Introduction to Speech and Natural Language Processing

Instructor: Tom Ko

Objectives

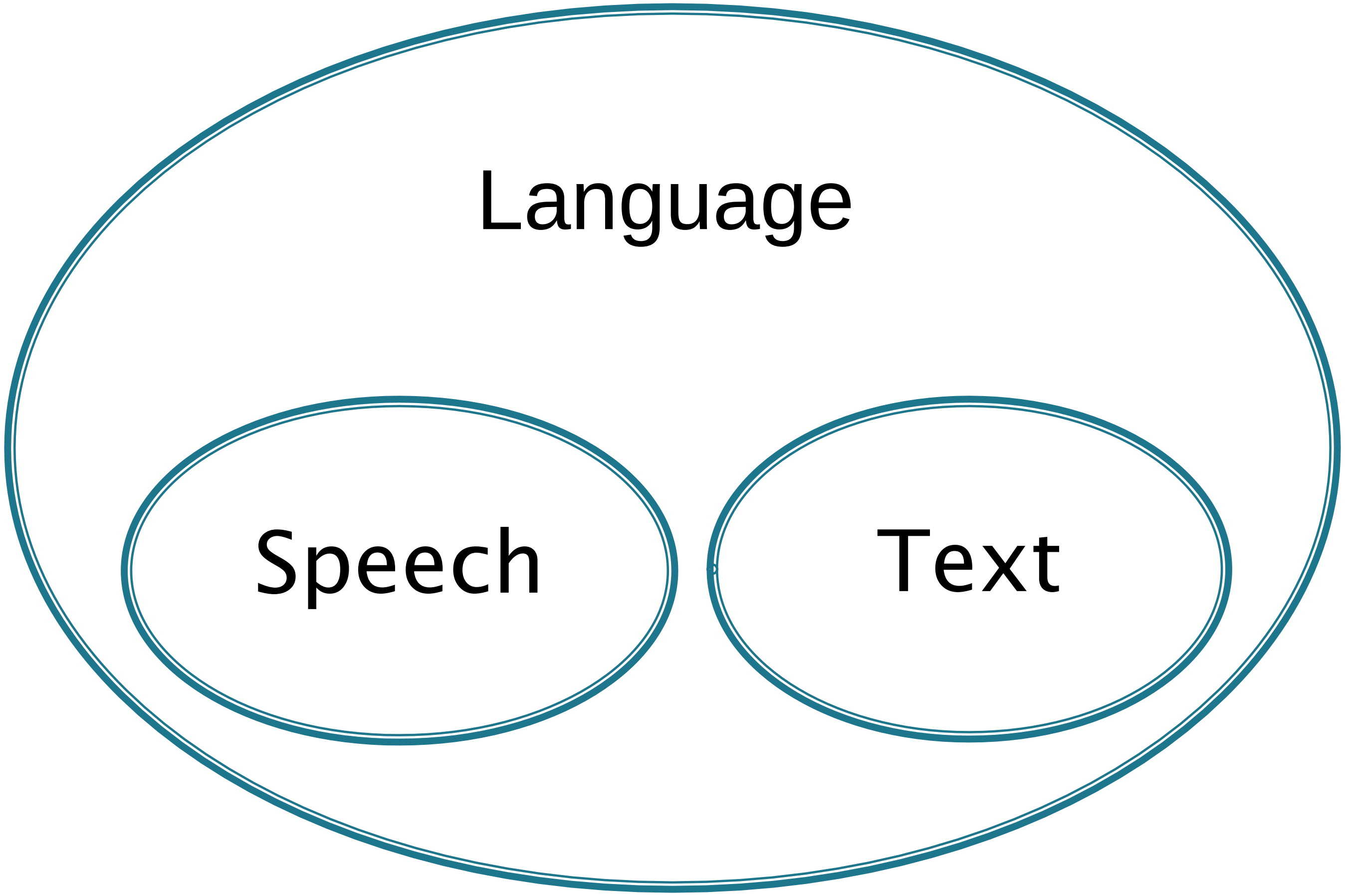
- Introduce speech related tasks
- Introduce NLP tasks
- Understand automatic speech recognition from a top-down approach

Speech and language

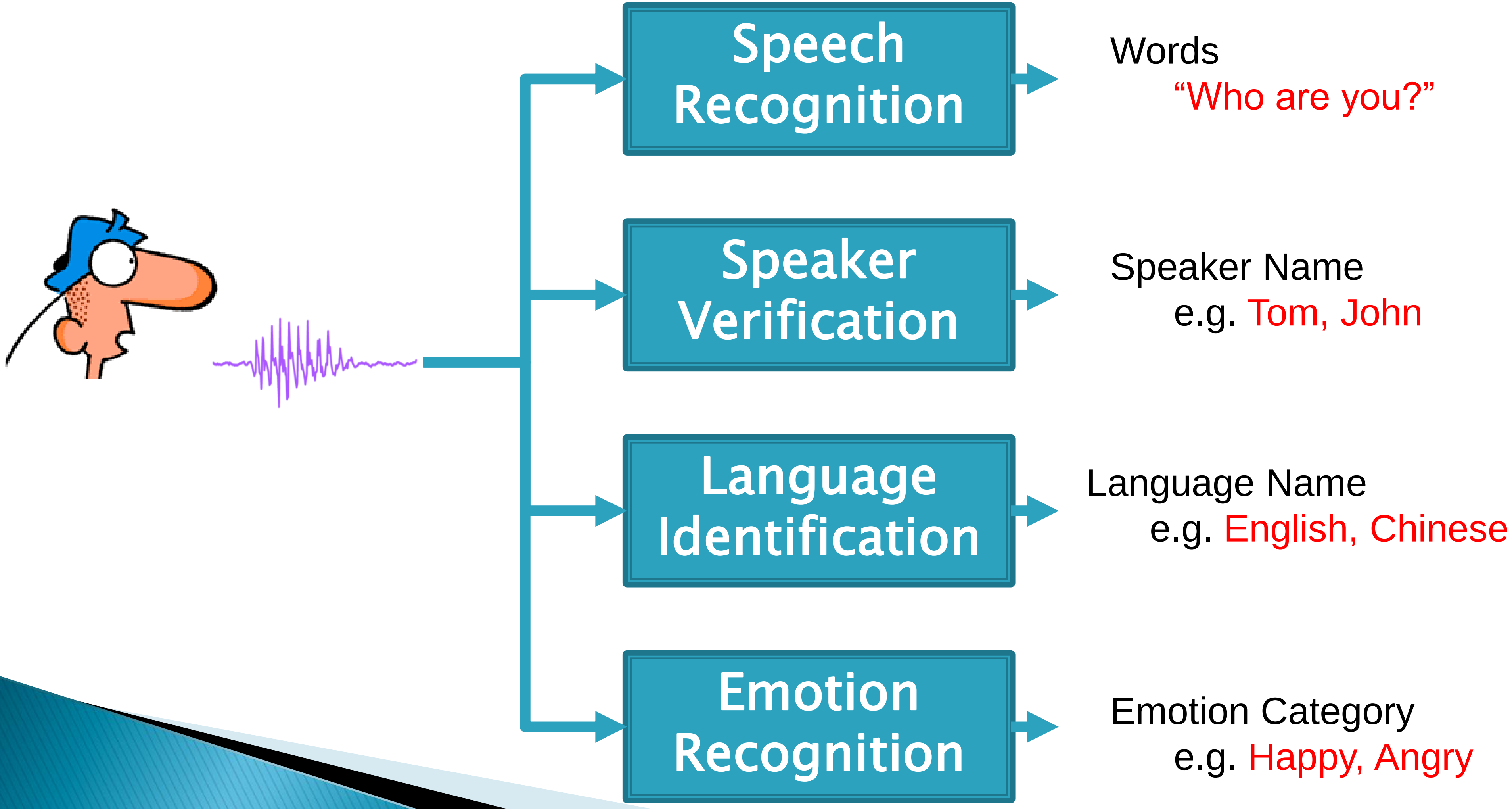
- **Speech** refers to the actual sound of spoken **language**.

- **Language** refers to a whole system of words and symbols, either written or spoken or both (except body language), for communication.

Speech and language



Major speech-related tasks



Speech Recognition

Words
“Who are you?”

Speaker Verification

Speaker Name
e.g. Tom, John

Language Identification

Language Name
e.g. English, Chinese

Emotion Recognition

Emotion Category
e.g. Happy, Angry

Major speech-related tasks

- The above tasks are all vocal-related. They have to make use of the information carried by the speech signal.

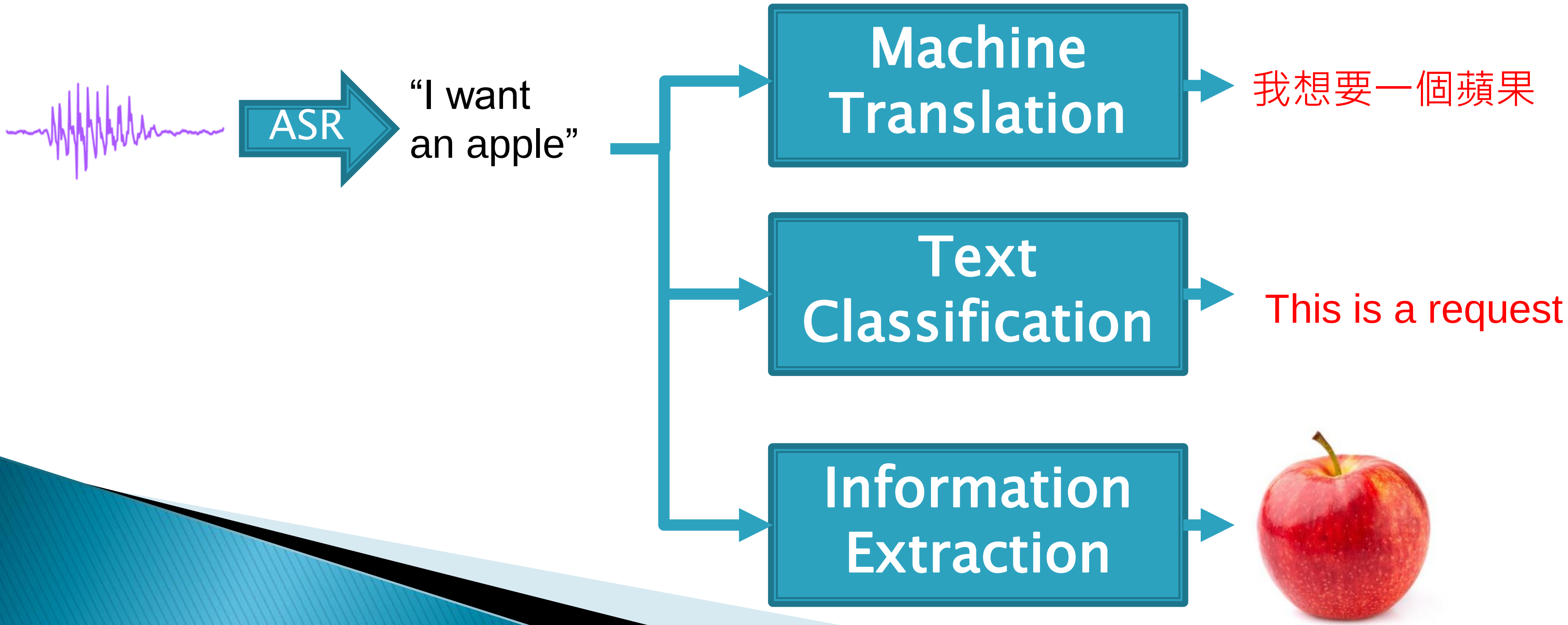
- Automatic speech recognition (ASR) is the most important task.

What makes ASR more difficult?

- Infinite number of classes
  - Infinite number of word combination
- Variable input and output length
- Out of vocabulary (OOV)
  - The words appearing in the test set may not appear in the training set.
- Sequence-to-sequence recognition

Relationship between ASR and NLP

- Automatic speech recognition (ASR)
- Natural language processing (NLP)



ASR

"I want an apple"

Machine Translation

我想要一個蘋果

Text Classification

This is a request

Information Extraction

NLP tasks

- They are mostly text-related tasks (no audio).
- The term “language understanding” itself is abstract.
  - What to understand?
- For human, they show their understanding by actions.
  - For machine, they show their understanding by concrete classification.

Confusion in human language

- Consider this sentence: “I am waiting for a man with a dog.”
  - Are you waiting for a man and a dog or waiting together with a dog ?
- If you mean the first one,
  - “I am waiting for a man and his dog.”
- Otherwise,
  - “I am waiting for a man together with my dog.”

- Another example: “He can complete the task which I assigned to him very quickly”.

- This kind of confusion is due to poor English writing.

Confusion in human language

- Once I was shopping in a mall, I am looking for a restaurant. I asked a lady.

- She pointed to a direction and said ”你往這邊一直走下去.”

- Should I walk straight to the end or go down one floor?

Confusion in human language

- “Please use mobile phones in the vestibule.”
  - *vestibule* ：門廳，門廊

- Does it mean “If I use mobile phone, I should use it in the vestibule.” ? Or does it mean a request?

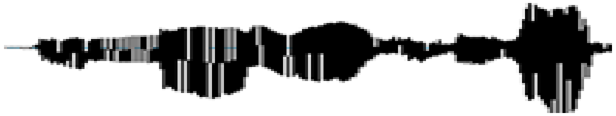
- The message is not only delivered by the text.

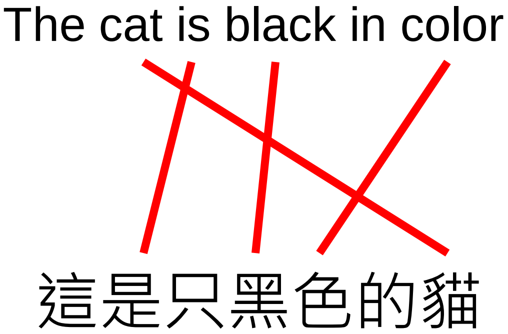
What makes NLP difficult?

- Confusion
  - Think about computer programming language
- Context
- Machine translation (MT) is regarded as one of the most representative task.

What makes MT even more difficult?

- Sequence-to-sequence recognition
- For ASR, the sequences are monotonic

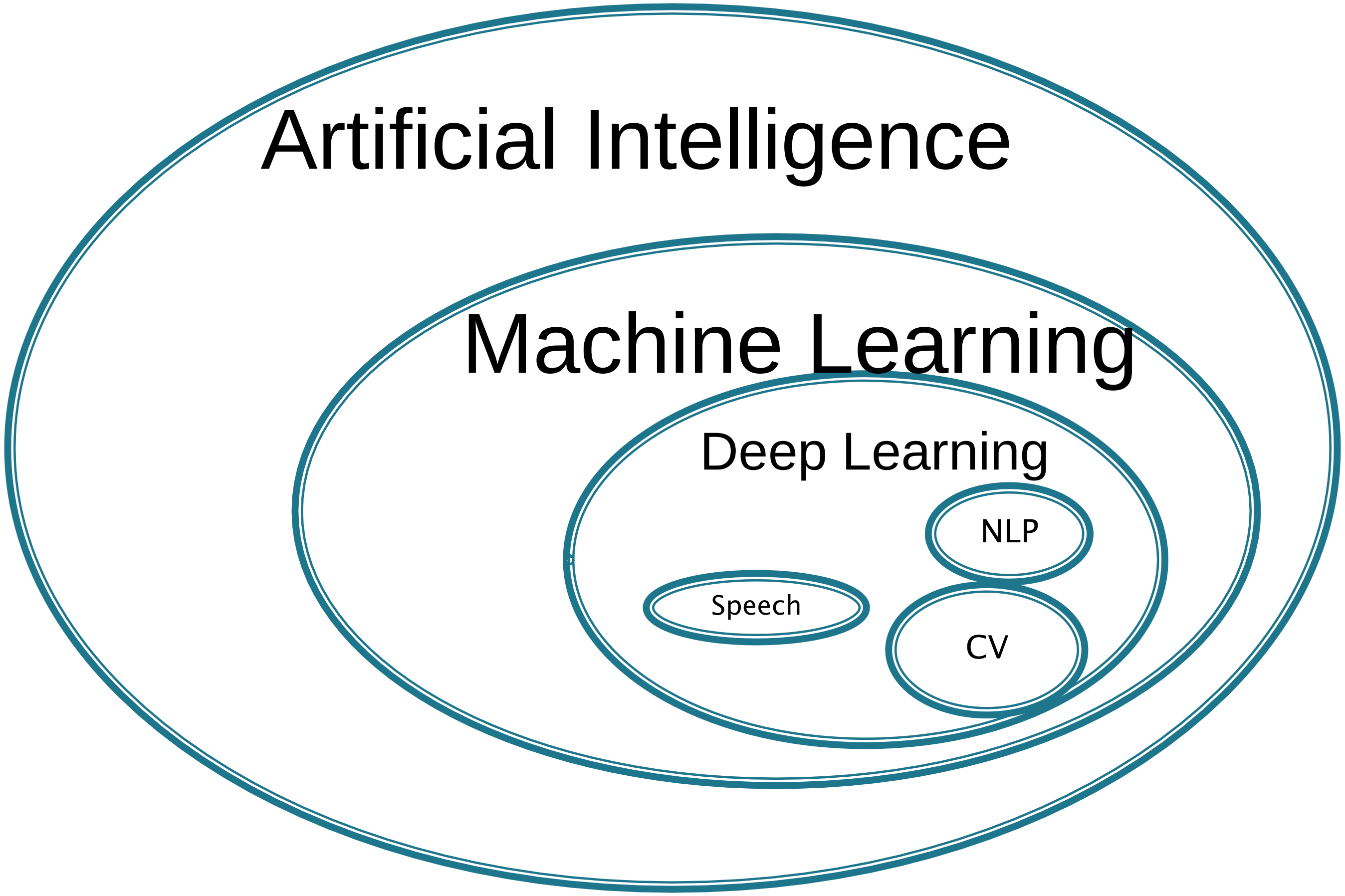




The cat is black in color

這是只黑色的貓

AI, ML and DL

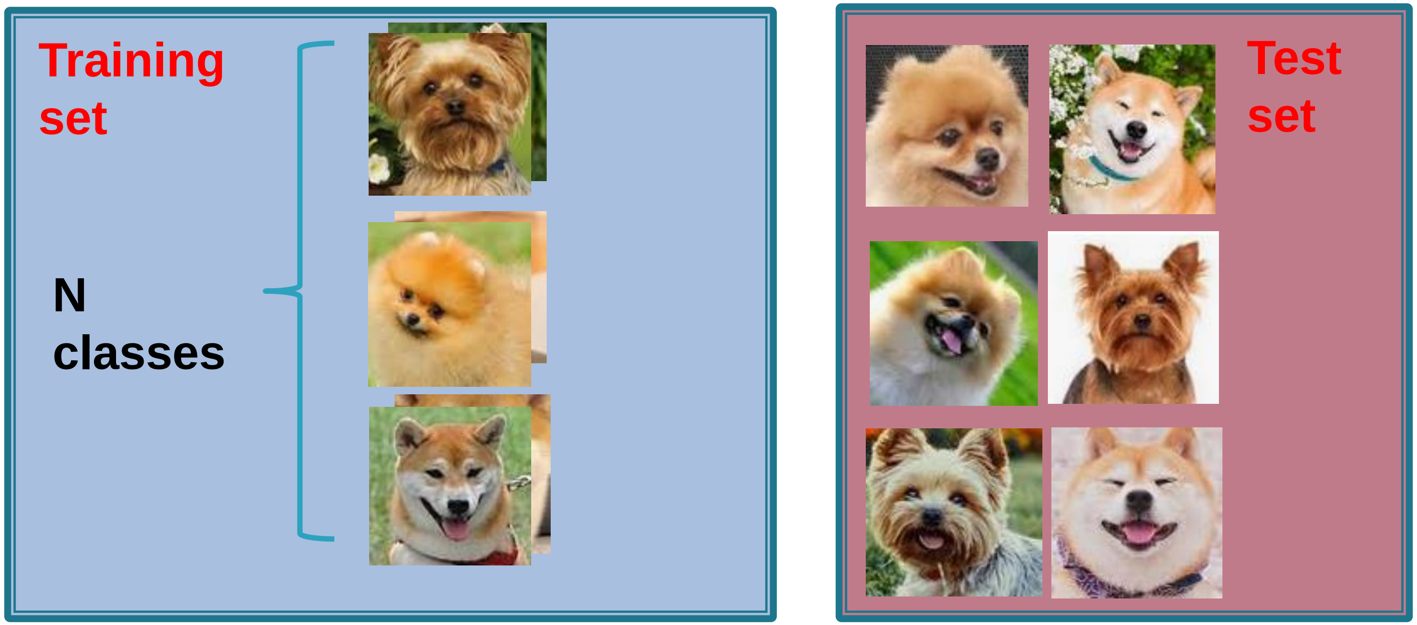

Artificial Intelligence
Machine Learning
Deep Learning
NLP
Speech
CV

It is all about classification

- Classification is a basic instinct of living organisms.
- Human can classify a lot of things in different domains.
- There are a lot of classification tasks which can be divided into different domains.

Classification task examples

Visual



Training set

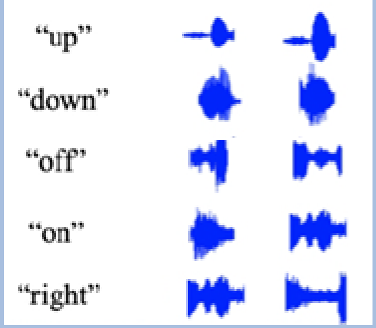
N classes

Test set

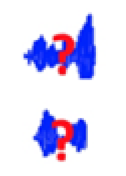
Classification task examples

Audio

Training set



Test set



Common in classification tasks

- Data
  - Training set, test set, development set
- Feature extraction
  - How to digitalize the input ?
- Variation and noise
- Model selection

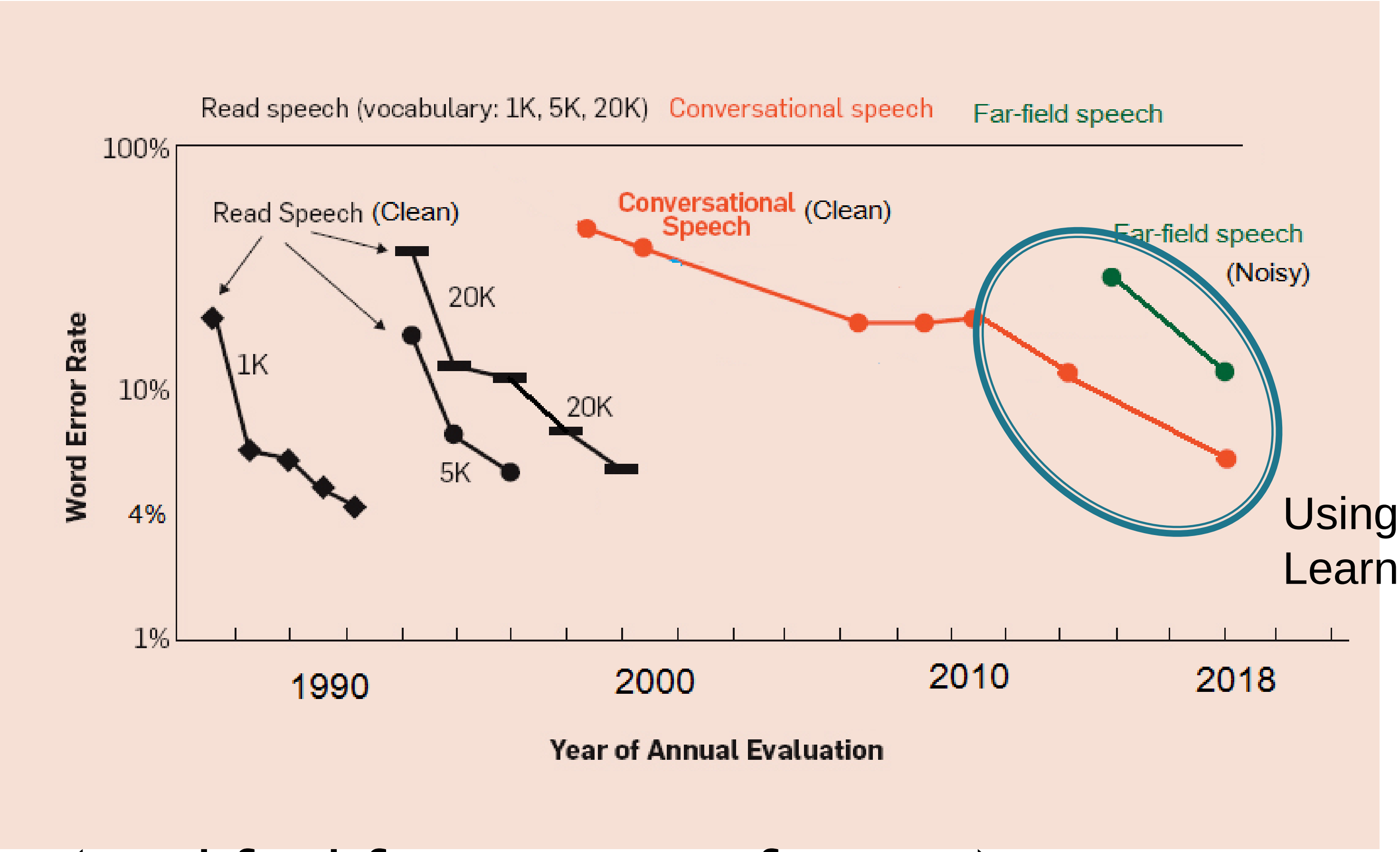
Supervised vs. unsupervised

- **Supervised learning**
  - The training data are labeled with their class.
- **Unsupervised learning**
  - The training data are unlabeled.

Generalization vs. overfitting

- Consider there are only 2 training utterances provided to a MT system
  - I want to eat something 我想吃東西
  - He wants to go to school 他想去學校
- After the training, does it know how to translate
  - He wants to eat something

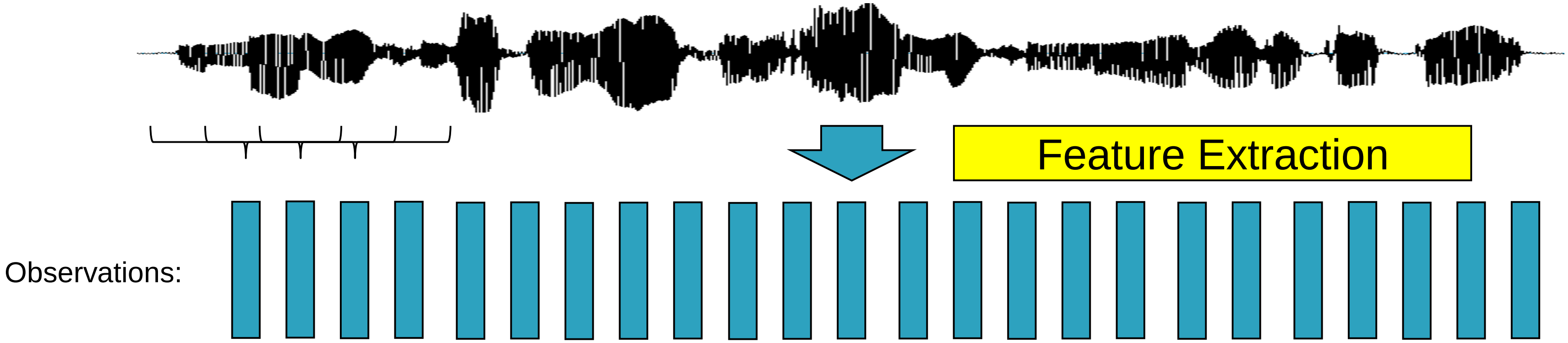
Historical Progress in ASR



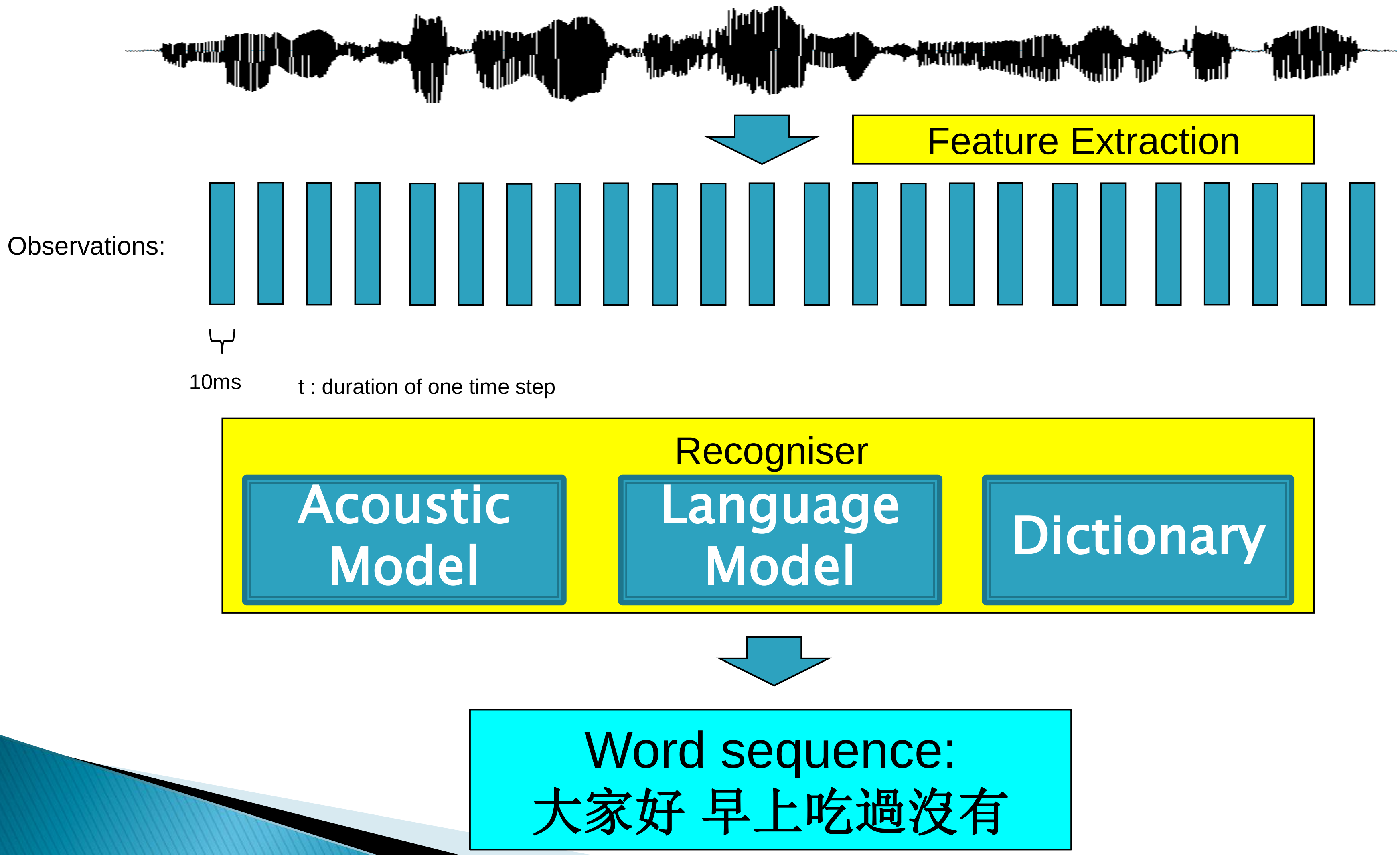
Using Deep Learning

(Modified from Microsoft News)

Overview of an ASR System



Overview of an ASR System



Feature Extraction

Observations:

10ms

t : duration of one time step

Recogniser

Acoustic Model

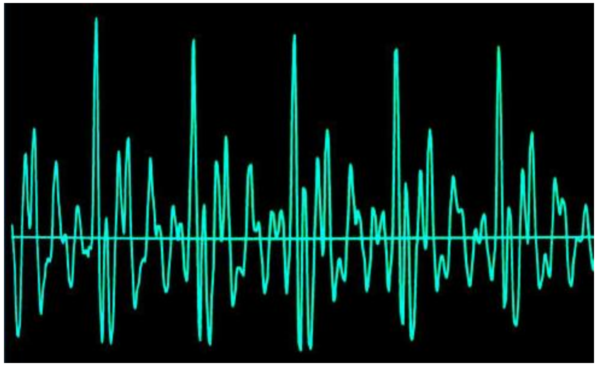
Language Model

Dictionary

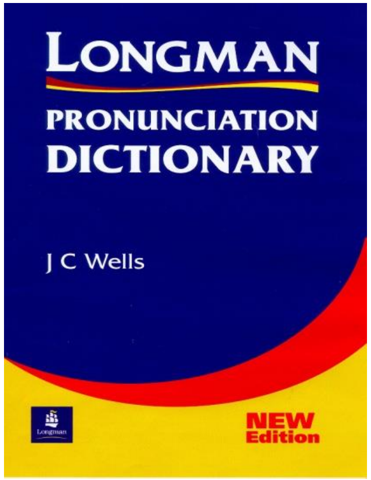
Word sequence:
大家好 早上吃過沒有

Components in an ASR system

Acoustic model



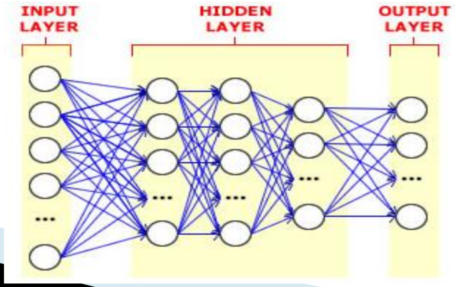
Dictionary



Language model

in order to

守株待兔

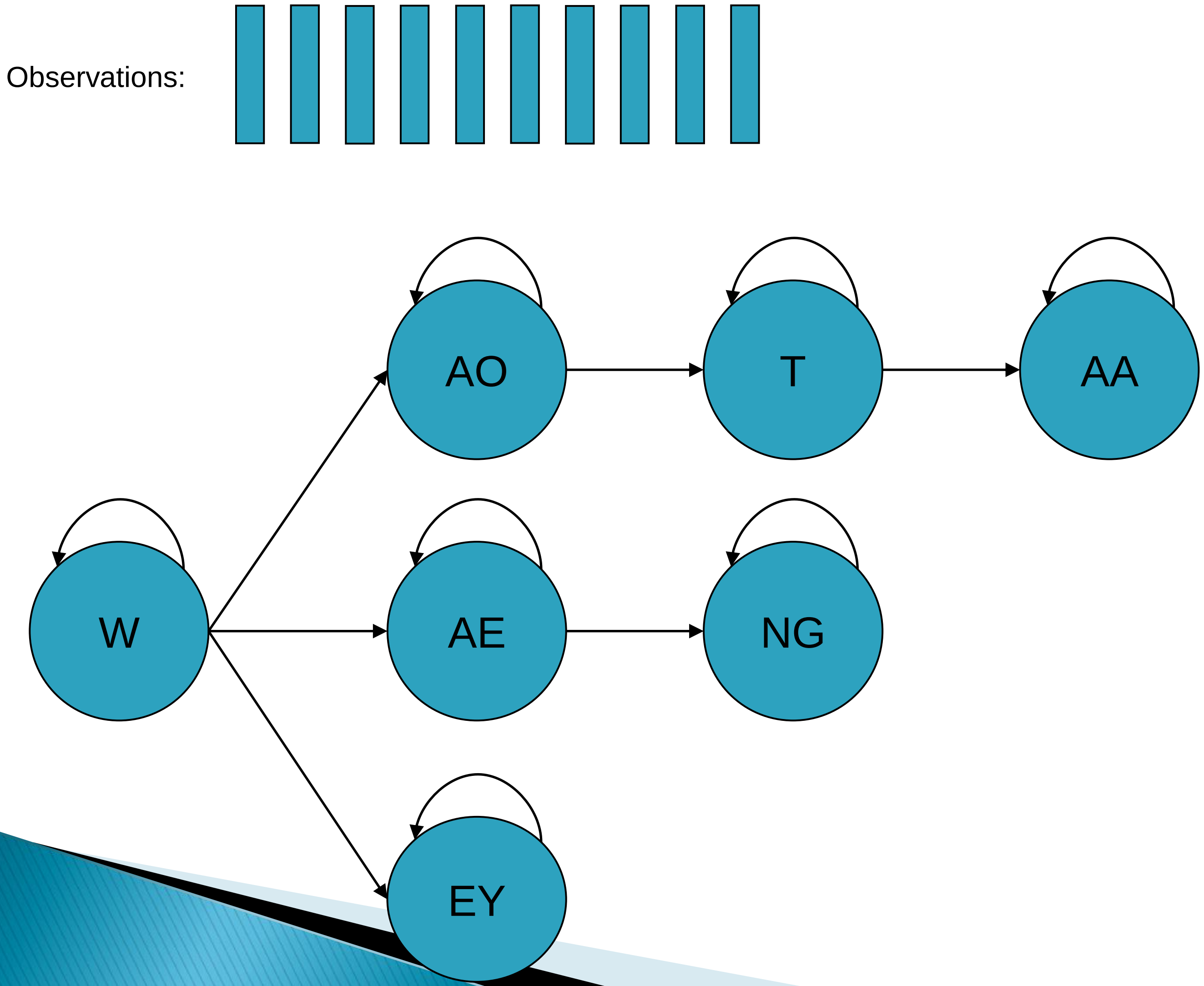


Dictionary and Phonemes

Dictionary	
Character / Word	Phonetic Transcription
我	*W AO*
你	*N IY*
他	*T AA*
早安	*Z AW AE N*

- Every language has its own set of phonemes
- Can’t pronounce “Sir” with Chinese phonemes

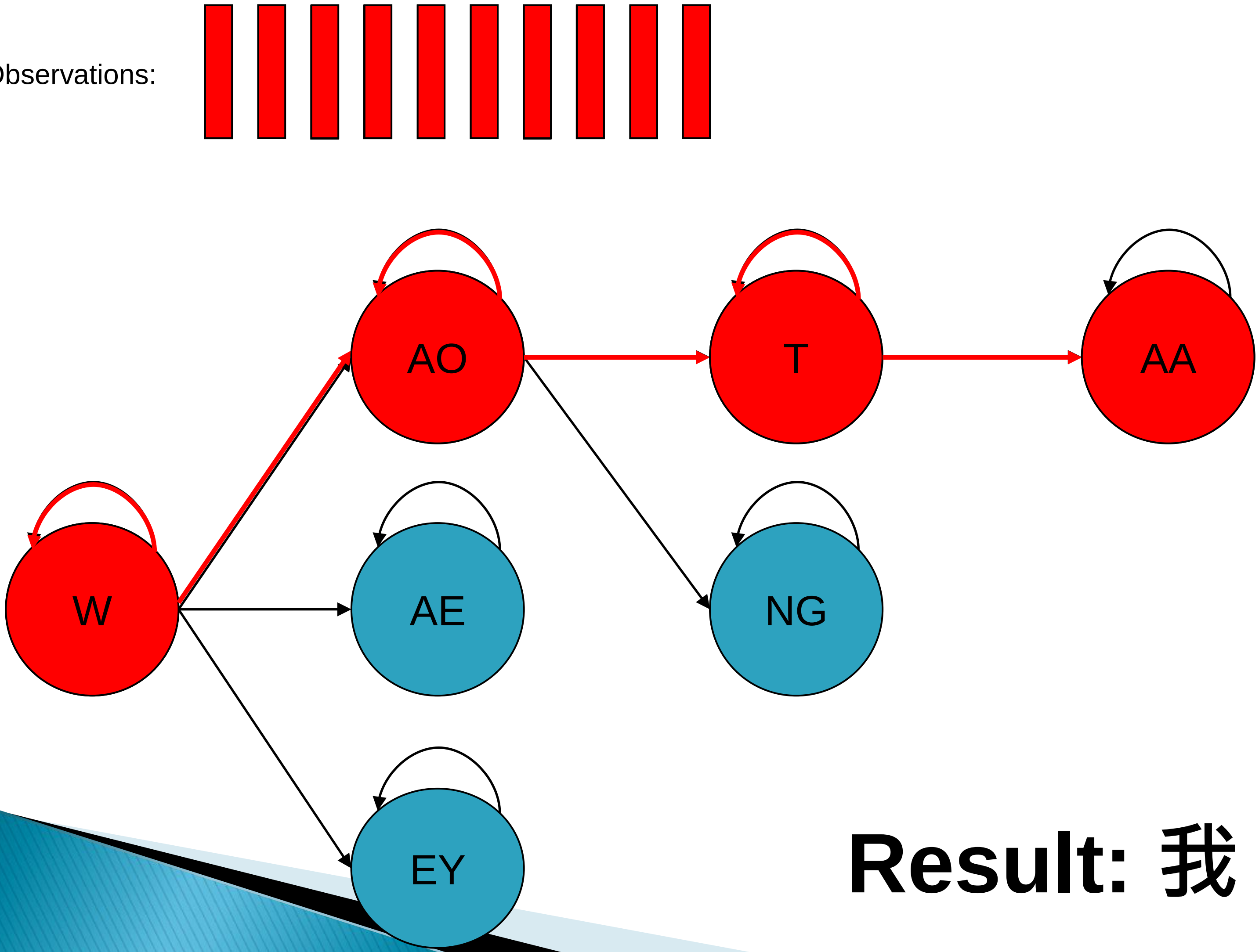
Hidden Markov Model (HMM)



Dictionary	
我	*W AO*
王	*W AE NG*
為	*W EY*
他	*T AA*
位	*W EY*

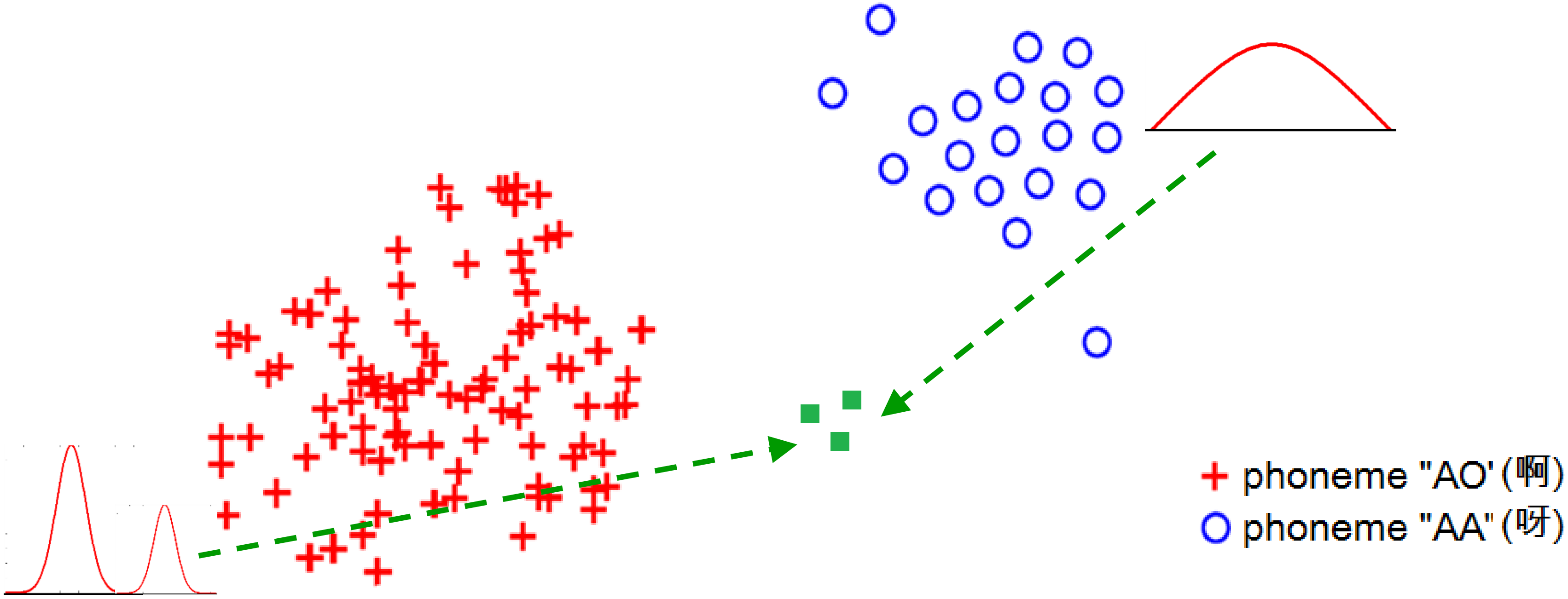
Hidden Markov Model (HMM)

Observations:

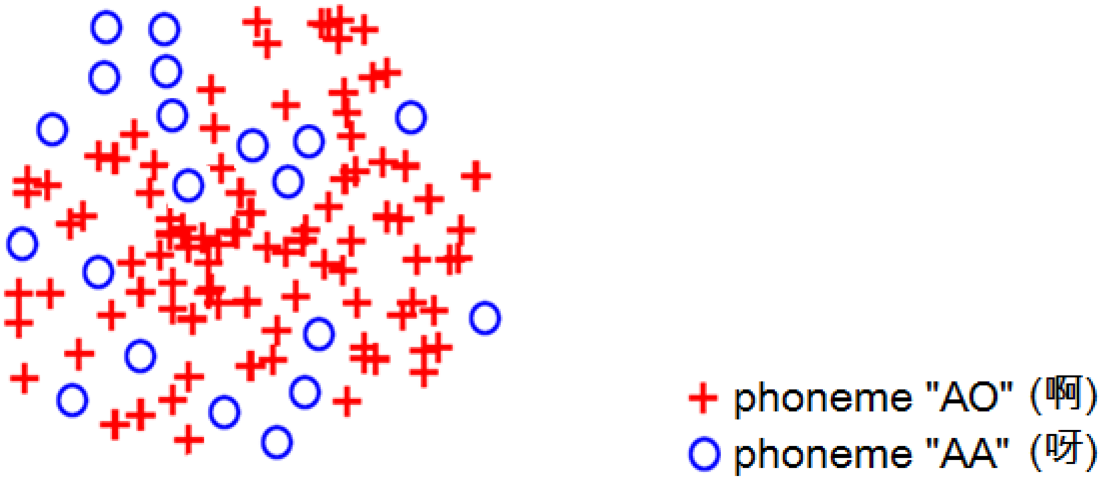


Result: 我 他

Conventional Way of Acoustic Modeling: Gaussian Mixture Modeling



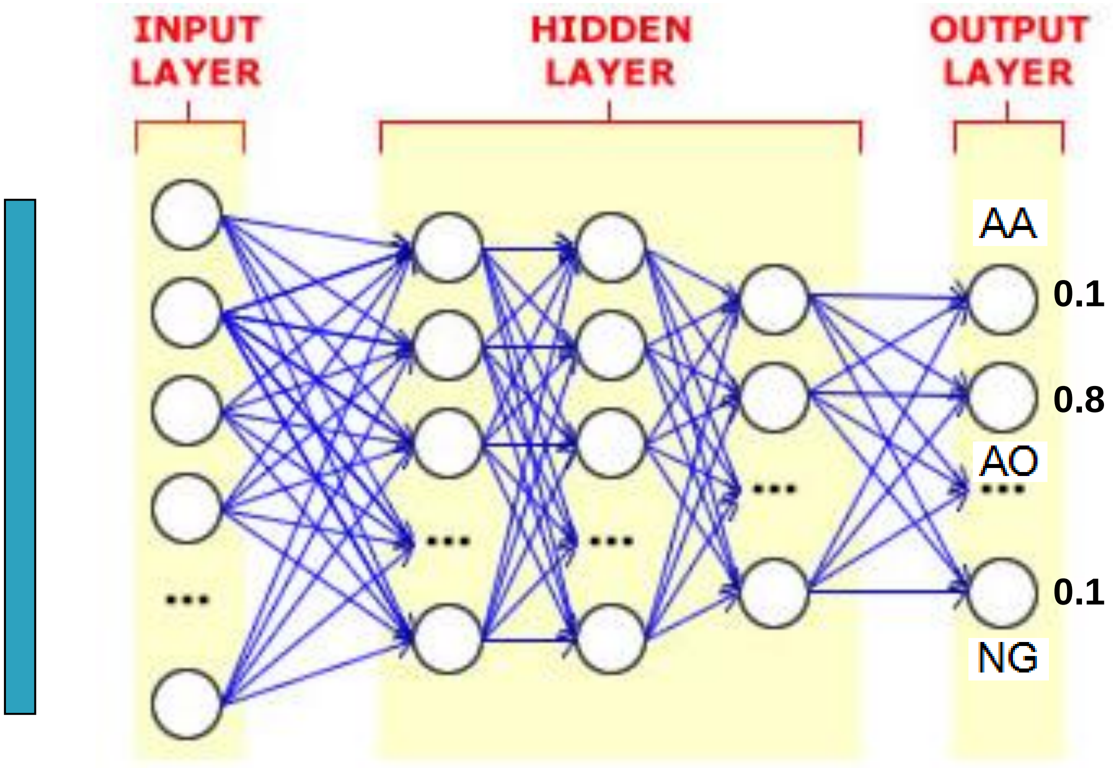
Difficult Cases



- A lot of confusion, resulting in recognition errors.

Modeling with Deep Neural Network

Observations:

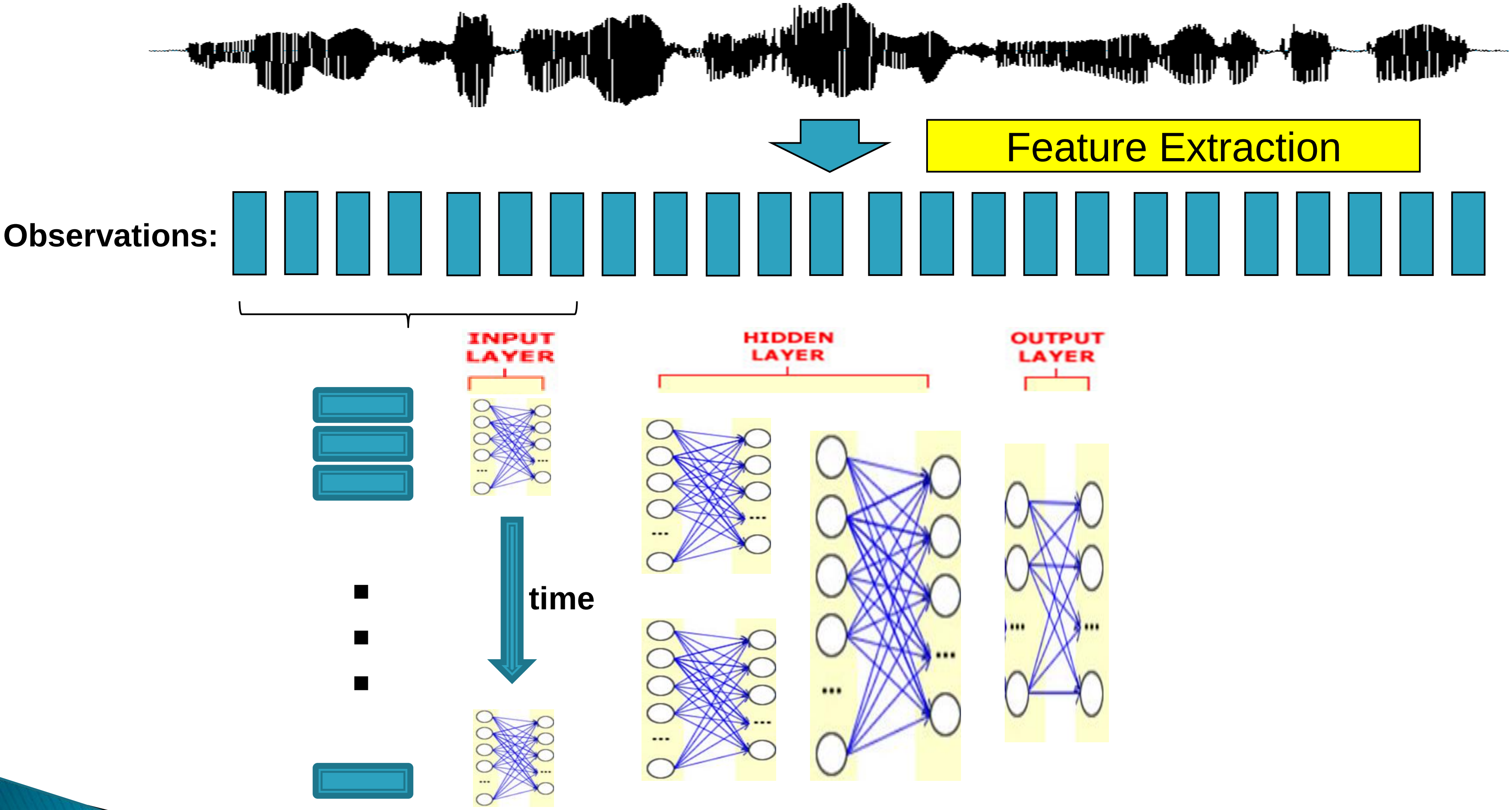


Common Choices of Acoustic Model

- Recurrent neural network
  - Long short term memory (LSTM)

- Non-recurrent neural network
  - Convolutional neural network (CNN)
  - Time-delay neural network (TDNN)

Time-delay DNN (TDNN)



- As good as RNN in modeling long range context dependencies but having shorter training time

State-of-the-art ASR performance

- Two major type of speech: Read speech and Conversational speech.

Speech Type	Vocab size	WER
Read	*5k chinese words*	*<3%*
Read	*20k chinese words*	*<5%*
Read (noisy)	*50k-100k chinese words*	*<10%*
Conversational	*50k-200k chinese words*	*<15%*
Conversational (noisy)	*50k-200k chinese words*	*<25%*

- The above figures assume that you have enough training data and under a close talking scenario.

Machine Translation

- To reverse the curse of Babel (Bible, Genesis 11:1-9)

Why is MT so hard?

- Typology
  - It means systematic cross-linguistic similarities and differences
  - Morphological difference
    - Number of morphemes per word
    - Whether the morphemes have clean boundaries
  - Structural difference
    - SVO (Subject-Verb-Object) languages: English, Mandarin, French
    - SOV (Subject-Object-Verb) languages: Japanese
    - VSO (Verb-Subject-Object) languages: Arabic, Hebrew

Why is MT so hard?

- Lexical divergences
  - In English, the word *bass* can mean a kind of fish or a kind of music instrument. For other languages, they are usually represented by different words.
  - *I know the answer* vs. *I know John*
  - Lexical gap
    - Japanese does not have a word for *privacy*
    - English does not have a word for *簫*

Rule-based MT (Classical MT)

- It relies on countless built-in linguistic rules and millions of bilingual dictionaries for each language pair.
- Need to be familiar with both languages (the source and the target)

Rule-based MT approaches

- Direct approach
  - Chinese: 守 株 待 兔
  - English: defend the tree and wait for a rabbit
- Transfer approach
  - To overcome the structural differences.
  - English: waiting for a rabbit under a tree
- Interlingua approach
  - English: a lazy living style

Jokes in MT

- While I am watching a movie:
  - 阿拉丁: 只是有時候, 我覺得我....
  - 公主: 被困
  - 公主: 就像你無法逃避你的出生
  - 阿拉丁: 對
  - 公主:隱性馬可夫模型
  - Hmm => 隱性馬可夫模型

Statistical MT

- Learn from the training data.
- It provides good quality when large and qualified corpora are available.

Speech technology

- Speech technology is a mixture of
  - Probability and Statistics
  - Signal Processing
  - Linguistic
  - Pattern Classification
  - Machine Learning
  - Artificial Intelligence
  - Deep Learning