CS 489/698: Introduction to Natural Language Processing

Lecture 1: Introduction & Fundamentals

Instructor: Freda Shi

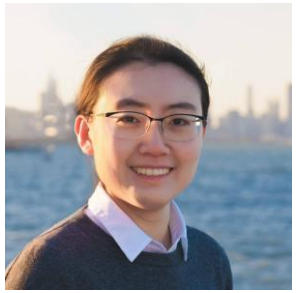
fhs@uwaterloo.ca

January 5th, 2026

UNIVERSITY OF WATERLOO

Instructors and Websites

- **Instructors:** Freda Shi (before mid-term) and Victor Zhong (after mid-term)



- **Office Hours** (Freda)**:** Friday 5-6pm until February 11th at DC 2522.
- **Office Hours** (Victor): TBA February 23rd onwards.

- **Course Website:** https://waterloo-nlp.github.io/intro-to-nlp-material
- **Discussion Forum:** https://piazza.com/uwaterloo.ca/winter2026/cs489698
- **LEARN** (for assignments)**:** https://learn.uwaterloo.ca/d2l/home/1220061

Teaching Assistants

Arthur Chen
haonan.chen@

Kath Choi
ymchoi@

Yifan Jiang
yifan.jiang@

Michael Ogezi
mogezi@

Sheng Yao
s57yao@

- **TA Office Hours:** To be announced before the assignment due dates.

Books

There will be no official textbooks.
However, we recommend the following ones for reference:

- Jurafsky and Martin, Speech and Language Processing, 3rd edition
  https://web.stanford.edu/~jurafsky/slp3/
- Jacob Eisenstein, Introduction to Natural Language Processing
  https://cseweb.ucsd.edu/~nnakashole/teaching/eisenstein-nov18.pdf
- Yoav Goldberg. A Primer on Neural Network Models for Natural Language Processing
  https://u.cs.biu.ac.il/~yogo/nnlp.pdf

Grade Breakdown

There will be no exams for this course.

- **Assignments (50%, 25% each)**
  Due on February 9th and March 2nd, respectively. There will be a major programming practice and some question-answering in each assignment.
- **Course Project (50% for CS 489, 25% for CS 698)**
  Developing an NLP system for a real-world application. Due on April 6th. Mandatory for passing the course.
  - **Mid-term check-in (10% for CS 489, 5% for CS 698)**: Due on February 23rd.
- **Individual Project (25% for CS 698 only)**
  An NLP-related research project of your own choice.
  - **Project proposal**: Due on February 23rd. Feedback will be made available in a week.

Lateness Policy

All assignments are due at 11:59 PM Eastern Time (Waterloo time) on the specified due date.

A universal 72-hour grace period applies to all assignment and project dues:

- Submissions within the grace period will be accepted without penalty.
- No additional extensions will be granted beyond the grace period.

- Assignment (excluding the final project) extensions will only be considered in cases of medical emergency, with:
  - A valid Verification of Illness Form (VIF) submitted, and that
  - The VIF covers the original due date and the entire grace period.
  - The extension application must be submitted before the original due date.

GenAI Policy

You may use GenAI tools however you can to assist your learning and growth, but remember these tools can’t be fully trusted and should never replace your own understanding. It is essential to carefully verify the content they generate, especially when working with detailed numbers or using AI to refine your write-ups.

Ultimately, you are fully responsible for the accuracy and integrity of all work you submit in this course.

Collaboration and Citation Policy

- **Permitted Collaboration**: Discussion of assignments with fellow students is encouraged, but all submitted work (including code and written solutions) must be completed independently.
- **Citation Requirements**: All external sources must be properly cited, including
  - Academic references and literature.
  - Generative AI tools (e.g., ChatGPT) if used beyond grammar check.
  - Any other resources consulted during assignment completion.
- **Student Responsibilities**:
  - Verify the accuracy of any AI-generated content.
  - Ensure all submitted work reflects your own understanding.
  - Provide complete and accurate citations for all sources.

Academic Integrity and Intellectual Property

Property of UW:

- Lecture content, spoken and written (and any audio/video recording thereof).
- Lecture handouts, presentations, and other materials prepared for the course (e.g., PowerPoint slides).
- Questions or solution sets from various types of assessments (e.g., assignments, quizzes, tests, final exams).
- Work protected by copyright (e.g., any work authored by the instructor or TA or used by the instructor or TA with permission of the copyright owner).

Sharing intellectual property without the intellectual property owner's permission is a violation of intellectual property rights.

Should I take the course?

You will find the course helpful if

- You are curious about the fundamentals and detailed implementations of language models.
- You are looking to apply state-of-the-art NLP techniques to your own goals.
- You are interested in NLP and computational linguistics research.

This course will not cover

- Basics of Python programming, probability, and algorithms.
- System and architectures for efficient language model training and deployment.
- GPU programming (except for a few high-level basics).

Prerequisites

The formal prerequisite for this course is to ensure you have sufficient mathematical and programming maturity.

The following background knowledge is strongly recommended:

- Basic knowledge of calculus, linear algebra, and probability.
- Python programming proficiency.
- Fundamentals of algorithms, e.g., basic complexity analysis.
- Understanding of basic data structures, e.g., lists, stacks, queues, trees, and graphs.

What is Natural Language Processing (NLP)?

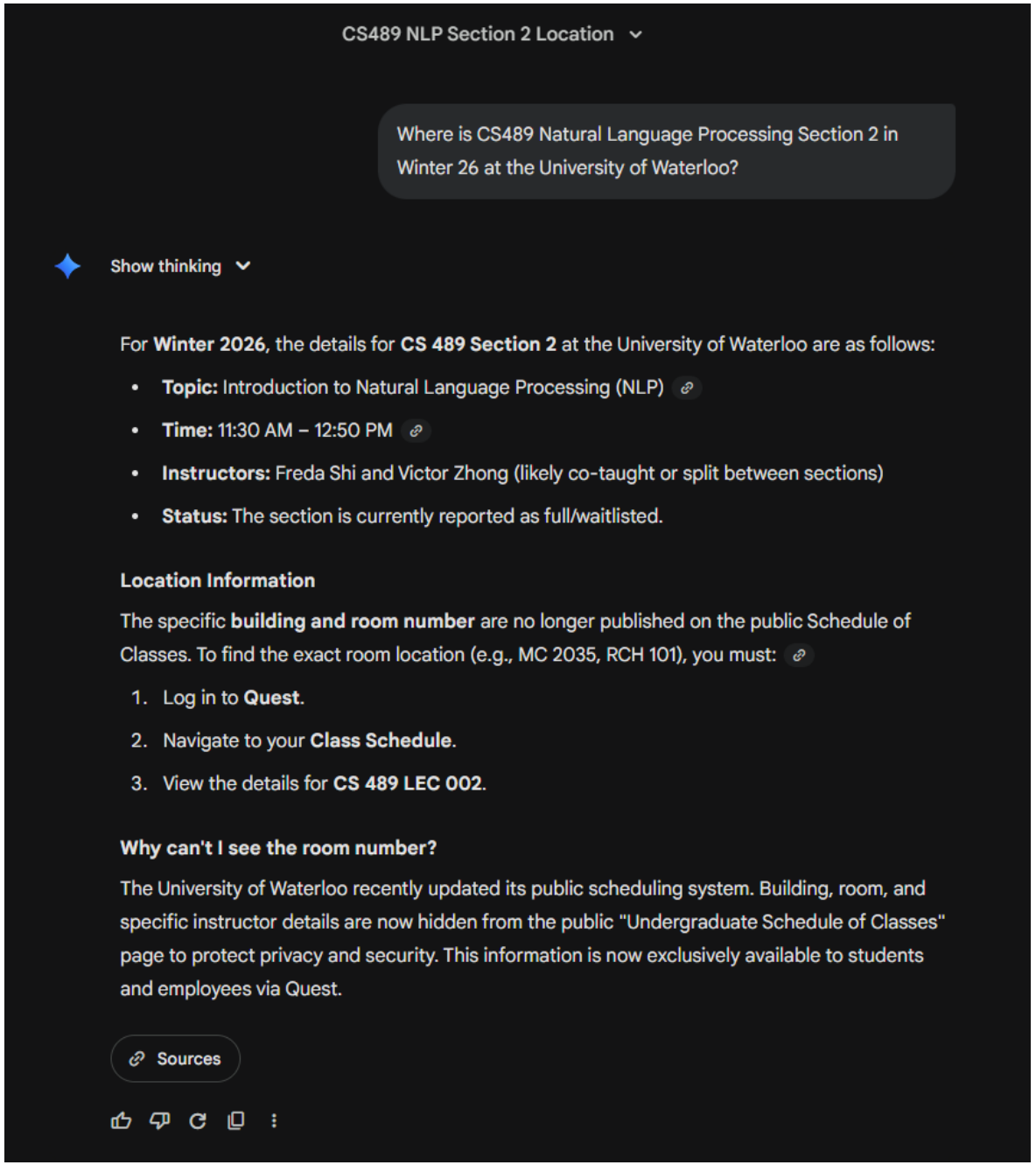
A cat is sitting next to a *dax*.

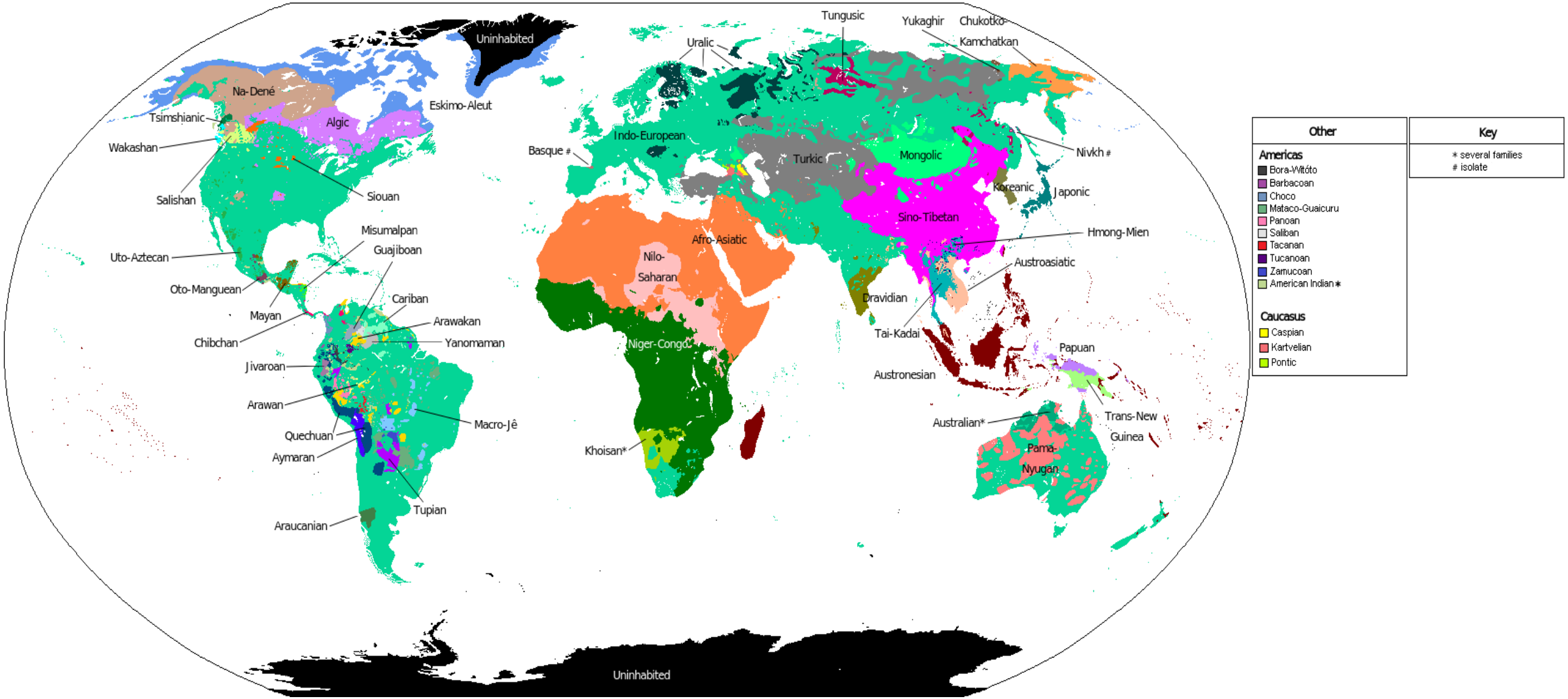
Q: What is a *dax*?

Goal: understand natural language with computational models.

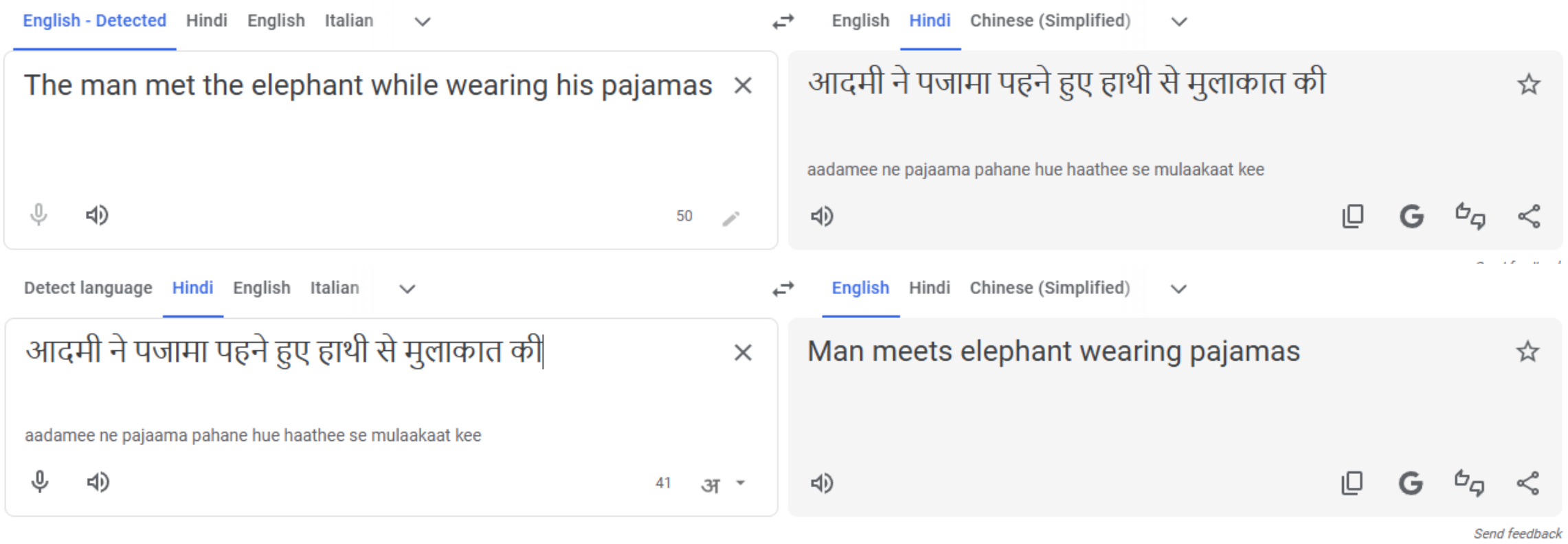
End systems that we want to build:

- Simple: Text classification, grammatical error correction, ...
- Complex: Translation, question answering, speech recognition, ...
- Unknown: Human-level comprehension (is this just NLP?)


CS489 NLP Section 2 Location
Where is CS489 Natural Language Processing Section 2 in Winter 26 at the University of Waterloo?
Show thinking
For Winter 2026, the details for CS 489 Section 2 at the University of Waterloo are as follows:
Topic: Introduction to Natural Language Processing (NLP)
Time: 11:30 AM – 12:50 PM
Instructors: Freda Shi and Victor Zhong (likely co-taught or split between sections)
Status: The section is currently reported as full/waitlisted.
Location Information
The specific building and room number are no longer published on the public Schedule of Classes. To find the exact room location (e.g., MC 2035, RCH 101), you must:
1. Log in to Quest.
2. Navigate to your Class Schedule.
3. View the details for CS 489 LEC 002.
Why can't I see the room number?
The University of Waterloo recently updated its public scheduling system. Building, room, and specific instructor details are now hidden from the public "Undergraduate Schedule of Classes" page to protect privacy and security. This information is now exclusively available to students and employees via Quest.
Sources


Uninhabited
Na-Dené
Eskimo-Aleut
Tsimshianic
Algic
Wakashan
Salishan
Siouan
Misumalpan
Guajiboan
Uto-Aztecan
Oto-Manguean
Mayan
Chibchan
Cariban
Arawakan
Yanomaman
Jivaroan
Arawan
Quechuan
Aymaran
Macro-Jê
Tupian
Araucanian
Tungusic
Yukaghir
Chukotko
Kamchatkan
Uralic
Indo-European
Basque #
Turkic
Mongolic
Nivkh #
Koreanic
Japonic
Sino-Tibetan
Hmong-Mien
Austroasiatic
Afro-Asiatic
Nilo-
Saharan
Dravidian
Tai-Kadai
Niger-Congo
Papuan
Austronesian
Trans-New
Guinea
Australian*
Pama-
Nyugan
Khoisan*
Uninhabited
Other
Americas
Bora-Witóto
Barbacoan
Choco
Mataco-Guaicuru
Panoan
Saliban
Tacanan
Tucanoan
Zamucoan
American Indian *
Caucasus
Caspian
Kartvelian
Pontic
Key
* several families
# isolate

Machine Translation



[Google Translation 2025-12]

Why is NLP difficult?

- Ambiguity: one form can have multiple meanings.

bank[1] [bangk] SHOW IPA

See synonyms for: bank / banked / banking on Thesaurus.com

noun

1. a long pile or heap; mass:
   a bank of earth;
   a bank of clouds.
2. a slope or acclivity.

SEE MORE

verb (used with object)

9. to border with or like a bank; embank:
   banking the river with sandbags at flood stage.
10. to form into a bank or heap (usually followed by *up*):
    to bank up the snow.

SEE MORE

verb (used without object)

15. to build up in or form banks, as clouds or snow.
16. *Aeronautics*. to tip or incline an airplane laterally.

SEE MORE

bank[2] [bangk] SHOW IPA

noun

1. an institution for receiving, lending, exchanging, and safeguarding money and, in some cases, issuing notes and transacting other financial business.
2. the office or quarters of such an institution.

SEE MORE

verb (used without object)

7. to keep money in or have an account with a bank:
   Do you bank at the Village Savings Bank?
8. to exercise the functions of a bank or banker.

SEE MORE

verb (used with object)

10. to deposit in a bank:
    to bank one's paycheck.

Why is NLP difficult?

- Ambiguity: one form can have multiple meanings.

She saw a cat with a telescope.

Who was with a telescope?

[Images generated with Nano Banana Pro]

Why is NLP difficult?

- Ambiguity: one form can have multiple meanings.
- Variability: multiple forms can share the same meaning.

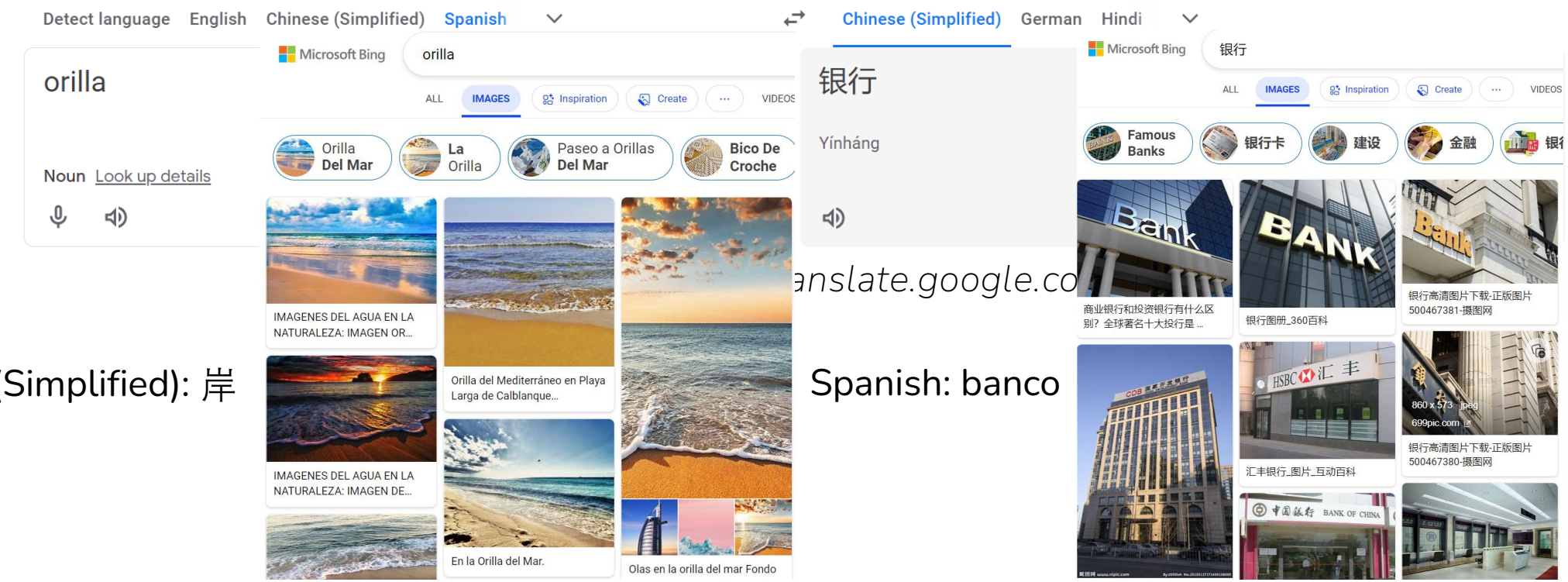
The cat is chasing the mouse.

The mouse is being chased by the cat.

[Image generated with Nano Banana Pro]

Why is NLP difficult?

- Ambiguity: one form can have multiple meanings.
- Variability: multiple forms can share the same meaning.
- Cross-lingual awareness, dialects, and accents.



Chinese (Simplified): 岸

Spanish: banco

Why is NLP difficult?

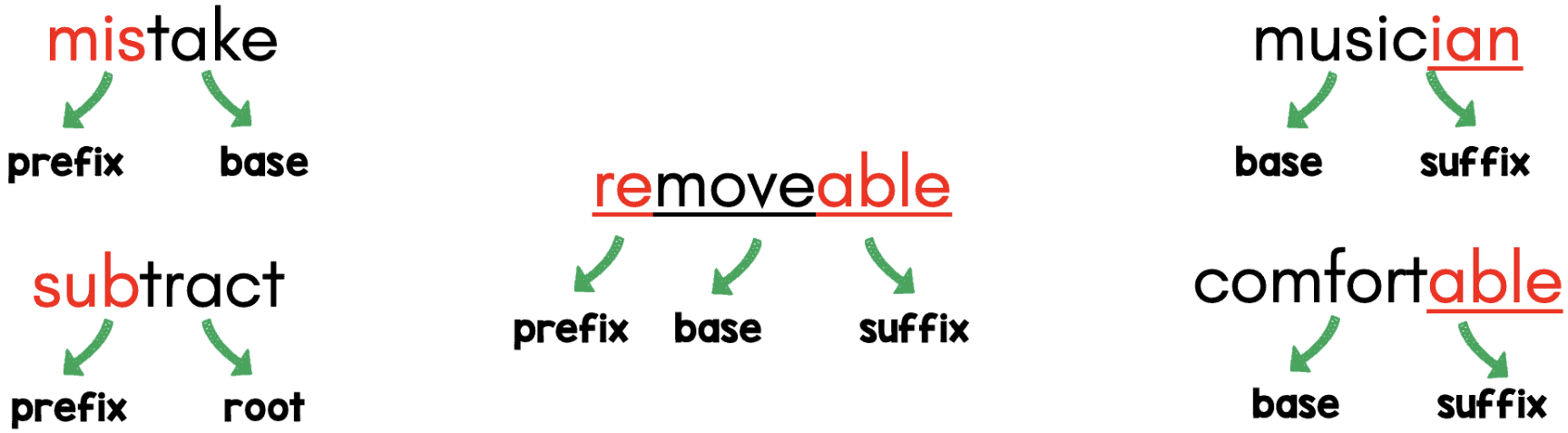
- Ambiguity: one form can have multiple meanings.
- Variability: multiple forms can share the same meaning.
- Cross-lingual awareness, dialects, and accents.
- Underlying meanings: politeness, humor, irony.

[Image generated with Nano Banana Pro]

Q: Do you have iced tea?
A1: No.
A2: We have iced coffee.

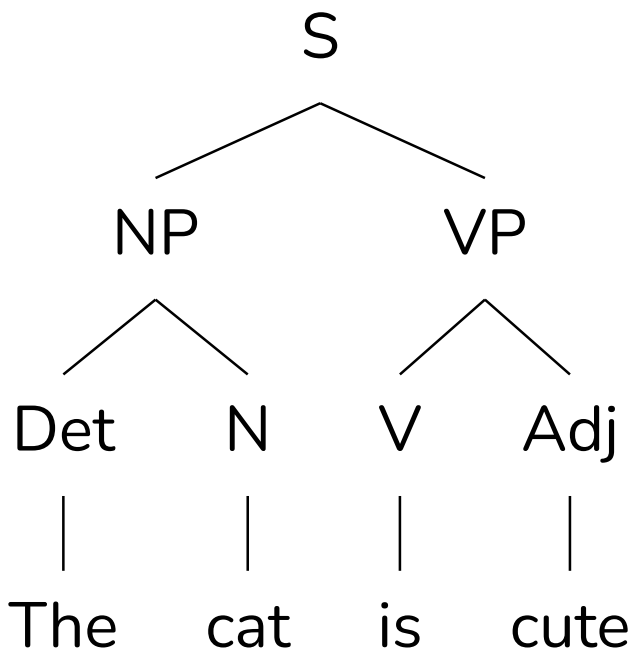
Roadmap: Levels of Language

- Morphology: What words/subwords are we dealing with?



Roadmap: Levels of Language

- Morphology: What words/subwords are we dealing with?
- Syntax: What phrases are we dealing with?



Roadmap: Levels of Language

- Morphology: What words/subwords are we dealing with?
- Syntax: What phrases are we dealing with?
- Semantics: What is the literal meaning?

What an excellent weather!

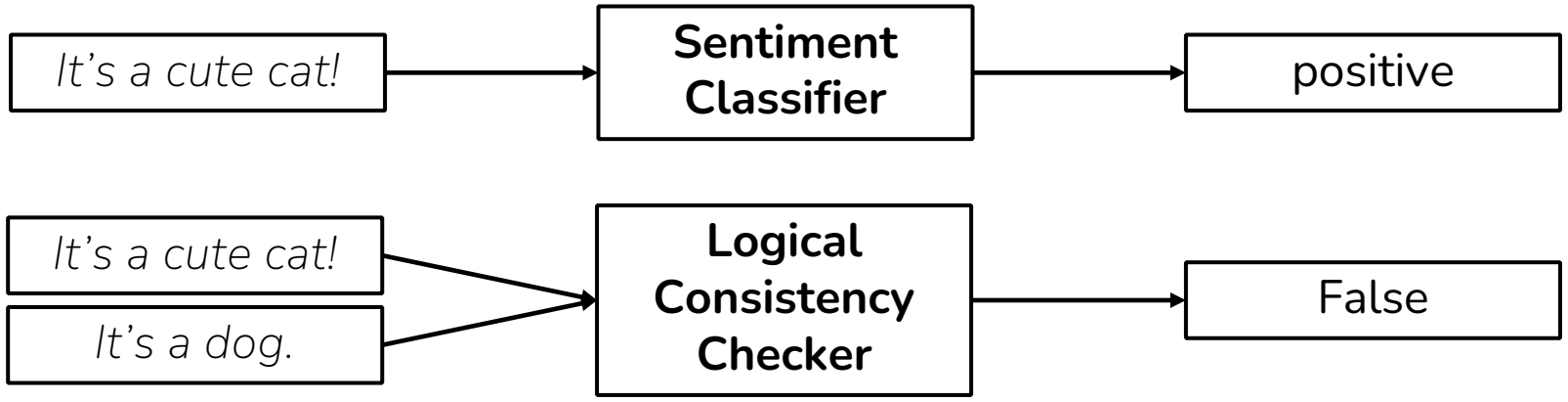
Roadmap: Levels of Language

- Morphology: What words/subwords are we dealing with?
- Syntax: What phrases are we dealing with?
- Semantics: What is the literal meaning?
- Pragmatics: What is the underlying meaning?

What an excellent weather!

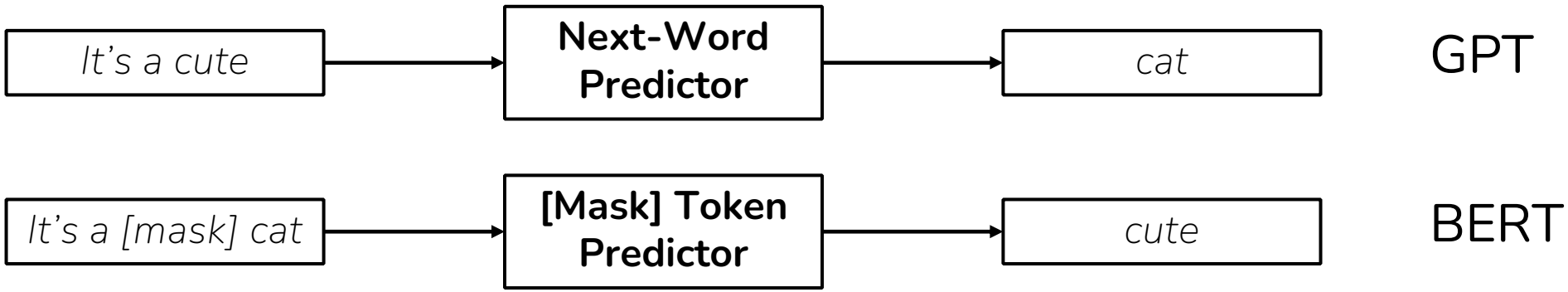
Roadmap: Modeling Approaches

- Classification



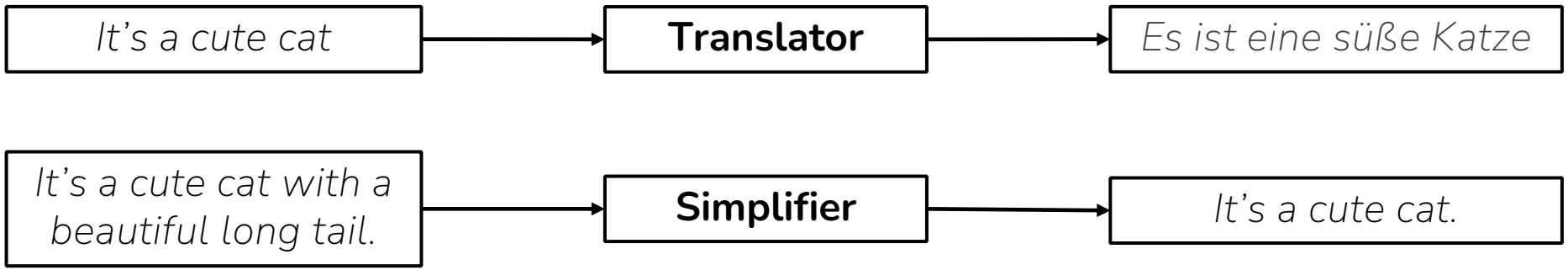
Roadmap: Modeling Approaches

- Classification
- Language modeling



Roadmap: Modeling Approaches

- Classification
- Language modeling
- Sequence-to-Sequence modeling



Brief History of NLP

1960s-1990s: Classical NLP

- Linguistic theories
- Manually-defined rules
- Small-scale datasets and experiments

1980s-2013: Statistical NLP

- Supervised machine learning with annotated data
- (Mostly) linear models with manually-defined features
- Linguistic knowledge used in annotation

2012-now: Neural NLP

- Neural network as primary architecture
- Learning representations from large corpora
- Less hand-crafting features/linguistic knowledge

2022-now: Large language models

- Pretrain a language model
- Use the pretrained language model for various tasks

Common Notations of (Discrete) Probability

$P(X = x)$ denotes the probability that random variable X takes on the value x.

Given two random variables X, Y.

Joint probability: $P(X = x, Y = y) = P(x, y)$

Marginal probability: $P(X = x) = P(x) = \sum_y P(x, y)$

X and Y are independent iff. $P(x, y) = P(x)P(y)$

Conditional probability: $P(\, X = x \mid Y = y \,) = P(\, x \mid y \,) = \frac{P(x,y)}{P(y)}$

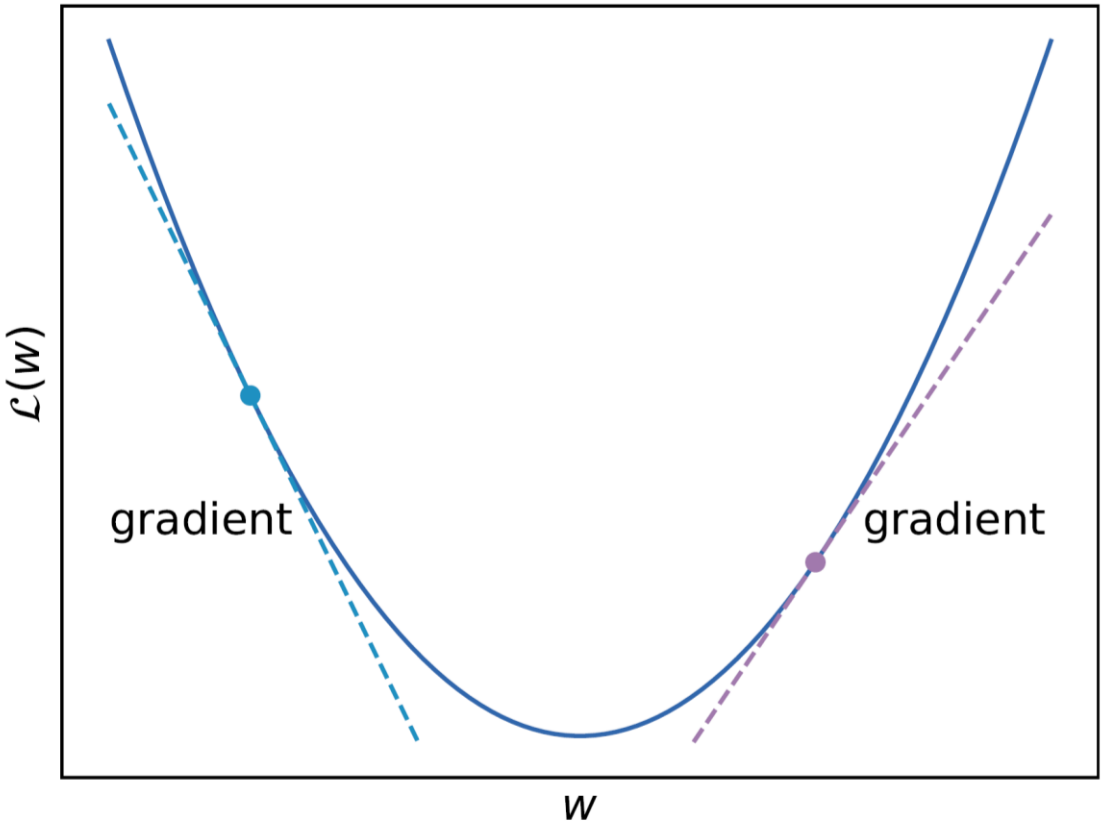
X and Y are independent $\Leftrightarrow P(x \mid y) = P(x) \Leftrightarrow P(\, y \mid x \,) = P(y)$

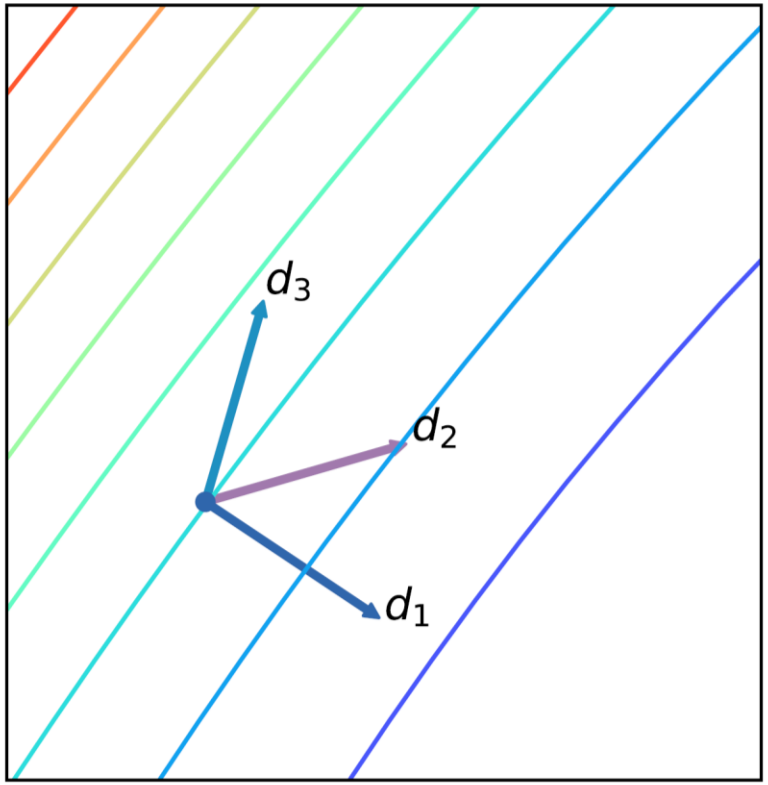
Expectation: $\mathbb{E}[X] = \mathbb{E}_{x \sim P}[x] = \sum_x x \, P(x)$

Finding the Optimum of a Function

Find the minimum of a convex function $\mathcal{L}(w; X, y)$.

Gradient descent: $w' \leftarrow w - \eta \nabla_w \mathcal{L}(w; X, y)$; η is the learning rate ("step size").





Next

Morphology, tokenization, word, subwords