Introduction to NLP, Language Models

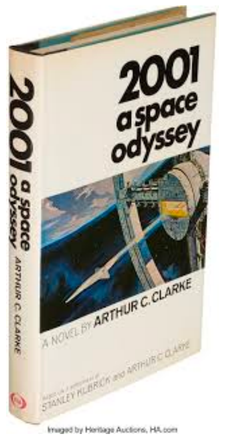
COMP3361 — Week 1

Lingpeng Kong
Department of Computer Science, The University of Hong Kong

Many materials from CSE517@UW, COMS W4705@Columbia, 11-711@CMU with special thanks!

“I'm sorry Dave, I'm afraid I can't do that.”

https://youtube.com/watch?v=ARJ8cAGm6JE



2001: A Space Odyssey

Talking machines, are we infinitely close or far away?

What can I help you with?

Play a good song.

Sorry, I couldn’t find ‘a good song’ in your music.

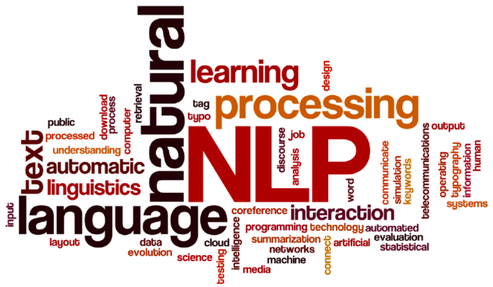
You need to do a better job understanding me.

Noted.

Yeah, make a note of that.

Here’s your note:

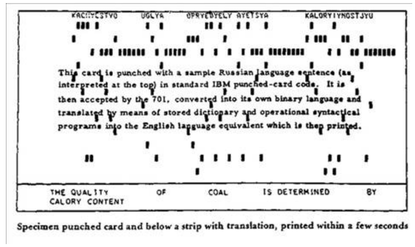
Demystify — what's inside?



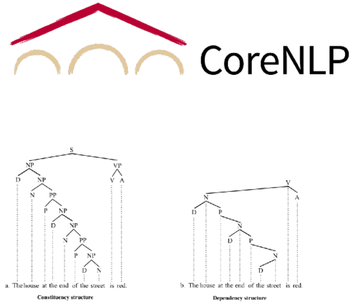
Symbolic NLP

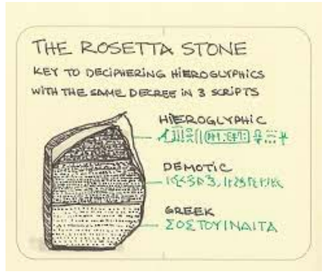
Statistical NLP

Neural NLP



Specimen punched card and below a strip with translation, printed within a few seconds





1950s - early 1990s

1990s - 2010s

2010s - present

What is NLP? Wait, what is language?

The abstraction of the real world — different languages take you to different worlds!

餃子

CX 844
Hong Kong -> New York

dumplings

嗩吶

CX 596
Hong Kong -> Osaka

チャルメラ

Something that makes sharp long voice, like screaming???

Do talking machines “understand”? Let’s play a game!

The cat is thrown out of the _____

door, window, dog

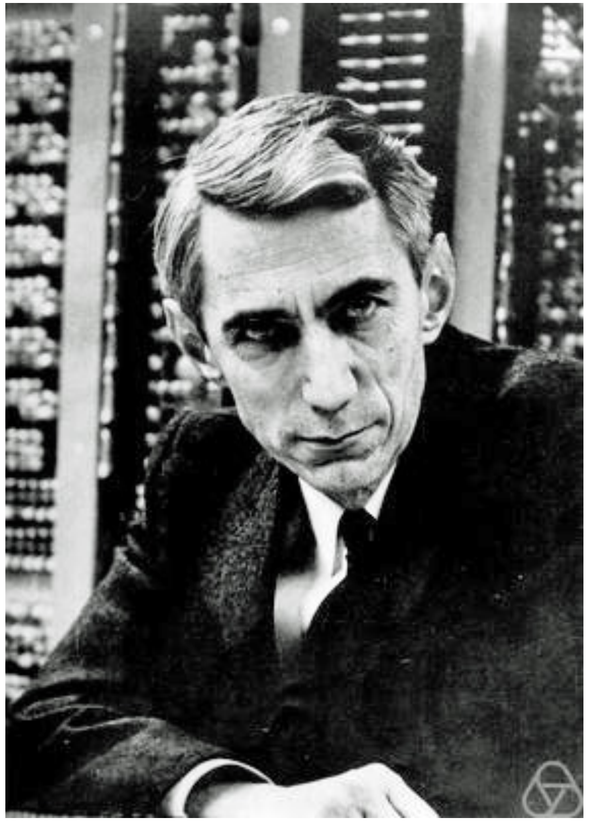
This year, I am going to do an internship in _____

Queen Mary Hospital, HSBC, Google, Amazon

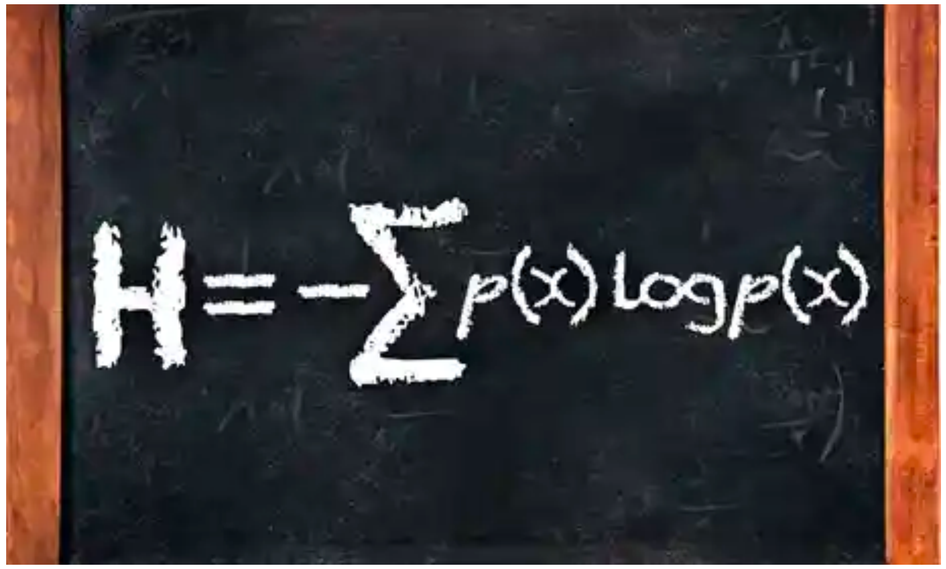
Majoring in computer science, this year, I am going to do an internship in _____

Queen Mary Hospital, HSBC, Google, Amazon

Shannon Game



Claude Elwood Shannon
(April 30, 1916 – February 24, 2001)



Information Theory; Entropy

Language models, and how to build it.

Dice, and how do we roll them
(probabilistic model)

Transformers, neural networks and many others
(powerful functions)

Generative Language Model

I am going to do an internship in Google

Making the dice

bag of words
(@Carnegie Mellon University)

1	Belief
2	Evidence
3	Reason
4	Claim
5	Think
6	Justify
7	Also
	…
99	Therefore
100	Google

Vocabulary

Generative Language Model

I

I

Generative Language Model

I am

am

Generative Language Model

I am going

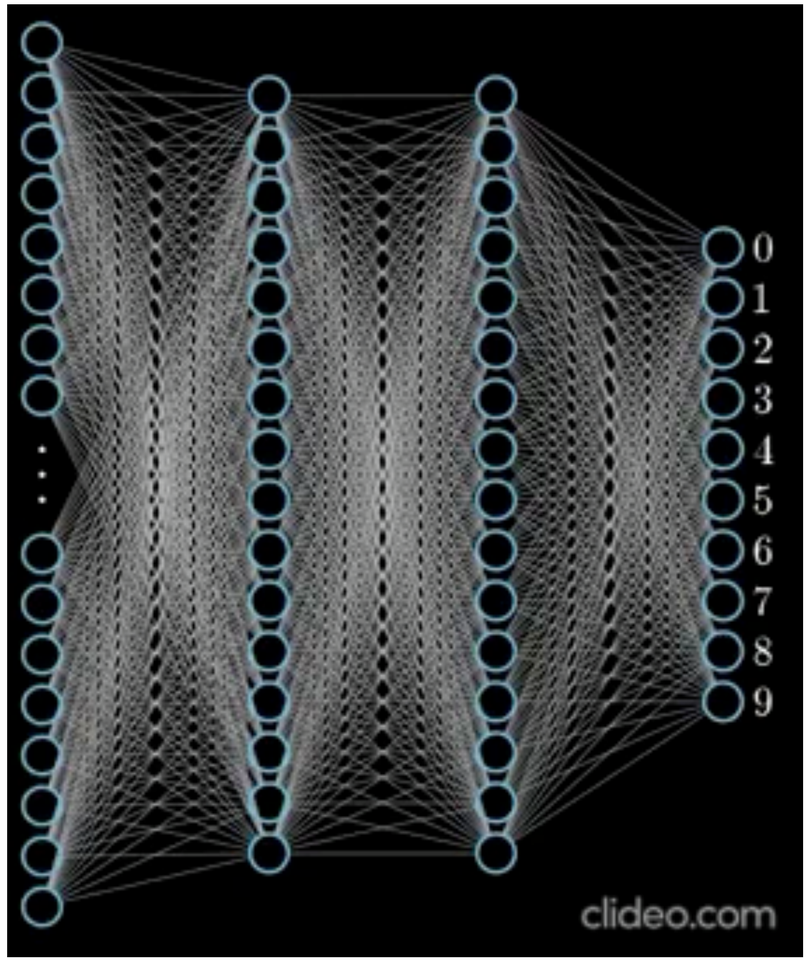
going

Generative Language Model

I am going to do an internship in Google

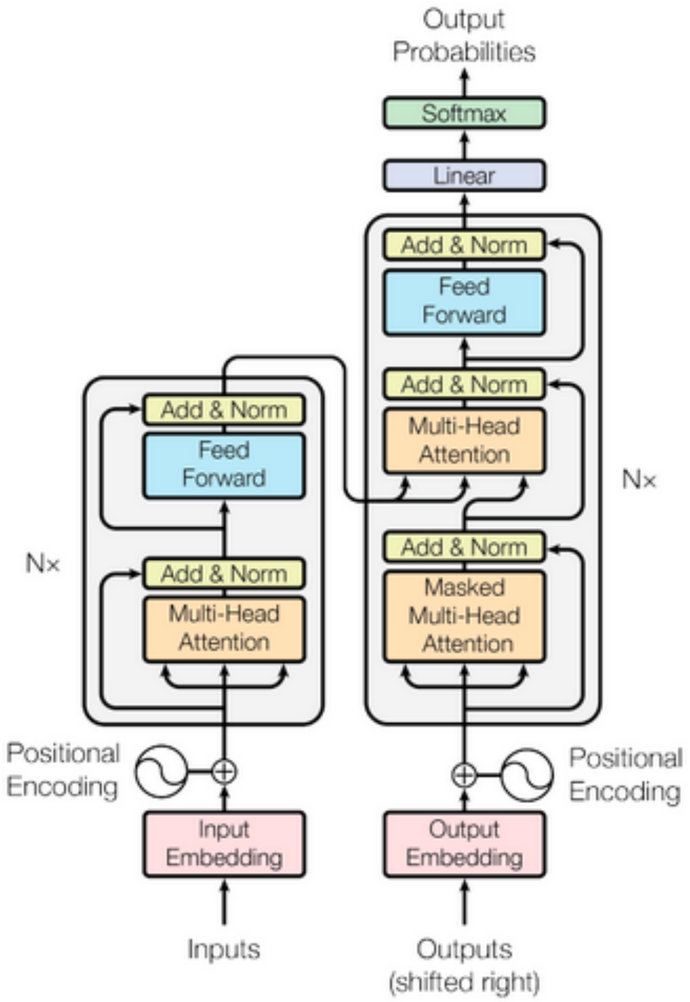
Google

Neuralize the dice!



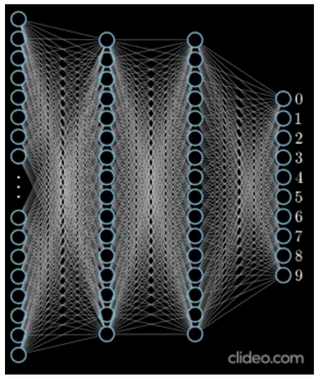
3Blue1Brown

Neural Networks (e.g. Transformers)



Generative Language Model

I am going to do an internship in Google



Google

Language models, and how to build it.

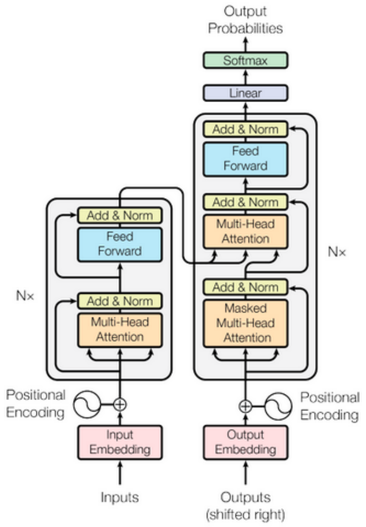
Dice, and how do we roll them
(probabilistic model)

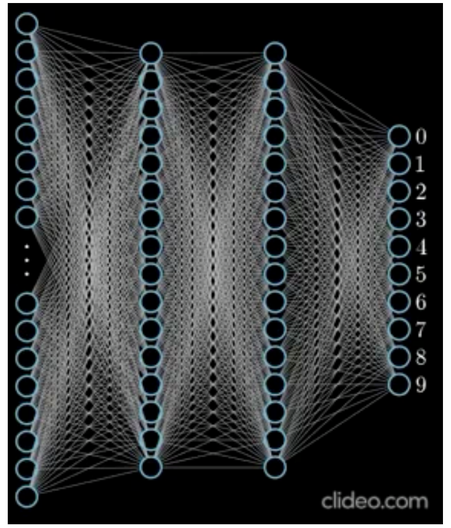
Transformers, neural networks and many others
(powerful functions)

$$p(\boldsymbol{x}) = \prod_i p(x_i|\boldsymbol{x}_{<i})$$

Learn

Parameterize





First problem — the language modeling problem

Given a finite vocabulary

$$\mathcal{V} = \{\text{belief}, \text{evidence}, \text{reason}, \text{claim}, \ldots \text{Google}, \text{therefore}\}$$

We have an infinite set of strings, $\mathcal{V}^\dagger$

<s> I am going to an internship in Google </s>
<s> an internship in Google </s>
<s> I am going going </s>
<s> Google is am </s>
<s> internship is going </s>

Formally:

$$p(x_1, x_2, \ldots x_n)$$
$$p(x_i \mid x_{i-1}, x_{i-2}, \ldots x_1)$$

Can we learn a "model" for this "generative process"? We need to "learn" a probability distribution:

$$\sum_{x \in \mathcal{V}^\dagger} p(x) = 1, \; p(x) \geq 0 \text{ for all } x \in \mathcal{V}^\dagger$$

Learn from what we've seen

The Language Modeling Problem

Given a *training sample* of example sentences, we need to "learn" a probabilistic model that assigns probabilities to every possible string:

p(<s> I am going to an internship in Google </s>) = 10^{-12}

p(<s> an internship in Google </s>) = 10^{-8}

p(<s> I am going going </s>) = 10^{-15}

…

It is a probability distribution p over strings, i.e., p is a function that satisfies

$$\sum_{x \in \mathcal{V}^\dagger} p(x) = 1,\ p(x) \geq 0 \text{ for all } x \in \mathcal{V}^\dagger$$

The Language Modeling Problem

<s> Sam I am </s>

<s> I am Sam </s>

<s> I do not like green eggs and ham </s>

Training Corpus

Is this a good model?

P(<s> Sam I am </s>) = 1/3

P(<s> I am Sam </s>) = 1/3

P(<s> I do not like green eggs and ham </s>) = 1/3

P(<s> I am </s>) = 0

P(<s> green Sam </s>) = 0

Naïve Model

The Language Modeling Problem

<s> Sam I am </s>

<s> I am Sam </s>

<s> I do not like green eggs and ham </s>

Training Corpus

The probability of the word “<s>” followed by the word “I”:

$$P(I \mid <s>) = 2/3$$

The Language Modeling Problem

<s> Sam I am </s>

<s> I am Sam </s>

<s> I do not like green eggs and ham </s>

Training Corpus

P(I | <s>) = 2/3

P(</s> | Sam) = 1/2

Naïve Model

The Language Modeling Problem

<s> Sam I am </s>

<s> I am Sam </s>

<s> I do not like green eggs and ham </s>

Training Corpus

P(I | <s>) = 2/3

P(am | I) = ?

P(</s> | Sam) = 1/2

Naïve Model

The Language Modeling Problem

<s> Sam I am </s>

<s> I am Sam </s>

<s> I do not like green eggs and ham </s>

Training Corpus

P(I | <s>) = 2/3

P(am | I) = 2/3

P(</s> | Sam) = 1/2

Naïve Model

The Language Modeling Problem

<s> Sam I am </s>

<s> I am Sam </s>

<s> I do not like green eggs and ham </s>

Training Corpus

$$P(\text{<s> Sam I am </s>}) = P(\text{Sam} \mid \text{<s>}) * P(\text{I} \mid \text{Sam}) * P(\text{am} \mid \text{I}) * P(\text{</s>} \mid \text{am})$$

Bi-gram Model

Course Logistics

Course Logistics

Website:

https://nlp.cs.hku.hk/comp3361

Prerequisites:

COMP3314 or COMP3340, MATH1853

We will assume a lot things from Machine Learning, Statistics, and Programming

This NLP course will be very difficult if you haven't taken these courses.

Assessment:

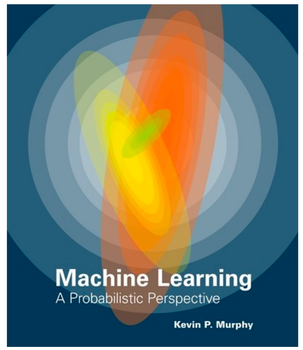
50% continuous assessment, 50% examination (final exam)

TA:

Qintong Li (https://qtli.github.io/)

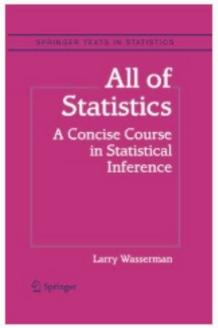
Course Logistics

We will assume a lot things from Machine Learning, Statistics, and Programming

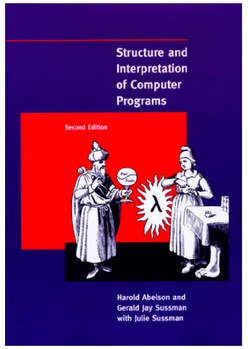


Supervised learning, unsupervised learning, regression, classification, loss function, neural networks, regularization …

(COMP3314B, Spring 2022)



Random variables, joint probability, conditional probability, Bayes' theorem …

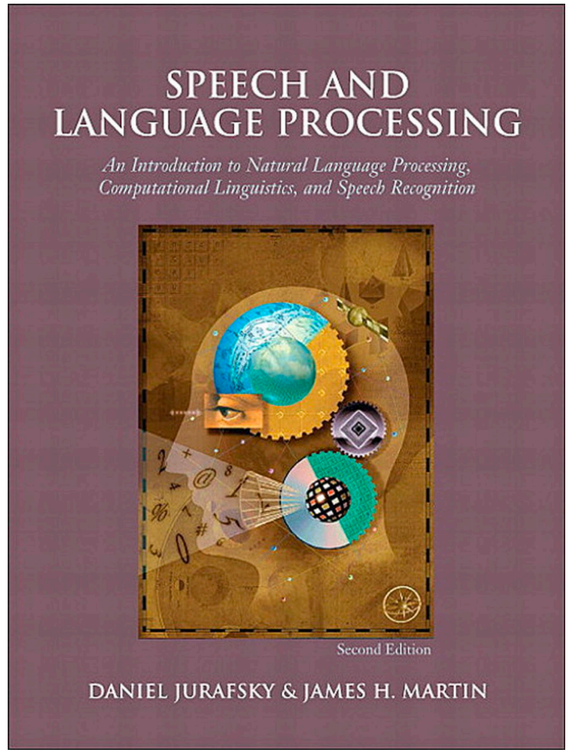


Data structures, dynamic programming, time/space complexity …

Course Logistics

Textbook recommendation (J&M):

https://web.stanford.edu/~jurafsky/slp3/



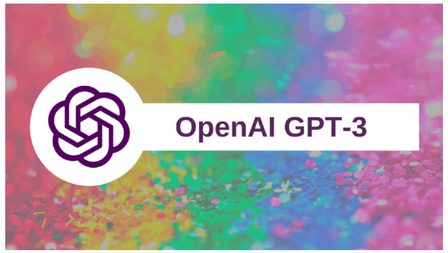
Assessments (in total ~4):

- Programming problems
- Problem sets

Honor code:

You are free to form study groups and discuss homeworks and projects. However, you must write up homeworks and code from scratch independently, and you must acknowledge in your submission all the students you discussed with.

What's Next?

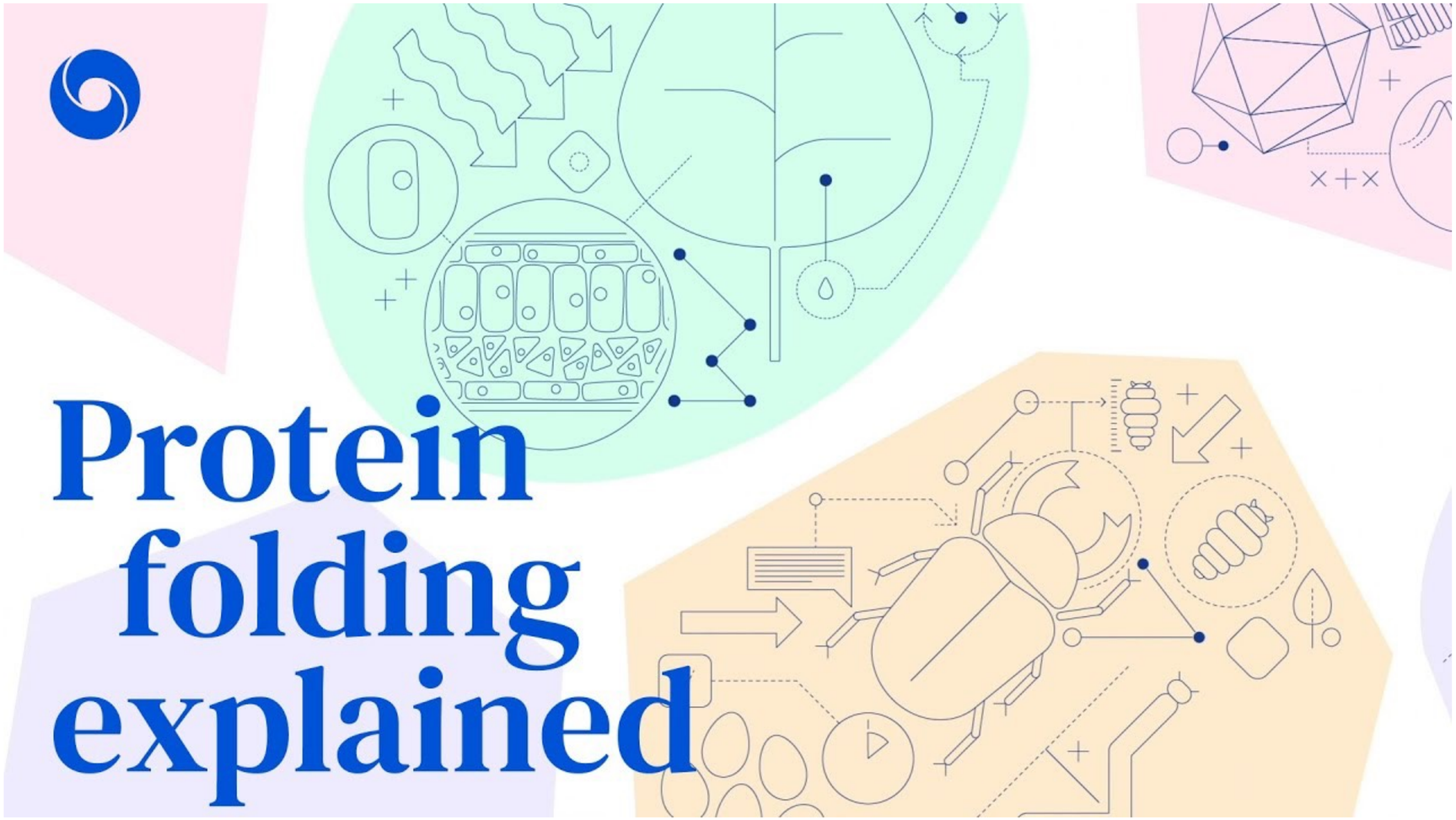


BERT, GPT-3, Word2vec, Glove, T5 ...

N-Gram Models, Hidden Markov Models ...

LSTMs, Recurrent Neural Networks, MLP, Transformers ...

It’s not about human language. It’s the language of life!



https://www.youtube.com/watch?v=KpedmJdrTpY

Google DeepMind

AlphaFold (Jumper et al, 2021)