CSC485/2501 Introduction to Computational Linguistics

Lecture 1

The Introduction

Announcement: Friday's Tutorial

- Shared PyTorch review tutorial with CSC401.
- Friday, Sep 6:
  - 10-11 BA1180.
  - 11-12 BA1190.
  - 12-13 ES B142.
  - Any session that works the best for you.

What is Computational Linguistics?

- The study that let computers understand **language**.
- Also known as Natural Language Process (NLP).

Nuclear bomb!

No that’s impossible.

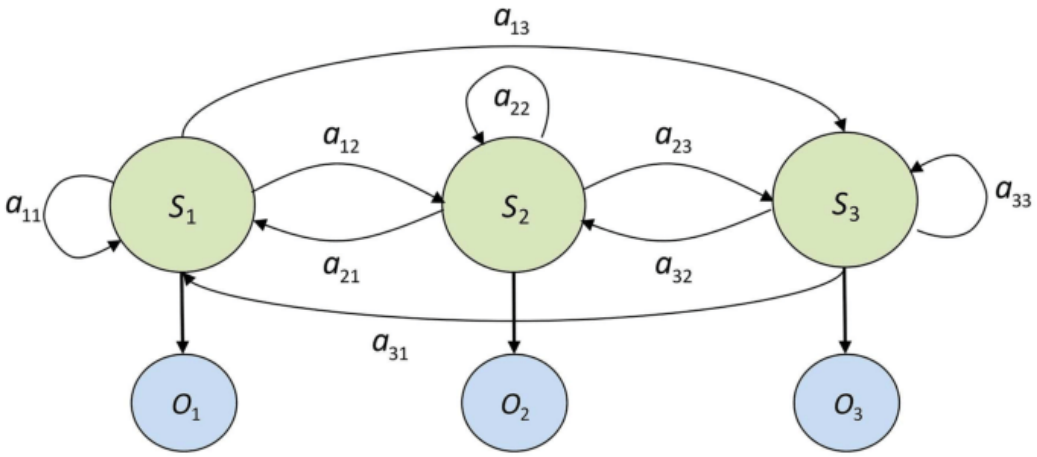
H bomb!

No that’s impossible.

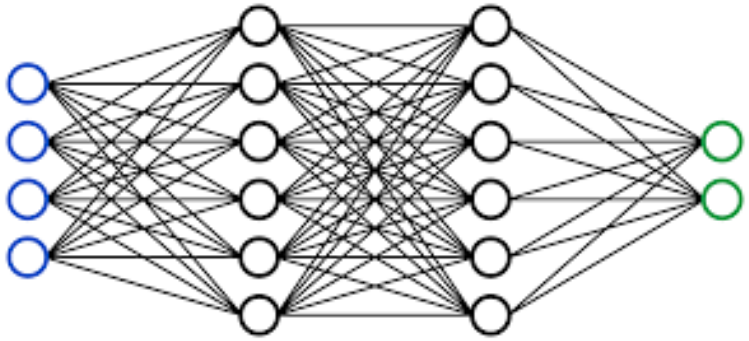
Symbolic vs.
Statistical

"I think we are forced to conclude that... probabilistic models give no particular insight into some of the basic problems of syntactic structure."

— *Syntactic Structures*. Chomsky (1957).



Hidden Markov Model

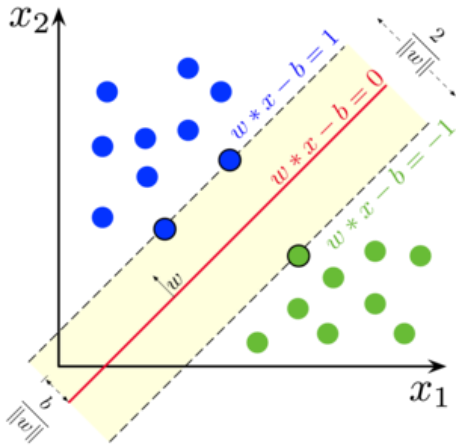


Neural Network

Random Number Generator*

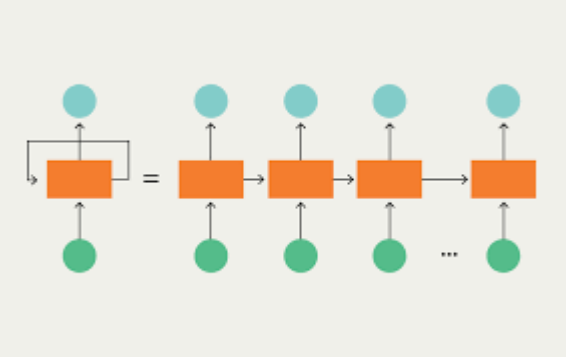
What Machine Learning Architecture?

Symbolic vs. Statistical

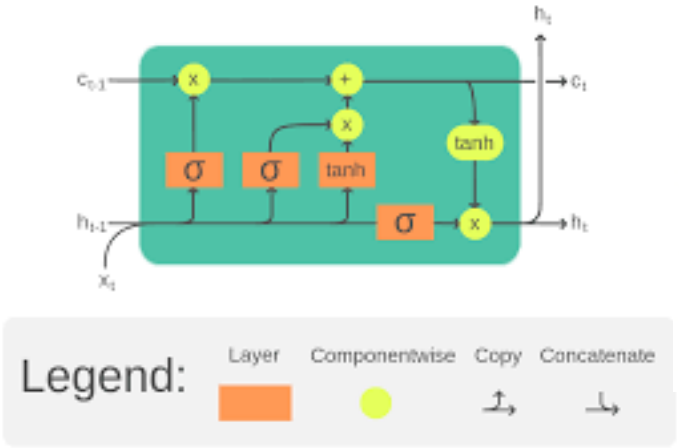


Support Vector Machine

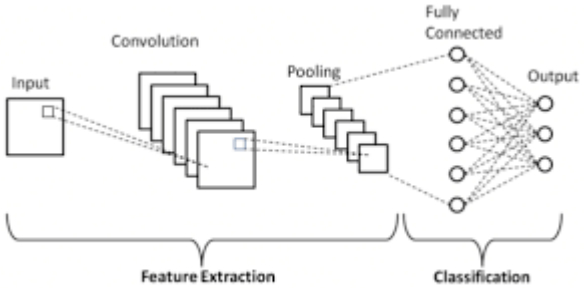
*Random Number Generator is not a real ML architecture.



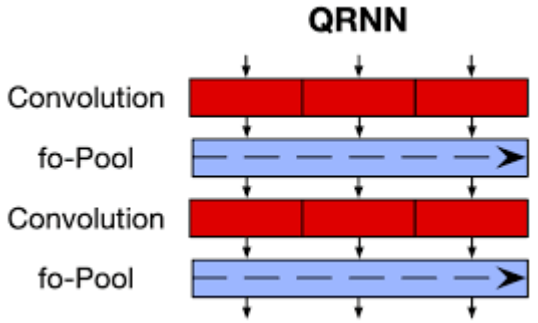
RNN



LSTM



CNN?

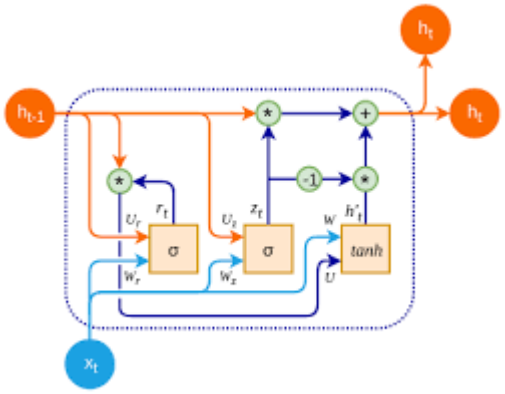


QRNN: CNN + RNN

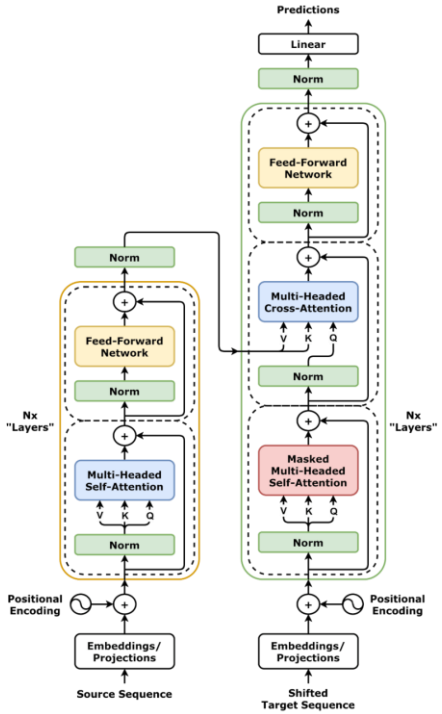
What LM Architecture?

What Machine Learning Architecture?

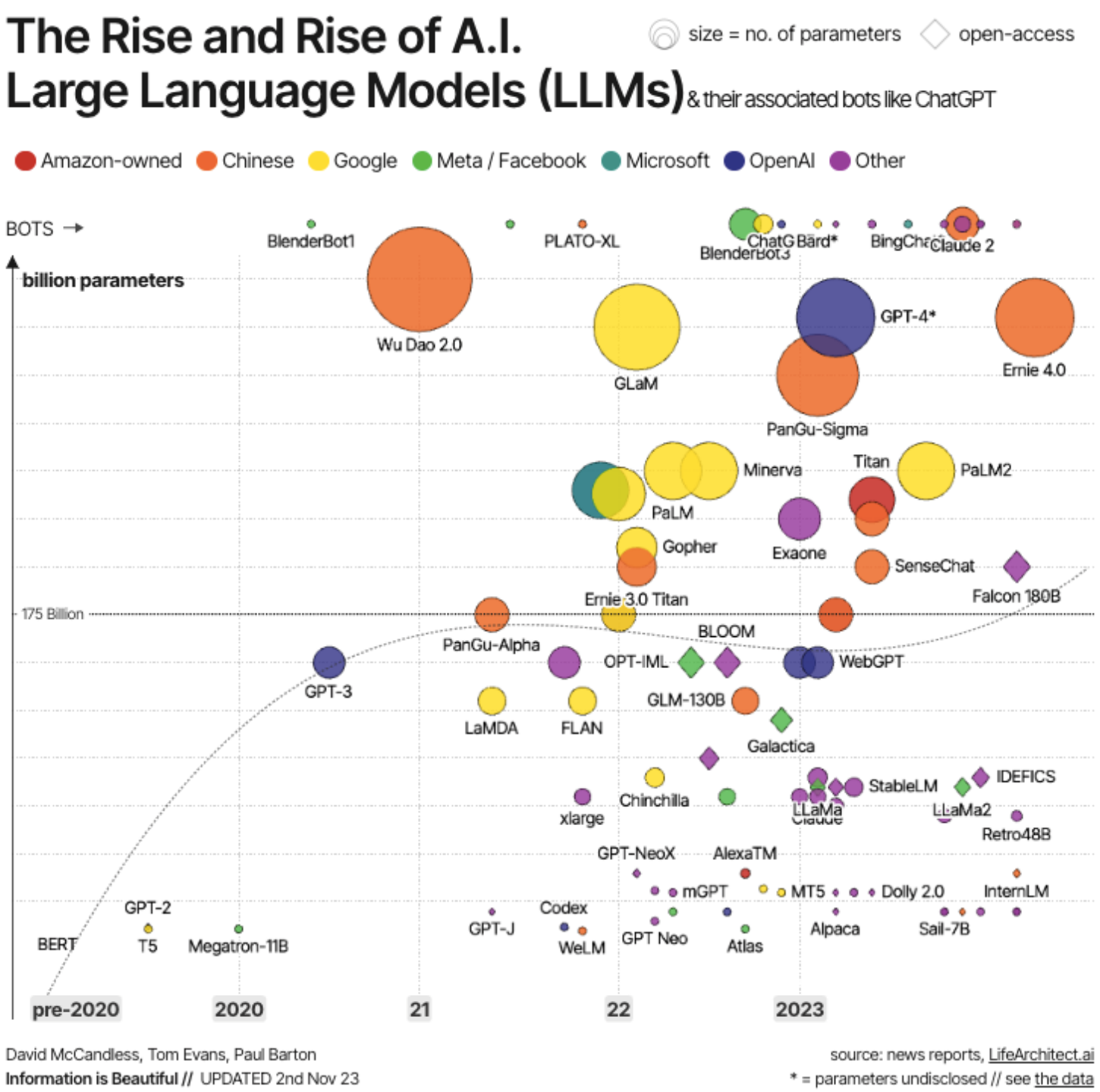
Symbolic vs. Statistical



GRU



👑 Transformers

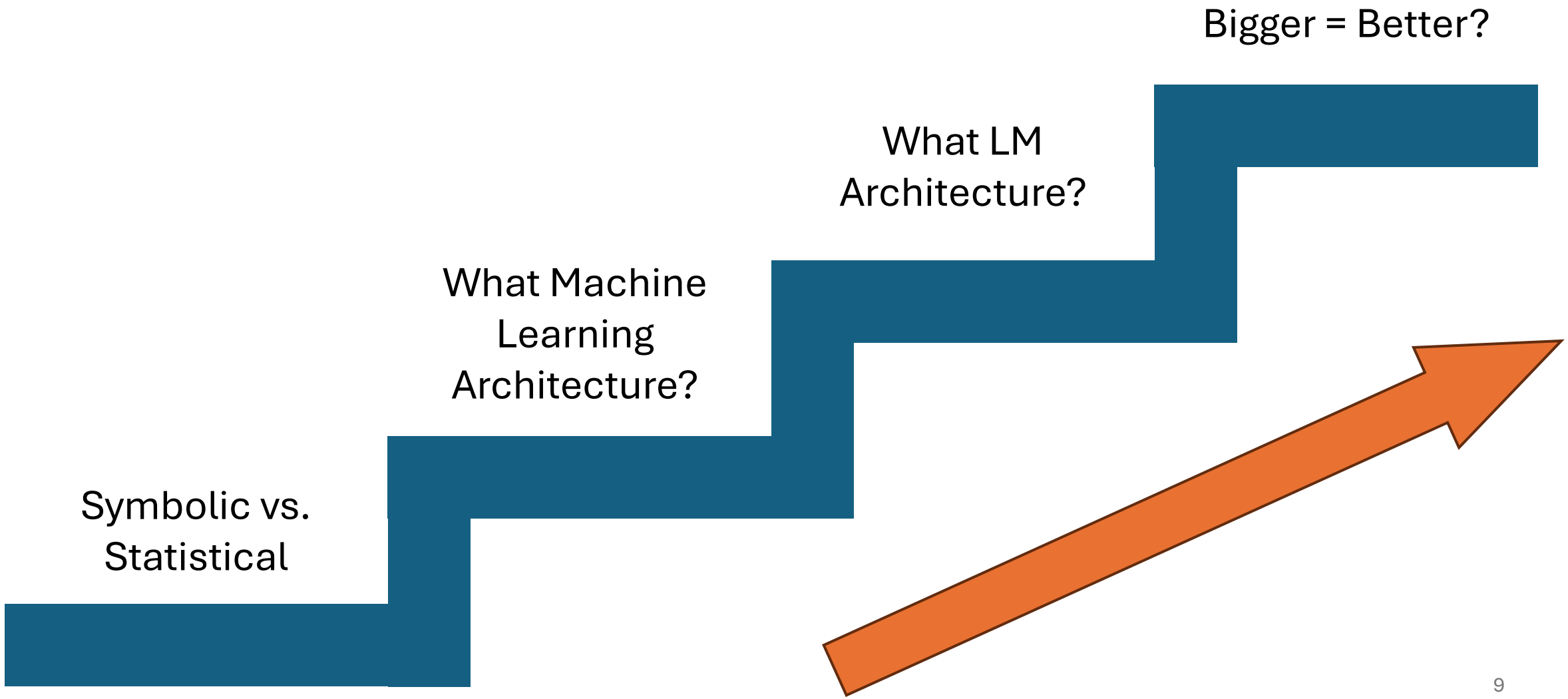


Bigger = Better?

at LM

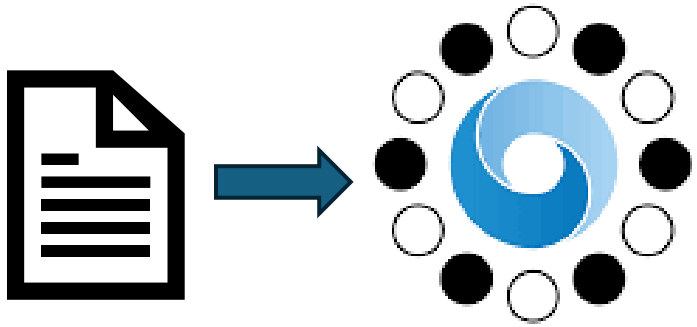
ecture?

What's Next?


Bigger = Better?
What LM Architecture?
What Machine Learning Architecture?
Symbolic vs. Statistical

AlphaGo

RL for natural language tasks?



- Hard to design reward.
  - Sparse
  - No clear objective
- Large search space.
- …
- RLHF: alignment but not problem solving.

Do we understand Human Language Processing?

- We still don’t know.
  - What is language.
  - What is a word.
  - What is a sentence.
  - Why human can speak language.
  - …

- Build better machine models of language from psycholinguistic inspirations.
- Not finding pseudo-psycholinguistic cues in these machine models.


May You
Live In
Interesting
Times
BIENNALE ARTE
2019
11.05—24.11
VENEZIA
GIARDINI/ARSENALE

- **Input:**
  - Spoken
  - Written
  - (Image)

- **Output:**
  - An action
  - A document or artifact
  - Some chosen text or speech
  - Some newly composed text or speech
  - …

Intelligent Language Processing Applications

- Document applications
- Searching for documents by meaning
- Summarizing documents
- Answering questions
- Extracting information
- Content and authorship analysis
- Helping language learners
- Helping people with disabilities
- …

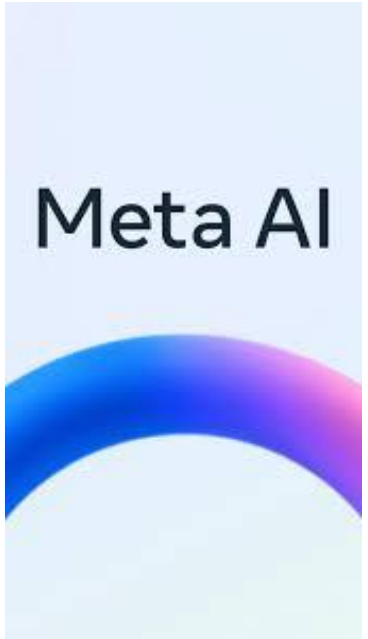

ChatGPT


Gemini


Claude


Grok


Copilot


Meta AI


文心一言


ChatGLM

Example: Early detection of Alzheimer's

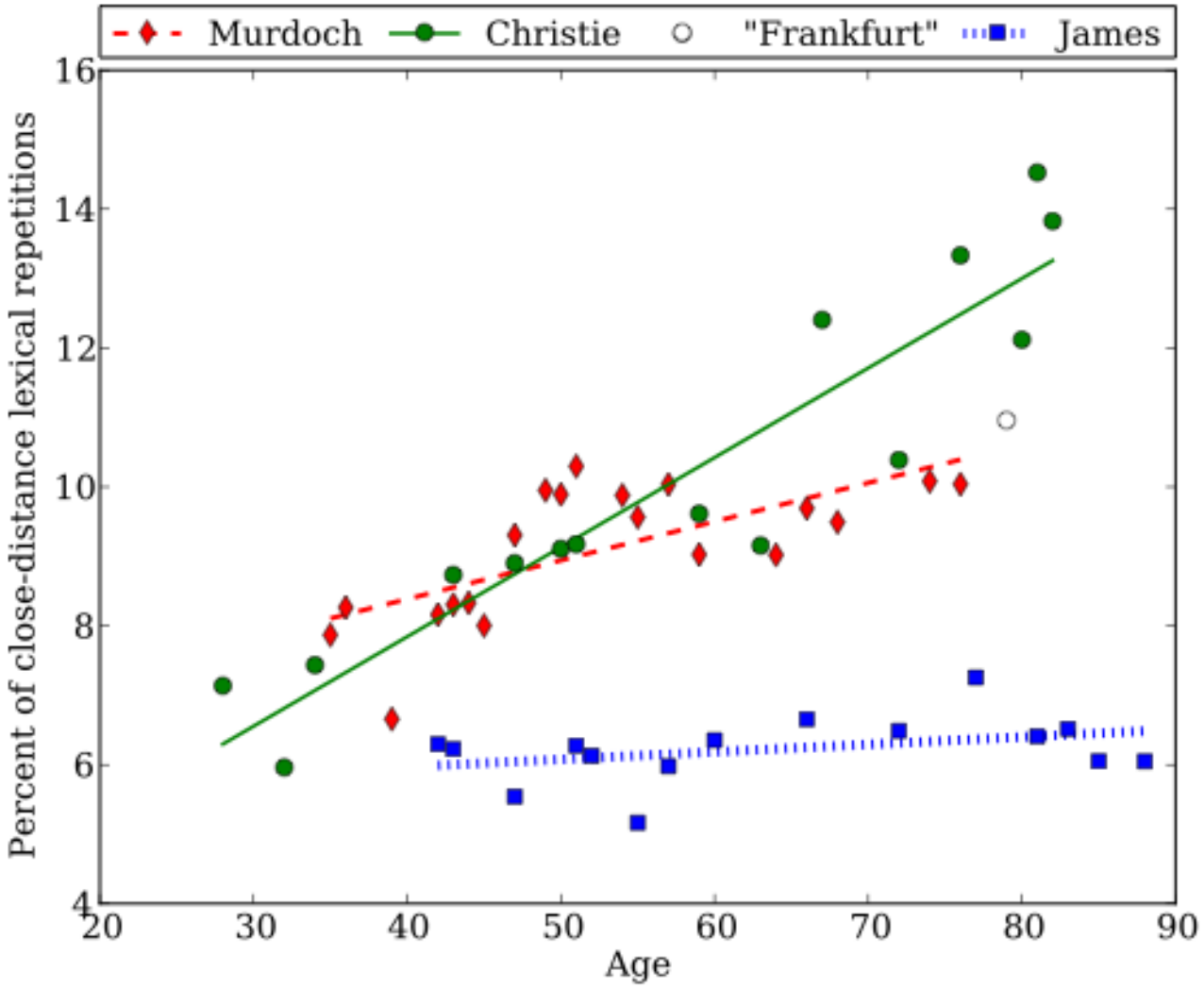
- Look for deterioration in complexity of vocabulary and syntax.
- Study: Compare three British writers

Iris Murdoch
Died of Alzheimer's

P.D. James
No Alzheimer's

Agatha Christie
Suspected Alzheimer's

Increase in short-distance word repetition



Rise, $p < .01$

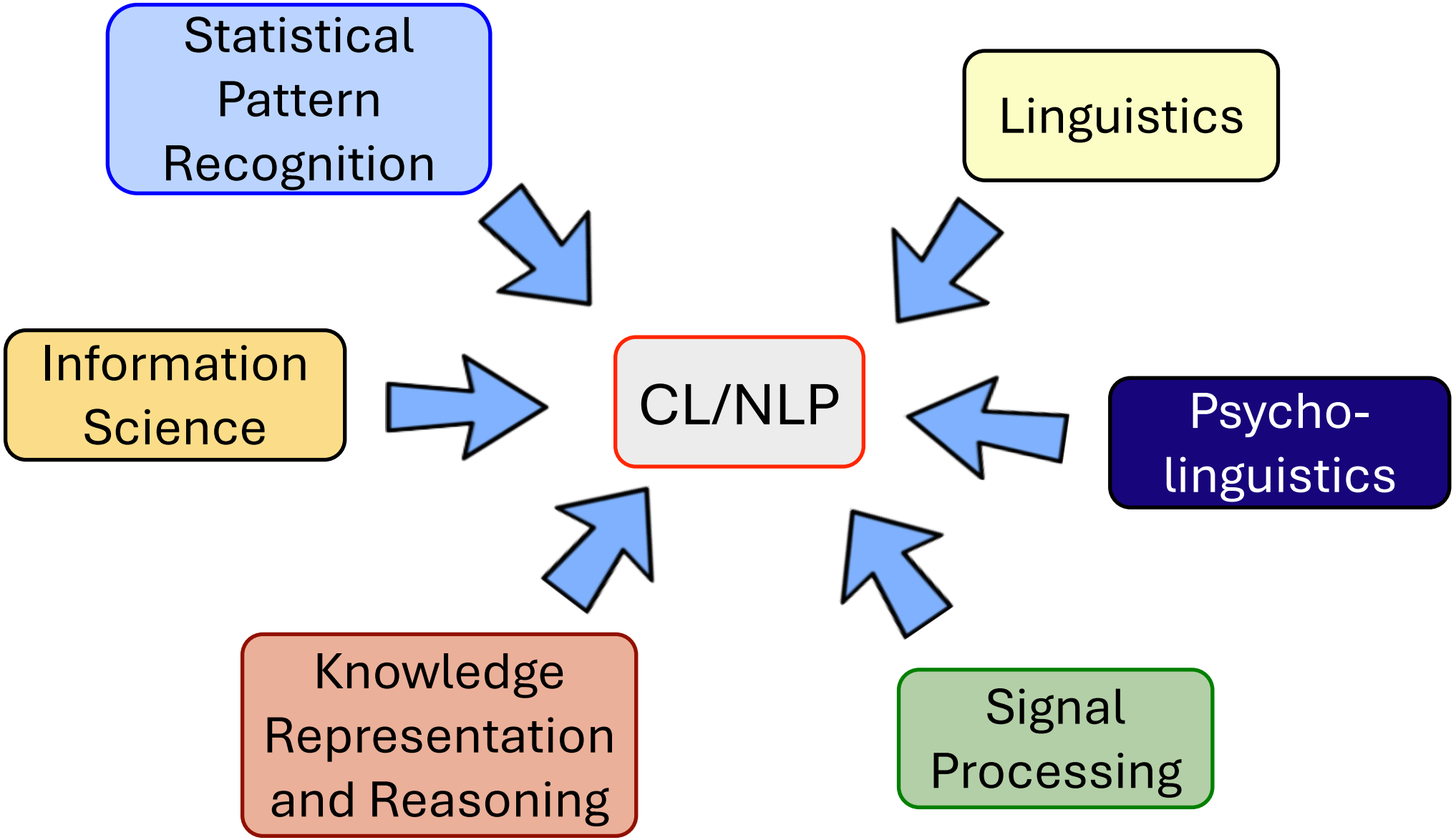
Iris Murdoch
Died of Alzheimer's

Rise, $p < .01$

Agatha Christie
Suspected Alzheimer's

n.s.

P.D. James
No Alzheimer's


Statistical Pattern Recognition
Linguistics
Information Science
CL/NLP
Psycho-linguistics
Knowledge Representation and Reasoning
Signal Processing

Focus of this Course

- Grammars & Parsing
- Understanding Language Models
- ChatGPT and Large Language Model
  - In-context learning (ICL)
  - Zero-shot learners (ZSL)
- Resolving Syntactic Ambiguity
- Determining Argument Structure
- Lexical Semantics, Word Sense
- Compositional Semantics
- Question Answering
- Understanding Pronoun Reference

Current Methods:

- Integrating statistical knowledge into grammars and parsing algorithms.
- Using text corpora as sources of linguistic knowledge.
- Interpreting and understanding the mechanisms that drive large language models.

Not Included

- NLP in applications: 
  - information retrieval (IR) and extraction (IE), intelligent web searching, and machine translation (MT)…
- Engineering details of Transformers and LLMs: 
  - Decoder LM, Prompt Engineering, RLHF, and PEFT…
- Graph-theoretic and spectral methods. 
- Speech recognition and synthesis. 
- Cognitively based methods. 
- Semantic inference,  semantic change/drift. 
- Understanding dialogues and conversations. 
- Bias, “fake news” detection, ethics in NLP. 





CSC 401/2511	CSC 2517	CSC 2518	CSC 2540	CSC 2611

Course Organization

Item	CSC 485	CSC 2501
A1: Dependency Parsing	30%	25%
A2: Word Sense Disambiguation and Language Models Interpretation	30%	25%
A3: Building Grammar Rules: Symbolic Machine Translation	30%	25%
Quizzes (one per week)	10%	10%
Course Content Survey	1% bonus	1% bonus
5 Essays (paper review, CSC2501 only)		3% x 5 = 15%

Quiz

- One quiz per week (13 quizzes in total).
- Appear in random places in lectures.
- Best 10 quizzes.
- Marks:
  - 1: correct and on time.
  - 0.5: on time.
- Quercus.
- Due Friday 5pm.

Course Content Survey

- Let me know what sub-topics you want me to cover in class.
  - Model editing, unsupervised parsing…
- Any feedback any time.
- If finished before reading week: 1 free bonus mark!

QUERCUS
UNIVERSITY OF TORONTO

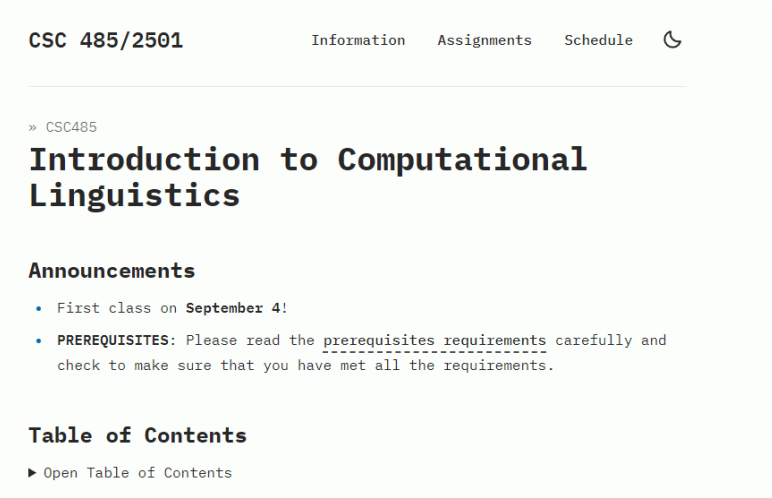
- Quizzes.
- Essays (CSC 2501).
- Non-public material.

MarkUs

- Assignments.

piazza

- Q&A with me and the TAs.
- Discussion.
- Assignment clarification.

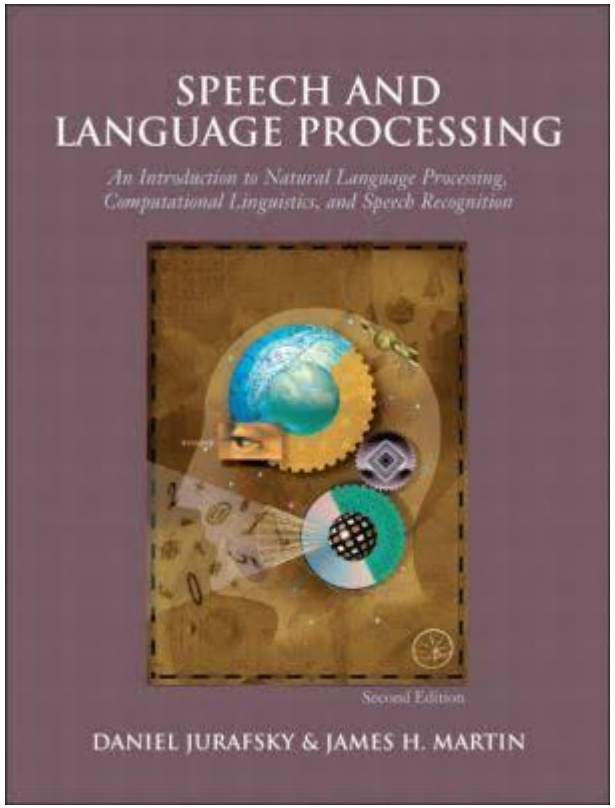


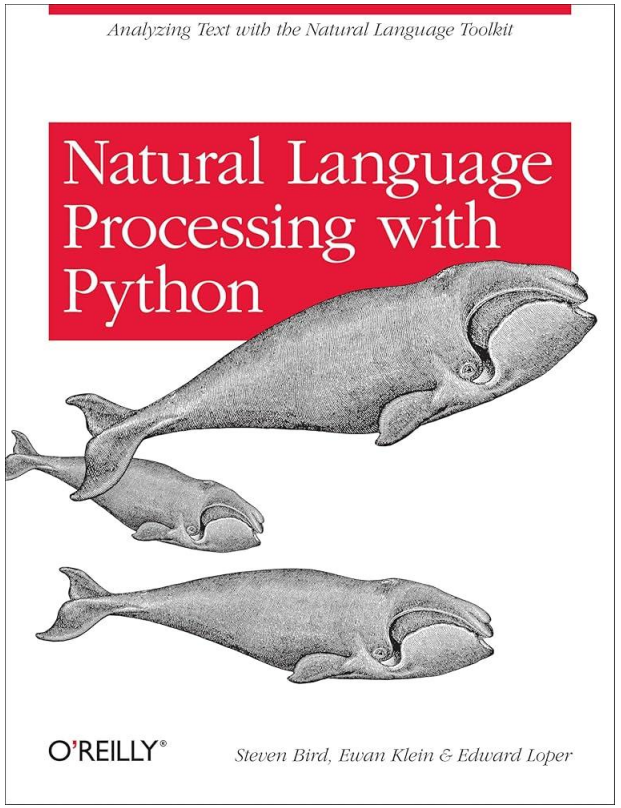
Course Website:

- Course Information.
- Slides, materials...

Quercus:
https://q.utoronto.ca/courses/354460
MarkUs:
https://markus.teach.cs.toronto.edu/markus/courses/21
Piazza:
https://piazza.com/class/lyd9xmcoghi3as
Course Website:
https://www.cs.toronto.edu/~niu/csc485/

Textbooks





- Speech and Language Processing.
  - Jurafsky and Martin.
  - Edition 3 (and edition 2).
- Natural Language Processing with Python.
  - Bird, Klein and Loper

Free versions online for both books.

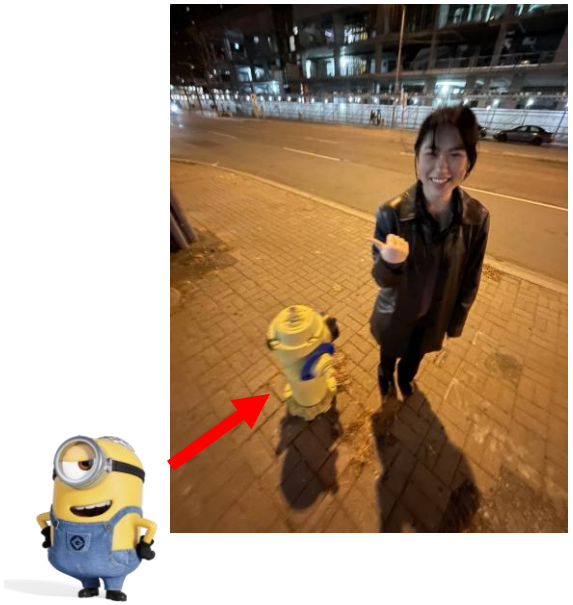
Online Resources & AI Writing Assistance

- Do NOT post any assignments online.
- Do NOT use any code generated by any AI assistance
  - ChatGPT, Copilot, etc.
- The work you submit **must be your own**.

- AI writing assistance is allowed only for refining the English grammar and/or spelling of text that you have already written.

Important Dates

- Office Hours:
  - Where? Here.
  - When? 11-12, between the two lectures.
- No late submission allowed.
  - Except in case of documented medical or other emergencies.
- Assignments: Due on 5pm, A1/A2: Thursday, A3: Wednesday.
- Quizzes: Friday 5pm.
- Essays: Monday before class. E3 due on a Wednesday.
- Remark Requests: submit within 7 days after the release of your mark.

	Zixing Zhao	Gavin (Yushi) Guan	Jinyue Feng
			
Research Area	Human-Computer Interaction, Applications of AI for stories.	Machine Learning and Systems (What my group does). Database for 3D Vision/Language for 3D Asset Retrieval (What I actually do).	Causal Reasoning in Language Models; Model Interpretability
Interest	sewing, plants.	Snowboarding, cleaning up my cats' hair at home.	Barista Skills; Board Games.
Fun Fact!	Terrible at taking pictures for other people (always shaky and blurry).	Have been building 2500km mileage on Toronto Bikeshare for the last year!	My cat eats rice.

	Bindu Dash	Devan Srinivasan	Jinman Zhao
			
Research Area	Social Reading Technology, Application of AI for stories.	TBD; prospective new researcher!	Grammar, syntax, mathematical linguistics, fairness in NLP.
Interest	Dance, Reading Novels.	MMA, Volleyball, Football (soccer).	Rock climbing, sleeping.
Fun Fact!	Survived an attack by a peacock when he was 8 years old.	I am semi-mildly uncomfortable around pigeon 🦚	3 adorable cats.

Don’t Forget!

- Shared PyTorch review tutorial with CSC401.
- Friday, Sep 6:
  - 10-11 BA1180.
  - 11-12 BA1190.
  - 12-13 ES B142.
  - Any session that works the best for you.

Quiz

- Which of the following neural architecture is the foundation of all major large language models?
  a) RNN
  b) LSTM
  c) Transformer
  d) Encoder-decoder

Submit your answers on Quercus. Due Friday 5pm.