

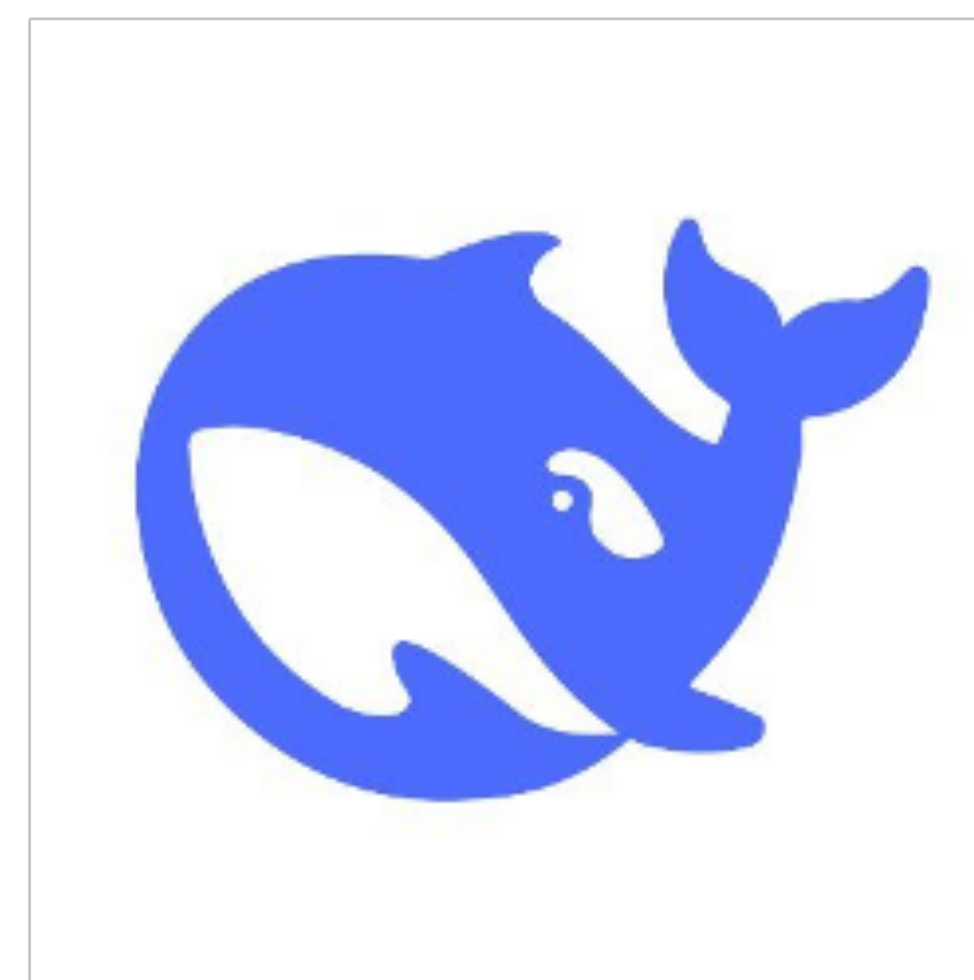
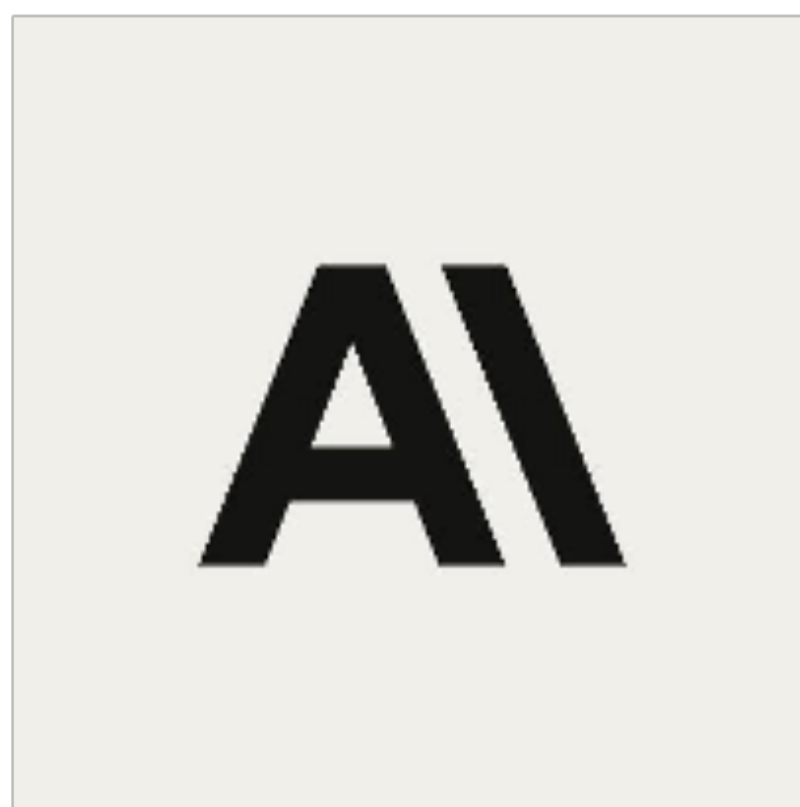
Introduction to NLP, Language Models

COMP7607 — Week 1

Lingpeng Kong

Department of Computer Science, The University of Hong Kong

Who's who in AI?



Who's who in AI?



OpenAI



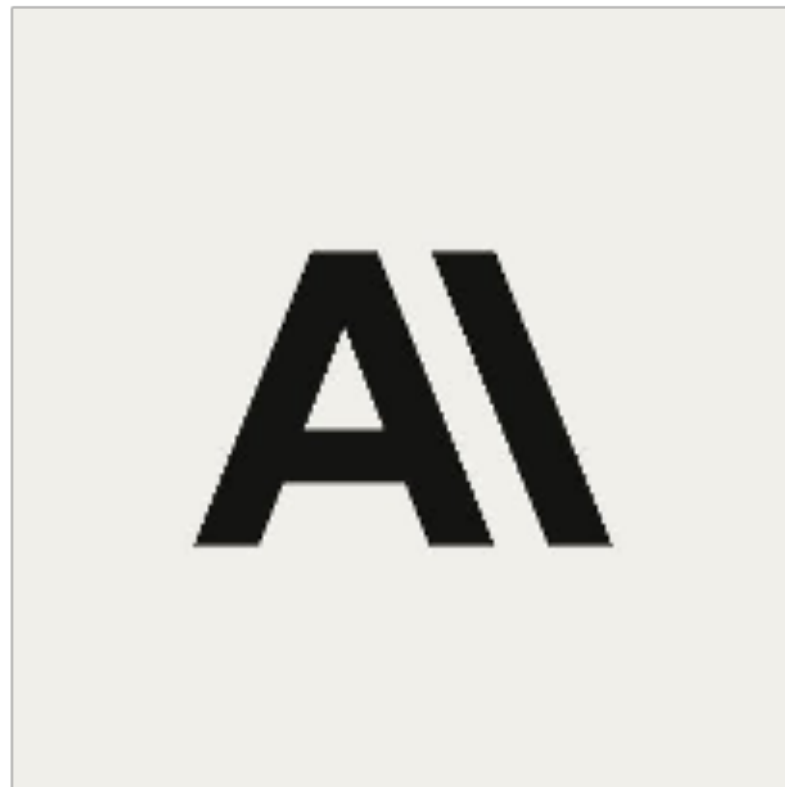
Google DeepMind



Meta



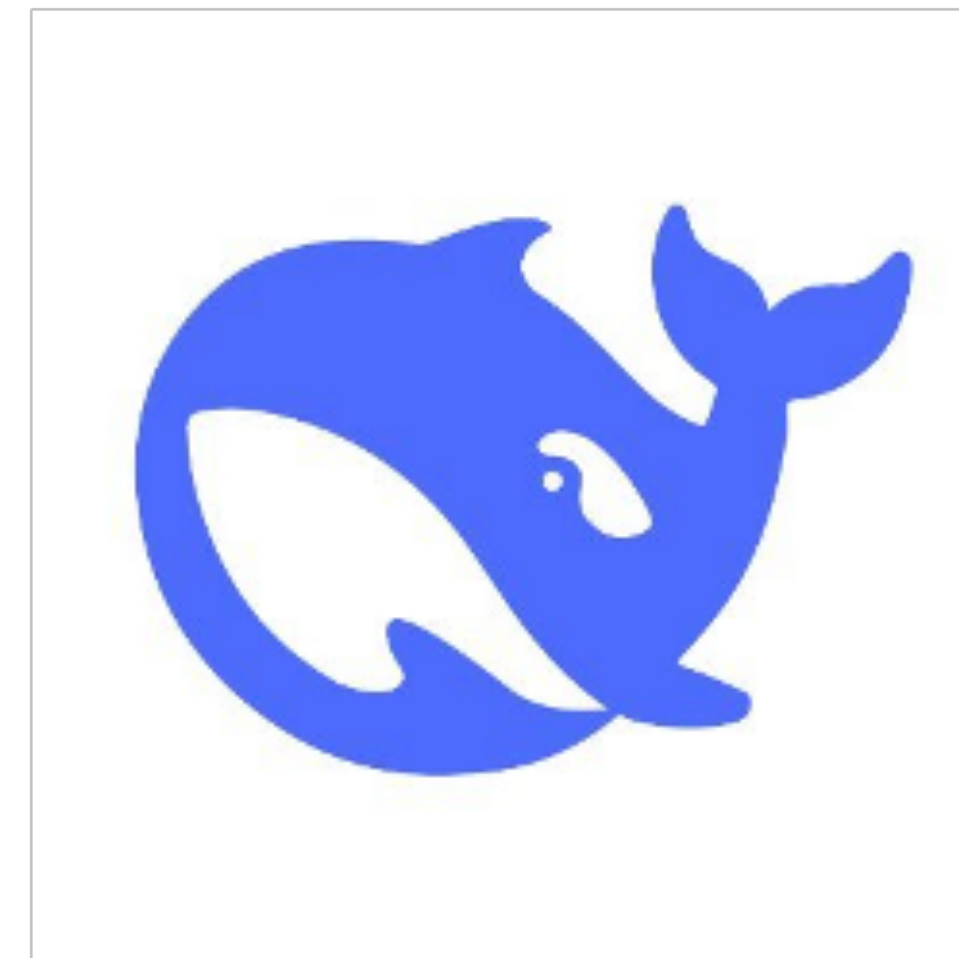
xAI



Anthropic

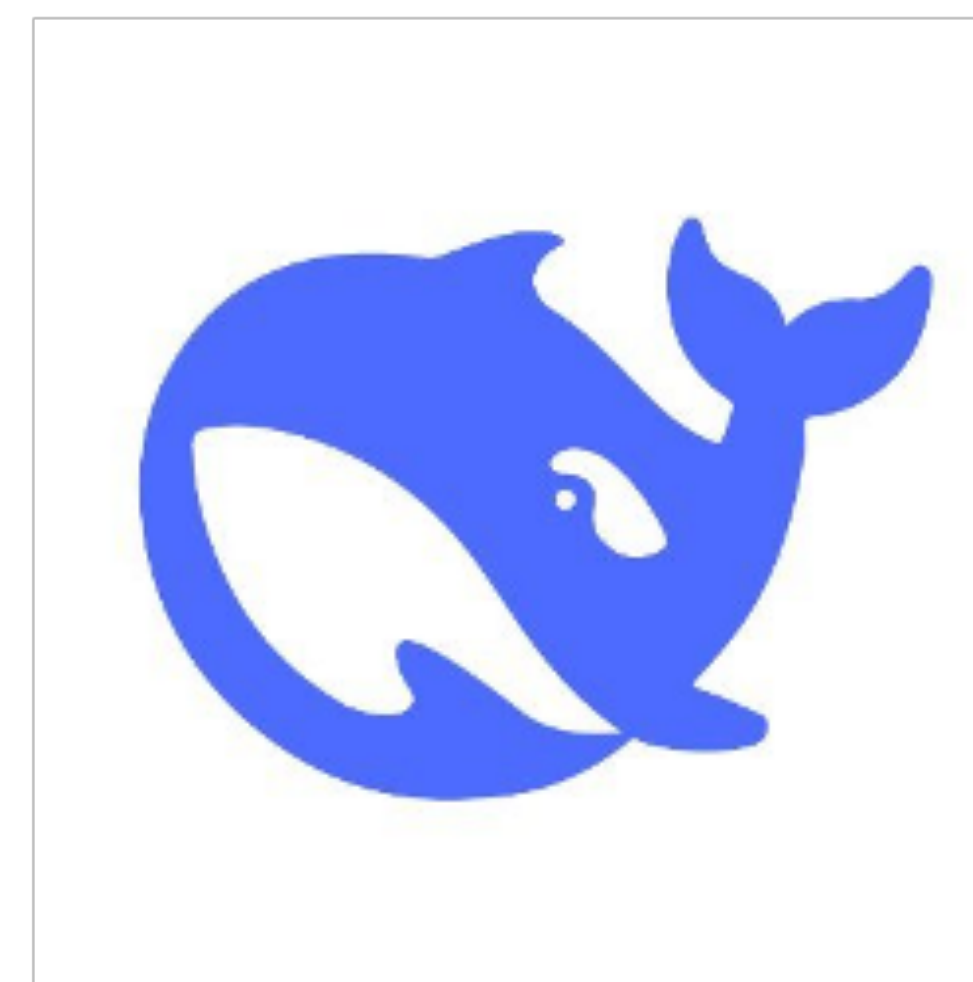
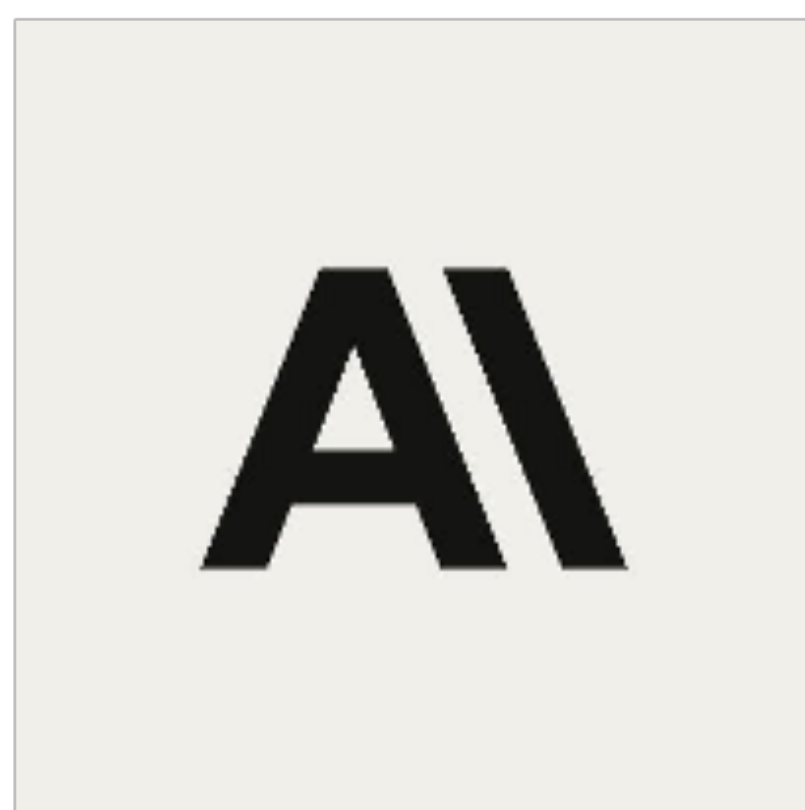


Alibaba



DeepSeek

How about their models?



How about their models?



GPT4, GPT-o1...



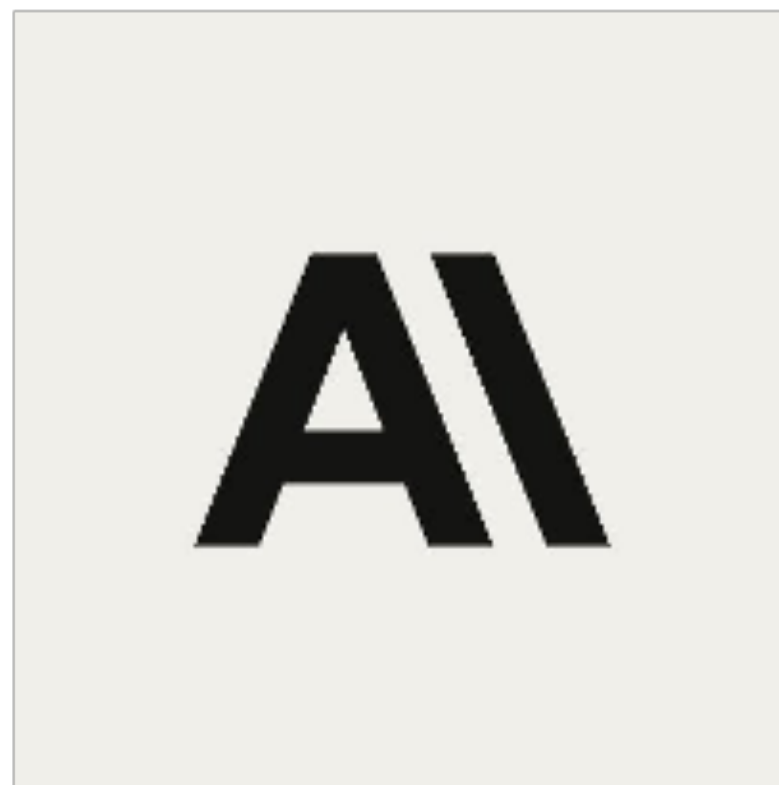
Gemini



Llama



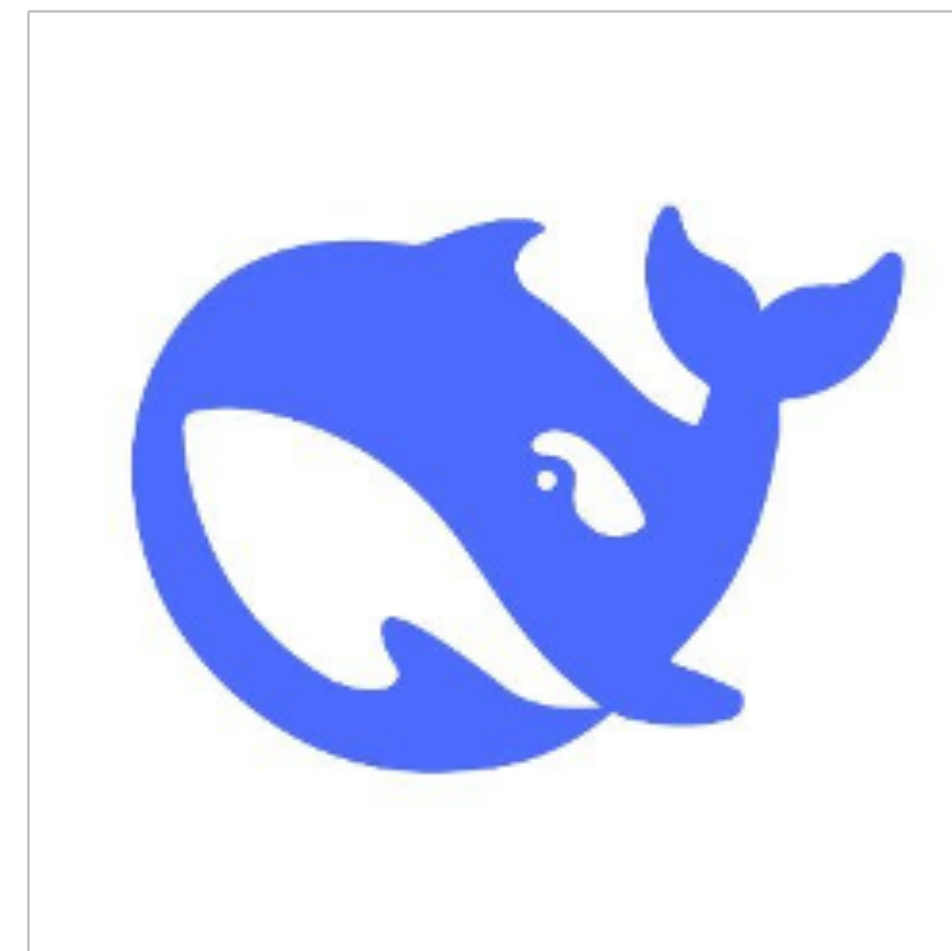
Grok



Claude



QWen



DeepSeek-R1

What will AI do to us?



! ChatGPT ↳ Inbox x



Professor Ian Holliday, Vice-President and Pro-Vice-Chancellor (Teaching and Learning) <hcm00101@hku.hk>

Fri, Feb 17, 2023, 4:12 PM



to mail.service ▾

Message from Professor Ian Holliday, Vice-President and Pro-Vice-Chancellor (Teaching and Learning)

Dear colleagues and students

You'll have heard of ChatGPT, one of many large language models sweeping the internet and shaking up global education.

We've started to consider the implications of AI-based tools for teaching and learning at HKU and plan to launch a broad-based campus debate involving both teachers and students. Since the implications are certain to be significant, however, it will take a while for us to settle on a long-term policy. Meanwhile, we're in the middle of a teaching semester with little room for manoeuvre as courses and assessments take place each week. We therefore need to adopt a short-term policy. This is it: [As an interim measure, we prohibit the use of ChatGPT or any other AI-based tool for all classroom, coursework and assessment tasks at HKU. Exemptions require written permission from course instructors. Students cannot provide themselves or other students with exemptions.](#)

Suspected violations of this interim policy will be treated as potential plagiarism cases. At HKU, plagiarism is defined as 'the use of another person's work (including but not limited to any materials, creations, ideas and data) as if one's own without due acknowledgement, whether or not such work has been published and regardless of the intent to deceive'. Making unacknowledged use of ChatGPT or another AI-based tool, treated for these purposes as 'another person', falls squarely within this definition. Teachers who suspect ChatGPT or another AI-based tool has been used may call a student in to discuss their work, set a supplementary oral examination, require a supplementary in-hall examination, or adopt other measures. Our full procedures for handling suspected plagiarism cases are set out here: https://intraweb.hku.hk/reserved_1/learn/plagiarism/Policy-on-Student-Plagiarism-in-UG-and-TPG.pdf.

Teachers with queries may contact colleagues in CETL and TELI. These two central units will also organize workshops on ChatGPT and other AI-based tools. Students with queries may consult their teachers. To be clear, our current policy is that, absent written permission, ChatGPT and other AI-based tools cannot be used for any credit-bearing activity at HKU.

Best wishes, Ian

Professor Ian Holliday
Vice-President (Teaching and Learning)
The University of Hong Kong

Feb 17, 2023: “ ... we prohibit the use of ChatGPT or any other AI-based tool for all classroom, coursework and assessment tasks at HKU. Exemptions require written permission from course instructors. Students cannot provide themselves or other students with exemptions.”

! Policy on Use of Generative Artificial Intelligence for Teaching and Learning Σ Inbox x 🖨 🔗



Professor Ian Holliday, Vice-President and Pro-Vice-Chancellor (Teaching and Learning) <hcm00101@hku.hk>

Fri, Jun 23, 2023, 6:35 PM



to mail.service ▾



Message from Professor Ian Holliday, Vice-President and Pro-Vice-Chancellor (Teaching and Learning)

Dear colleagues

Four months ago we put a temporary halt to student use of ChatGPT and generative AI for all classroom, coursework and assessment tasks at HKU. This was in response to student concern about unregulated use of AI tools particularly for assessment tasks. We placed no constraints on teachers and many experimented with generative AI during their S2 courses.

At the same time, we committed to launching a broad campus debate on the implications of AI-based tools for T&L at HKU. The Generative AI Task Force was formed with both staff and student members, CETL hosted two workshops at the start of March, and then a series of five GAITF policy-driven explorations in May. Many hundreds of teachers and students participated f2f or online. In parallel, we rapidly opened up free staff access to ChatGPT and other generative AI tools.

Earlier this month, GAITF condensed a wide-ranging debate into a fairly brief UG/TPG policy paper that was sent for consultation to the Faculties and several central units. It was also circulated to Academic Board and the Teaching and Learning Quality Committee. It has now been endorsed by Senate and can be viewed [here](#) (with HKU portal login).

Generative AI has many implications for T&L at HKU and ultimately will reshape pedagogy in almost all courses. However, the most pressing impacts fall in the sphere of assessment. For sure students in S1, 2023-24 will share the concerns expressed by students in S2, 2022-23 about academic integrity, equity and fairness. We therefore need to work on this during the summer.

S1 teaching will start 10 weeks today. The Senate paper calls on teachers to review their assessment tasks before then. It assumes that there are currently no reliable means for detecting plagiarism and notes that some forms of assessment are thus especially vulnerable to generative AI. It asks the Faculties to create fast-track course revision approval procedures by the end of this month.

The Teaching and Learning Innovation Centre (TALIC), which on July 1 will amalgamate CETL, TELI and T&LEMU, stands ready to support teachers through email, phone and WhatsApp helplines, a drop-in advisory service and a teachers' guide. Faculties have already received advice on potentially problematic assessment tasks. If you're not sure what to do about your course assessments, please email TALIC at talic@hku.hk.

I wouldn't normally send out such a long email. It's simply a time of great change in T&L at HKU and we need to do everything we can to stay abreast of it all so we can continue to deliver for our students. Grateful thanks in advance for your assistance with reviewing your courses and revising your assessment tasks!

Best wishes for the summer – Ian

Professor Ian Holliday
VP/T&L

July 23, 2023: “ ... We placed **no constraints** on teachers and many experimented with generative AI during their S2 courses.”

What will AI do to us?



INTJ

INNOVATIVE, INDEPENDENT,
STRATEGIC, LOGICAL,
RESERVED, INSIGHTFUL.
DRIVEN BY THEIR OWN
ORIGINAL IDEAS TO
ACHIEVE IMPROVEMENTS.

INTP

INTELLECTUAL, LOGICAL,
PRECISE, RESERVED,
FLEXIBLE, IMAGINATIVE.
ORIGINAL THINKERS WHO
ENJOY SPECULATION AND
CREATIVE PROBLEM
SOLVING.

ENTJ

STRATEGIC, LOGICAL,
EFFICIENT, OUTGOING,
AMBITIOUS, INDEPENDENT.
EFFECTIVE ORGANIZERS OF
PEOPLE AND LONG-RANGE
PLANNERS.

ENTP

INVENTIVE, ENTHUSIASTIC,
STRATEGIC, ENTERPRISING,
INQUISITIVE, VERSATILE.
ENJOY NEW IDEAS AND
CHALLENGES, VALUE
INSPIRATION.

INFJ

IDEALISTIC, ORGANIZED,
INSIGHTFUL, DEPENDABLE,
COMPASSIONATE, GENTLE.
SEEK HARMONY AND
COOPERATION; ENJOY
INTELLECTUAL
STIMULATION.

INFP

SENSITIVE, CREATIVE,
IDEALISTIC, PERCEPTIVE,
CARING, LOYAL.
VALUE INNER HARMONY
AND PERSONAL GROWTH,
FOCUS ON DREAMS AND
POSSIBILITIES.

ENFJ

CARING, ENTHUSIASTIC,
IDEALISTIC, ORGANIZED,
DIPLOMATIC, RESPONSIBLE.
SKILLED COMMUNICATORS
WHO VALUE CONNECTION
WITH PEOPLE.

ENFP

ENTHUSIASTIC, CREATIVE,
SPONTANEOUS,
OPTIMISTIC, SUPPORTIVE,
PLAYFUL..
VALUE INSPIRATION, ENJOY
STARTING NEW PROJECTS,
SEE POTENTIAL IN OTHERS.

ISTJ

RESPONSIBLE, SINCERE,
ANALYTICAL, RESERVED,
REALISTIC, SYSTEMATIC.
HARDWORKING AND
TRUSTWORTHY WITH
SOUND PRACTICAL
JUDGEMENT.

ISFJ

WARM, CONSIDERATE,
GENTLE, RESPONSIBLE,
PRAGMATIC, THOROUGH.
DEVOTED CARETAKERS
WHO ENJOY BEING
HELPFUL TO OTHERS.

ESTJ

EFFICIENT, OUTGOING,
ANALYTICAL, SYSTEMATIC,
DEPENDABLE, REALISTIC.
LIKE TO RUN THE SHOW
AND GET THINGS DONE IN
AN ORDERLY FASHION.

ESFJ

FRIENDLY, OUTGOING,
RELIABLE, CONSCIENTIOUS,
ORGANIZED, PRACTICAL.
SEEK TO BE HELPFUL AND
PLEASE OTHERS, ENJOY
BEING ACTIVE AND
PRODUCTIVE.

ISTP

ACTION-ORIENTED,
LOGICAL, ANALYTICAL,
SPONTANEOUS, RESERVED,
INDEPENDENT.
ENJOY ADVENTURE,
SKILLED AT
UNDERSTANDING THINGS.

ISFP

GENTLE, SENSITIVE,
NURTURING, HELPFUL,
FLEXIBLE, REALISTIC.
SEEK TO CREATE A
PERSONAL ENVIRONMENT
THAT IS BOTH BEAUTIFUL
AND PRACTICAL.

ESTP

OUTGOING, REALISTIC,
ACTION-ORIENTED,
CURIOUS, VERSATILE,
SPONTANEOUS.
PRAGMATIC PROBLEM
SOLVERS AND SKILLFUL
NEGOTIATORS.

ESFP

PLAYFUL, ENTHUSIASTIC,
FRIENDLY, SPONTANEOUS,
TACTFUL, FLEXIBLE.
HAVE A STRONG COMMON
SENSE, ENJOY HELPING
PEOPLE IN TANGIBLE WAYS.



Table 1: MBTI types for LLMs.

	Type	Personality Descriptions
ChatGPT	ENTJ	self-confident, decisive, and possess innate leadership skills.
GPT-4*	INTJ	experts skilled in achieving their own goals.
Bloom7b	ISTJ	pragmatic, responsible, values tradition and loyalty.
BaiChuan7b	ENFP	smart, curious, and imaginative.
BaiChuan13b	INFP	highly adaptable and idealistic
OpenLlama7b	INFJ	has strong insight into people and adheres to one's own values.



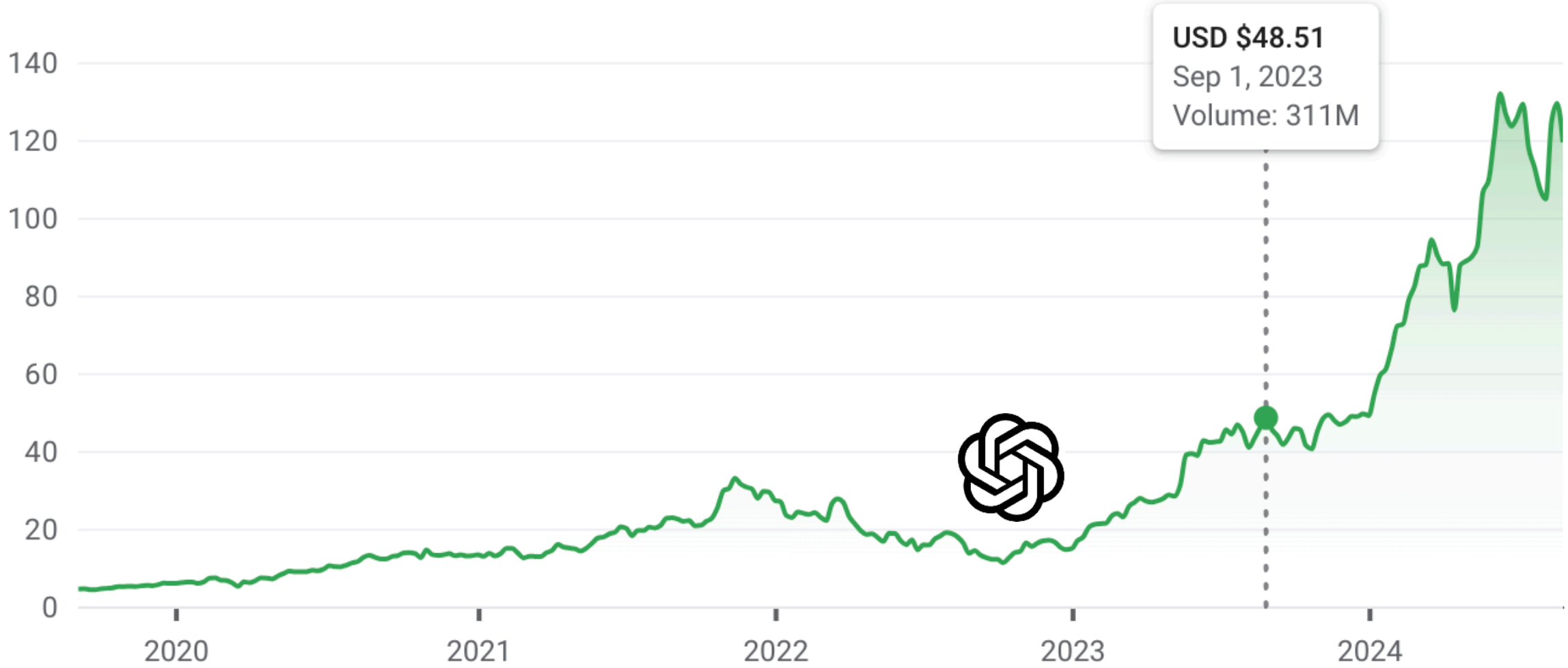
What's happening?

NVIDIA Corp

\$119.37 ↑ 2,570.47% +114.90 5Y

After Hours: **\$119.23** (↓ 0.12%) **-0.14**
Closed: Aug 30, 7:59:56 PM UTC-4 · USD · NASDAQ · Disclaimer

1D 5D 1M 6M YTD 1Y 5Y MAX



What's happening?

Chinchilla Optimal LLM Training Cost				
Model	Size (# Parameters)	Tokens	GPU	Optimal Training Compute Cost
1 Billion Parameter	1 Billion	20 Billion	A100	\$ 294
10 Billion Parameter	10 Billion	205 Billion	A100	\$ 29,866
67 Billion Parameter	67 Billion	1,500 Billion	A100	\$ 1,398,601
175 Billion Parameter	175 Billion	3,700 Billion	A100	\$ 9,348,291
280 Billion Parameter	280 Billion	5,900 Billion	A100	\$ 24,038,462
520 Billion Parameter	520 Billion	11,000 Billion	A100	\$ 83,284,771
1 Trillion Parameter	1,000 Billion	21,200 Billion	A100	\$ 308,372,183
10 Trillion Parameter	10,000 Billion	216,200 Billion	A100	\$ 28,935,185,185

Should we have predicted all this?

Training compute (FLOPs) of milestone Machine Learning systems over time
n = 121



Figure 1: Trends in $n = 121$ milestone ML models between 1952 and 2022. We distinguish three eras. Notice the change of slope circa 2010, matching the advent of Deep Learning; and the emergence of a new large-scale trend in late 2015.

Should we have predicted all this?

Training compute (FLOPs) of milestone Machine Learning systems over time

$n = 102$

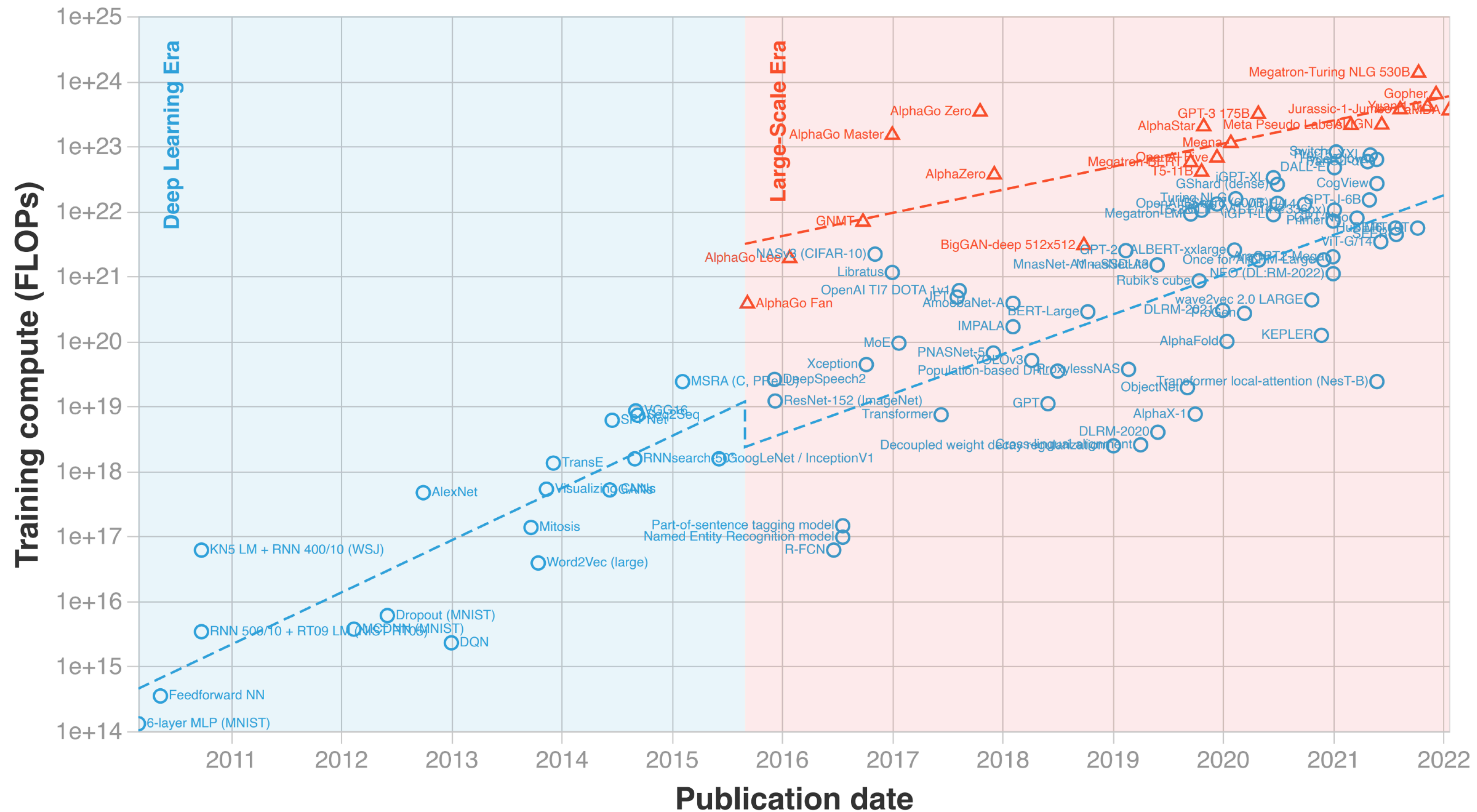
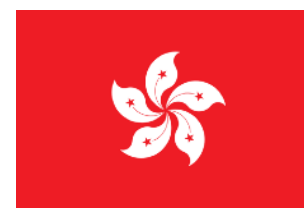
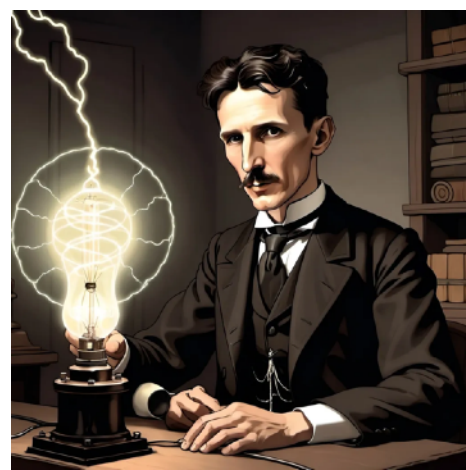


Figure 3: Trends in training compute of $n=102$ milestone ML systems between 2010 and 2022. Notice the emergence of a possible new trend of large-scale models around 2016. The trend in the remaining models stays the same before and after 2016.

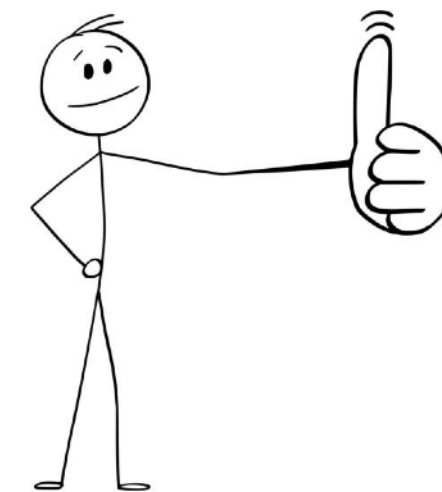
Will the trend continue?



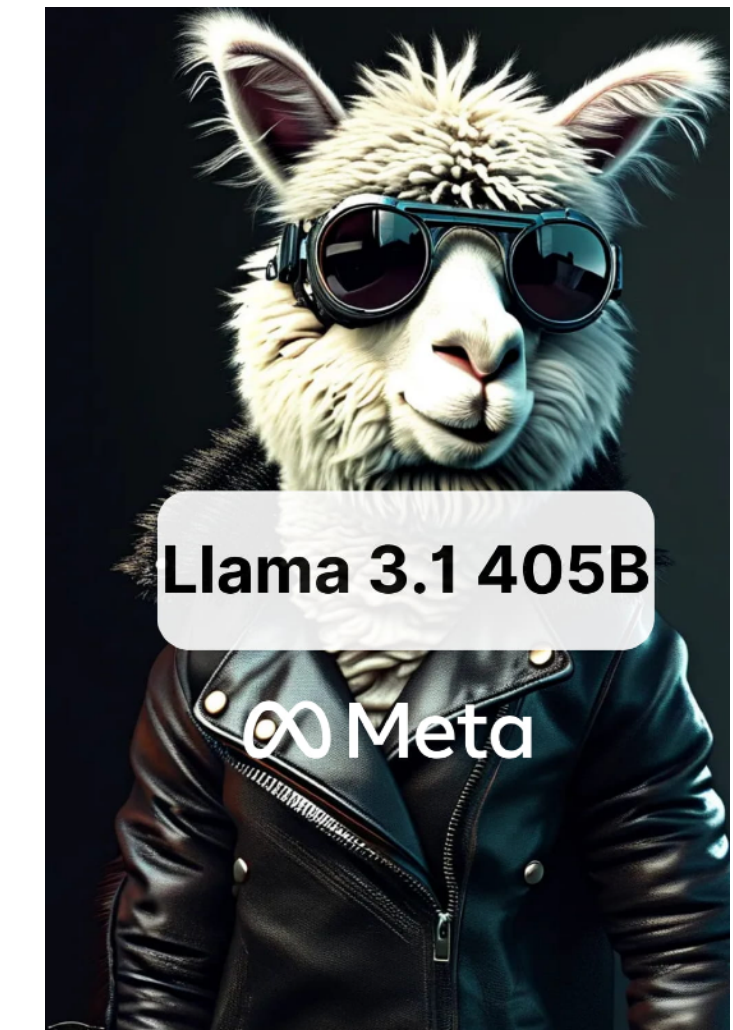
H100 1 Hour \approx 0.7 kWh



Hong Kong: 0.14 TWh / day



\sim 1 Million Llama 3.1 405B

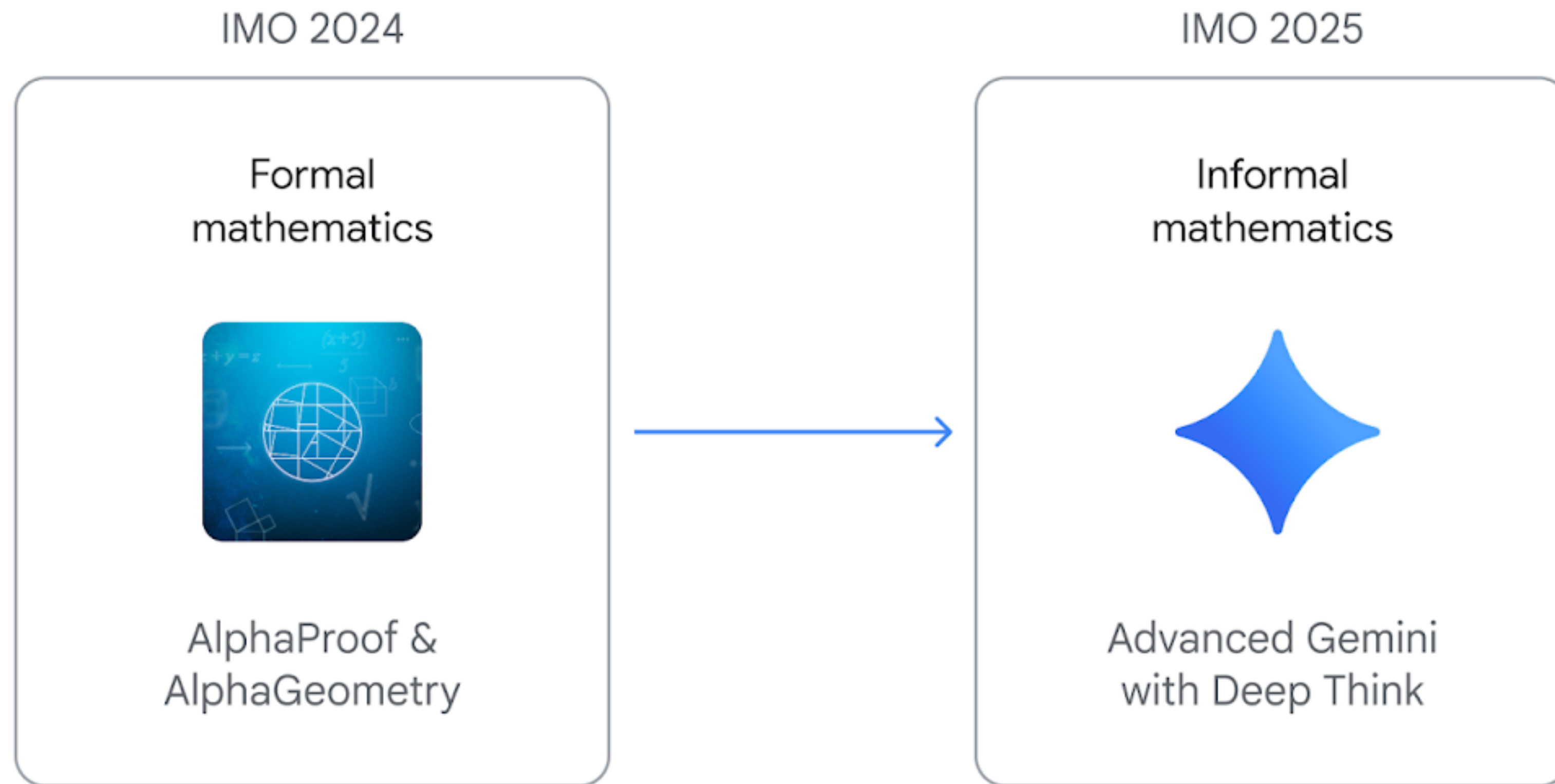


Inference Running on 8 H100 GPU



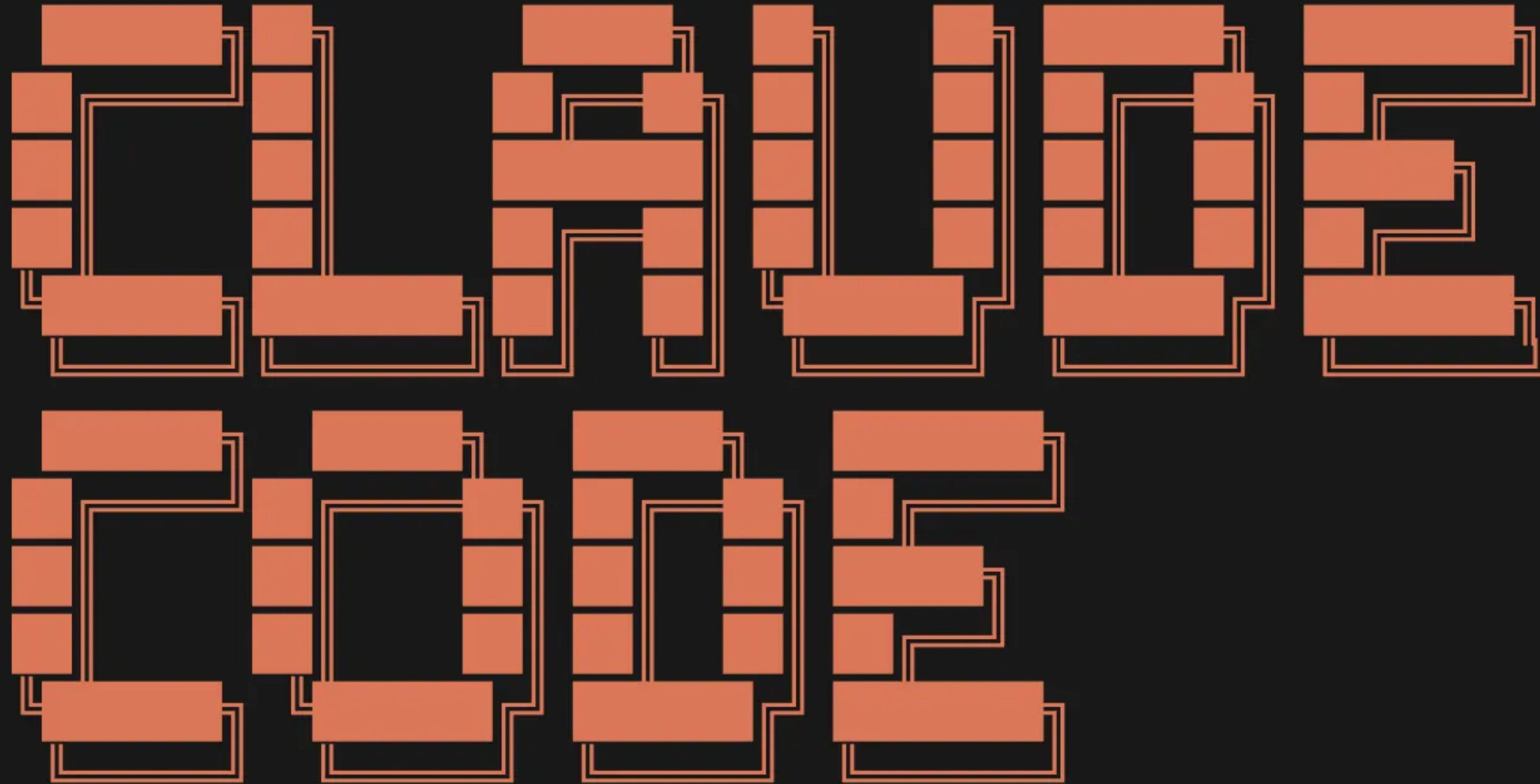
\sim 4 Million iPhones in Hong Kong

Testing the Upper Limits of Reasoning Ability



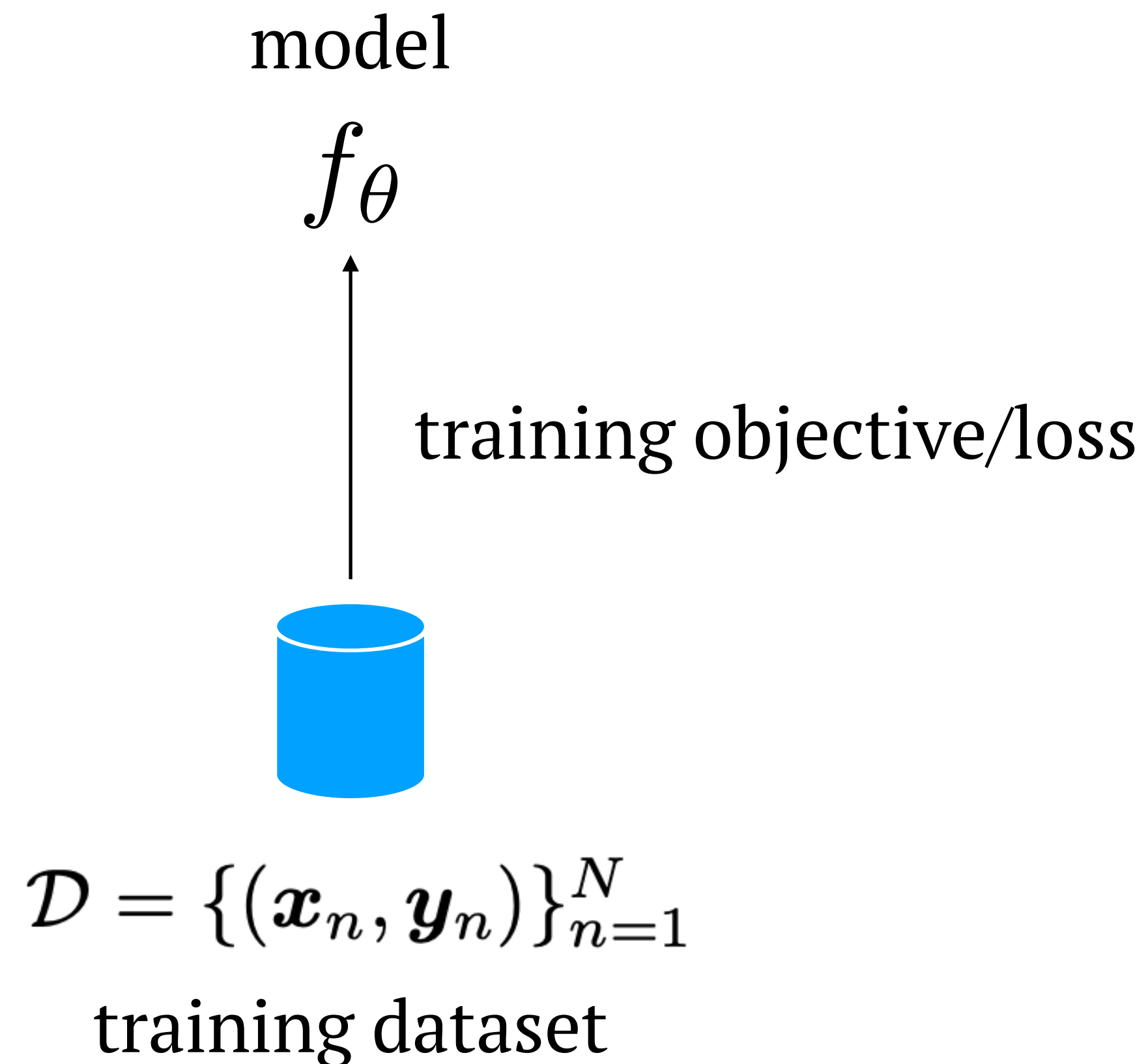


* Welcome to Claude Code

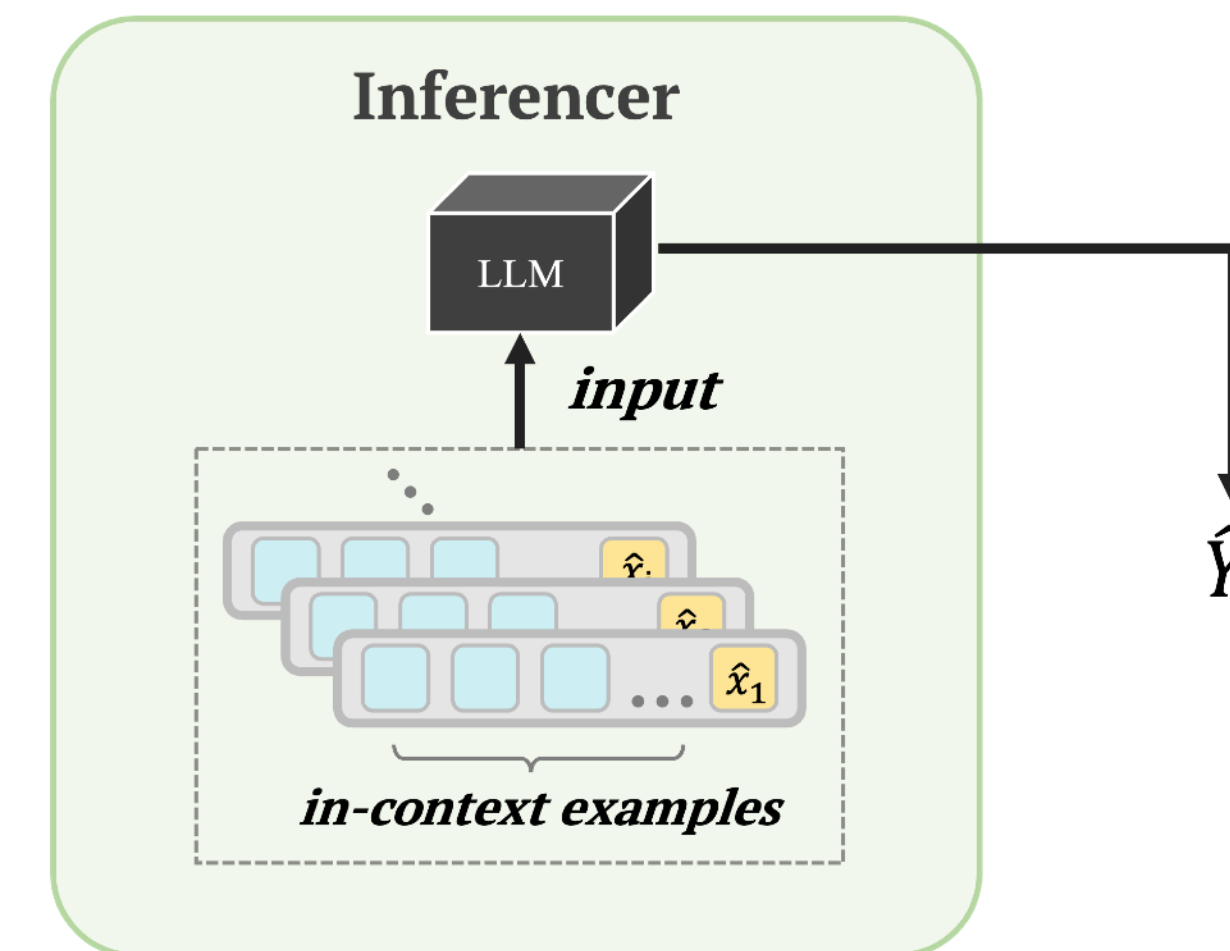
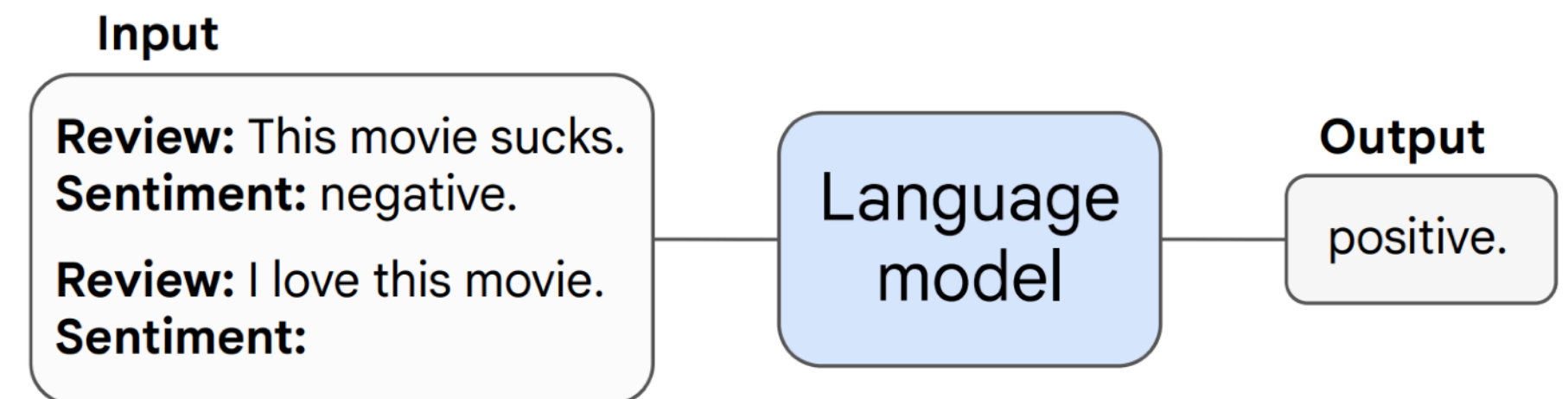


Press **Enter** to continue

Machine Learning (Today)



Supervised Learning



Zero-shot / In-Context Learning

Why not every AI is that powerful?

What can I help you with?

Play a good song.

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

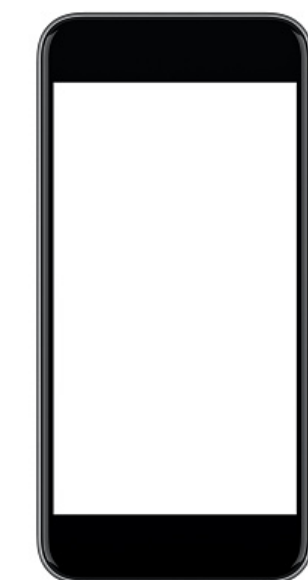


You need to do a better job understanding me.

Noted.

Yeah, make a note of that.

Here's your note:



What is NLP? Wait, what is language?

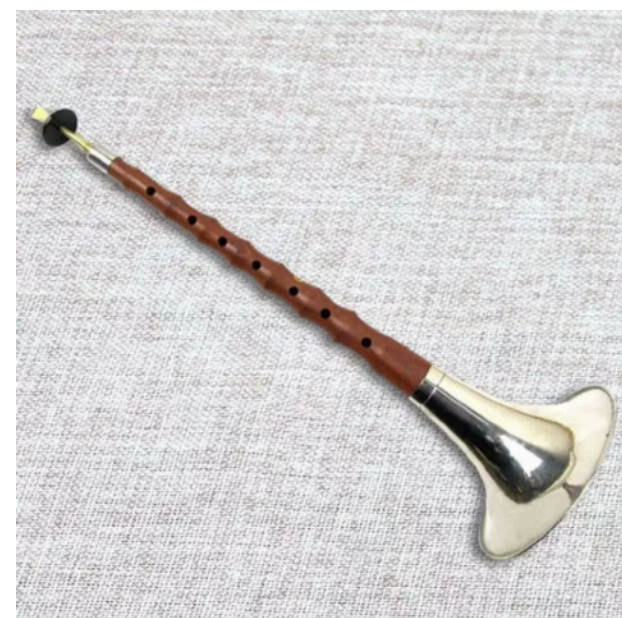
The abstraction of the real world – different languages take you to different worlds!



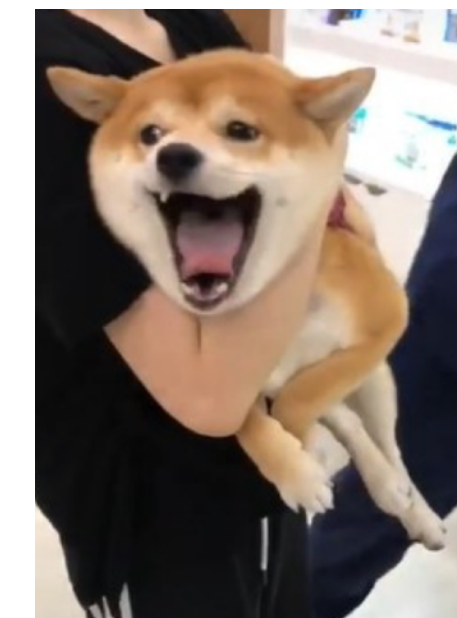
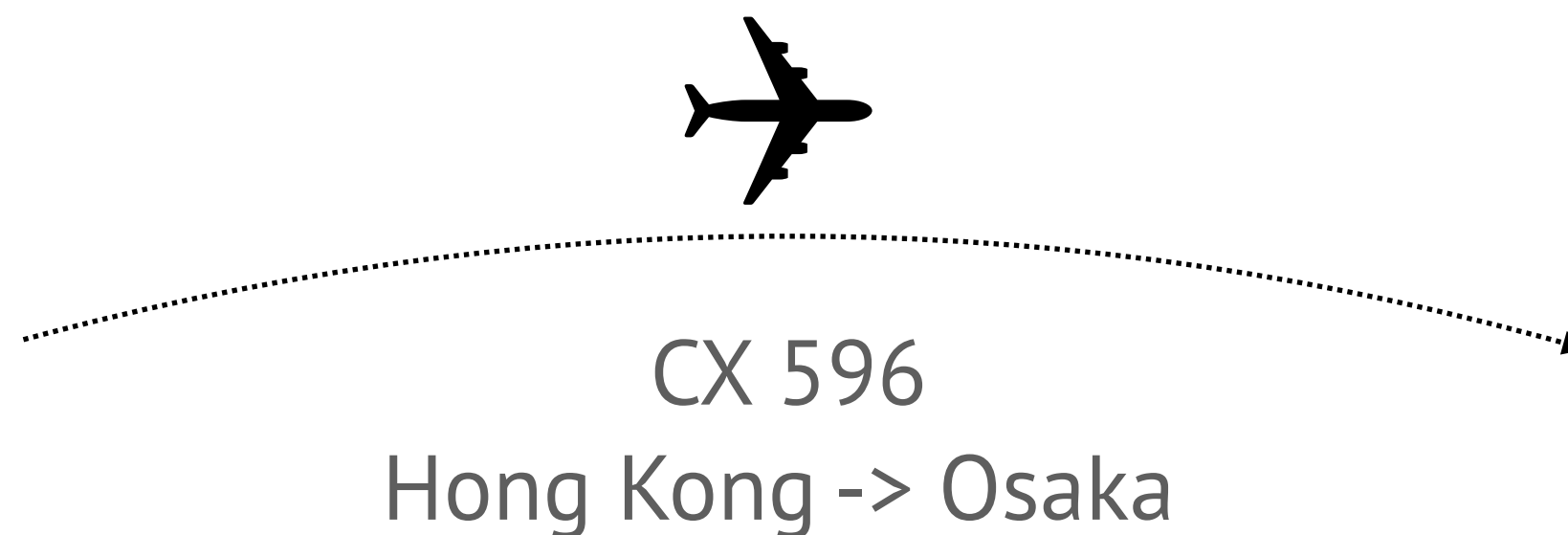
餃子



dumplings



唄



チャルメラ

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

Do AI “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)

A photograph of a chalkboard with the mathematical formula for entropy written in white chalk. The formula is $H = -\sum p(x) \log p(x)$. The chalkboard is dark, and the formula is the central focus. The board is framed by wooden borders on the left and right sides.

Information Theory; Entropy

Language models, and how to build it.



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



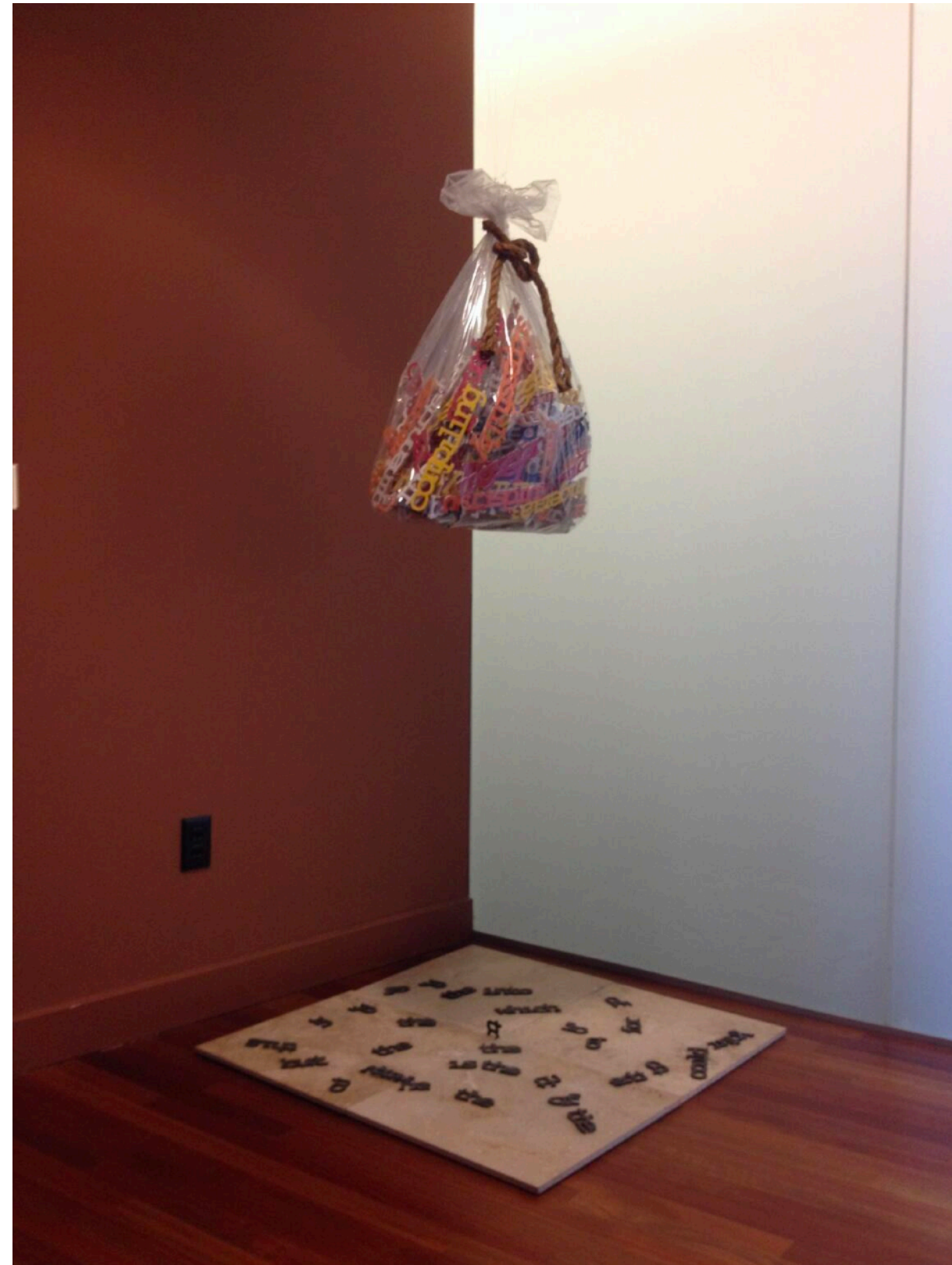
Transformers, neural networks and many others
(powerful functions, and inside configurations)

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

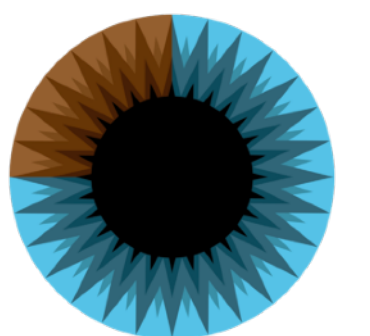
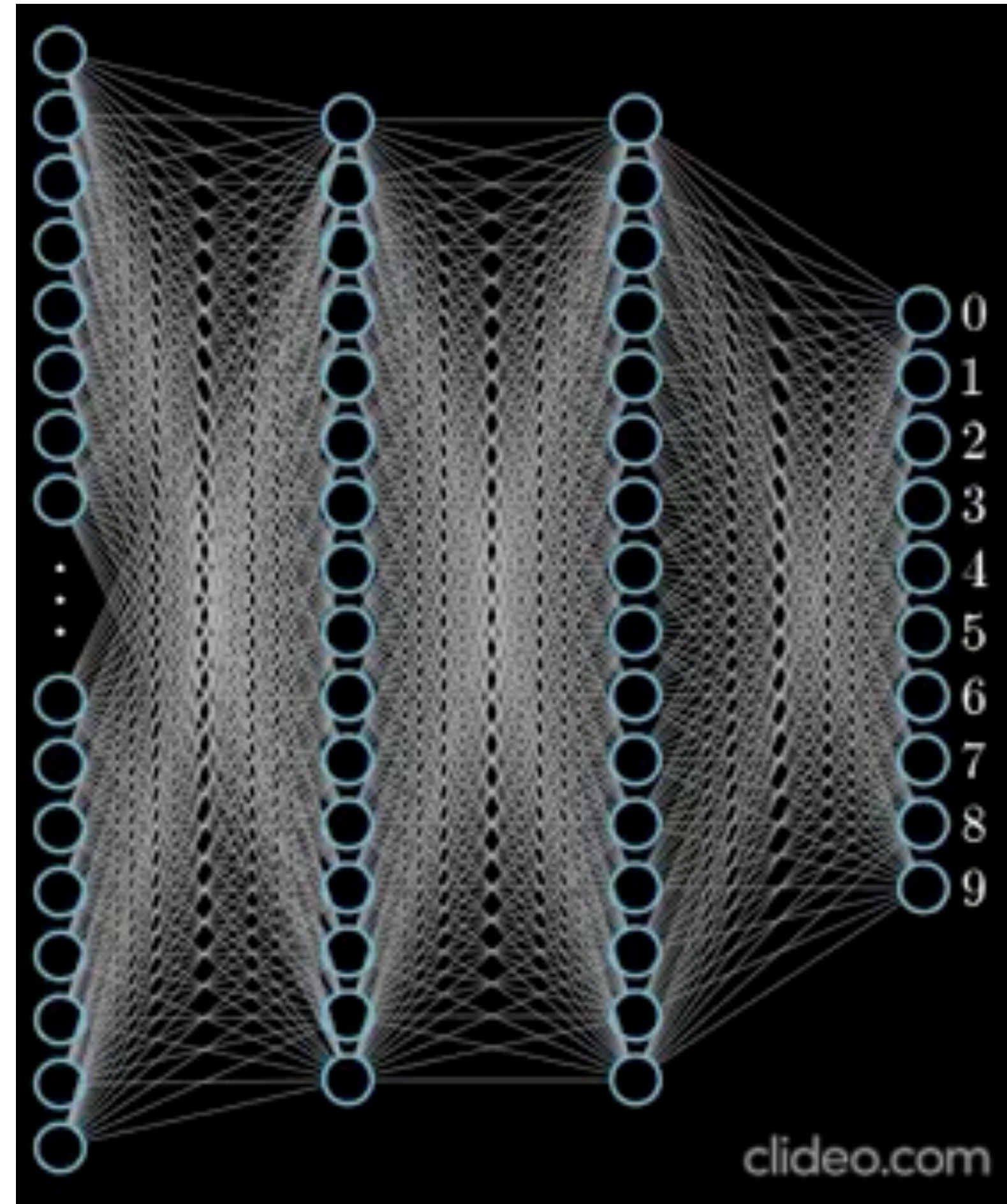
Generative Language Model

I am going to do an internship in Google

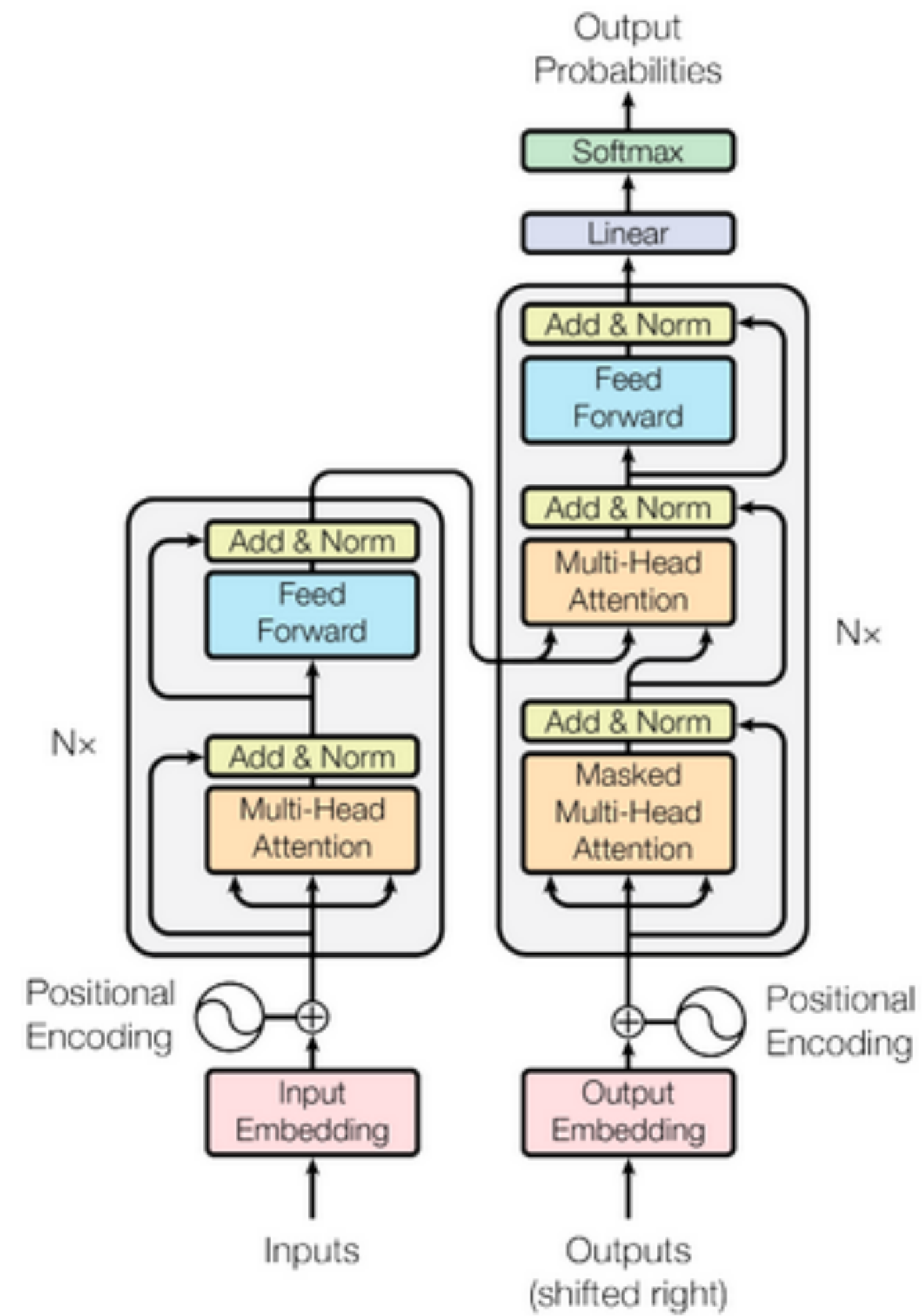


Google

Neutralize the dice!

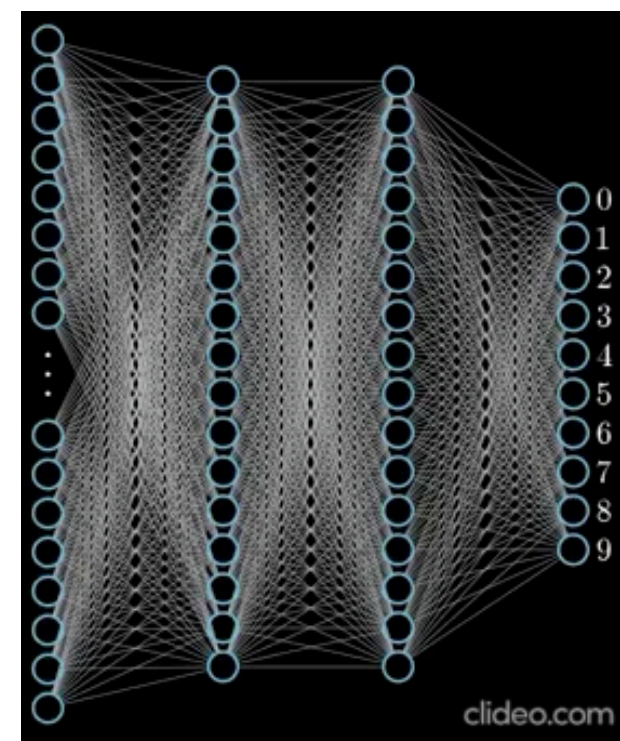


Neural Networks (e.g. Transformers)



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)



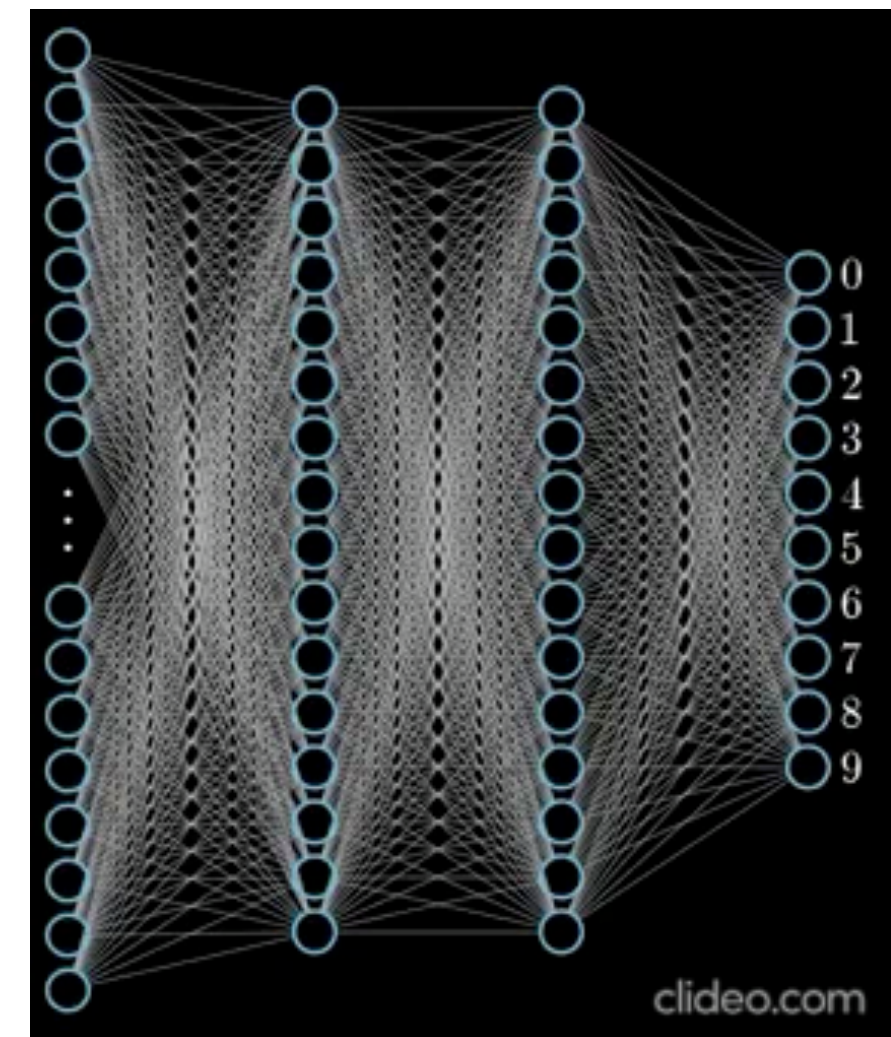
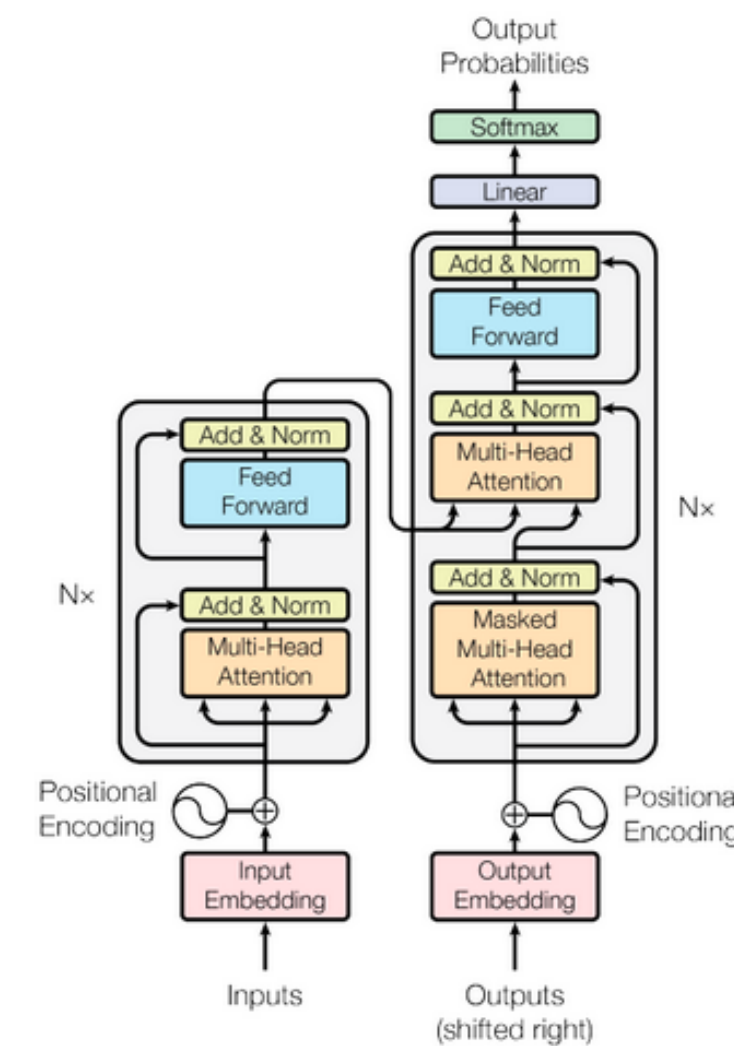
Transformers, neural networks and many others
(powerful functions)

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

Learn



Parameterize



First problem – the language modeling problem

Given a finite vocabulary

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

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

$\langle s \rangle$ I am going to an internship in Google $\langle /s \rangle$

$\langle s \rangle$ an internship in Google $\langle /s \rangle$

$\langle s \rangle$ I am going going $\langle /s \rangle$

$\langle s \rangle$ Google is am $\langle /s \rangle$

$\langle s \rangle$ internship is going $\langle /s \rangle$

Formally:

$$p(x_1, x_2, \dots, x_n)$$

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

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:

$$p(\langle s \rangle \text{ I am going to an internship in Google } \langle /s \rangle) = 10^{-12}$$

$$p(\langle s \rangle \text{ an internship in Google } \langle /s \rangle) = 10^{-8}$$

$$p(\langle s \rangle \text{ I am going going } \langle /s \rangle) = 10^{-15}$$

...

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

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

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?

$$P(\text{<s> Sam I am </s>}) = 1/3$$

$$P(\text{<s> I am </s>}) = 0$$

$$P(\text{<s> I am Sam </s>}) = 1/3$$

$$P(\text{<s> green Sam </s>}) = 0$$

$$P(\text{<s> I do not like green eggs and ham </s>}) = 1/3$$

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”:

$$P(I \mid \langle s \rangle) = 2/3$$

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

$$P(I \mid \langle s \rangle) = 2/3$$

$$P(\langle /s \rangle \mid \text{Sam}) = 1/2$$

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

$$P(I \mid \langle s \rangle) = 2/3$$

$$P(\text{am} \mid I) = ?$$

$$P(\langle s \rangle \mid \text{Sam}) = 1/2$$

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

$$P(I \mid \langle s \rangle) = 2/3$$

$$P(\text{am} \mid I) = 2/3$$

$$P(\langle s \rangle \mid \text{Sam}) = 1/2$$

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/comp7607-fall2025>

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% final project

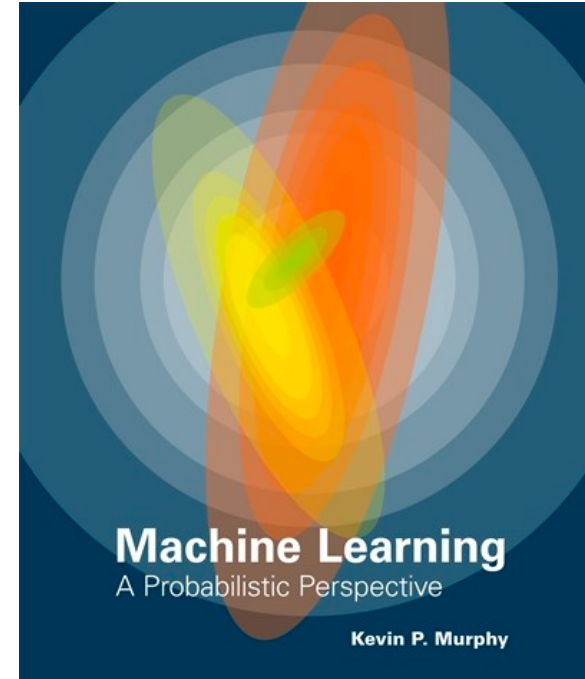
TA:

For 7607A: Gao Peifeng, Wang Jiannan

For 7607B: Xiong Jing, Xie Yuhan

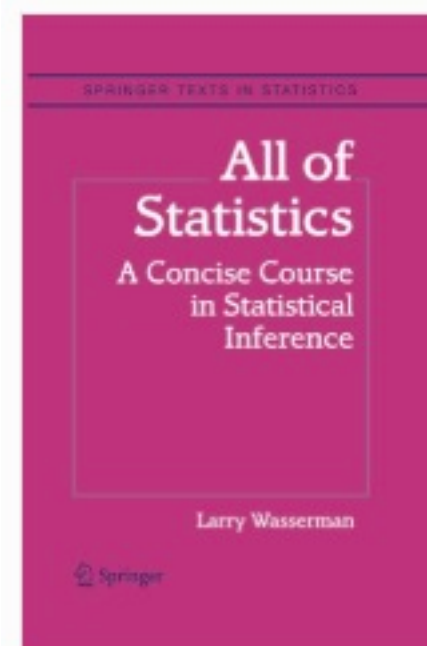
Course Logistics

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

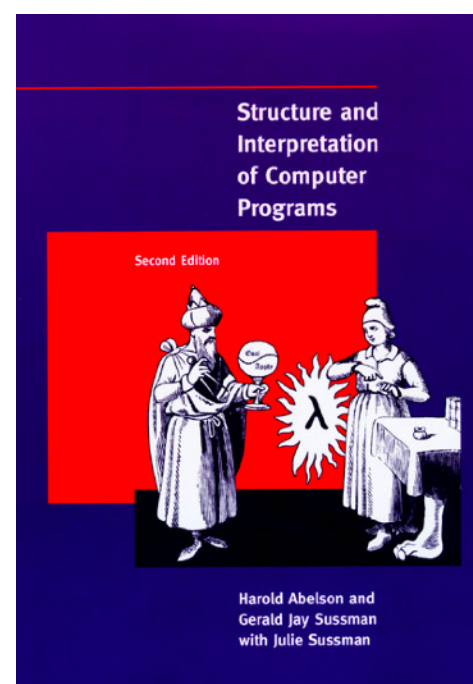


Supervised learning, unsupervised learning, regression, classification, loss function, neural networks, regularization ...

(COMP3314)



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/>

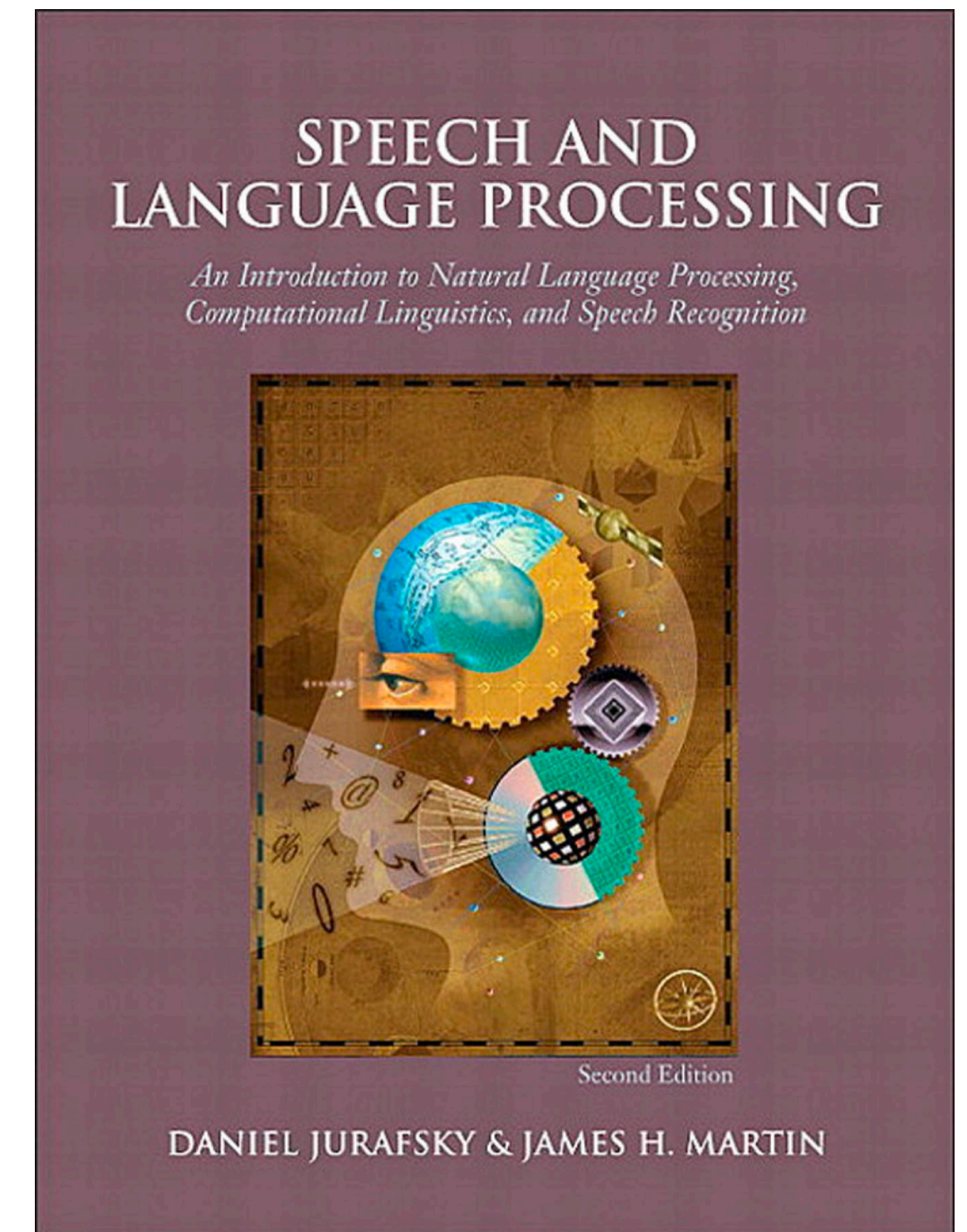
Assessments (in total ~3):

Programming problems

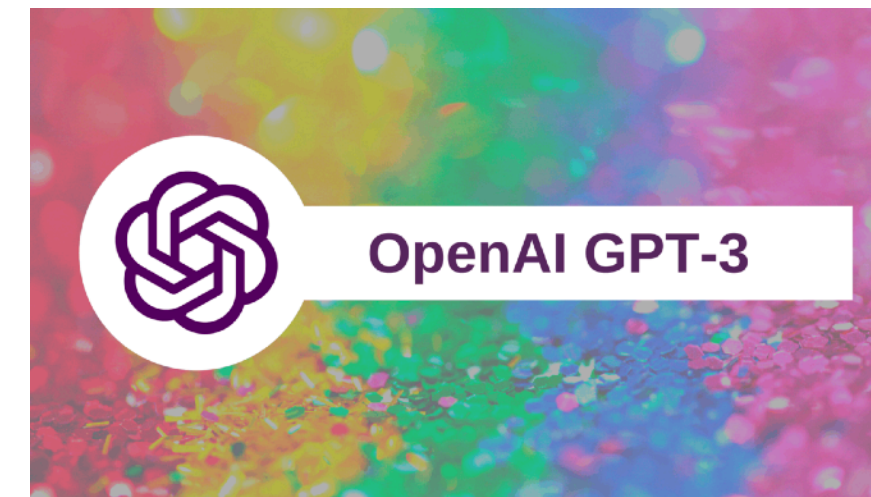
Problem sets

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 ...

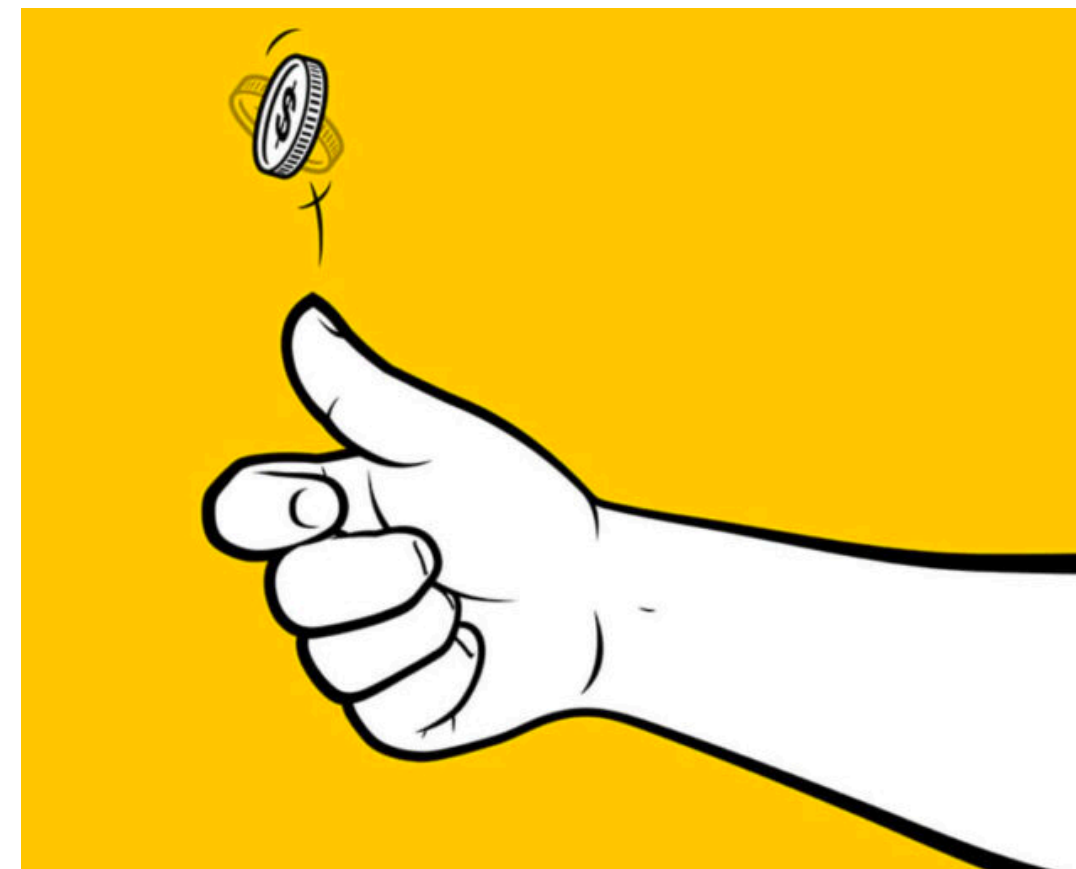
What is probability?



$$P(\text{head}) = 50\%$$

Frequentist

Bayesian



events (repeated trials)



uncertainty (ignorance)

Probability

Joint probability:

$$\Pr(A \wedge B) = \Pr(A, B)$$

Probability of a union of two events:

$$\Pr(A \vee B) = \Pr(A) + \Pr(B) - \Pr(A \wedge B)$$

Conditional probability:

$$\Pr(B|A) \triangleq \frac{\Pr(A, B)}{\Pr(A)}$$

If A and B are independent events:

$$\Pr(A, B) = \Pr(A) \Pr(B)$$

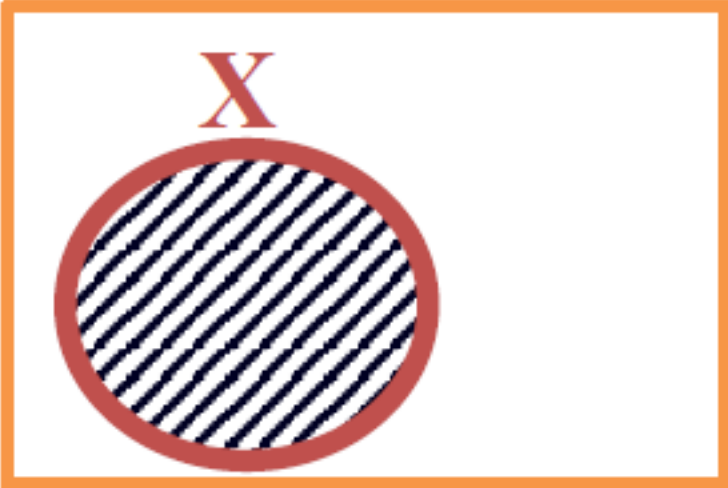
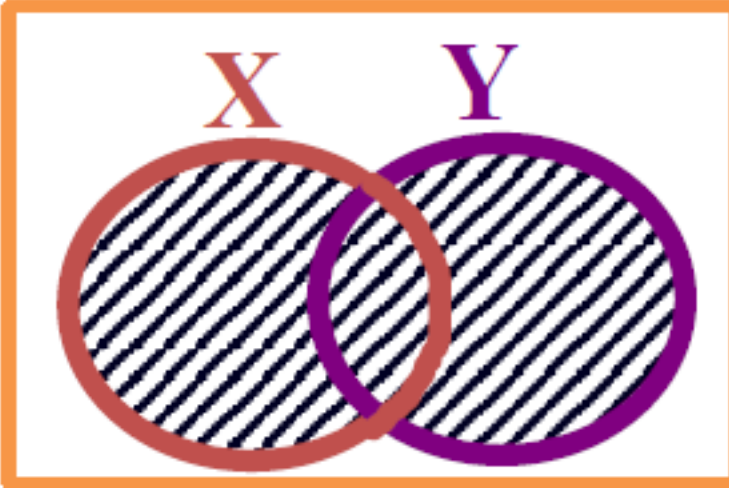
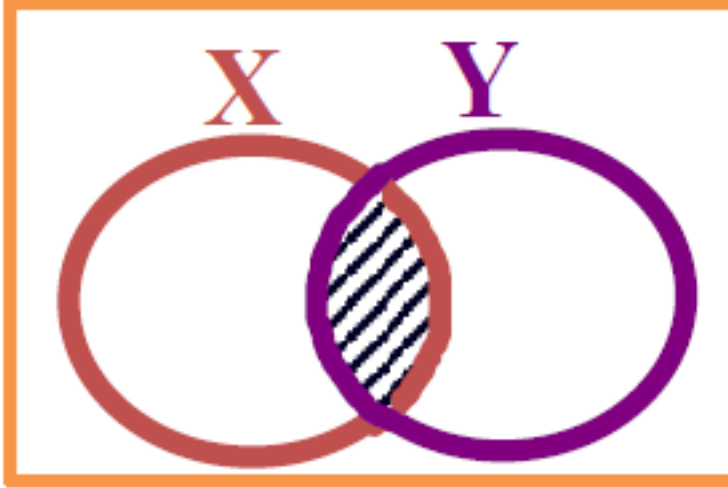

If the events are mutually exclusive:

$$\Pr(A \vee B) = \Pr(A) + \Pr(B)$$

Conditional independence:

$$\Pr(A, B|C) = \Pr(A|C) \Pr(B|C)$$

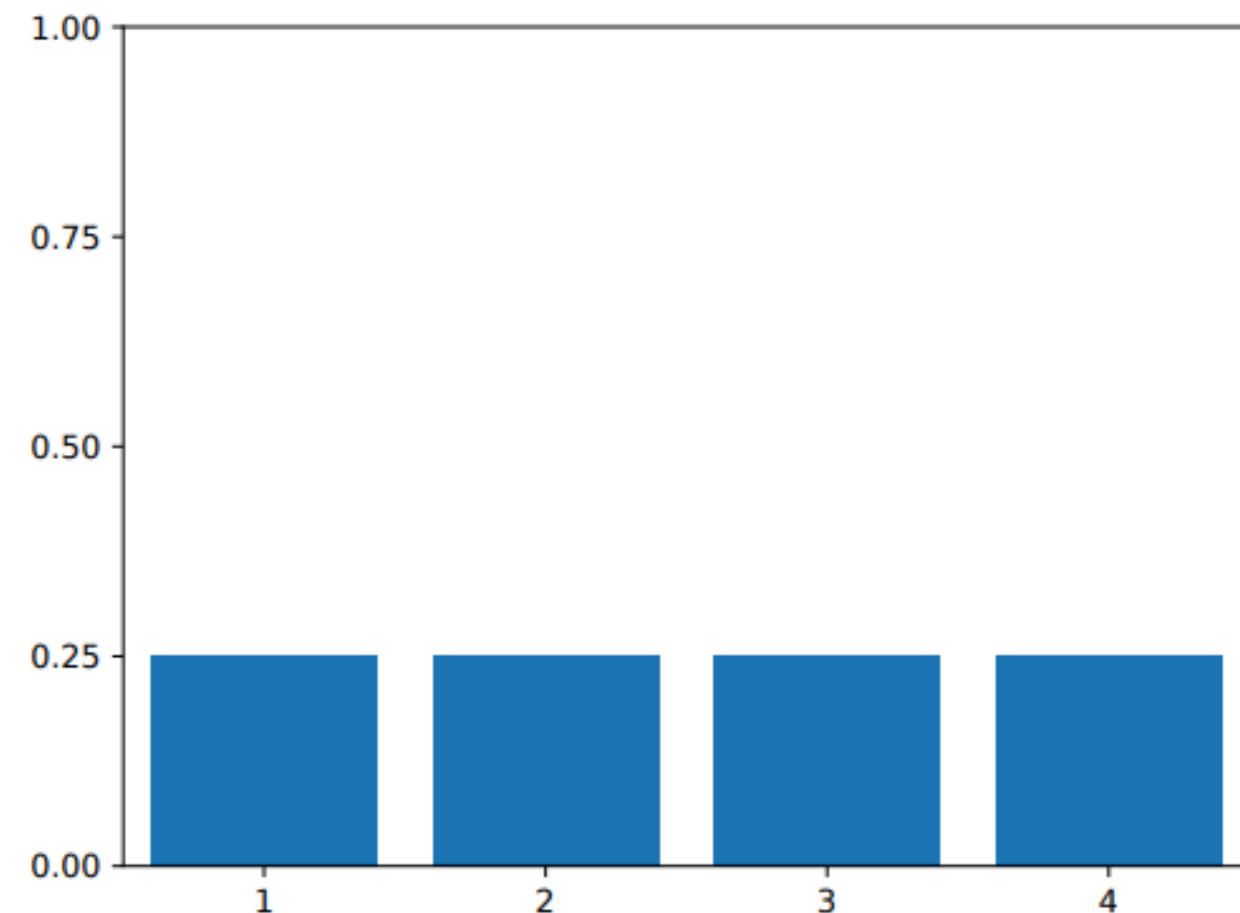
Probability

Marginal	Union	Joint	Conditional
$P(X)$ The probability of X occurring	$P(X \cup Y)$ The probability of X or Y occurring	$P(X \cap Y)$ The probability of X and Y occurring	$P(X Y)$ The probability of X occurring given that Y has occurred
			

Random Variables

X represents some unknown quantity of interest

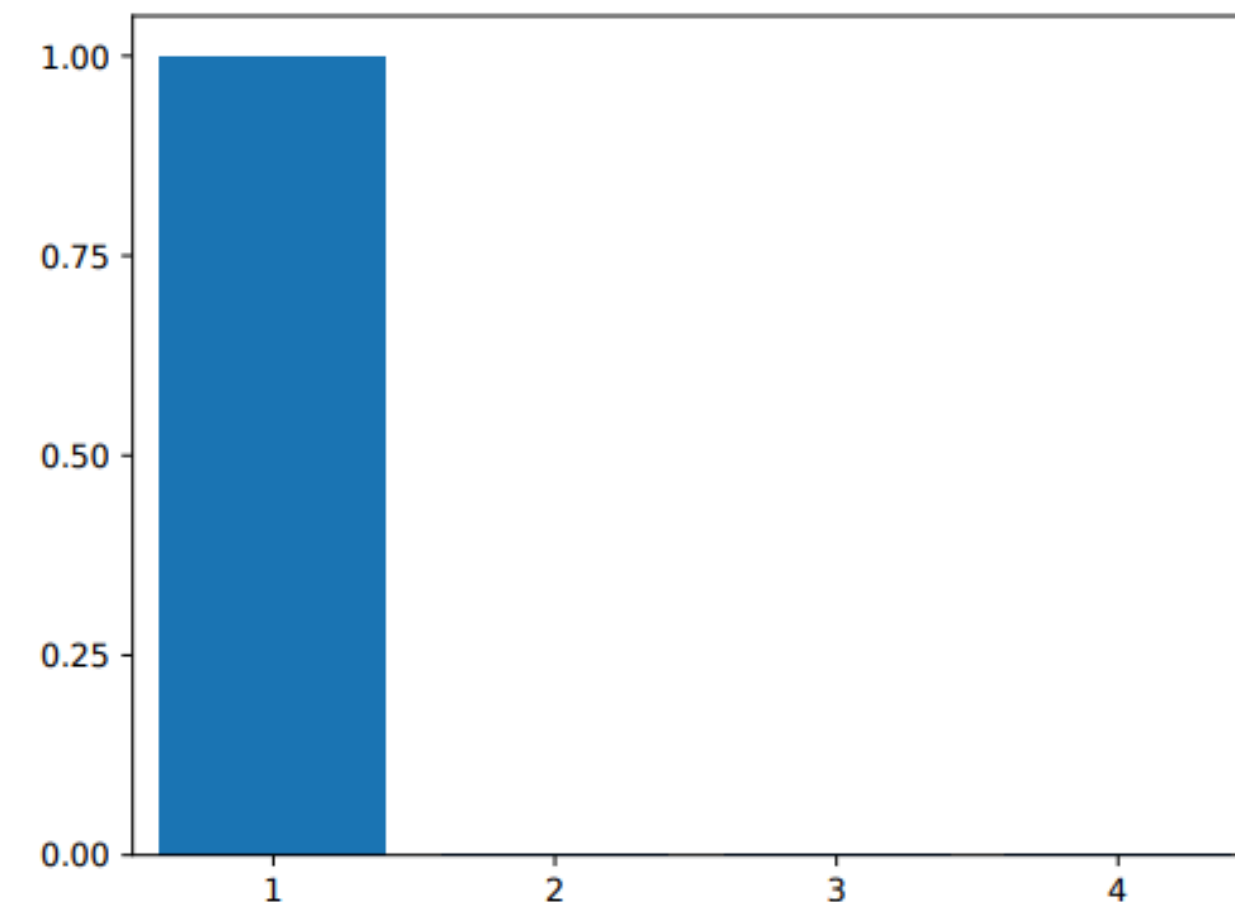
We call it a **random variable** if the value of X is unknown and/or could change.



(a)

uniform distribution

probability mass function (pmf):

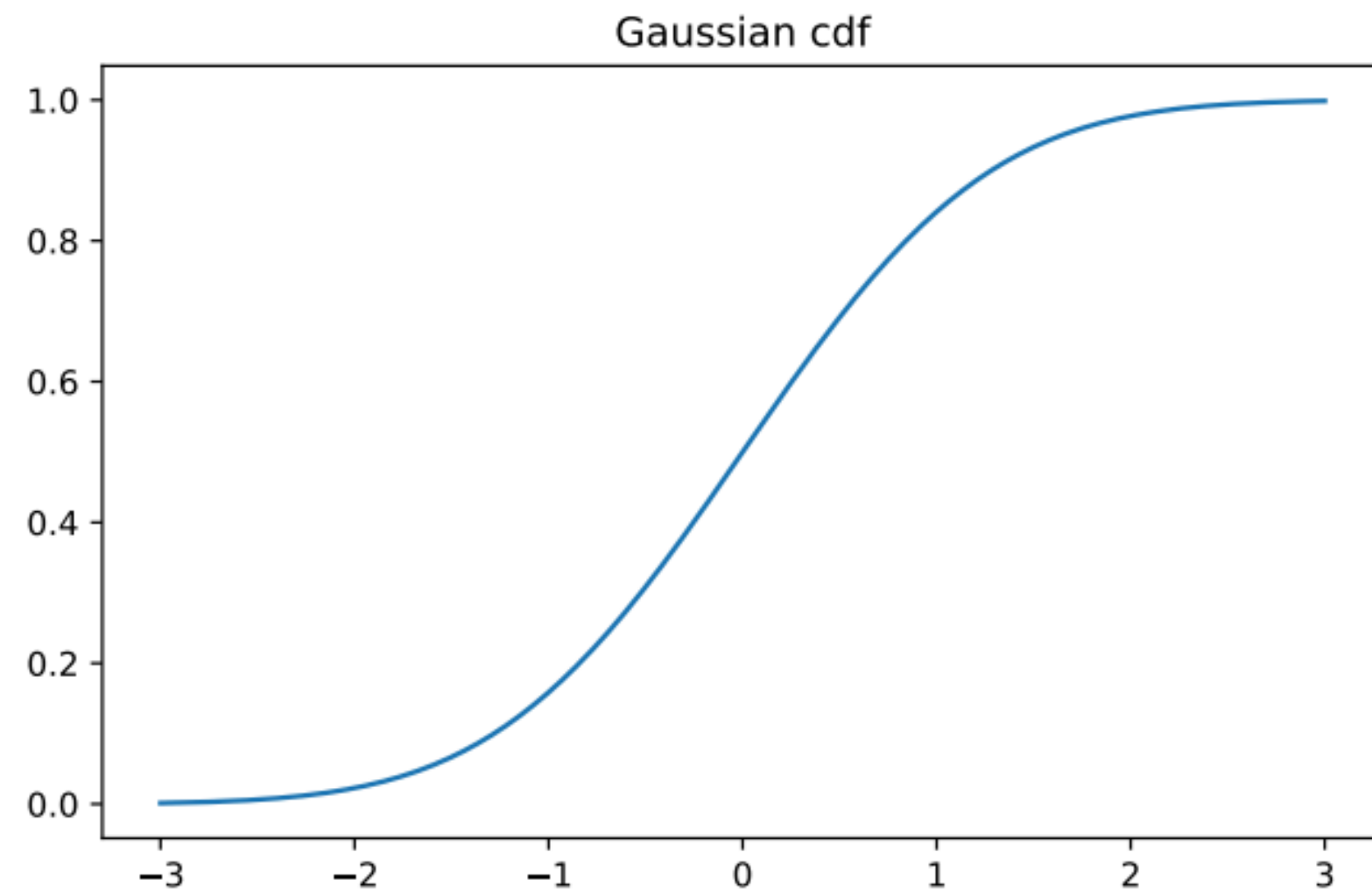


(b)

degenerate distribution
(delta function)

$$p(x) \triangleq \Pr(X = x)$$

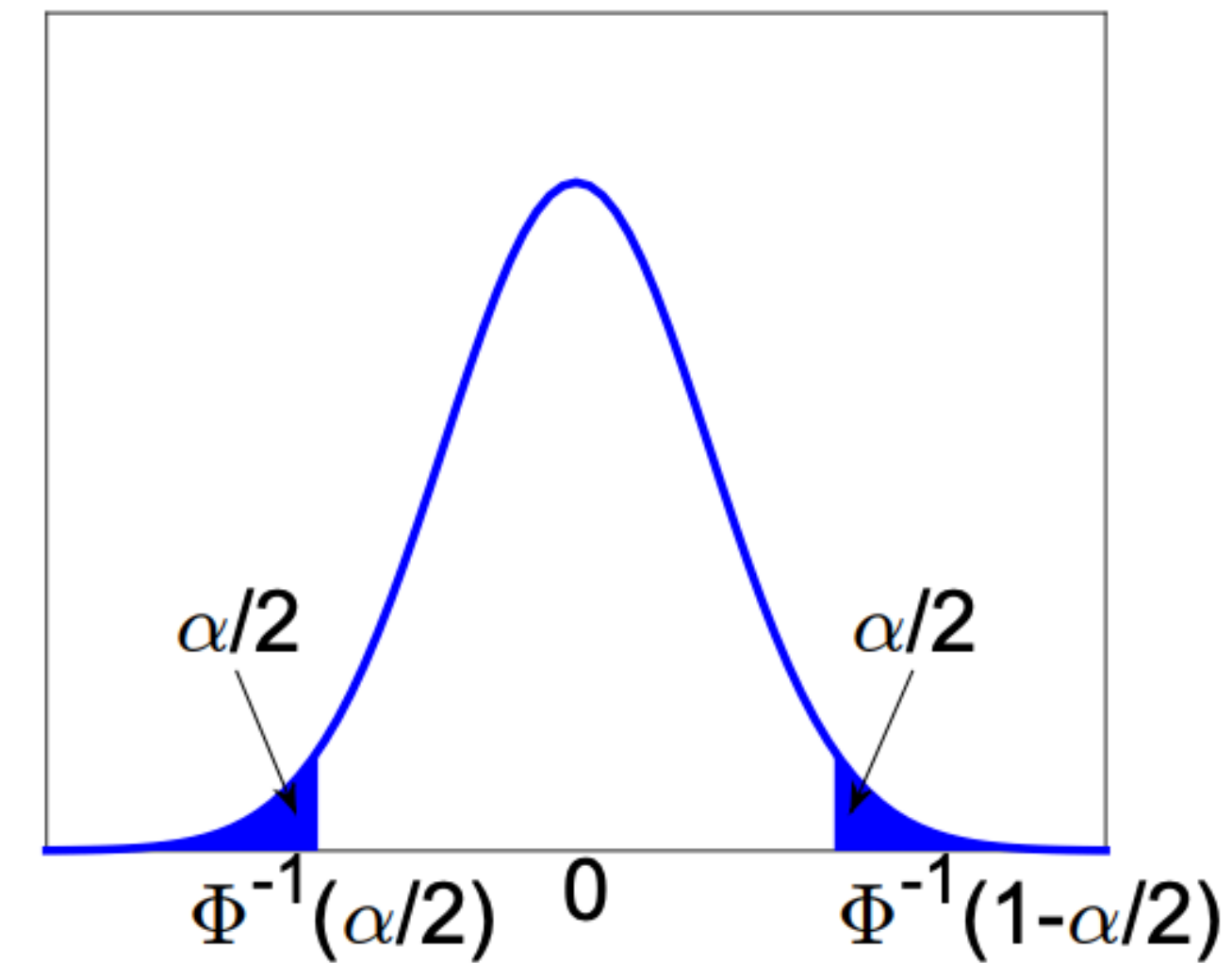
Continuous Random Variables



(a)

cumulative distribution function (cdf)

$$P(x) \triangleq \Pr(X \leq x)$$



(b)

probability density function (pdf)

$$p(x) \triangleq \frac{d}{dx} P(x)$$

$$\Pr(a < X \leq b) = \int_a^b p(x) dx = P(b) - P(a)$$

$$\Pr(x \leq X \leq x + dx) \approx p(x) dx$$

Sets of Related Random Variables

$p(X, Y)$	$Y = 0$	$Y = 1$
$X = 0$	0.2	0.3
$X = 1$	0.3	0.2

marginal distribution

$$p(X = x) = \sum_y p(X = x, Y = y)$$

conditional distribution

$$p(Y = y|X = x) = \frac{p(X = x, Y = y)}{p(X = x)}$$

$$p(x, y) = p(x)p(y|x)$$

chain rule

$$p(\mathbf{x}_{1:D}) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)p(x_4|x_1, x_2, x_3) \dots p(x_D|\mathbf{x}_{1:D-1})$$

Bayes' rule

$$p(H = h|Y = y) = \frac{p(H = h)p(Y = y|H = h)}{p(Y = y)}$$

prior distribution — what we know about possible values of H before we see any data

$$p(H)$$

observation distribution — possible outcomes Y we expect to see if H = h

$$p(Y|H = h)$$

likelihood — evaluate the observation distribution at a point corresponding to the actual observations, y

$$p(Y = y|H = h)$$

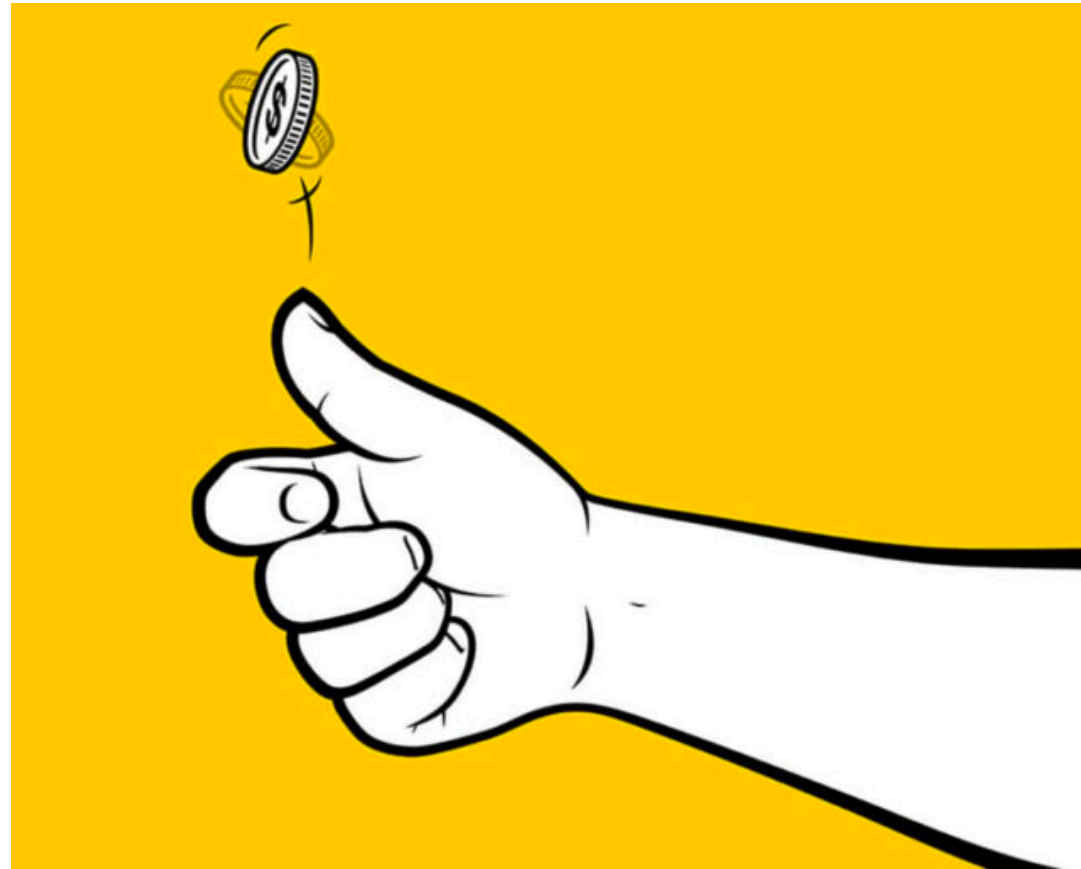
marginal likelihood

$$p(Y = y) = \sum_{h' \in \mathcal{H}} p(H = h')p(Y = y|H = h') = \sum_{h' \in \mathcal{H}} p(H = h', Y = y)$$

posterior — our new belief state about the possible value of H

$$p(H = h|Y = y)$$

A Simple Question?



What's the probability of head up?

$P(\text{Heads})$

0.8?

A Simple Question?

$$P(\text{Heads}) = \theta$$

$$P(\text{Tails}) = 1 - \theta$$

$$D = \{\text{head, head, head, head, tail}\}$$

α_H total number of heads

α_T total number of tails

$$P(D|\theta) = P(\alpha_H, \alpha_T | \theta) = \theta^{\alpha_H} (1 - \theta)^{\alpha_T}$$

Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data

$$\begin{aligned}\hat{\theta} &= \arg \max_{\theta} P(D | \theta) \\ &= \arg \max_{\theta} \log P(D | \theta)\end{aligned}$$

Set derivative to zero

$$\frac{\partial \log P(D | \theta)}{\partial \theta} = 0$$

Maximum Likelihood Estimation

Set derivative to zero

$$\frac{\partial \log P(D | \theta)}{\partial \theta} = 0$$

$$\frac{\partial \log \theta^{\alpha_H} (1 - \theta)^{\alpha_T}}{\partial \theta} = 0$$

$$\hat{\theta}_{\text{MLE}} = \frac{\alpha_H}{\alpha_H + \alpha_T} = \frac{4}{4 + 1} = 0.8$$

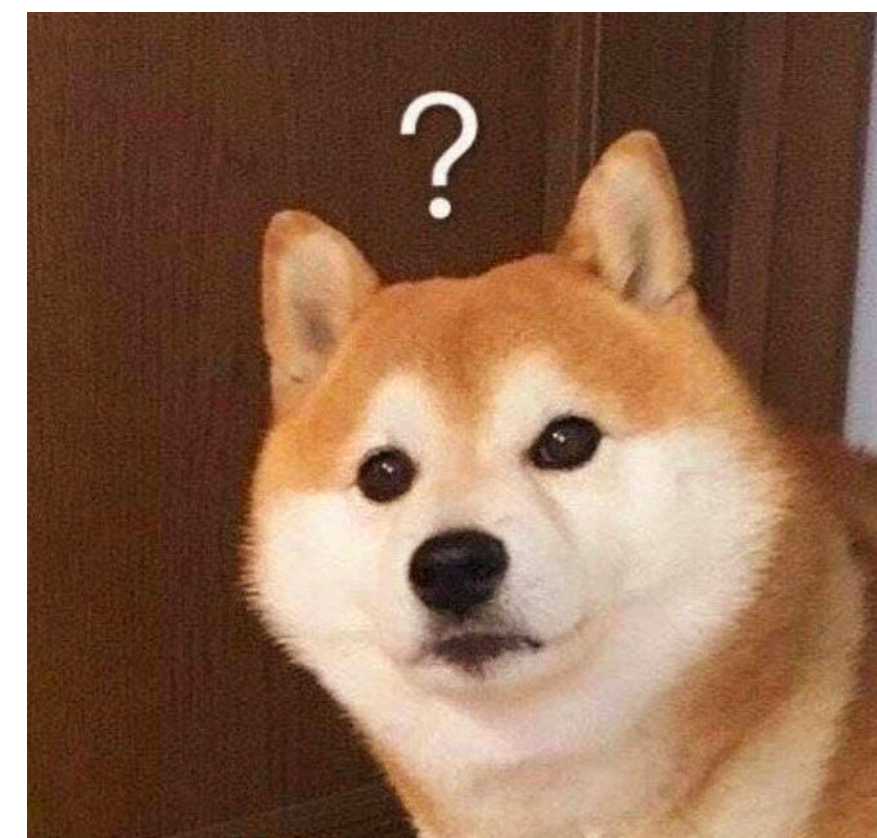
Maximum Likelihood Estimation

If you flip the coin 5 times, get 4 times head

$$\hat{\theta}_{\text{MLE}} = \frac{\alpha_H}{\alpha_H + \alpha_T} = \frac{4}{4 + 1} = 0.8$$

If you flip the coin 5000 times, get 4000 times head

$$\hat{\theta}_{\text{MLE}} = \frac{\alpha_H}{\alpha_H + \alpha_T} = \frac{4000}{4000 + 1000} = 0.8$$



Maximum Likelihood Estimation

MLE: Choose θ that maximizes the probability of observed data

$$\begin{aligned}\hat{\theta} &= \arg \max_{\theta} P(D | \theta) \\ &= \arg \max_{\theta} \log P(D | \theta)\end{aligned}$$

Set derivative to zero

$$\frac{\partial \log P(D | \theta)}{\partial \theta} = 0$$

$$\hat{\theta}_{\text{MLE}} = \frac{\alpha_H}{\alpha_H + \alpha_T}$$


Bayesian Learning

$$P(\theta | D) = \frac{P(D | \theta)P(\theta)}{P(D)}$$

posterior

$$\arg \max_{\theta} P(\theta | D) = \arg \max_{\theta} \frac{P(D | \theta)P(\theta)}{P(D)}$$

$$= \arg \max_{\theta} P(D | \theta)P(\theta)$$



$P(D | \theta)$

likelihood

$P(\theta)$

prior

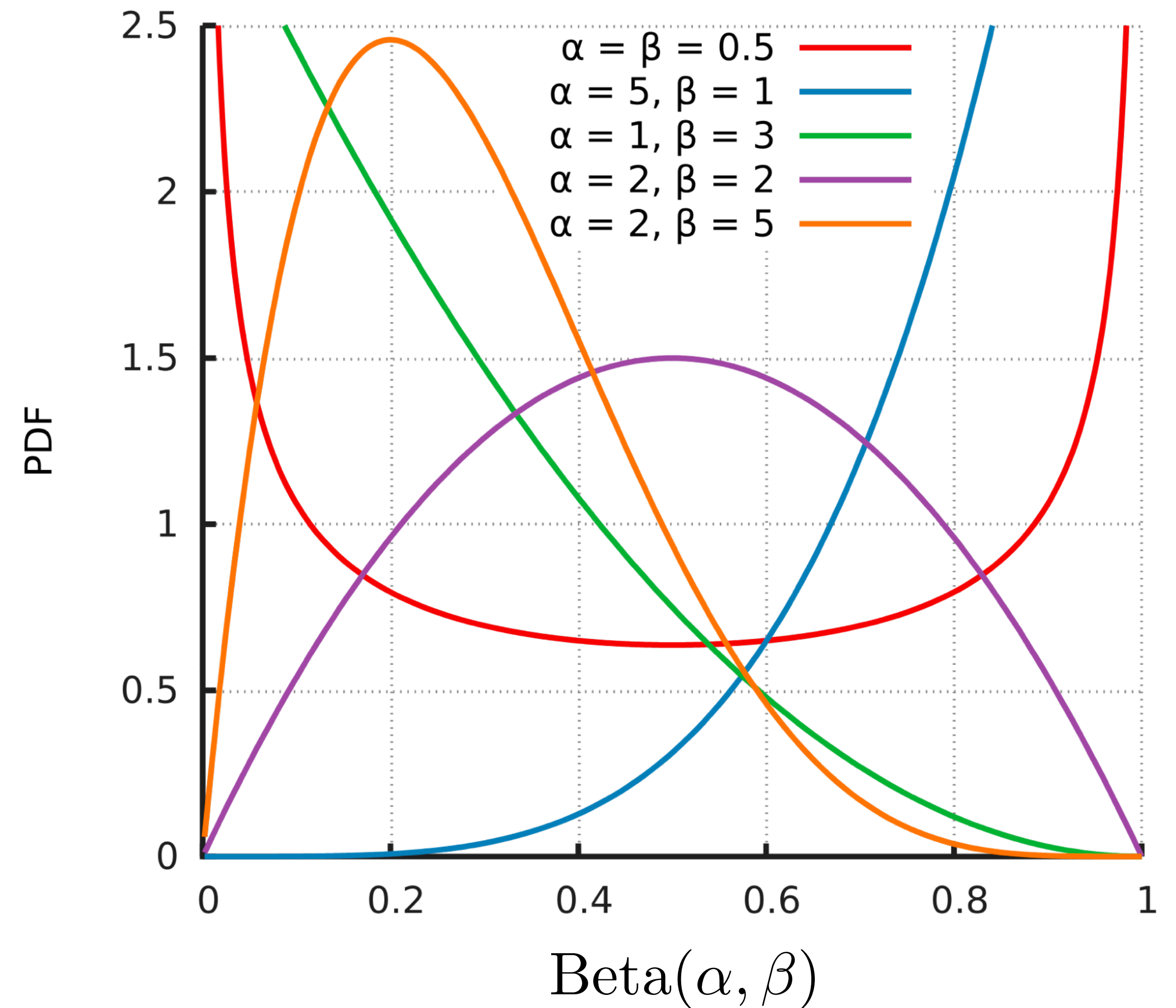
Bayesian Learning

prior distribution

$$P(\theta) \sim \text{Beta}(\beta_H, \beta_T)$$

Beta Distribution

$$P(\theta) = \frac{\theta^{\beta_H - 1} (1 - \theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)}$$



Bayesian Learning

prior distribution

$$P(\theta) \sim \text{Beta}(\beta_H, \beta_T)$$

$$P(\theta) = \frac{\theta^{\beta_H-1} (1-\theta)^{\beta_T-1}}{B(\beta_H, \beta_T)}$$

likelihood function

$$P(D|\theta) = P(\alpha_H, \alpha_T | \theta) = \theta^{\alpha_H} (1-\theta)^{\alpha_T}$$

posterior distribution

$$P(\theta | D) \propto \theta^{\alpha_H+\beta_H-1} (1-\theta)^{\alpha_T+\beta_T-1}$$

$$P(\theta | D) \sim \beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

Estimating Parameters

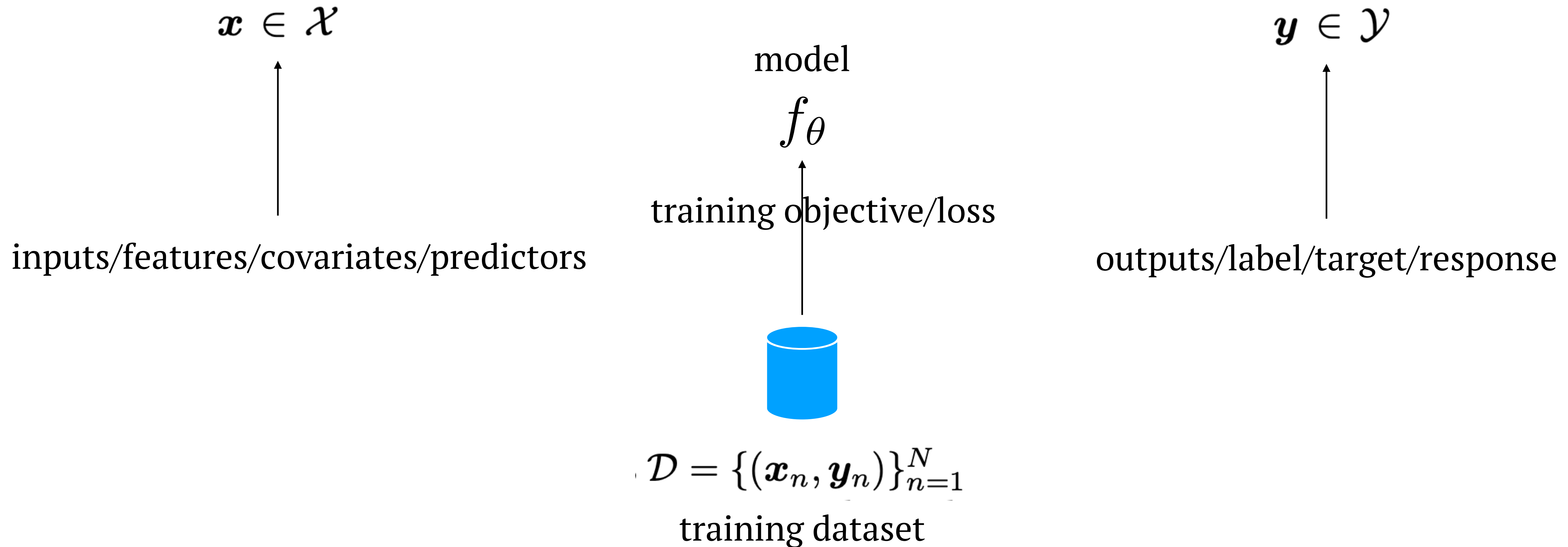
MLE: Maximum Likelihood Estimate, choose θ that maximizes the probability of observed data

$$\hat{\theta} = \arg \max_{\theta} P(D | \theta)$$

MAP: Maximum a Posteriori, choose θ that is most probable given prior probability and the data

$$\hat{\theta} = \arg \max_{\theta} P(\theta | D) = \arg \max_{\theta} \frac{P(D | \theta)P(\theta)}{P(D)}$$

Recap: Supervised Learning in a Nutshell

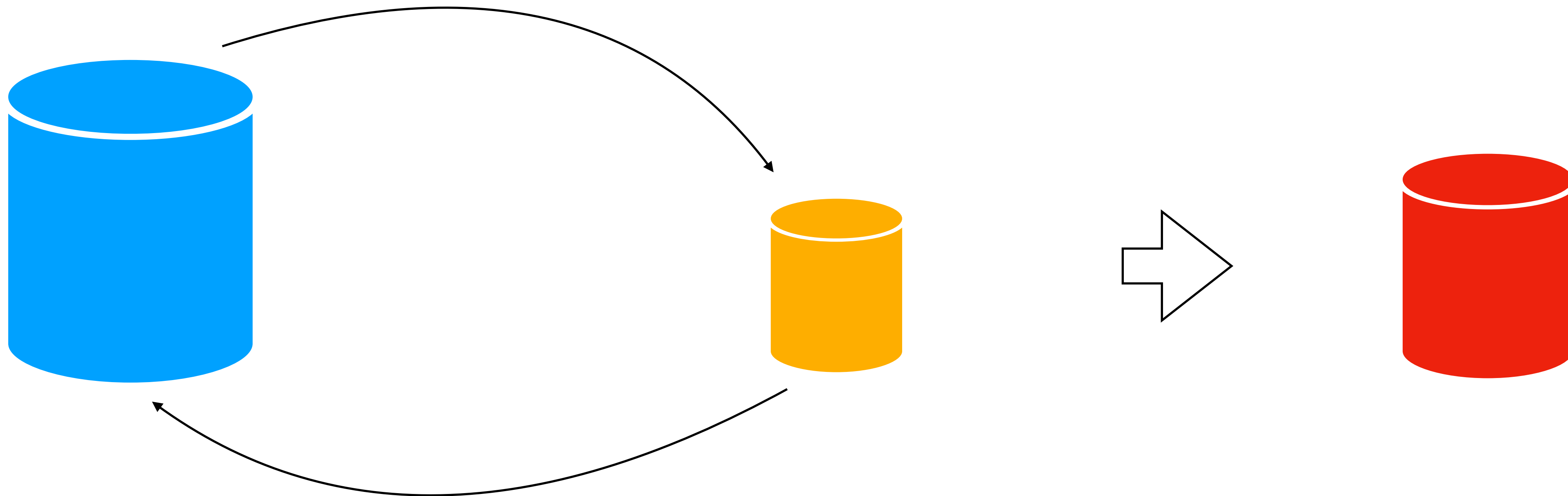


Case Study: Play Golf

	Outlook	Temperature	Humidity	Windy	Play Golf
0	Rainy	Hot	High	FALSE	No
1	Rainy	Hot	High	TRUE	No
2	Overcast	Hot	High	FALSE	Yes
3	Sunny	Mild	High	FALSE	Yes
4	Sunny	Cool	Normal	FALSE	Yes
5	Sunny	Cool	Normal	TRUE	No
6	Overcast	Cool	Normal	TRUE	Yes
7	Rainy	Mild	High	FALSE	No
8	Rainy	Cool	Normal	FALSE	Yes
9	Sunny	Mild	Normal	FALSE	Yes
10	Rainy	Mild	Normal	TRUE	Yes
11	Overcast	Mild	High	TRUE	Yes
12	Overcast	Hot	Normal	FALSE	Yes
13	Sunny	Mild	High	TRUE	No

Train / Dev / Test Set

model selection



$$\mathcal{D}_{\text{train}} = \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^{N_{\text{train}}}$$

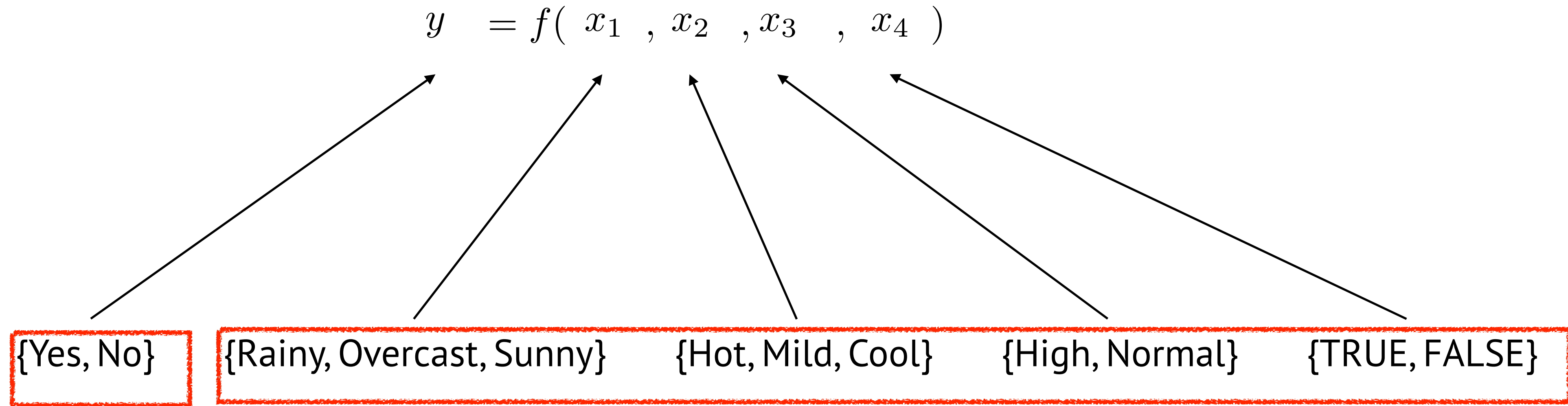
$$\mathcal{D}_{\text{dev}} = \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^{N_{\text{dev}}}$$

$$\mathcal{D}_{\text{test}} = \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^{N_{\text{test}}}$$

Case Study: Play Golf

	x_1	x_2	x_3	x_4	y
	Outlook	Temperature	Humidity	Windy	Play Golf
0	Rainy	Hot	High	FALSE	No
1	Rainy	Hot	High	TRUE	No
2	Overcast	Hot	High	FALSE	Yes
3	Sunny	Mild	High	FALSE	Yes
4	Sunny	Cool	Normal	FALSE	Yes
5	Sunny	Cool	Normal	TRUE	No
6	Overcast	Cool	Normal	TRUE	Yes
7	Rainy	Mild	High	FALSE	No
8	Rainy	Cool	Normal	FALSE	Yes
9	Sunny	Mild	Normal	FALSE	Yes
10	Rainy	Mild	Normal	TRUE	Yes
11	Overcast	Mild	High	TRUE	Yes
12	Overcast	Hot	Normal	FALSE	Yes
13	Sunny	Mild	High	TRUE	No

Case Study: Play Golf

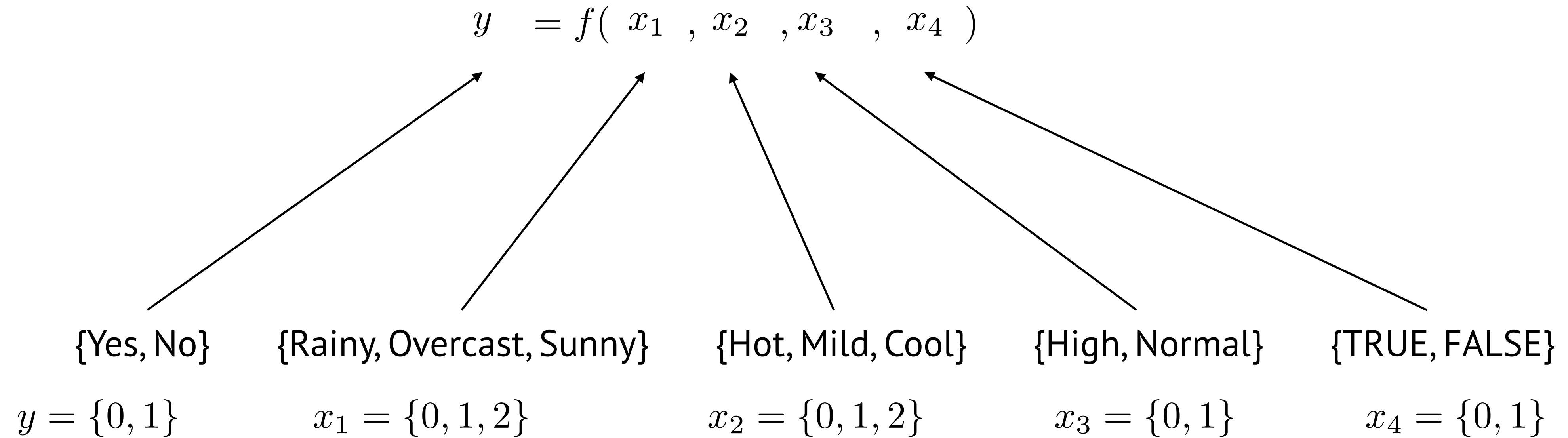


Binary Classification Problem

Discrete Features

What's in our toolbox?

Case Study: Play Golf



	Outlook	Temperature	Humidity	Windy	Play Golf
0	Rainy	Hot	High	FALSE	No

$$(\mathbf{x}^{(i)} = (0, 0, 0, 1), \mathbf{y}^{(i)} = 1)$$

Perceptron

$$f(\mathbf{x}; \boldsymbol{\theta}) = \mathbb{I}(\mathbf{w}^\top \mathbf{x} + b > 0)$$

Inference / Prediction: $(\mathbf{x}^{(i)} = (0, 0, 0, 1), \mathbf{y}^{(i)} = 1)$ $\boldsymbol{\theta} = \{\mathbf{w} = (3, -4, 5, -2), b = 1\}$

Learning / Training:



$$\mathcal{D}_{\text{train}} = \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^{N_{\text{train}}}$$

Update Rules

Objective / Loss

$\hat{\boldsymbol{\theta}}$

Perceptron

$$f(\mathbf{x}; \boldsymbol{\theta}) = \mathbb{I}(\mathbf{w}^\top \mathbf{x} + b > 0)$$

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t (\hat{y} - y) \mathbf{x}$$

$$(\mathbf{x}^{(i)} = (0, 0, 0, 1), \mathbf{y}^{(i)} = 1) \quad \boldsymbol{\theta} = \{\mathbf{w} = (3, -4, 5, -2), b = 1\} \quad \eta_t = 1$$

$$\hat{y} =$$

Supervised Learning

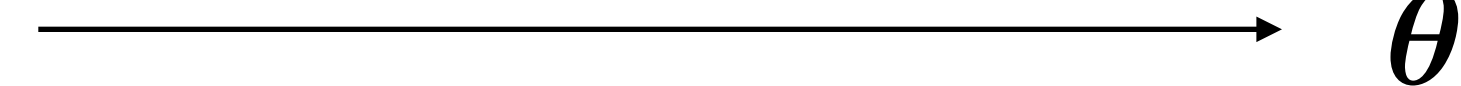
Model / Function $y = f(x; \theta)$

Learning / Training:



$$\mathcal{D}_{\text{train}} = \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^{N_{\text{train}}}$$

Objective / Loss



$\hat{\theta}$

Probabilistic Models

How does this happen?

They all sampled from a distribution!



$$\mathcal{D}_{\text{train}} = \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^{N_{\text{train}}}$$

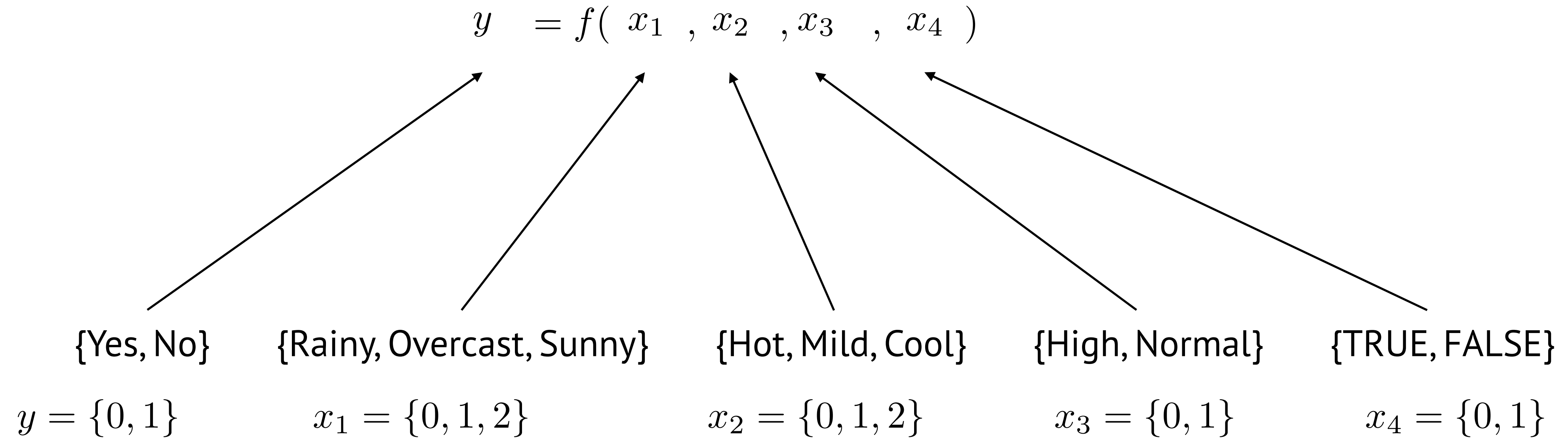
$p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta})$ Model / Function

$$p(\mathcal{D}_{\text{train}} \mid \boldsymbol{\theta}) = \prod_{n=1}^{N_{\text{train}}} p(\mathbf{y}^{(n)} \mid \mathbf{x}^{(n)}, \boldsymbol{\theta})$$

Objective / Loss

Learning / Training $\longrightarrow \hat{\boldsymbol{\theta}}$

Case Study: Play Golf



	Outlook	Temperature	Humidity	Windy	Play Golf
0	Rainy	Hot	High	FALSE	No

$$(\mathbf{x}^{(i)} = (0, 0, 0, 1), \mathbf{y}^{(i)} = 1)$$

Logistic Regression

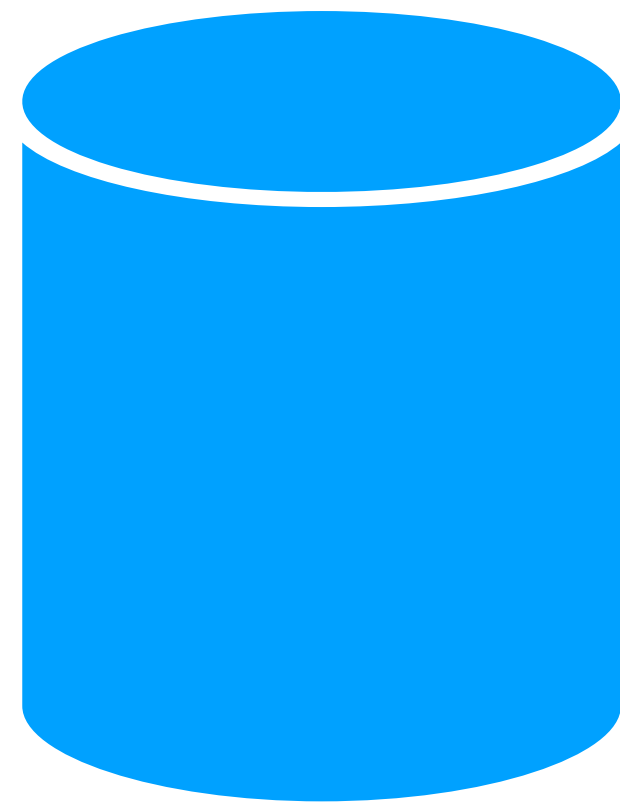
$$p(y | \mathbf{x}, \boldsymbol{\theta}) = \text{Ber}(y | \sigma(\mathbf{w}^\top \mathbf{x} + b))$$

Inference / Prediction: $(\mathbf{x}^{(i)} = (0, 0, 0, 1), \mathbf{y}^{(i)} = 1)$ $\boldsymbol{\theta} = \{\mathbf{w} = (3, -4, 5, -2), b = 1\}$

Learning / Training:

Objective / Loss:

$$p(\mathcal{D} | \boldsymbol{\theta}) \longrightarrow \arg \min_{\boldsymbol{\theta}} [-\log P(\mathcal{D} | \boldsymbol{\theta})]$$



→ $\hat{\boldsymbol{\theta}}$

$$-\sum_{n=1}^{N_{\text{train}}} [(y^{(n)}) \log p(y = 1 | \mathbf{x}^{(n)}, \boldsymbol{\theta}) + (1 - y^{(n)}) \log p(y = 0 | \mathbf{x}^{(n)}, \boldsymbol{\theta})]$$

Update rule:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t (\sigma(\mathbf{w}_t \mathbf{x} + b) - y) \mathbf{x}$$

$$\mathcal{D}_{\text{train}} = \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^{N_{\text{train}}}$$

Probabilistic Language Models

Assign a probability to a sentence

$P(\text{"I am going to school"}) > P(\text{"I are going to school"})$

Grammar Checking

I had some coffee this morning.

$P(\text{"我今早喝了一些咖啡"}) > P(\text{"我今早吃了一些咖啡"})$

Machine translation

$P(\text{"Can we put an elephant into the refrigerator? No, we can't.}) > P(\text{"Can we put an elephant into the refrigerator? Yes, we can.})$

Question Answering

Probabilistic Language Models

$$\mathcal{V} = \{\text{the, dog, laughs, saw, barks, cat, \dots}\}$$

A sentence in the language is a sequence of words

$$x_1 x_2 \dots x_n$$

For example

the dog barks STOP

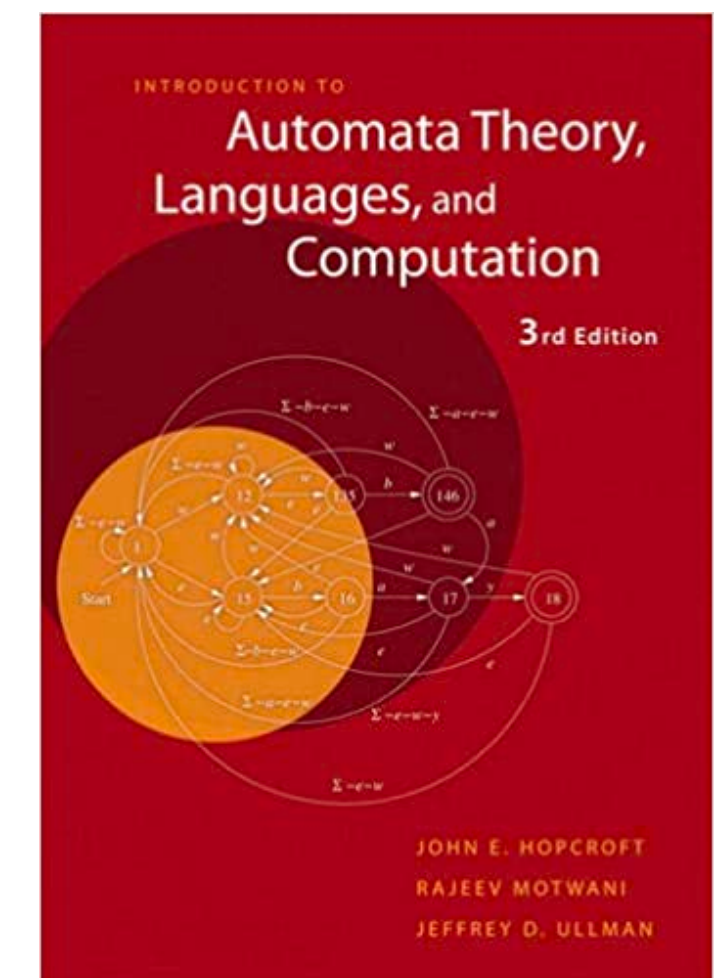
the cat saw the dog STOP

...

Definition (Language Model)

1. For any $\langle x_1 \dots x_n \rangle \in \mathcal{V}^\dagger$, $p(x_1, x_2, \dots, x_n) \geq 0$

2. In addition,
$$\sum_{\langle x_1 \dots x_n \rangle \in \mathcal{V}^\dagger} p(x_1, x_2, \dots, x_n) = 1$$



(Hopcroft, Motwani, Ullman)

A (very bad) method for learning a LM

Number of times the sentence $x_1 \dots x_n$ is seen in the training corpus

$$c(x_1 \dots x_n)$$

Total number of sentences in the training corpus N

$$p(x_1 \dots x_n) = \frac{c(x_1 \dots x_n)}{N}$$

Why this is very bad?

Markov Models

Consider a sequence of random variables X_1, X_2, \dots, X_n , each take any value in \mathcal{V}

The joint probability of a sentence is

$$\begin{aligned} &P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) \\ &= P(X_1 = x_1) \prod_{i=2}^n P(X_i = x_i | X_1 = x_1, \dots, X_{i-1} = x_{i-1}) \end{aligned}$$



$$= P(X_1 = x_1) \prod_{i=2}^n P(X_i = x_i | X_{i-1} = x_{i-1})$$

First-order Markov Assumption

Trigram Language Models

A trigram language model consists of a finite set \mathcal{V} , and a parameter $q(w \mid u, v)$

For each trigram u, v, w , such that $w \in \mathcal{V} \cup \{\text{STOP}\}$, $u, v \in \mathcal{V} \cup \{*\}$.

$q(w \mid u, v)$ can be interpreted as the probability of seeing the word w immediately after the bigram (u, v) .

For any sentence $x_1 \dots x_n$, where $x_i \in \mathcal{V}$ for $i = 1 \dots (n - 1)$, and $x_n = \text{STOP}$

$$p(x_1 \dots x_n) = \prod_{i=1}^n q(x_i \mid x_{i-2}, x_{i-1})$$

where we define $x_0 = x_{-1} = *$

Trigram Language Models

For example, for the sentence

the dog barks STOP

$$p(\text{the dog barks STOP}) = q(\text{the} \mid *, *) \times q(\text{dog} \mid *, \text{the}) \times q(\text{barks} \mid \text{the, dog}) \times q(\text{STOP} \mid \text{dog, barks})$$

Problem solved? How can we find $q(w \mid u, v)$

Parameters (of the model)

$$q(w \mid u, v)$$

How many parameters?

How to “estimate” them from training data?

Trigram Language Models

Parameters (of the model)

$$q(w \mid u, v)$$

How many parameters?

$$|\mathcal{V}|^3$$

How to “estimate” them from training data?

$$q(w \mid u, v) = \frac{c(u, v, w)}{c(u, v)}$$

$$q(\text{barks} \mid \text{the, dog}) = \frac{c(\text{the, dog, barks})}{c(\text{the, dog})}$$

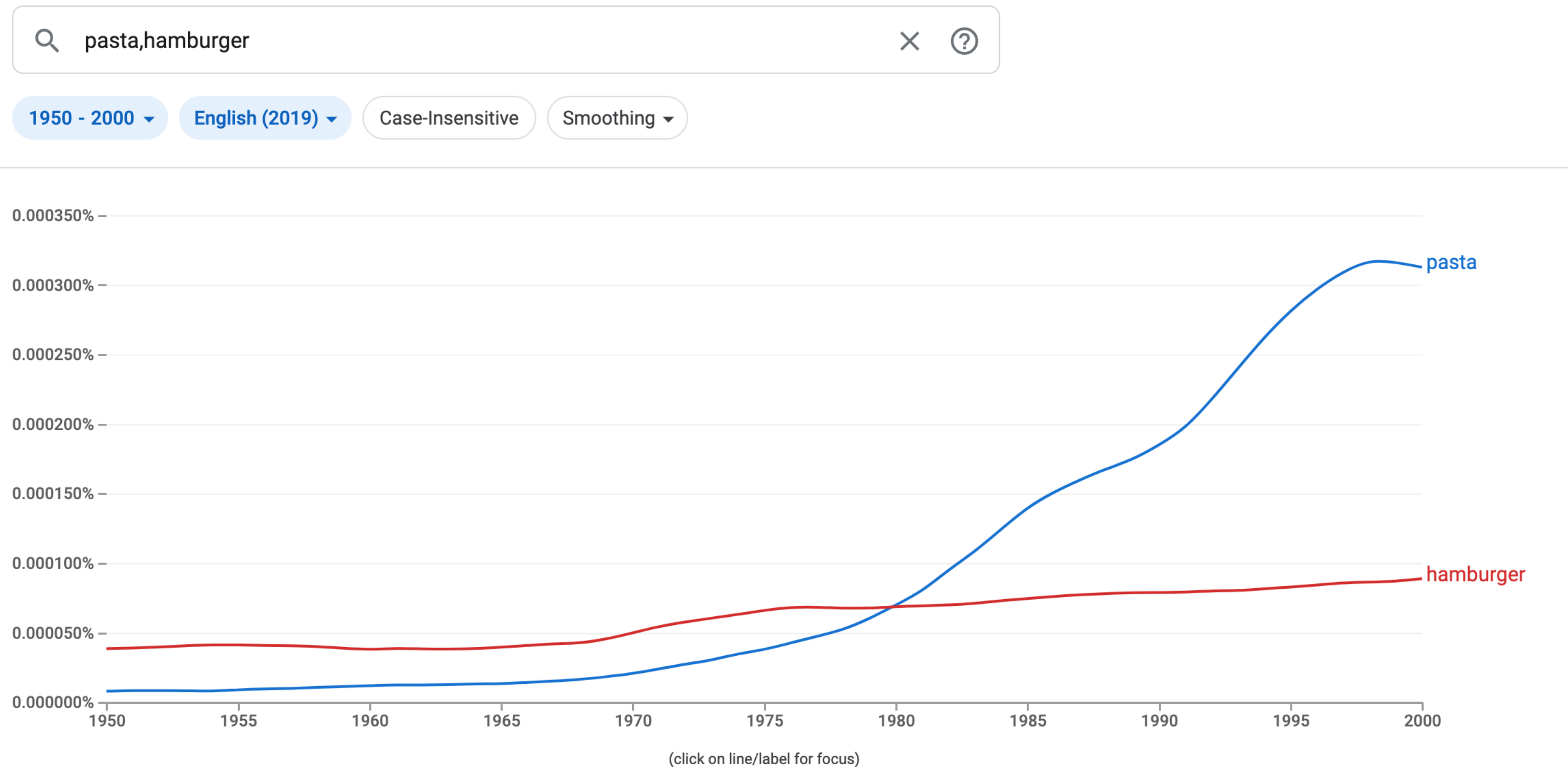
Trigram Language Models

How to “estimate” them from training data?

$$q(w | u, v) = \frac{c(u, v, w)}{c(u, v)}$$

$$q(\text{barks} | \text{the, dog}) = \frac{c(\text{the, dog, barks})}{c(\text{the, dog})}$$

N-gram counts!



Pasta v.s. Hamburger ([Google Books Ngram Viewer](#))

Sparse Data Problems

Maximum likelihood estimate:

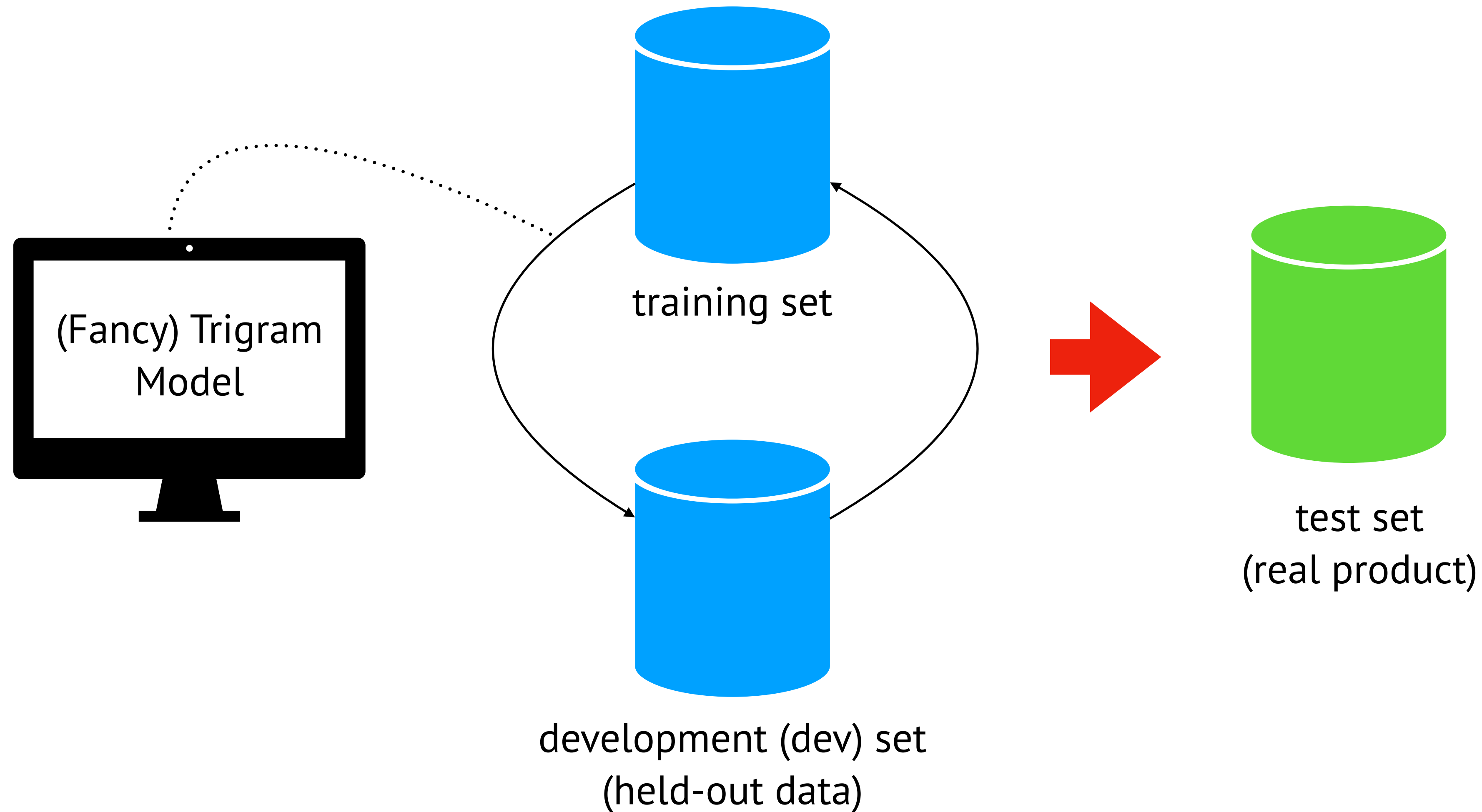
$$q(w \mid u, v) = \frac{c(u, v, w)}{c(u, v)}$$

$$q(\text{barks} \mid \text{the, dog}) = \frac{c(\text{the, dog, barks})}{c(\text{the, dog})}$$

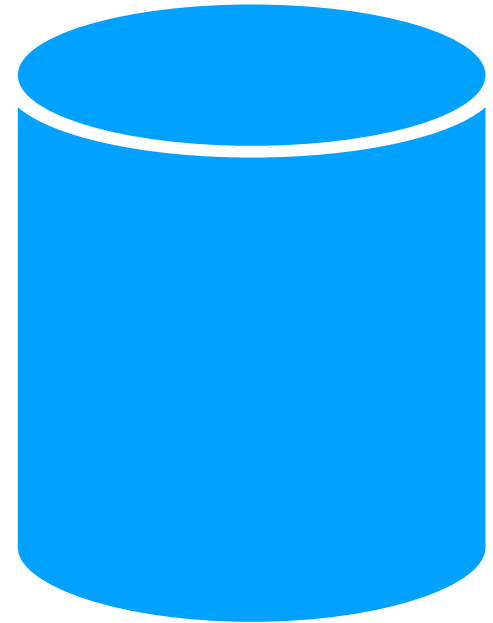
$|\mathcal{V}|^3$

Say vocabulary size is 20000. We have $8 * 10^{12}$ parameters!!

Evaluating Language Models: Perplexity



Evaluating Language Models: Perplexity



development (dev) set
(held-out data)

...

$x^{(i)}$ the cat laughs STOP

$x^{(i+1)}$ the dog laughs at the cat STOP

...

We can compute the probability it assigns to the entire set of test sentences

$$\prod_{i=1}^m p(x^{(i)})$$

The **higher** this quantity is, the better the language model is at modeling unseen sentences.

Evaluating Language Models: Perplexity

The **higher** this quantity is, the better the language model is at modeling unseen sentences.

$$\prod_{i=1}^m p(x^{(i)})$$

Perplexity on the test corpus is derived as a direction transformation of this.

$$\text{ppl} = 2^{-l}$$

$$l = \frac{1}{M} \sum_{i=1}^m \log_2 p(x^{(i)})$$

M is the total length of the sentences in the test corpus.

What if the model estimate $q(w | u, v) = 0$ and the trigram appears in the dataset?

Wait, why we love this number in the first place?

Let the model predicts $q(w | u, v) = 1/N$

$$l = \frac{1}{M} \sum_{i=1}^m \log_2 p(x^{(i)})$$

$$\text{ppl} = 2^{-l} = N$$

A uniform probability model — The perplexity is equal to the vocabulary size!

Perplexity can be thought of as the effective vocabulary size under the model!
For example, the perplexity of the model is 120 (even though the vocabulary size is say 10,000), then this is roughly equivalent to having an effective vocabulary of 120.

How much your language model updates the uniform guess!

Bayes Factors:

<https://www.youtube.com/watch?v=lG4VkPoG3ko>

Smoothing for Language Models

If the model estimate $q(w | u, v) = 0$ and the trigram appears in the test data, ppl goes up to infinity.

When we have **sparse** statistics:

P(w | denied the)

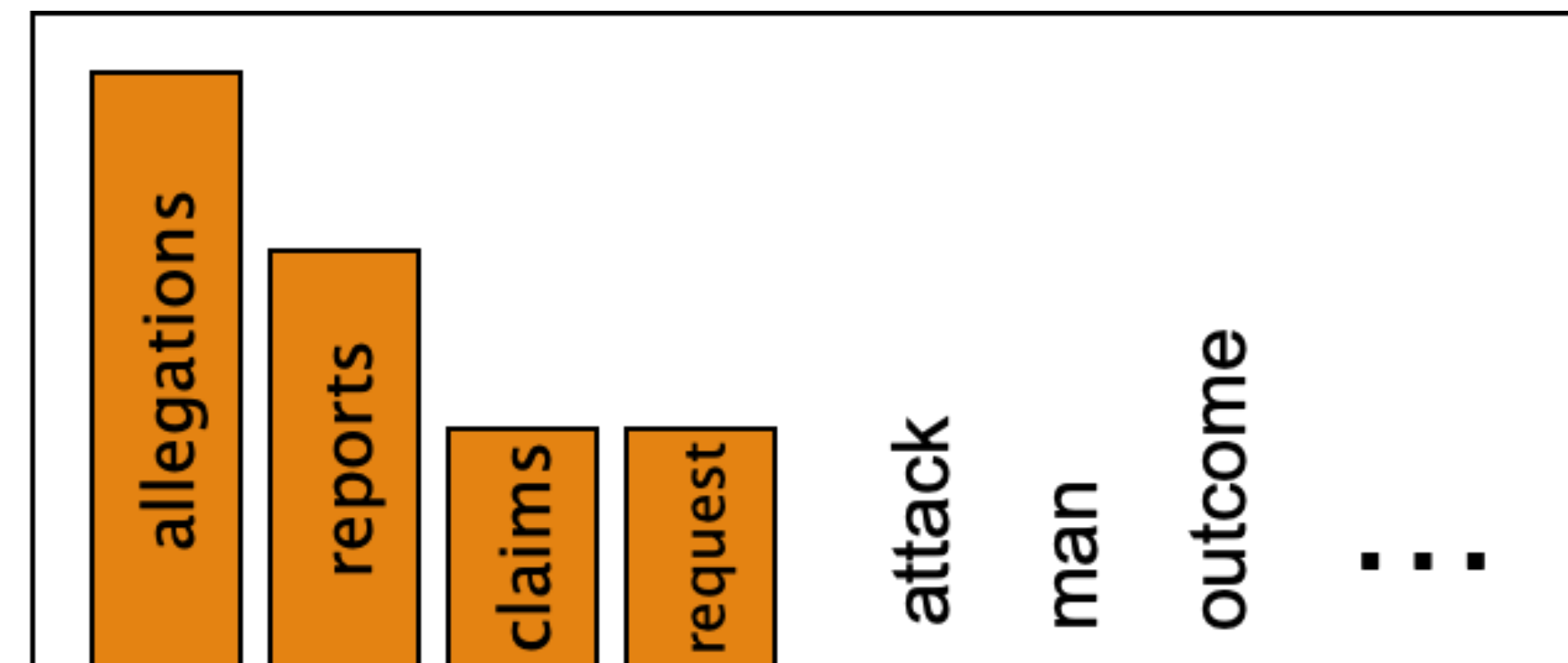
3 allegations

2 reports

1 claims

1 request

7 total



Steal probability mass to generalize better:

P(w | denied the)

2.5 allegations

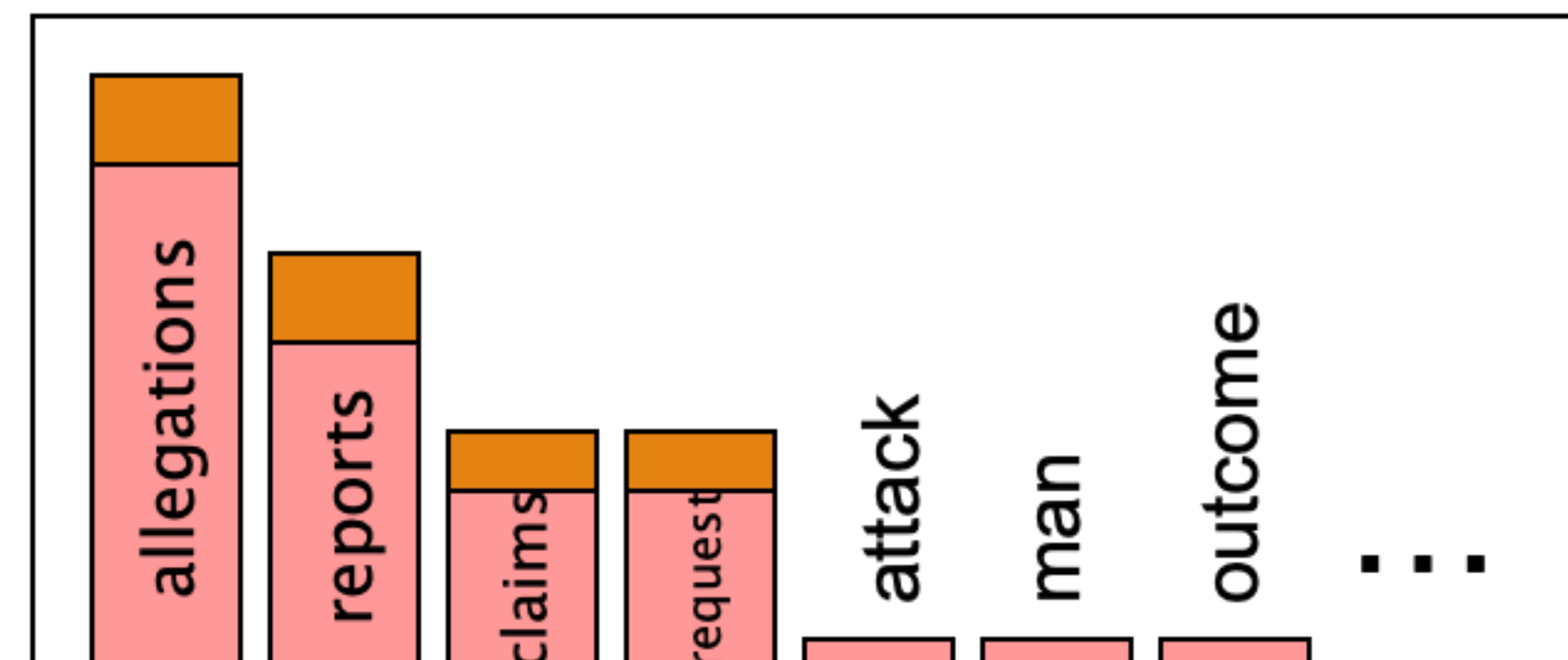
1.5 reports

0.5 claims

0.5 request

2 other

7 total



Example from Dan Klein

Add-one (Laplace) smoothing

Considering a bigram model here, pretend we saw each word one more time than we did.

MLE estimate:

$$q_{\text{MLE}}(w_i | w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}$$

Add-one smoothing:

$$q_{\text{Laplace}}(w_i | w_{i-1}) = \frac{c(w_{i-1}, w_i) + 1}{c(w_{i-1}) + |\mathcal{V}|}$$

Linear Interpolation (Stupid Backoff)

Trigram Model, Bigram Model, Unigram Model

Trigram maximum-likelihood estimate: $q(w_i | w_{i-2}, w_{i-1}) = \frac{c(w_{i-2}, w_{i-1}, w_i)}{c(w_{i-2}, w_{i-1})}$

Bigram maximum-likelihood estimate: $q(w_i | w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}$

Unigram maximum-likelihood estimate: $q(w_i) = \frac{c(w_i)}{c(\cdot)}$

Which one suffers from the data sparsity problem the most?
Which one is more accurate?

Linear Interpolation (Stupid Backoff)

$$q(w_i | w_{i-2}, w_{i-1}) = \lambda_1 \times q_{\text{ML}}(w_i | w_{i-2}, w_{i-1}) \\ + \lambda_2 \times q_{\text{ML}}(w_i | w_{i-1}) \\ + \lambda_3 \times q_{\text{ML}}(w_i)$$

where $\lambda_1 + \lambda_2 + \lambda_3 = 1$, and $\lambda_i \geq 0$ for all i .

How to choose the value of $\lambda_1, \lambda_2, \lambda_3$

Use the held-out corpus

Hyperparameters



maximize the probability of held-out data.

Markov Models in Retrospect

Consider a sequence of random variables X_1, X_2, \dots, X_n , each take any value in \mathcal{V}

The joint probability of a sentence is

$$\begin{aligned} &P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) \\ &= P(X_1 = x_1) \prod_{i=2}^n P(X_i = x_i | X_1 = x_1, \dots, X_{i-1} = x_{i-1}) \\ &= P(X_1 = x_1) \prod_{i=2}^n P(X_i = x_i | X_{i-1} = x_{i-1}) \end{aligned}$$



First-order Markov Assumption

$$P(X_i = x_i | X_1 = x_1, \dots, X_{i-1} = x_{i-1})$$

Is it possible to directly model this probability?