

# Course Overview & Introduction

COMP90042

Natural Language Processing

Lecture 1

Semester 1 2022 Week 1  
Jey Han Lau



# Prerequisites

- COMP90049 “Introduction to Machine Learning” or COMP30027 “Machine Learning”
  - Modules → Welcome → Machine Learning Readings
- Python programming experience
- No knowledge of linguistics or advanced mathematics is assumed
- Caveats – Not “vanilla” computer science
  - Involves some basic **linguistics**, e.g., syntax and morphology
  - Requires **maths**, e.g., algebra, optimisation, linear algebra, dynamic programming

# Expectations and outcomes

- Expectations
  - develop Python skills
  - keep up with readings
  - lecture/discussion board participation
- Outcomes
  - Practical familiarity with range of text analysis technologies
  - Understanding of theoretical models underlying these tools
  - Competence in reading research literature

# Assessment

- **Assignments** (25% total for 3 activities)
  - 2 programming exercises
    - Released in week 4 and 5; 1 week to complete
  - 1 peer review of project report
    - Released in week 11; 1.5 week to complete
- **Group Project** (35%)
  - Released in week 7; 5 weeks to complete
- **Exam** (40%)
  - 2 hours, open book
  - Covers content from lectures, workshop and prescribed reading
- **Hurdle** >50% exam (20/40), and >50% for assignments + project (30/60)

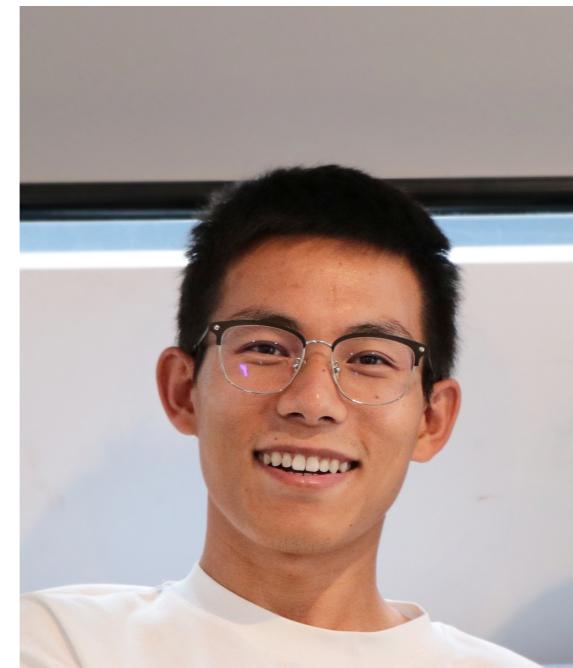
# Teaching Staff

Lecturer



Jey Han Lau

Head Tutor



Rongxin Zhu

# Tutors

- Zheng Wei Lim
- Chunhua Liu
- Yulia Otmakhova
- Viktoria Schram
- Gisela Vallejo
- Jun Wang

# Student Representative

- Keshav Prasath Asaithambi
  - ▶ [kprasath@student.unimelb.edu.au](mailto:kprasath@student.unimelb.edu.au)

# Recommended Texts

- Texts:
  - Jurafsky and Martin, [Speech and Language Processing](#), 3<sup>rd</sup> ed., Prentice Hall. draft
  - Eisenstein; [Natural Language Processing](#), Draft 15/10/18
  - Goldberg; [A Primer on Neural Network Models for Natural Language Processing](#)
- Recommended for learning python:
  - Steven Bird, Ewan Klein and Edward Loper, [Natural Language Processing with Python](#), O'Reilly, 2009

# Contact hours

- **Lectures**
  - Mon 15:15-16:15 Old Arts Public Lecture Theatre / Online
  - Tue 13:00-14:00 Old Arts Public Lecture Theatre / Online
- **Workshops: several across the week**
  - Worksheets & programming exercises
- **Method of contact — ask questions on the Canvas discussion board**

# Hybrid Lectures

- Trialing hybrid lectures for the first few weeks
  - Module > How to Join Livestreamed Lectures
  - Use the chat to ask questions!

≡ COMP90042\_2022\_SM1

2022 Semester 1

Home

Subject Overview

Modules

Assignments

Discussions

Grades

Lecture Capture

Zoom

Chat

Natural Language Processing (COMP90042\_2022\_SM1) ↗

COMP90042

Natural Language Processing



Dear Students,

Welcome to COMP90042, 2022 semester 1. We hope that together we can have a journey of discovery and wonder throughout this semester.

Some notes before week 1:

- We will have two 1 hour lectures every Monday from 15:15-16:15 and Tuesday from 13:00-14:00. These lectures will be held at the [Old Arts Public Lecture Theatre](#) ↗ and also live-streamed via [Lecture Capture](#) and I'd encourage all students to attend. Recordings of these lectures will also be posted on

# Hybrid Lectures

- Gauge interest, participation rate and feasibility
- Preliminary version (v1) of lecture slides have been published (**Modules > Lectures > Slides**)
- Lecture slides may be updated from time to time
- Lecture recordings will be available after each lecture

# Discord

- <https://discord.gg/jUGaWZr9Xt>
- An informal space for COMP90042 students
- Not actively monitored by teaching team →  
**do not** use it to ask content-related questions
- Make friends! You'll need them for the group project

# Python

- Making extensive use of python
  - workshops feature programming challenges
  - provided as interactive ‘notebooks’
    - Modules → Using Jupyter Notebook and Python
  - assignment and project in python
- Using several great python libraries
  - NLTK (basic text processing)
  - Numpy, Scipy, Matplotlib (maths, plotting)
  - Scikit-Learn (machine learning tools)
  - keras, pytorch (deep learning)

# Python

- New to Python?
  - Expected to pick this up during the subject, on your own time
  - Learning resources on worksheet

# Natural Language Processing

- Interdisciplinary study that involves linguistics, computer science and artificial intelligence.
- Aim of the study is to understand how to design algorithms to process and analyse human language data.

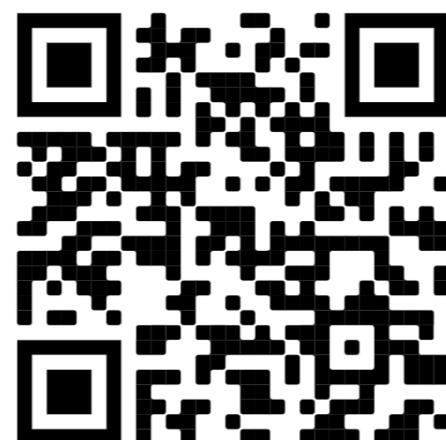
# Why process text?

- Masses of information ‘trapped’ in unstructured text
- How can we find or analyse this information?
- Let computers automatically reason over this data?
- First need to understand the structure, find important elements and relations, etc...
- Over 1000s of languages....

# Language Generation Demo

# Why are you interested in NLP?

[PollEv.com/jeyhanlau569](https://PollEv.com/jeyhanlau569)



# Motivating Applications (Sci-fi)

- Intelligent conversational agent, e.g. TARS in Interstellar (2014)
  - ▶ <https://www.youtube.com/watch?v=wVEfFHzUby0>
  - ▶ Speech recognition
  - ▶ Natural language understanding
  - ▶ Speech synthesis

# Motivating Applications (Real-world)

- IBM ‘Watson’ system for Question Answering
  - QA over large text collections
    - Incorporating information extraction, and more
  - [https://www.youtube.com/watch?v=II-M7O\\_bRNg](https://www.youtube.com/watch?v=II-M7O_bRNg)
- Research behind Watson is *not* revolutionary
  - But this is a transformative result in the history of AI
  - Combines cutting-edge text processing components with large text collections and high performance computing

English – detected ▾ ↔ Chinese (Simplified) ▾

Today we are having a lecture on natural language processing × 今天我们要进行自然语言处理的讲座 Jīntiān wǒmen yào jīnxíng zìrán yǔyán chǔlǐ de jiǎngzuò

Feedback

[Open in Google Translate](#)

🔍 google translate|

🔍 google translate english to spanish

🔍 google translate audio

🔍 google translate english to french

🔍 google translate website

🔍 google translate statistics

🔍 translate to hindi

🔍 translate to english

🔍 inside google translate

who is the first australian prime minister

All News Images Videos Maps More Settings Tools

About 78,100,000 results (1.18 seconds)

Prime Minister of Australia (1)

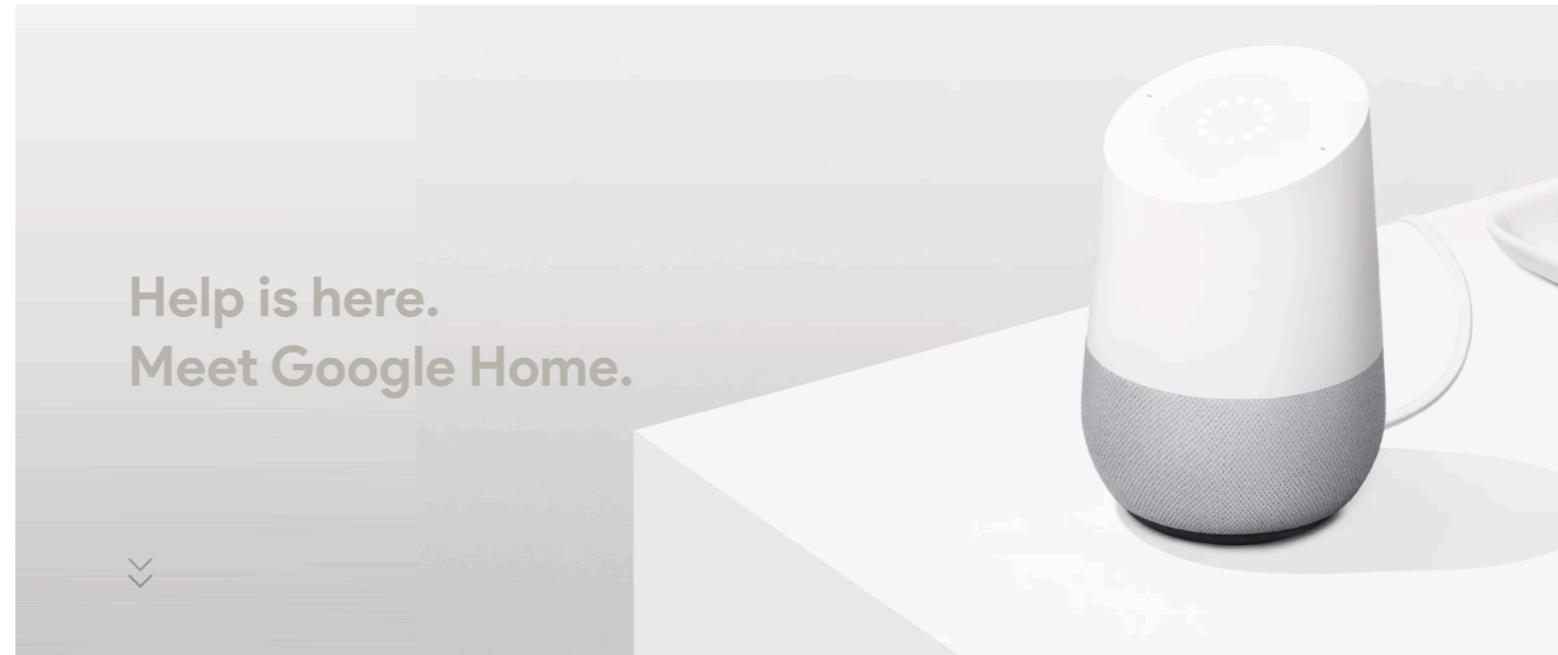
Edmund Barton



Australia's first prime minister, **Edmund Barton** at the central table in the House of Representatives in 1901.

[en.wikipedia.org/wiki/Prime\\_Minister\\_of\\_Australia](https://en.wikipedia.org/wiki/Prime_Minister_of_Australia)

[Prime Minister of Australia - Wikipedia](#)



Help is here.  
Meet Google Home.

# Course Overview

- **Word, sequences, and documents**
  - Text preprocessing
  - Language models
  - Text classification
- **Structure learning**
  - Sequence tagging (e.g. part-of-speech)
- **Deep learning for NLP**
  - Feedforward and recurrent models

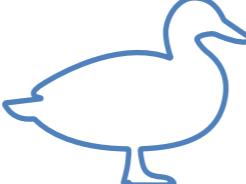
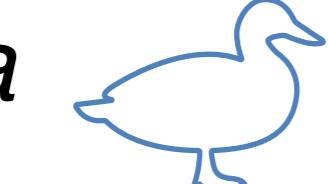
# Course Overview

- **Semantics**
  - How words form meaning
- **Syntax**
  - How words are arranged
- **Applications**
  - Machine translation
  - Information extraction
  - Question answering

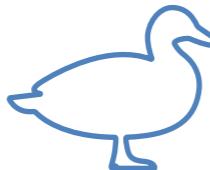
# Are Machines Intelligent Yet?

- Alan Turing, famously proposed the **Turing test**, to assess whether a machine is intelligent
- Alan Turing predicted in 1950 that by 2000 a machine with 10 gigabytes of memory has 30% of fooling the human interrogator.
- The smartest conversational agent we have today are far away from being truly intelligent...

# Challenges of Language: Ambiguity

- *I made her duck:*
  - *I cooked  for her*
  - *I cooked  belonging to her*
  - *I caused her to quickly lower her head or body*
  - *I waved my magic wand and turned her into a *
- Why so many possible interpretations?

# Challenges of Language: Ambiguity

- *Duck* can mean:
  - Noun: 
  - Verb: move head or body quickly down (e.g. to dodge something)
- *Her* can be a dative pronoun (i.e. indirect object to a verb) or possessive pronoun
- *Make* is syntactically ambiguous:
  - Transitive (takes one object: *duck*)
  - Ditransitive (1st object: *her*; 2nd object: *duck*)
  - Can take a direct object and verb: object (*her*) is caused to perform the verbal action (*duck*)

What are other challenges that made language processing difficult?

[PollEv.com/jeyhanlau569](https://PollEv.com/jeyhanlau569)

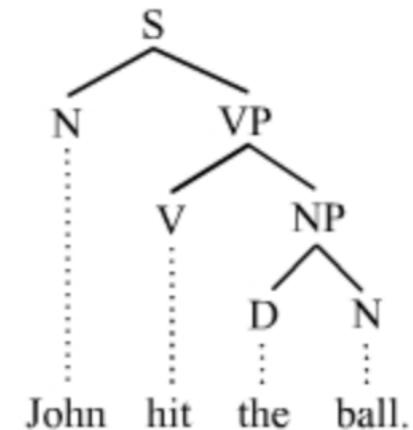


# A brief history of NLP: 1950s

- "Computing Machinery and Intelligence", Alan Turing
  - Turing test: measure machine intelligence via a conversational test
- "Syntactic Structures", Noam Chomsky
  - Formal language theory: uses algebra and set theory to define formal languages as sequences of symbols
  - *Colourless green ideas sleep furiously*
    - Sentence doesn't make sense
    - But its grammar is fine
    - Highlights the difference between semantics (meaning) and syntax (sentence structure)

# 1960-1970s

- Symbolic paradigm
  - Generative grammar
    - Discover a system of rules that generates grammatical sentences
  - Parsing algorithms
- Stochastic paradigm
  - Bayesian method for optical character recognition and authorship attribution
- First online corpus: Brown corpus of American English
  - 1 million words, 500 documents from different genres (news, novels, etc)



# 1970-1980s

- Stochastic paradigm
  - Hidden Markov models, noisy channel decoding
  - Speech recognition and synthesis
- Logic-based paradigm
  - More grammar systems (e.g. Lexical functional Grammar)
- Natural language understanding
  - Winograd's SHRDLU
  - Robot embedded in a toy blocks world
  - Program takes natural language commands (*move the red block to the left of the blue block*)
  - Motivates the field to study semantics and discourse

# 1980-1990s

- Finite-state machines
  - Phonology, morphology and syntax
- Return of empiricism
  - Probabilistic models developed by IBM for speech recognition
  - Inspired other data-driven approaches on part-of-speech tagging, parsing, and semantics
  - Empirical evaluation based on held-out data, quantitative metrics, and comparison with state-of-the-art

# 1990-2000s: Rise of Machine Learning

- Better computational power
- Gradual lessening of the dominance of Chomskyan theories of linguistics
- More language corpora developed
  - Penn Treebank, PropBank, RSTBank, etc
  - Corpora with various forms of syntactic, semantic and discourse annotations
- Better models adapted from the machine learning community: support vector machines, logistic regression

# 2000s: Deep Learning

- Emergence of very deep neural networks (i.e. networks with many many layers)
- Started from the computer vision community for image classification
- Advantage: uses raw data as input (e.g. just words and documents), without the need to develop hand-engineered features
- Computationally expensive: relies on GPU to scale for large models and training data
- Contributed to the AI wave we now experience:
  - Home assistants and chatbots

# Future of NLP

- Are NLP problems solved?
  - ▶ Machine translation still is far from perfect
  - ▶ NLP models still can't reason over text
  - ▶ Not quite close to passing the Turing Test
    - Amazon Alexa Prize: <https://www.youtube.com/watch?v=WTGuOg7GXYU>