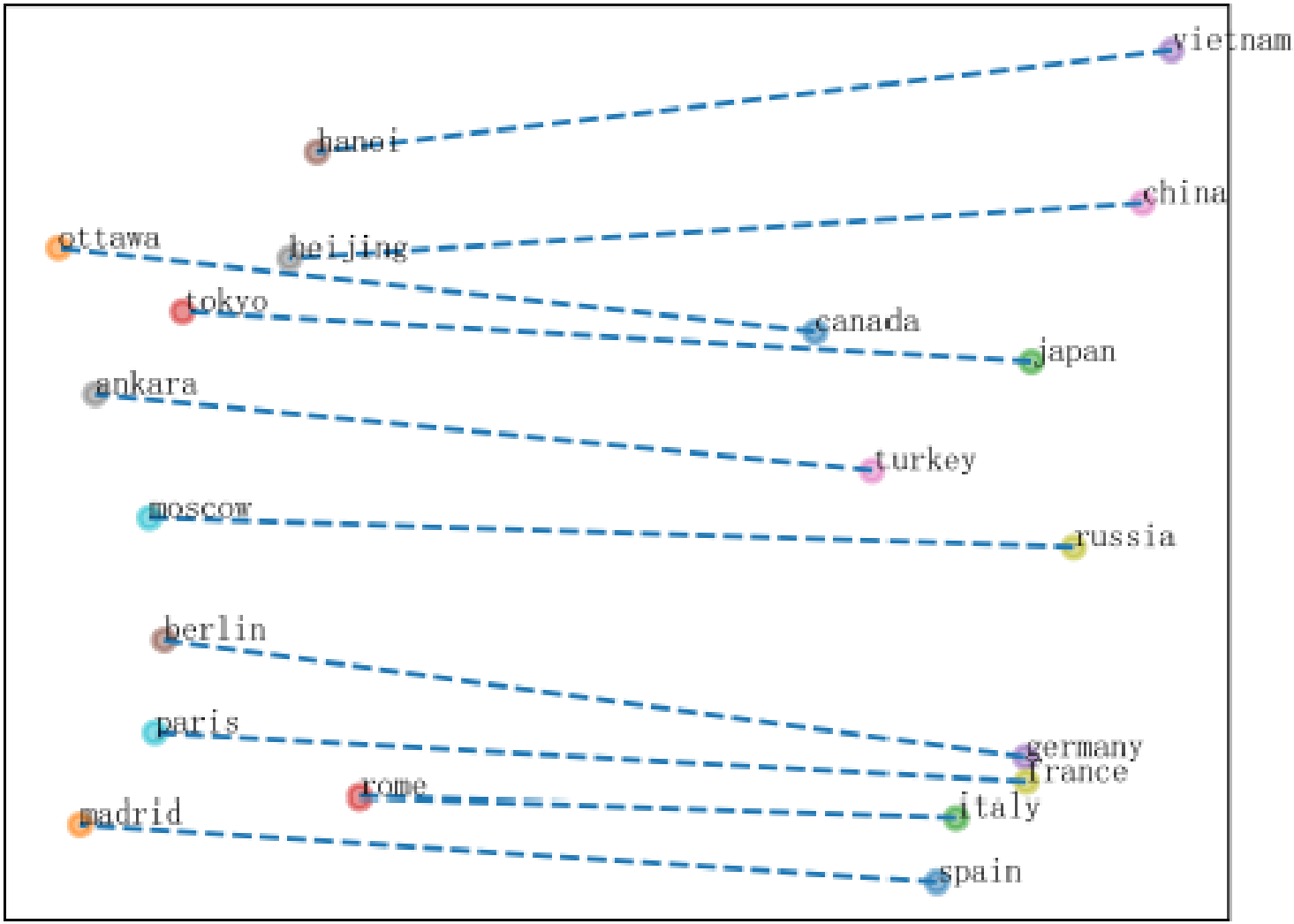
- https://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles.xml.bz2
- 13+ G file size, 4,000,000+ articles
- Preprocess, 2G+ memory required

```
1 # Remove words occuring less than 20 times, and words
      occuring in more than 10% of the documents. (keep_n is
       the vocabulary size)
2 wiki.dictionary.filter_extremes(no_below=20, no_above
      =0.1,^^lkeep_n=100000)
```

Train Word2Vec

Use Gensim for 24 hors training on virtual server

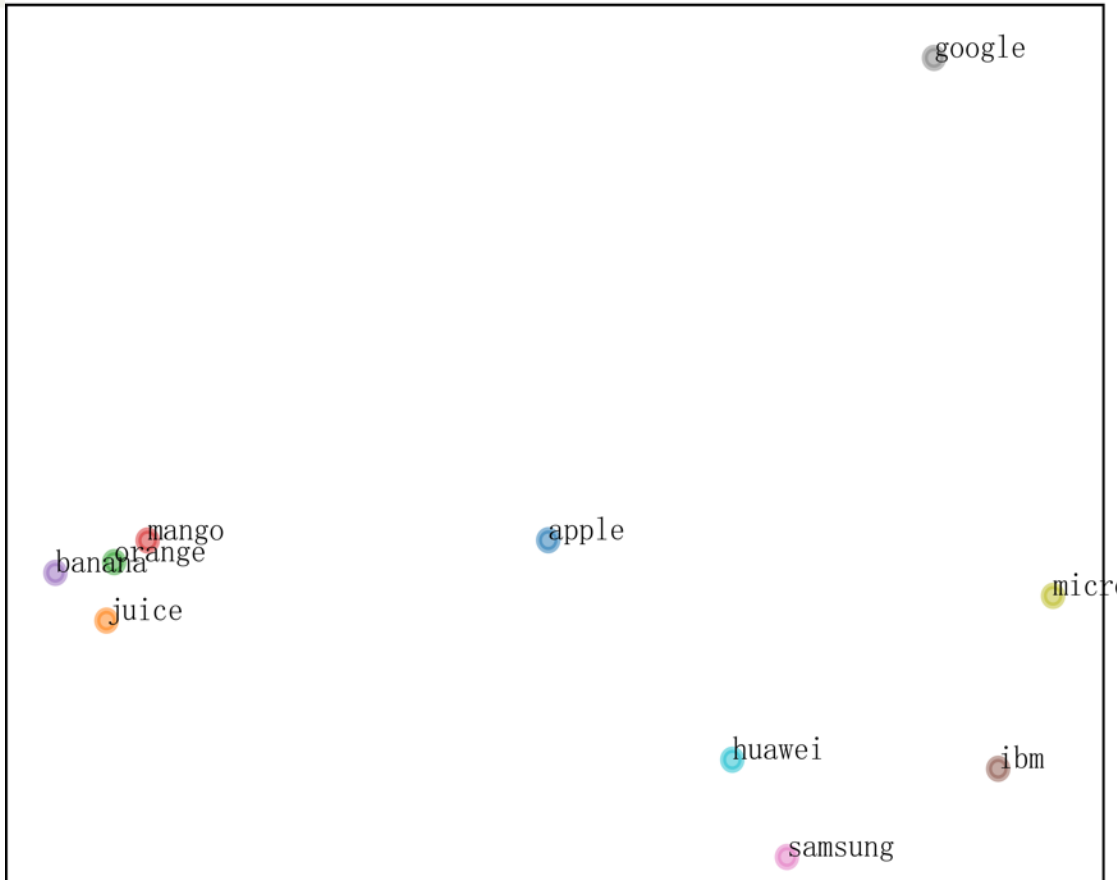
Word Analogy: Country-Captain Paris



Word2Vec Result

- Word similarity:

```
1 model = gensim.models.Word2Vec.load(os.path.join(MODEL_PATH
      , 'word2vec.model'), mmap='r')  # load large array
2 print(model.similarity('apple', 'mango'), model.similarity(
      'apple', 'ibm'))
3 0.465574139702 0.531822866913
```



Syntax testing

Superlative adjectives

- good : best = big : biggest
- good : best = bad : worst
- good : best = cold : coldest
- good : best = easy : easiest

Present and past tense

- walking : walked = increasing : increased
- walking : walked = falling : fell
- walking : walked = going : went
- walking : walked = hitting : hit
- walking : walked = hiding : hid

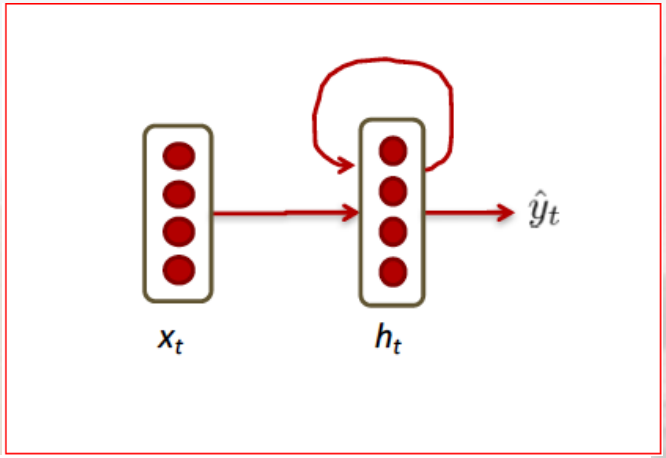
Recurrent Neural Network (RNN)

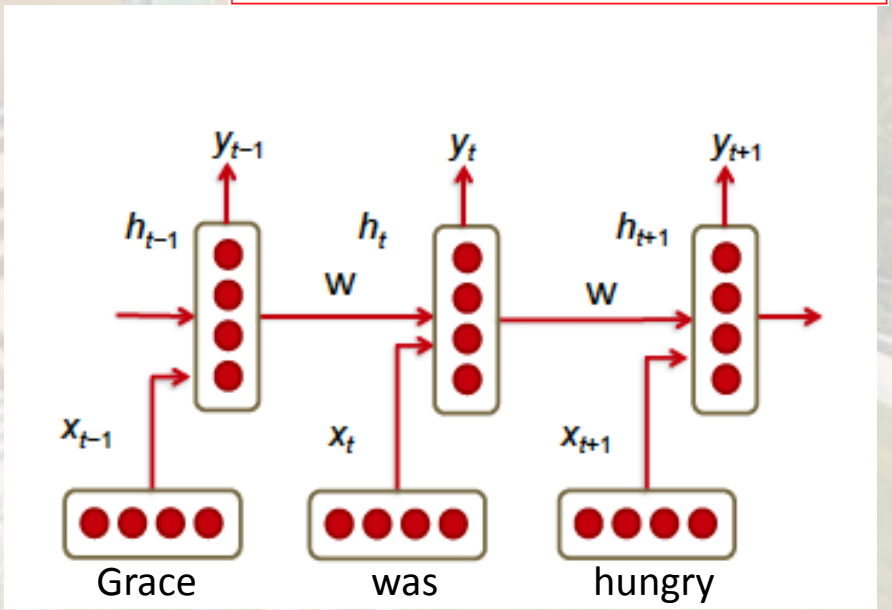
Language is a sequence (not a bag) of words.

- The food is good, not bad at all.
- The food is bad, not good at all.

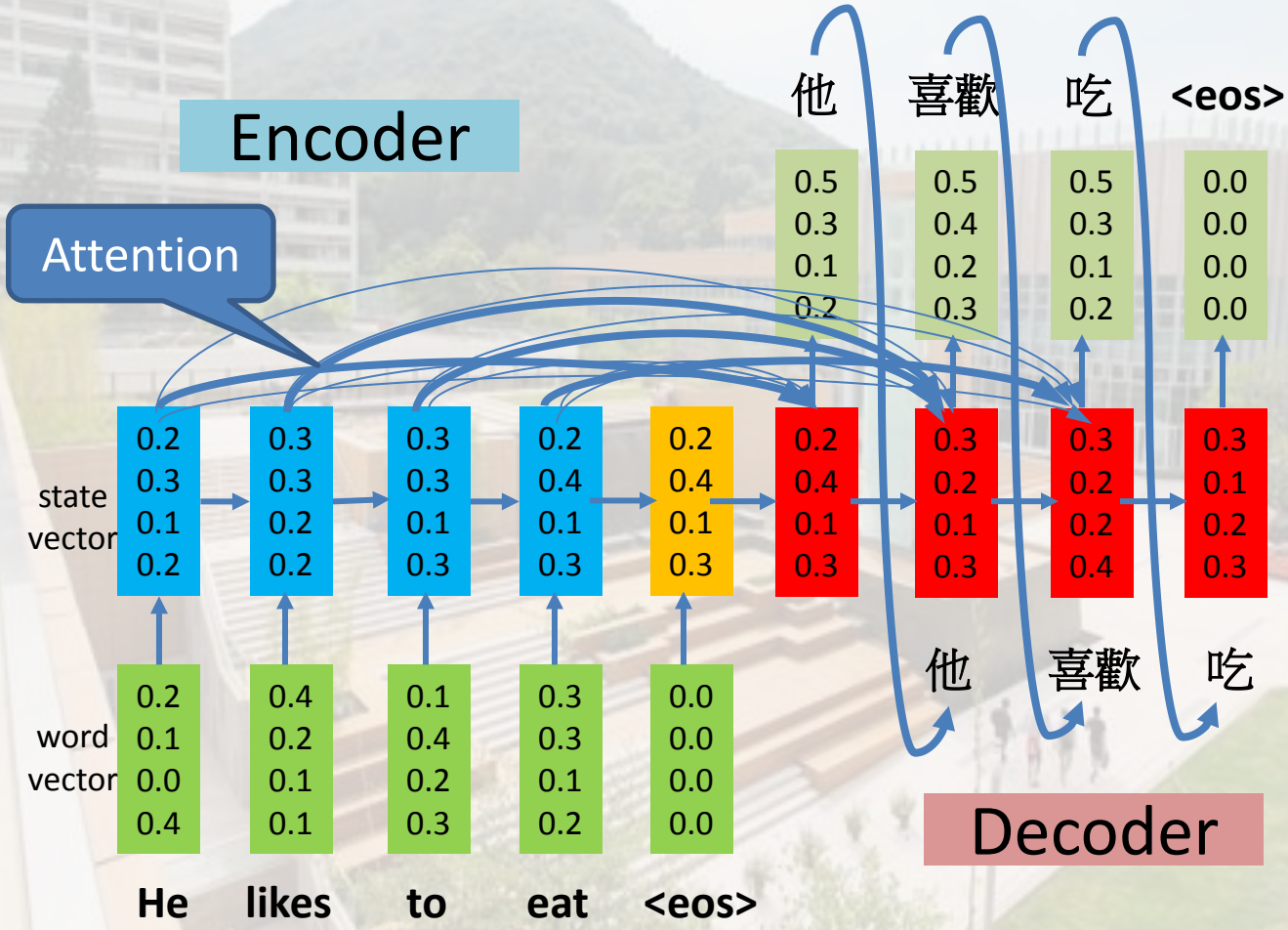
RNN is used to build language model

1. maintain word order
2. share parameters across the sequence
3. keep track of long-term dependencies





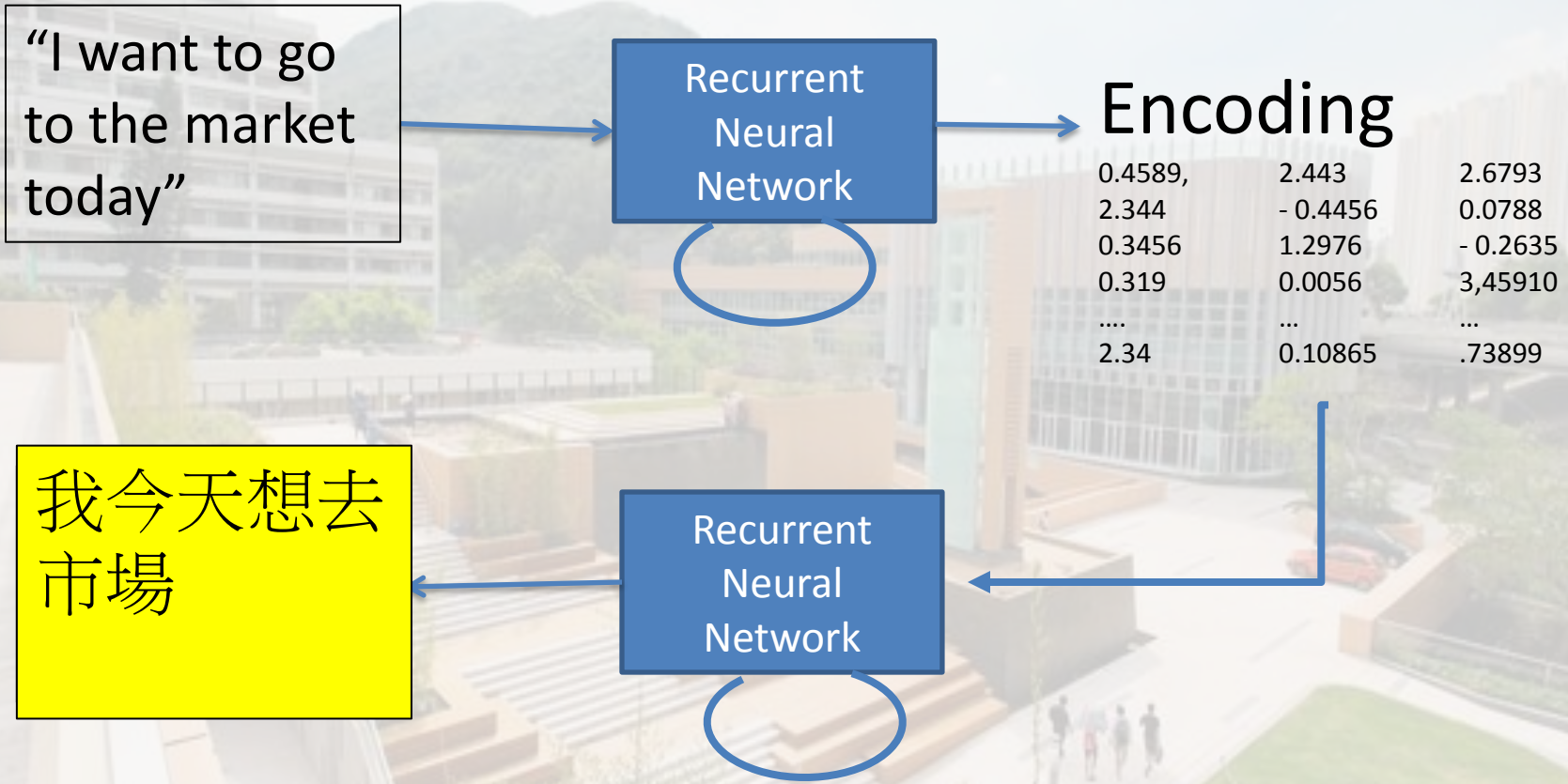
A BiLSTM Encoder and LSTM with Attention Decoder



Our Translation Advantages

- Focused on financial documents
- Confidentiality – run on local servers
- Glossary for different fields
- Similar subsections and styles,
  e.g., introduction of companies, CVs, financial statement, declarations, …
- Training data:
  about 900 IPO documents and 2.1 paired sentences
- Got ITF support and industrial interest.

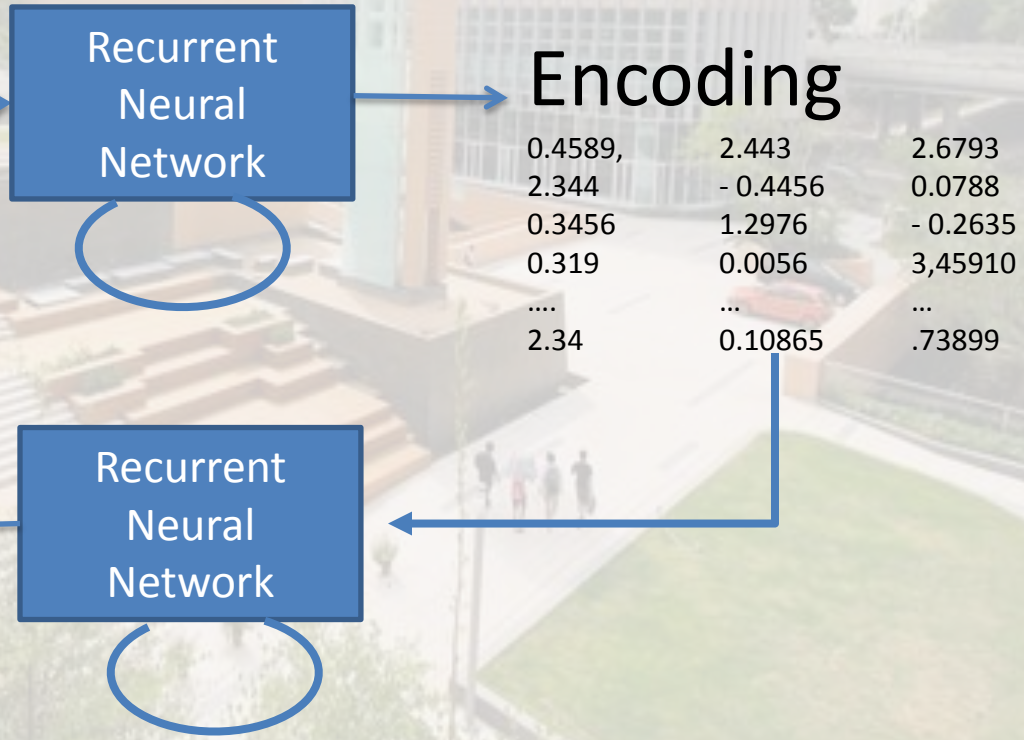
Machine translation



Automatic Answering System

Google Chat logs
Employees and Tech Support Team:

Connection refused or something like that.



May I know the version of network connect?

Picture Caption

Input: picture

Output : image caption

Recurrent Neural Network

Encoding

0.4589,	2.443	2.6793
2.344	- 0.4456	0.0788
0.3456	1.2976	- 0.2635
0.319	0.0056	3,45910
….	…	…
2.34	0.10865	.73899

Recurrent Neural Network

梁振英和
曾蔭權坐在
一盆花兩旁

Language Model for Style

All work of Shakespeare → Language model (RNN) →

KING LEAR:
O, if you were a feeble sight, the courtesy of your law,
Your sight and several breath, will wear the gods
With his heads, and my hands are wonder'd at the deeds,
So drop upon your lordship's head, and your opinion
Shall be against your honour.

Base on statistics, the next word is generated according to probability

G

- Deep Learn
  and regress

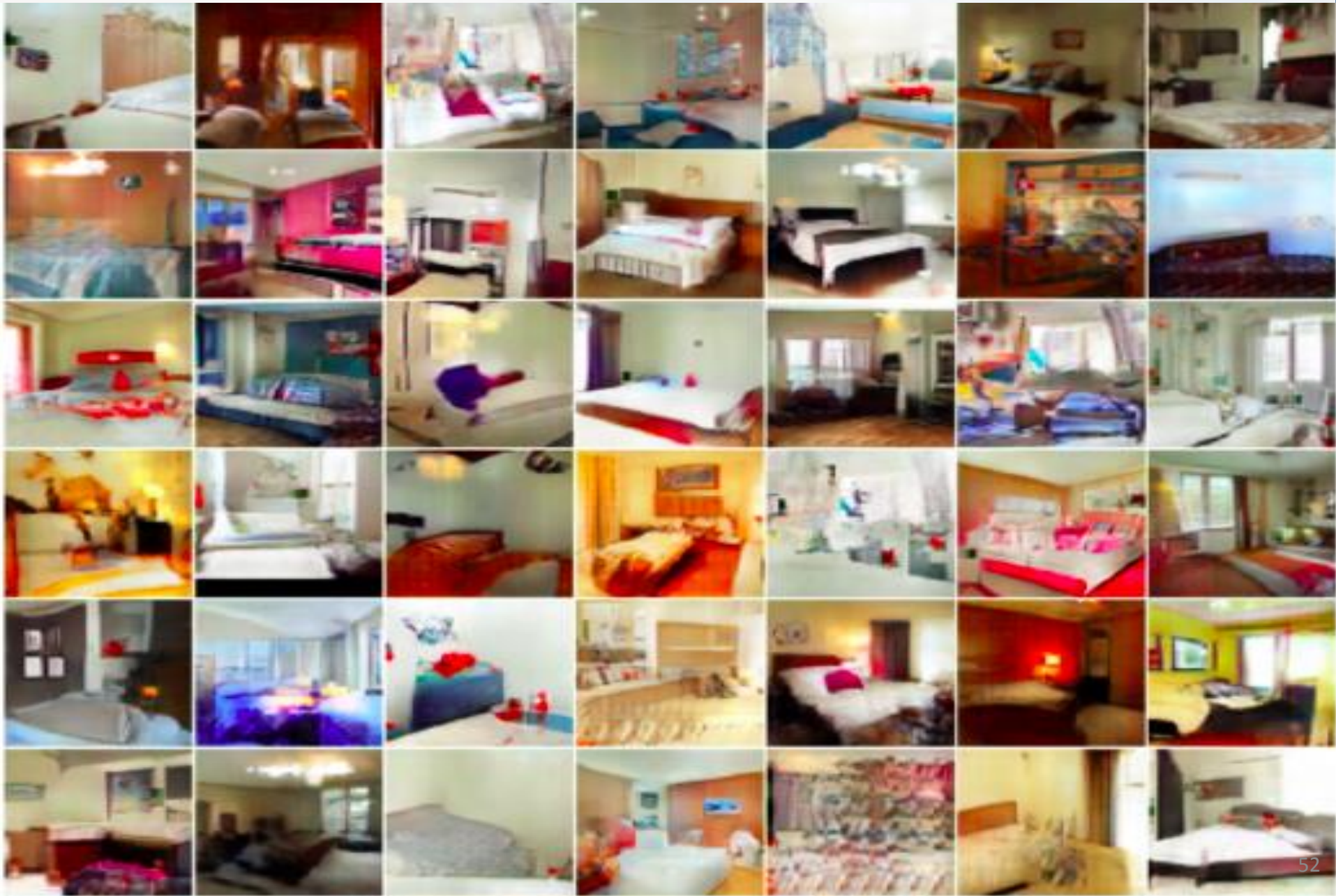
Other exampl

- Image
- Speech
- Handwritin

37:29 The righteous shall inherit the land, and leave it for an inheritance unto the children of Gad according to the number of steps that is linear in *b*.

hath it not been for the singular taste of old Unix, “new Unix” would not exist.

Bedrooms created by Generative models



What projects can benefit from DL?

Problems with an hypothesis (data relationship)

- Existence of a pattern
- Large amount of data

Supervised learning for classification or regression

Input data with labels

Hypothesis: label is linked to a pattern

Ex: box office prediction, image recognition, spam emails, …

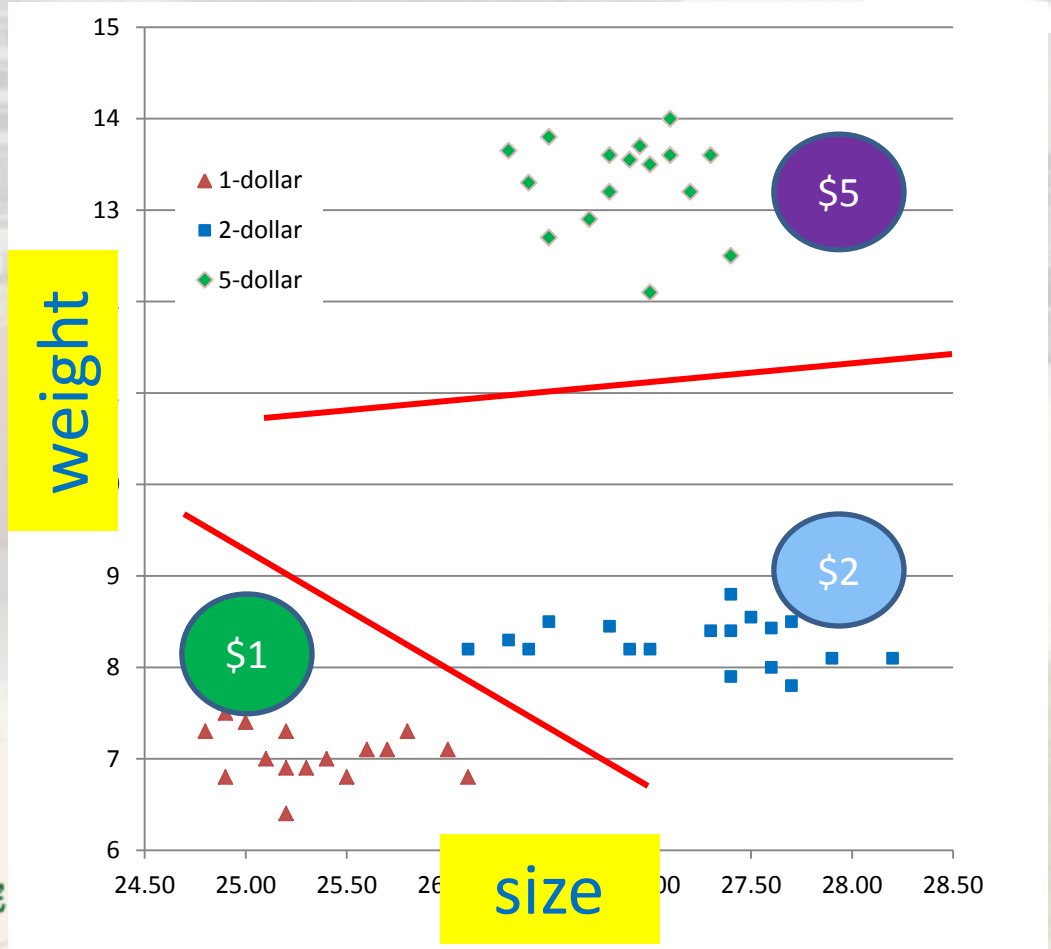
Unsupervised learning to learn about the data

Input raw data

Coin Classification

Training Data: Supervised learning

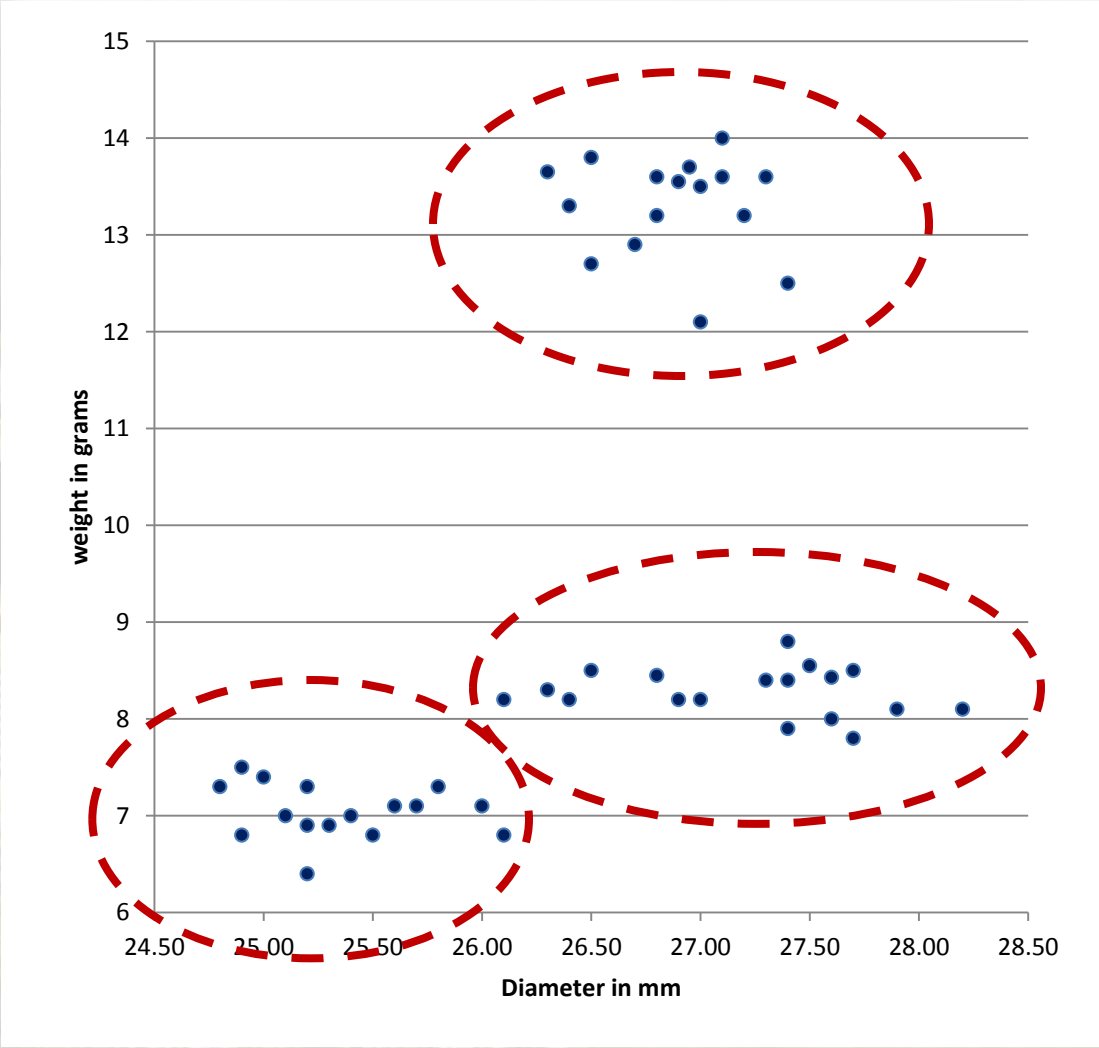
(<size, weight >, type)



Unsupervised Learning

Learn from raw data and look for patterns

- Grouping exists



Where to go?

- Supervised learning
  - More data, better the results
  - Limited by the amount labeled data
- Unsupervised learning (unlimited data)
  - Bigger model (more layers and neurons) better the results
  - Limited by ability to process data (raw)
- Reinforcement learning
  - choose action to maximize expected long term reward (example: AlphaGo)

THANKS
Q & A

恒生管理學院
HANG SENG
MANAGEMENT COLLEGE

Language models

- "My cat sits on the mat" is assigned high probability because a completely valid sentence, syntactically and semantically.
- "car cold swim red the" has low probability

Unigram:

$P(w_1, w_2, \ldots, w_n) = P(w_1)P(w_2)\ldots P(w_n)$

Bigrams:

$P(w_1, w_2, \ldots, w_n) = P(w_1)P(w_2|w_1)\ldots P(w_n|w_{n-1})$

Trigrams:

$P(w_1, w_2, \ldots, w_n) = P(w_1)P(w_2|w_1)\ldots P(w_n|w_{n-1}w_{n-2})$


带好孩子
Take the child
小心落水
Fall into water carefully

上海市公安局城市轨道交通分局
Urban Mass Transportation Branch Shanghai Public Security Bureau
www.engrish.com

Replace the battery watch band

Term-document

- Term-document (Matrix A)
  - Rows: terms/words 术语/词语
  - Columns: documents 文档
  - $A_{i,j}$ = frequency of the word *i* occurs in document *j*
- Bag of words hypothesis (in information retrieval)
  - document – bag of words (order does not matter)
  - words indicate document topic 词语表示文档主题
  - documents of similar topics use similar words
    相似题目的文件使用相似的词
  - their document vectors will be similar 文档向量相似
- Focus: document similarity

Document 1: My husband and I have three cats and two dogs, and I consider myself a cat person *and* a dog person. In fact, cats and dogs often come into our lives in unexpected ways….

Document 2: I like cats and dogs. My cat sits on the mat. My dog like to play with my cat….

Document 3: Hong Kong is one of the world's most significant financial centres, with the highest Financial Development Index score and consistently ranks as the world's most competitive and freest economic entity…

Terms	Document 1	Document 2	Document 3
cat / cats	3	3	0
Dog / dogs	3	2	0
Hong Kong	0	0	1
financial	0	0	2
I / my / our / myself	5	4	0
economy	0	0	0

Word-context (similarity)

- Word-context – generalization of term-document
  - Rows: words
  - Columns: context (document, phrases, sentences, paragraphs, chapters, books, …)
- Word similarity Focus shifted from rows to columns;
- Distributional hypothesis (in linguistics)
  - words in similar contexts have similar meanings
    相似语境中的词语具有相似的含义 (Harris 1954)
  - "You shall know a word by the company it keeps."
    物以類聚 (Firth 1957)
- Hypothesis: Related words often appear in same documents 相关单词经常出现在同一文档中

"economic", "money", "finance", "banks", likely appear together
"money", "shrimp", "liver", "north" seldom appear together

Document 1: My husband and I have three cats and two dogs, and I consider myself a cat person *and* a dog person. In fact, cats and dogs often come into our lives in unexpected ways….

Document 2: I like cats and dogs. My cat sits on the mat. My dog like to play with my cat….

Document 3: Hong Kong is one of the world's most significant financial centres, with the highest Financial Development Index score and consistently ranks as the world's most competitive and freest economic entity…

"Dog" and "cat" are similar

Terms	Document 1	Document 2	Document 3
cat / cats	3	3	0
Dog / dogs	3	2	0
Hong Kong	0	0	1
financial	0	0	2
I / my / our / myself	5	4	0
economy	0	0	0

Co-occurrence Matrix

Disadvantages of term-document matrix:

- size NxM is very large
- Number of documents, M, is not fixed

Window-based Co-occurrence Matrix (affinity matrix)

– no of times each word appeared within a window of particular size around word of interest.

I like cats and dogs. My dog likes my cat.
My cat and my dog sit on the mat.

Window size =1

X =

	I	like	cat	and	dog	my	sit	on	the	mat
I	0	1	0	0	0	0	0	0	0	0
like	1	0	1	0	1	1	0	0	0	0
cat	0	1	0	2	0	2	0	0	0	0
and	0	0	2	0	1	1	0	0	0	0
dog	0	1	0	1	0	2	1	0	0	0
my	0	1	2	1	2	0	0	0	0	0
sit	0	0	0	0	1	0	0	1	0	0
on	0	0	0	0	0	0	1	0	1	0
the	0	0	0	0	0	0	0	1	0	1
mat	0	0	0	0	0	0	0	0	1	0

I like cats and dogs. My dog likes my cat.
My cat and my dog sit on the mat.

"Dog" and "cat" are more similar than with "mat", "sit"…

size =2

X =

	I	like	cat	and	dog	my	sit	on	the	mat
I	0	1	1	0	0	0	0	0	0	0
like	1	0	2	1	1	2	0	0	0	0
cat	1	2	0	2	1	3	0	0	0	0
and	0	1	2	0	2	2	0	0	0	0
dog	0	1	1	2	0	2	1	1	0	0
my	0	2	3	2	2	0	1	0	0	0
sit	0	0	0	0	1	1	0	1	1	0
on	0	0	0	0	1	0	1	0	1	1
the	0	0	0	0	0	0	1	1	0	1
mat	0	0	0	0	0	0	0	1	1	0

Singular Value Decomposition

$X_{NxN} = U_{NxN}\ S_{NxN}\ V_{NxN}$

$$\underset{N}{} \begin{matrix} N \\ N \end{matrix} \begin{bmatrix} \text{---}\ u_1\ \text{---} \\ \text{---}\ u_2\ \text{---} \\ \vdots \end{bmatrix} \quad \begin{matrix} N \\ \end{matrix} \begin{bmatrix} s_1 & 0 & \ldots. \\ 0 & s_2 & \ldots. \\ \vdots & \vdots & \end{bmatrix} \quad \begin{matrix} N \\ \end{matrix} \begin{bmatrix} | & | & \\ v_1 & v_2 & \ldots. \\ | & | & \end{bmatrix} \begin{matrix} \\ N \end{matrix}$$

Based on the largest k values, s_1, s_2, … s_k, use the submatrix U_{Nxk} for word embedding matrix

$$\begin{matrix} \\ N \end{matrix} \begin{matrix} k \\ \end{matrix} \begin{bmatrix} \text{---}\ u_1\ \text{---} \\ \text{---}\ u_2\ \text{---} \\ \vdots \end{bmatrix} \quad \begin{matrix} k \\ k \end{matrix} \begin{bmatrix} s_1 & 0 & \ldots. \\ 0 & s_2 & \ldots. \\ \vdots & \vdots & \end{bmatrix} \quad \begin{matrix} N \\ \end{matrix} \begin{bmatrix} | & | & \\ v_1 & v_2 & \ldots \\ | & | & \end{bmatrix}$$

Problems and Issues

- Matrix dimension changes because of new words
- Large and extremely sparse matrix (ignore words such as “the”, “has”, “is”, …)
- Quadratic cost to perform SVD

Another Approach based on iteration.

Continuous Bag of Words Models (CBOW)

- Predict center word "sits" from surrounding context {"My", "cat", "on", "the", "mat"}

w_{c-2} w_{c-1} w_{c+1} w_{c+2} w_{c+3}

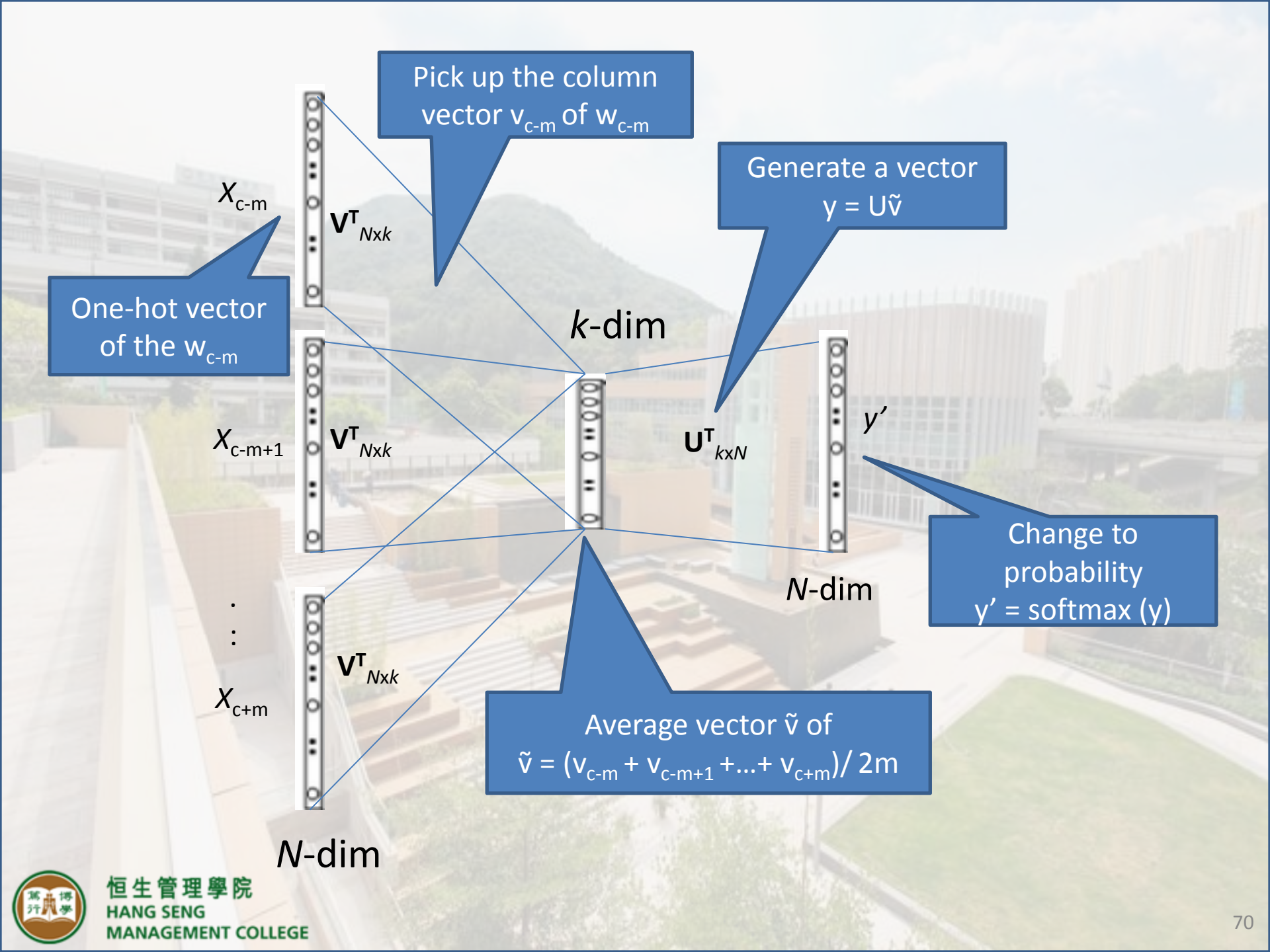
- Input word matrix $V \in R^{kxN}$,
  v_i: the column of V represents w_i,
  i.e., $v_i = Vx_i$ and x_i = one-hot vector of w_i

$$V = \begin{bmatrix} | & | & | & | & | \\ v_{c-1} & v_{c+3} & v_{c-2} & v_{c+1} & v_{c+2} \\ | & | & | & | & | \end{bmatrix} \begin{matrix} \\ k \\ \\ \end{matrix}$$

N

"cat" "mat" "My" "on" "the"

$$cat = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ 0 \\ 0 \\ \vdots \\ \vdots \\ \vdots \\ 0 \end{bmatrix} \quad mat = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ 1 \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix}$$


Pick up the column vector v_{c-m} of w_{c-m}
Generate a vector y = Uṽ
One-hot vector of the w_{c-m}
X_{c-m}
V^T_{Nxk}
k-dim
X_{c-m+1}
V^T_{Nxk}
U^T_{kxN}
y'
N-dim
Change to probability y' = softmax (y)
X_{c+m}
V^T_{Nxk}
Average vector ṽ of ṽ = (v_{c-m} + v_{c-m+1} +...+ v_{c+m})/ 2m
N-dim

Objective function for training

Train U and V to learn the generated probability y' from true (expected) probability y, which is one-hot vector.

- Objective function (cross entropy)

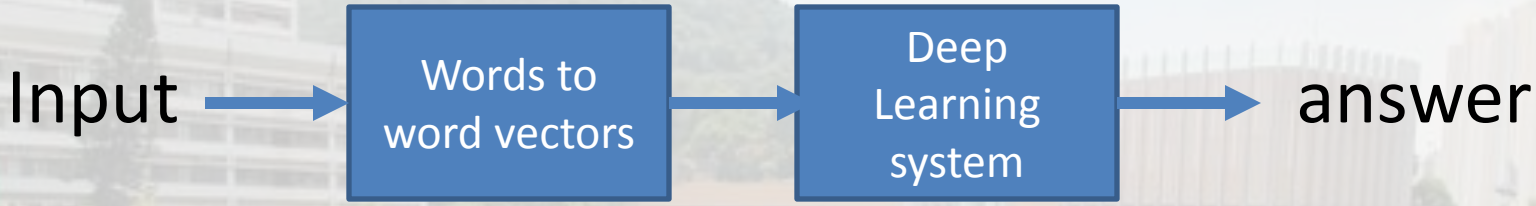
$$H(y',y) = -\sum_{j=1,\dots N} y_j \log(y'_j)$$
$$= -\, y_c \log(y'_c)$$
$$= -\log(y'_c)$$

Intuition:

- y' is perfect, i.e., $y'_c=1$, $H(y',y) = 0$ (no adjustment)
- y' is bad, ie., $y'_c=0.01$, $H(y',y) = -\log(0.01) = 4.6$

Evaluation of word vectors

Extrinsic evaluation



- slow and unclear the problematic part

Intrinsic evaluation

- Word Vector Analogies
- Analogy Evaluations
- Correlation Evaluation

HSMC Awarded HK$7million

恒生管理學院獲研資局700萬港元研究資助

Establishment of Deep Learning Research & Applications Centre 深度學習研究與應用中心

(i) support research through Deep Learning

(ii) provide Deep Learning environment (Big Data and Cloud Computing)

(iii) train Deep Learning specialists

3-Year project with outside collaborators

Vision and Image