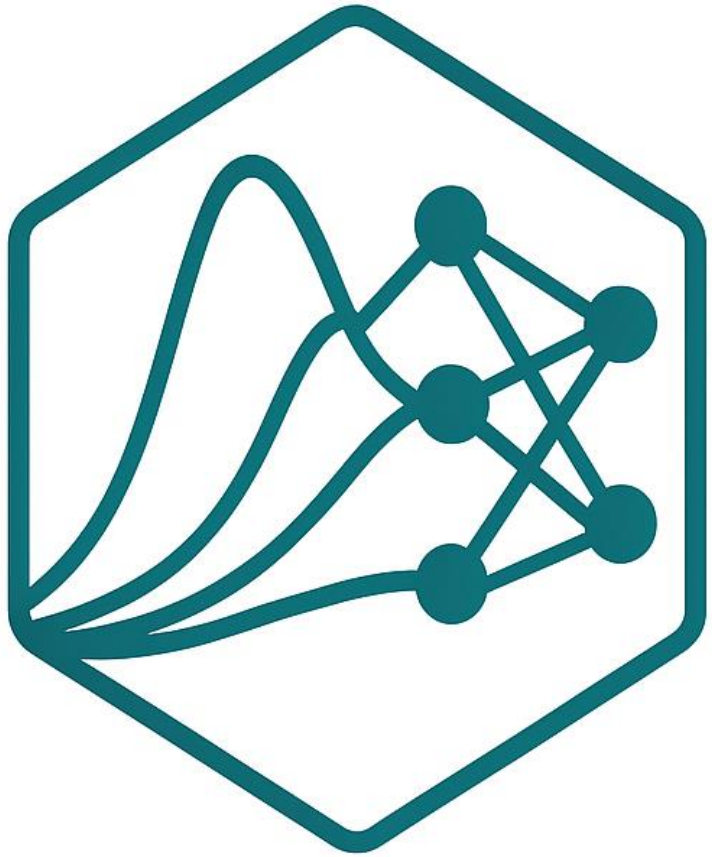


IPML

IMPRECISE
PROBABILISTIC
MACHINE LEARNING

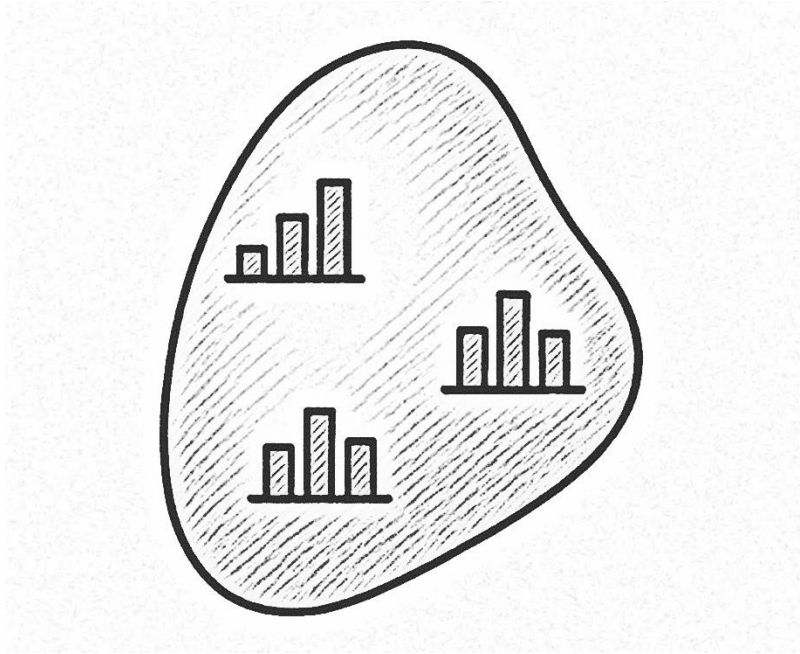
Lecture 4: Belief Function

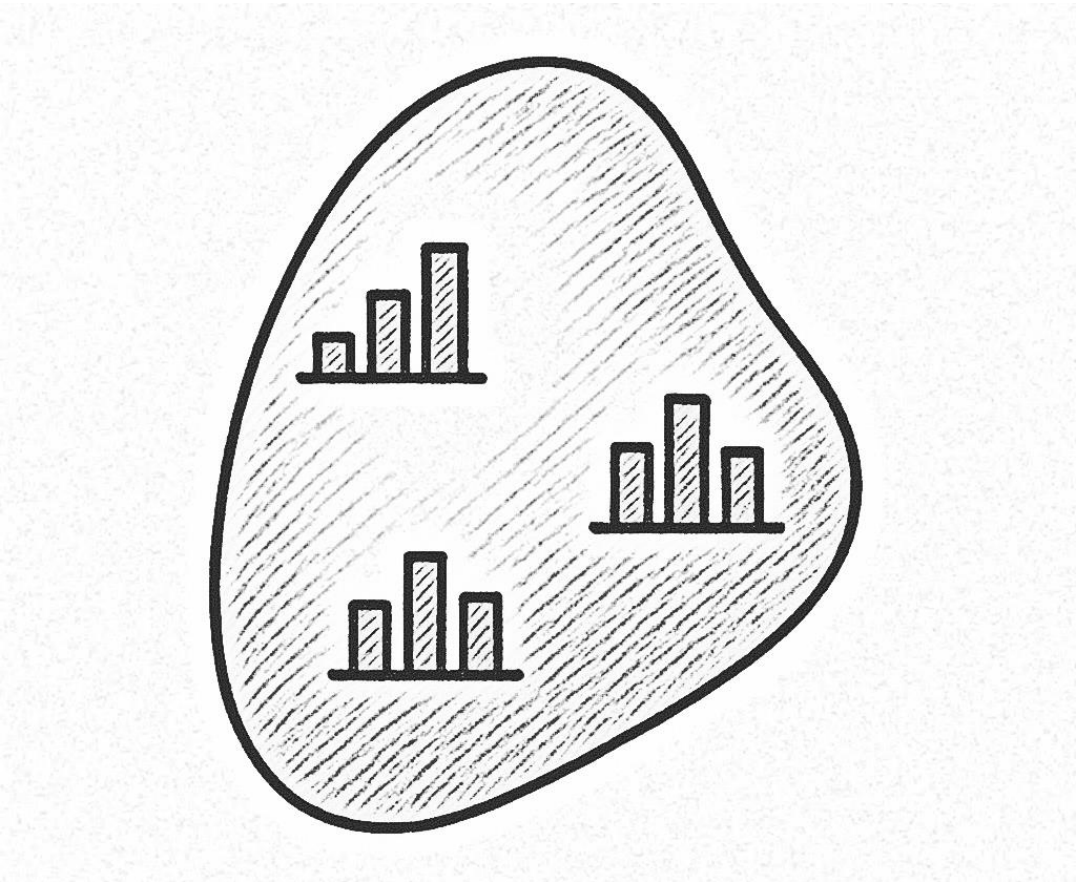
Krikamol Muandet

14 November 2025

Outline

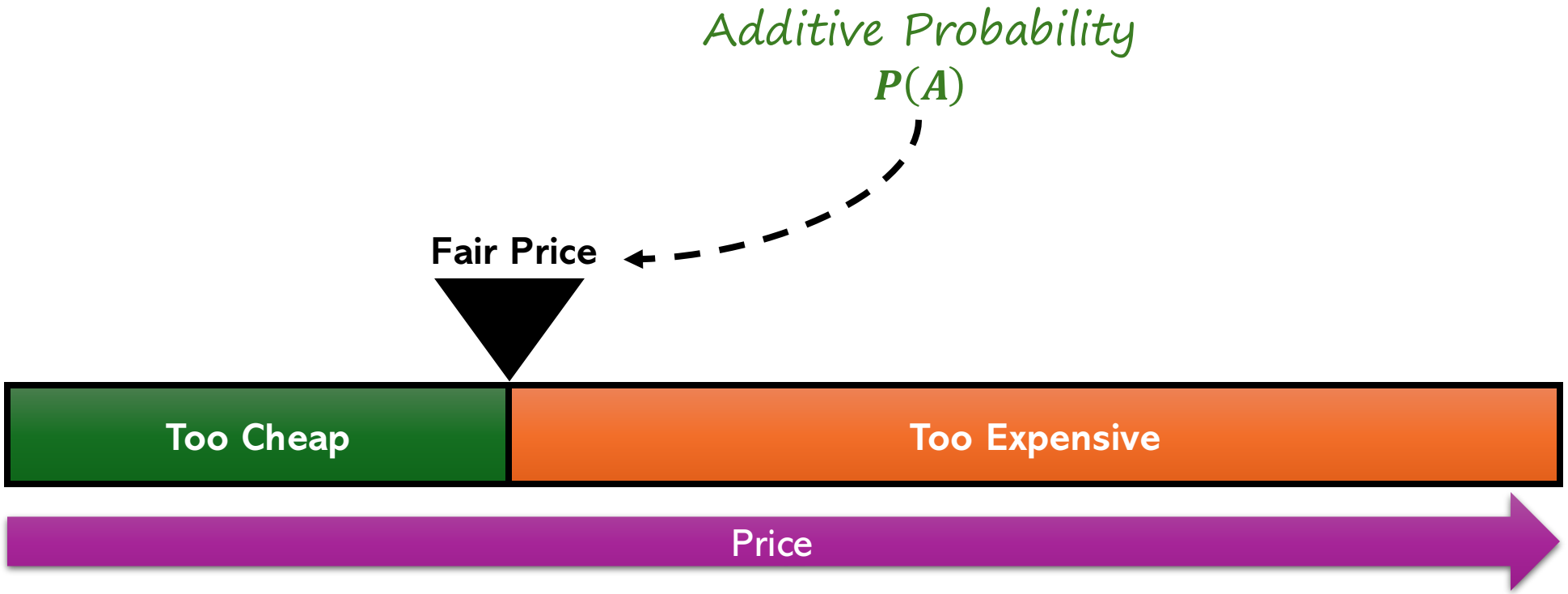
1. Recap on Imprecise Probability
2. Belief Functions
3. Dempster's Rule of Combination
4. Application to LLM



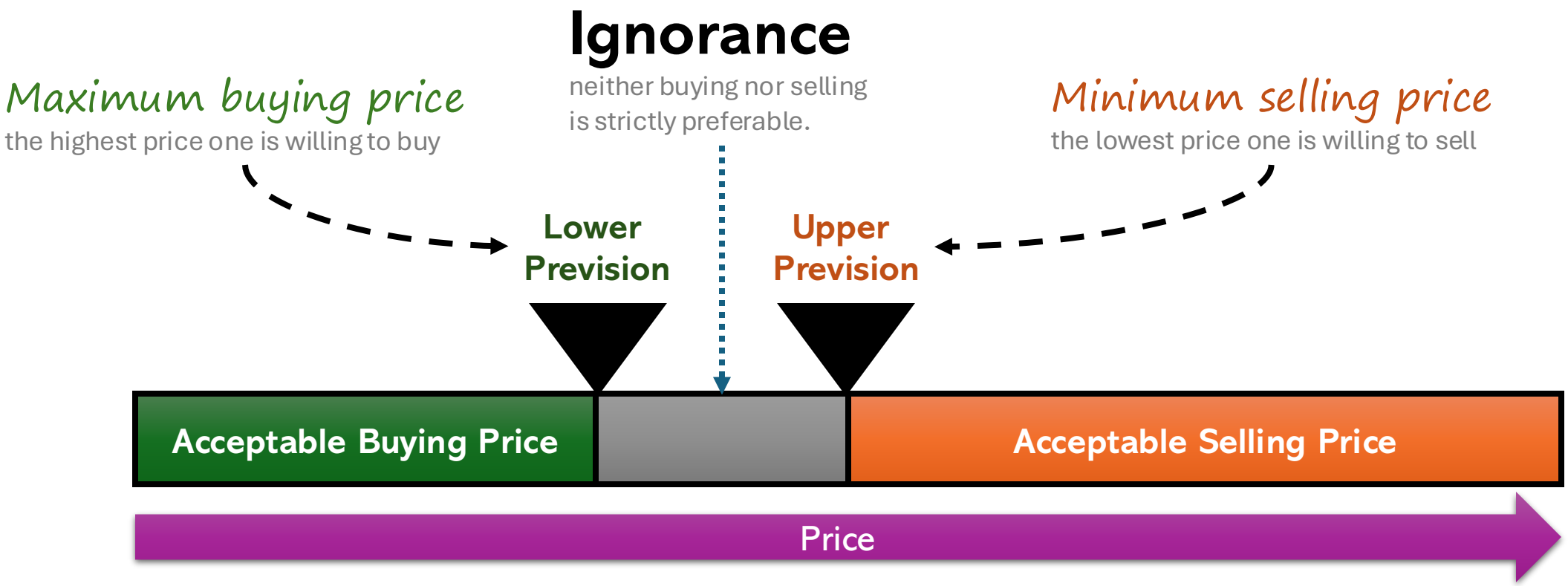


Recap on Imprecise Probability

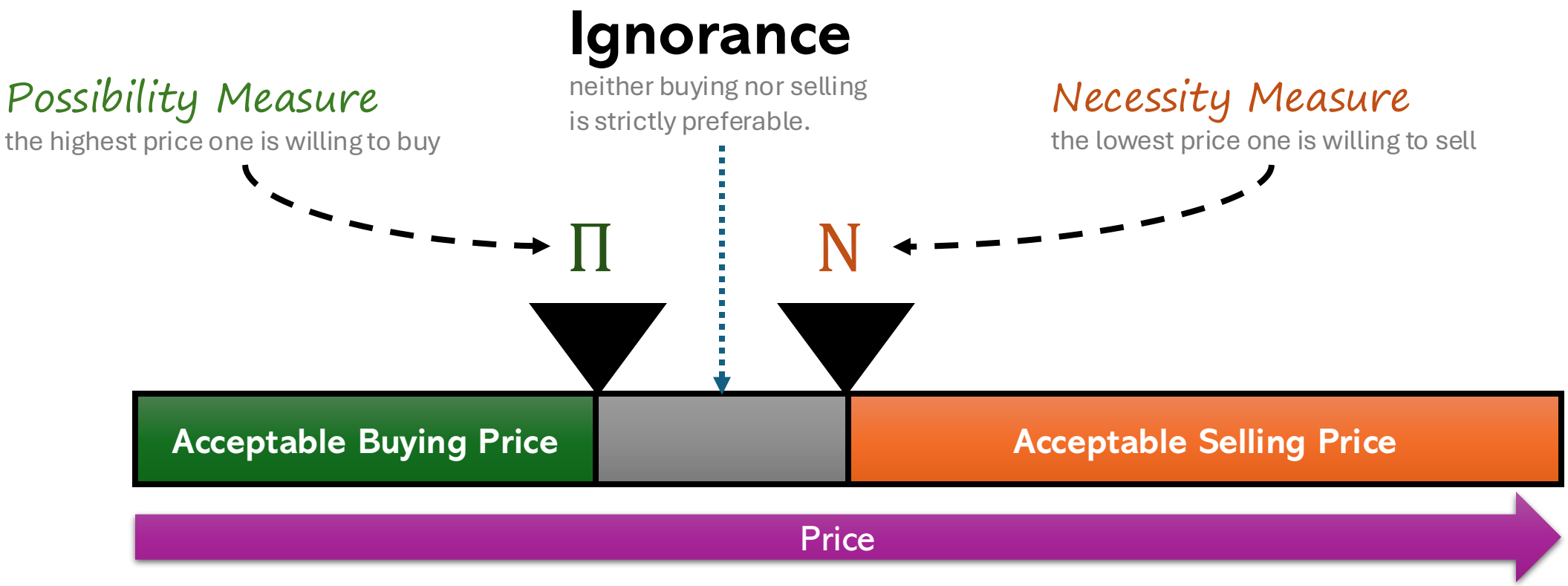
Betting Perspective



Betting Perspective



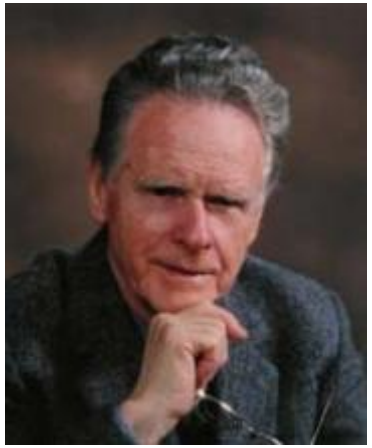
Betting Perspective



Belief Functions

Historical Background

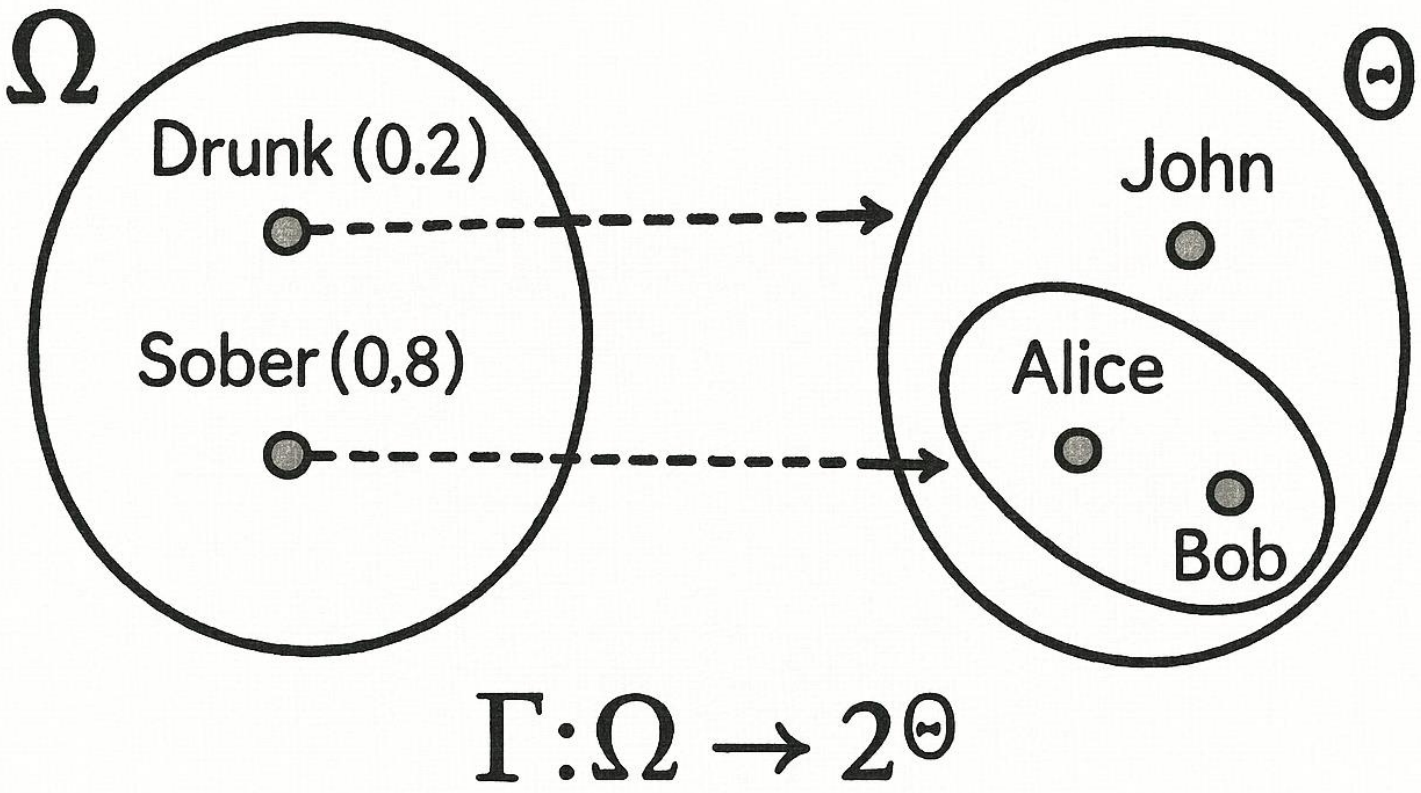
- Dempster-Shafer (DS) theory of belief functions was designed to deal with **imperfect information**.
- Originally introduced by Dempster in the 1960s for statistical inference through **a multi-valued mapping** and later expanded by Shafer in the 1970s into a general **theory of evidence**.
- Popularised and developed by Smets as the **transferable belief model (TBM)** in the 1980s.
- From the 1990s, growing number of applications in information fusion, knowledge representation, and machine learning.

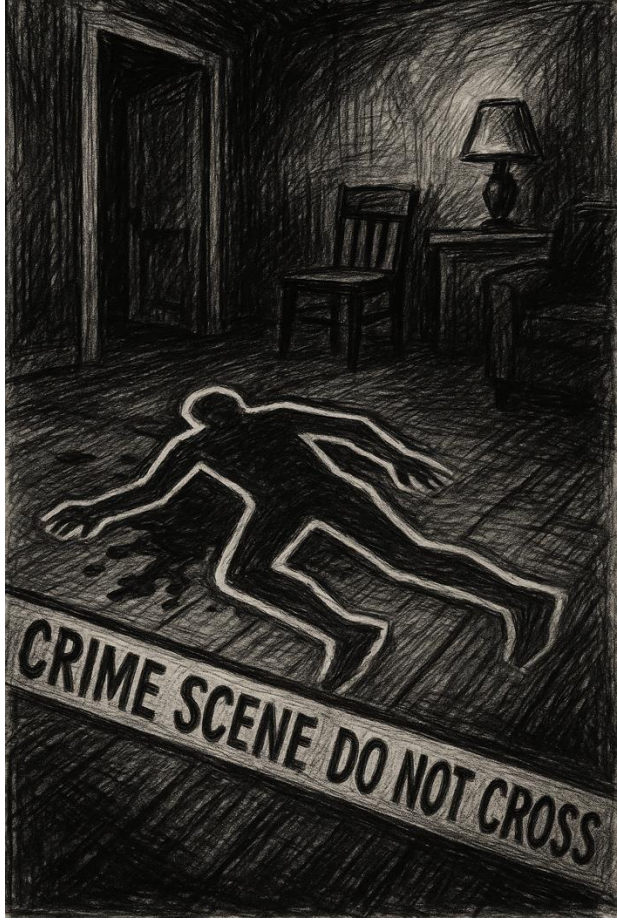


Arthur P. Dempster

Glenn Shafer

Who is the murderer?

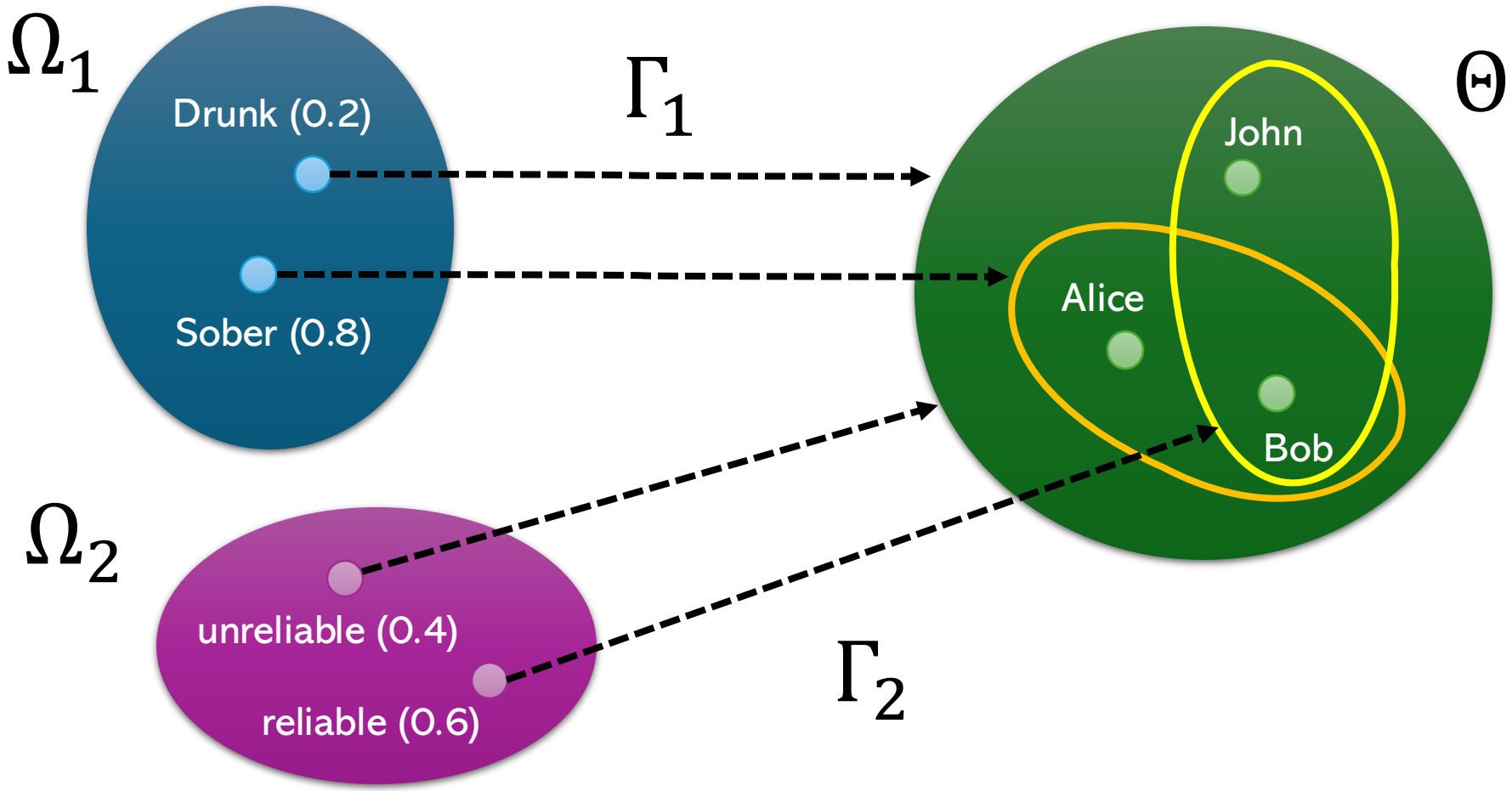


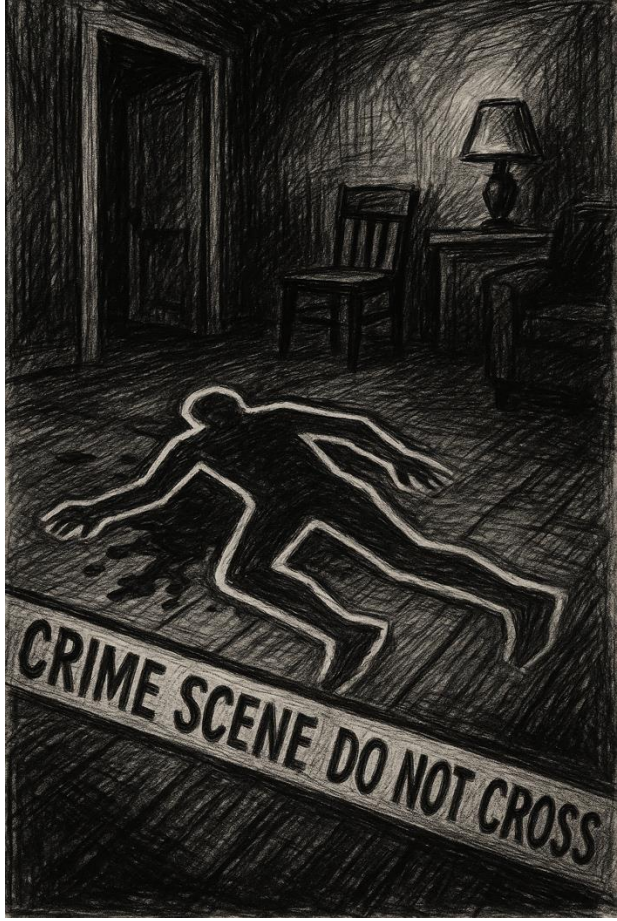


The probability P on Ω induces a **mass assignment** $m: 2^{\Theta} \rightarrow [0,1]$ on the subsets of Θ:

$$m(\{\text{Alice}, \text{Bob}\}) = 0.8$$
$$m(\Theta) = 0.2$$

Who is the murderer?





The probability P on Ω induces a **mass assignment** $m: 2^{\Theta} \rightarrow [0,1]$ on the subsets of Θ:

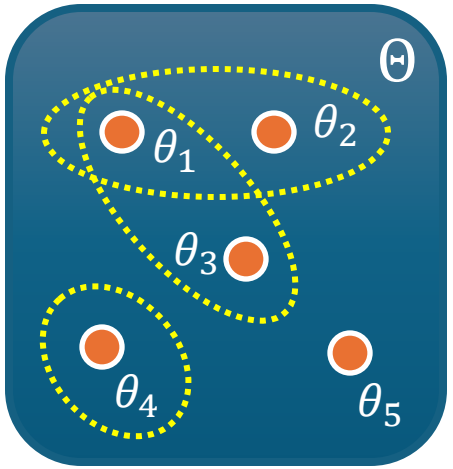
$m_1(\{\text{Alice}, \text{Bob}\}) = 0.8$
$m_1(\Theta) = 0.2$
$m_2(\{\text{John}, \text{Bob}\}) = 0.6$
$m_2(\Theta) = 0.4$

Belief Functions as Set Functions

- A **frame of discernment** $\Theta = \{\theta_1, \theta_2, \dots, \theta_n\}$
- A basic probability assignment, or **mass function**, over Θ is a set function m: $2^{\Theta} \rightarrow [0,1]$ such that

$$m(\emptyset) = 0, \qquad \sum_{A \subseteq \Theta} m(A) = 1$$

- The mass $m(A)$ measures the **belief** committed exactly to A.
- A set $A \in 2^{\Theta}$ for which $m(A) > 0$ is called a **focal set** of m.
- The union of focal sets is called a **core** of m.



$A_1 = \{\theta_1, \theta_2\}$
$A_2 = \{\theta_1, \theta_3\}$
$A_3 = \{\theta_4\}$
$m(A_1) = 0.40$
$m(A_2) = 0.35$
$m(A_3) = 0.25$

Interpretation: Random Code Semantics

- **Random code semantics (Shafer 1981)**: A coded message containing reliable information about a proposition "$X \in \Theta$"
- The message was encoded using a code in a set $\Omega = \{\omega_1, \dots, \omega_n\}$. Each code ω_i has a chance $P(\omega_i)$ of being selected.
- A **multi-valued mapping** $\Gamma : \Omega \to 2^{\Theta} \backslash \{\emptyset\}$ defines the meaning of the message as "$X \in \Gamma(\omega_i)$".
- The mass $m(A)$ is the probability that the meaning of the message is "$X \in A$":

$$m(A) = P(\{\omega \in \Omega : \Gamma(\omega) = A\}) = \sum_{i=1}^{n} P(\omega_i)\, \mathrm{I}(\Gamma(\omega_i) = A)$$

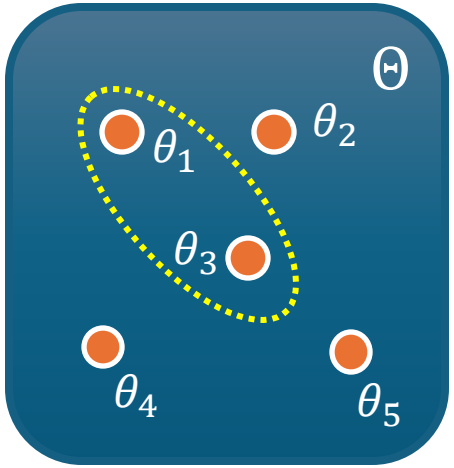
Interpretation: Random Set

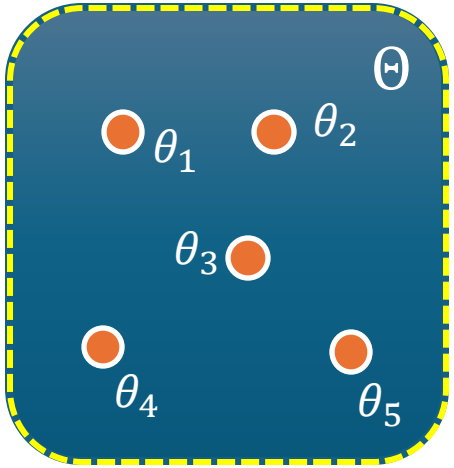
- A **random set** is defined via the tuple $(\Omega, 2^{\Omega}, P, \Gamma)$ where $(\Omega, 2^{\Omega}, P)$ is a probability space and Γ is a mapping from Ω to 2^{Θ}.
- Given any mass function $m: 2^{\Theta} \to [0,1]$, we can define the random set $(\Omega, 2^{\Omega}, P, \Gamma)$ with

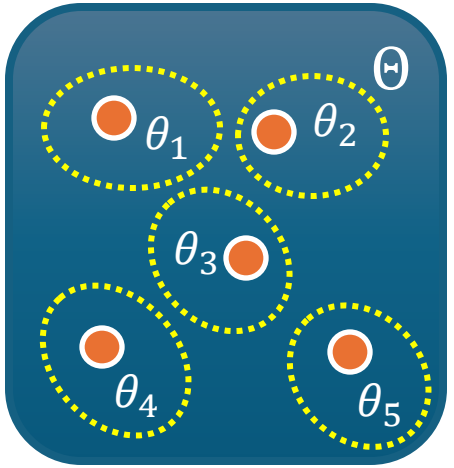
$$\Omega = 2^{\Theta}, \qquad P(\{A\}) = m(A), \qquad \Gamma(A) = A, \qquad A \subseteq \Theta$$

Special Cases

- **Logical mass function m_A**: If we believe that the truth is in $A \subseteq \Theta$ **for sure**, then $m_A = 1$.
- **Vacuous mass function m_Θ**: $m_\Theta = 1$ represents **total ignorance**.
- If all focal sets of m are singletons, m is said to be **Bayesian**. It is equivalent to a probability distribution.

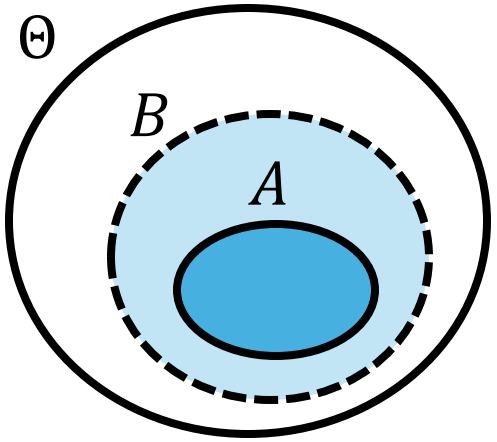




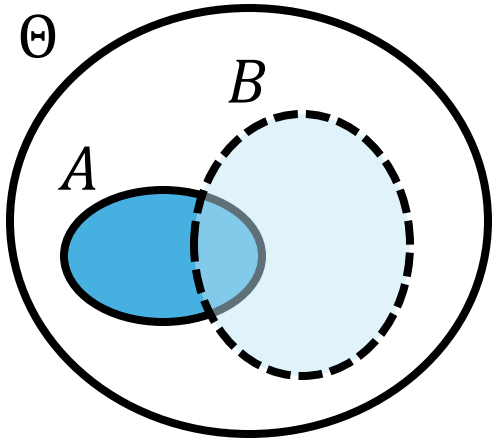


Certainty and Possibility

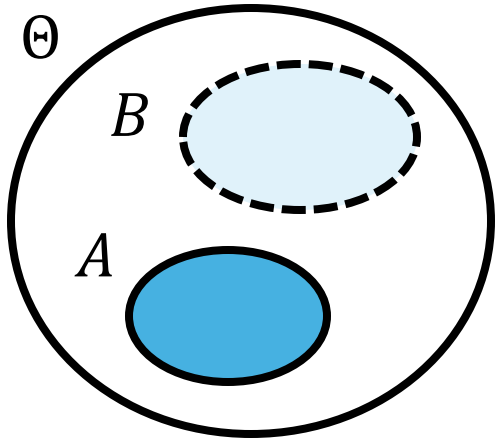
- Our evidence tells us that for some $A \in \Theta$, $X \in A$ **for sure and nothing more**. This is represented by the logical mass function m_A.
- What can we say about the proposition "$X \in B$" for some $B \subseteq \Theta$?



Certain: If $A \subseteq B$, we know for sure that $X \in B$. It supported/implied by the evidence.



Possible: If $A \cap B \neq \emptyset$, we cannot exclude that $X \in B$. It is consistent with the evidence.



Impossible: If $A \cap B = \emptyset$, the proposition "$X \in B$" is impossible It is inconsistent with the evidence.

Belief Function

- For an arbitrary mass function m with focal sets, $B_1, B_2, \ldots, B_n$, the proposition $X \in A$ is **supported by the evidence** whenever $B_i \subseteq A$.
- A **belief function** Bel induced by m is defined for all $A \subseteq \Theta$ as:

$$Bel(A) = \sum_{B \subseteq A} m(B)$$

$$= P(\{\omega \in \Omega : \Gamma(\omega) \subseteq A\})$$

- The **total degree of belief** supporting the fact that the true state is in A.

$$Bel(\emptyset) = 0, \qquad Bel(\Theta) = 1$$

Plausibility Function

- The proposition $X \in A$ is **consistent with the evidence** whenever $B_i \cap A \neq \emptyset$.
- A **plausibility function** Pl induced by m is defined for all $A \subseteq \Theta$ as:

$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B) = 1 - Bel(A^c)$$

- The number $Pl(A)$ is called the **plausibility** of A. It represents the **total sum of belief** that are not in contradiction with A.
- Elementary properties:
  - $Pl(\emptyset) = 0, \quad Pl(\Theta) = 1$
  - For all $A \subseteq \Theta$, $Bel(A) \leq Pl(A)$
  - For any $A \subseteq \Theta$, $Pl(A) = 1 - Bel(A^c)$

Who is the murderer?

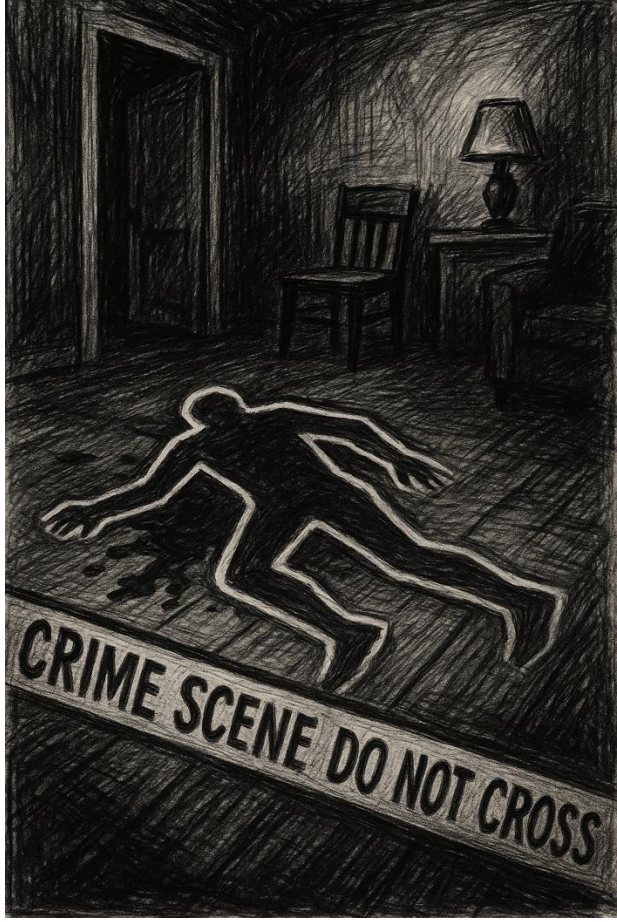
- $m(\{\text{Alice}, \text{Bob}\}) = 0.8, m(\Theta) = 0.2$

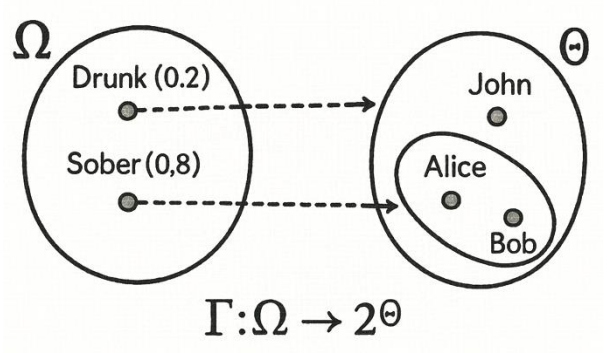
	A	A^c	Θ
Bel	0.8	0	1
Pl	1	0.2	1

- Observe that

$$Bel(\Theta) = Bel(A \cup A^c) \geq Bel(A) + Bel(A^c)$$
$$Pl(\Theta) = Pl(A \cup A^c) \leq Pl(A) + Pl(A^c)$$

- Bel is **superadditive** and Pl is **subadditive**.





Special Cases

- If the mass function m is Bayesian,

$$Bel(A) = Pl(A) = P(A) \text{ for } A \subseteq \Theta.$$

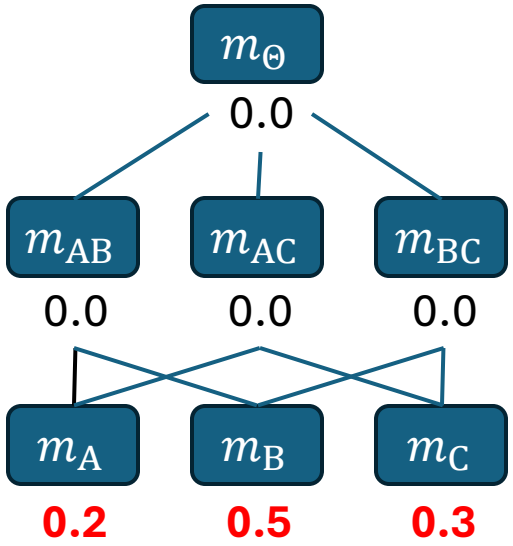
- If the focal sets of m are nested, the m is said to be **consonant**:
  - Pl becomes a **possibility measure**, i.e.,

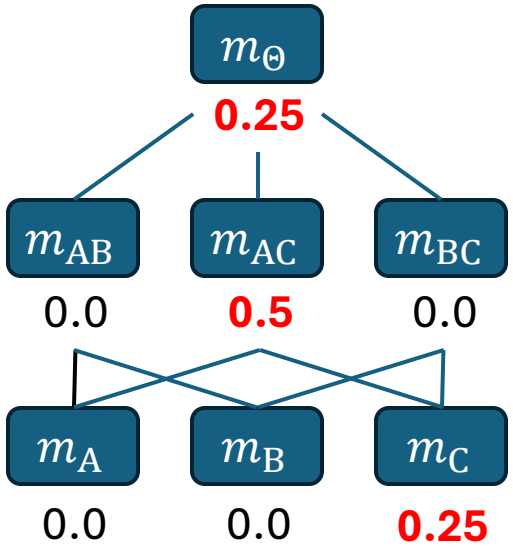
$$Pl(A \cup B) = \max\big(Pl(A), Pl(B)\big), \ \forall A, B \subseteq \Theta$$

  - Bel becomes a **necessity measure**, i.e.,

$$Bel(A \cap B) = \min\big(Bel(A), Bel(B)\big), \forall A, B \subseteq \Theta$$

- The **contour function** $pl(\theta) = Pl(\{\theta\})$ corresponds to the possibility distribution (membership function).





Characterisation of Belief Functions

- **Complete monotonicity**: For any $k \geq 2$ and for any family $\mathrm{A}_1, \ldots, \mathrm{A}_\mathrm{k} \subseteq \Theta$,

$$Bel\left(\bigcup_{i=1}^{k} A_i\right) \geq \sum_{\emptyset \neq I \subseteq \{1,\ldots,k\}} (-1)^{|I|+1} Bel\left(\bigcap_{i \in I} A_i\right)$$

- **Möbius inverse**: Any completely monotone set function Bel such that $Bel(\emptyset) = 0$ and $Bel(\Theta) = 1$ corresponds to a unique mass function m:

$$m(A) = \sum_{\emptyset \neq B \subseteq A} (-1)^{|A|-|B|} Bel(B), \qquad \forall \mathrm{A} \subseteq \Theta$$

Equivalent Representation

- For all $A \subseteq \Theta$,

$$Bel(A) = 1 - Pl(A^c)$$

$$m(A) = \sum_{\emptyset \neq B \subseteq A} (-1)^{|A|-|B|} Bel(B)$$

$$m(A) = \sum_{B \subseteq A} (-1)^{|A|-|B|+1} Pl(B^c)$$

Commonality Function

- For all $A \subseteq \Theta$, a **commonality number** is defined as

$$Q(A) = \sum_{B \supseteq A} m(B)$$

- The amount of mass which can move freely through the entire event A. It expresses **how surprising** it is to see A happens.

Example

- Let $\Theta = \{\theta_1, \theta_2, \theta_3\}$ with mass function $m(\{\theta_1\}) = 1/3$ and $m(\Theta) = 2/3$
  - $Bel(\{\theta_1\}) = m(\{\theta_1\}) = 1/3$, $Bel(\{\theta_2\}) = 0$, $Bel(\{\theta_3\}) = 0$
  - $Bel(\{\theta_1, \theta_2\}) = m(\{\theta_1\}) = 1/3$
  - $Bel(\{\theta_1, \theta_3\}) = m(\{\theta_1\}) = 1/3$
  - $Bel(\{\theta_2, \theta_3\}) = 0$
  - $Bel(\Theta) = m(\{\theta_1\}) + m(\Theta) = 1/3 + 2/3 = 1$

- For the event, $A = \{\theta_1, \theta_2\}$, we have

$$Bel(\{\theta_1, \theta_2\}) = m(\{\theta_1\}) + m(\{\theta_2\}) + m(\{\theta_1, \theta_2\}) = m(\{\theta_1\}) = \frac{1}{3}$$

$$Pl(\{\theta_1, \theta_2\}) = 1 - Bel(\{\theta_1, \theta_2\}^c) = 1 - Bel(\{\theta_3\}) = 1$$

$$Q(\{\theta_1, \theta_2\}) = \sum_{B \supseteq \{\theta_1, \theta_2\}} m(B) = m(\Theta) = \frac{2}{3}$$

Credal Set

- The set $\mathcal{P}(m)$ of probability measures compatible with m, i.e.,

$$\mathcal{P}(m) = \{\boldsymbol{P} : \ Bel(A) \leq P(A) \leq Pl(A), \forall A \subseteq \Theta\}$$

- The belief function is the **lower envelop** of $\mathcal{P}(m)$

$$Bel(A) = \min_{P \in \mathcal{P}(m)} \boldsymbol{P}(A), \quad \forall A \subseteq \Theta$$

- Not all lower envelops of sets of probability measures are belief functions.

Dempster’s Rule of Combination

Dempster's Rule

- The orthogonal sum of two mass functions m_1 and m_2 on Θ is the mass function $m_1 \oplus m_2$ defined as $(m_1 \oplus m_2)(\emptyset) = 0$ and

$$(m_1 \oplus m_2)(A) = \frac{1}{\kappa} \sum_{B \cap C = A} m_1(B) m_2(C), \quad \forall A \neq \emptyset$$

- The **degree of conflict** κ between m_1 and m_2 can be computed by

$$\kappa = \sum_{B \cap C = \emptyset} m_1(B) m_2(C)$$

- If $\kappa = 1$, m_1 and m_2 are **not combinable**.

Applications in LLM

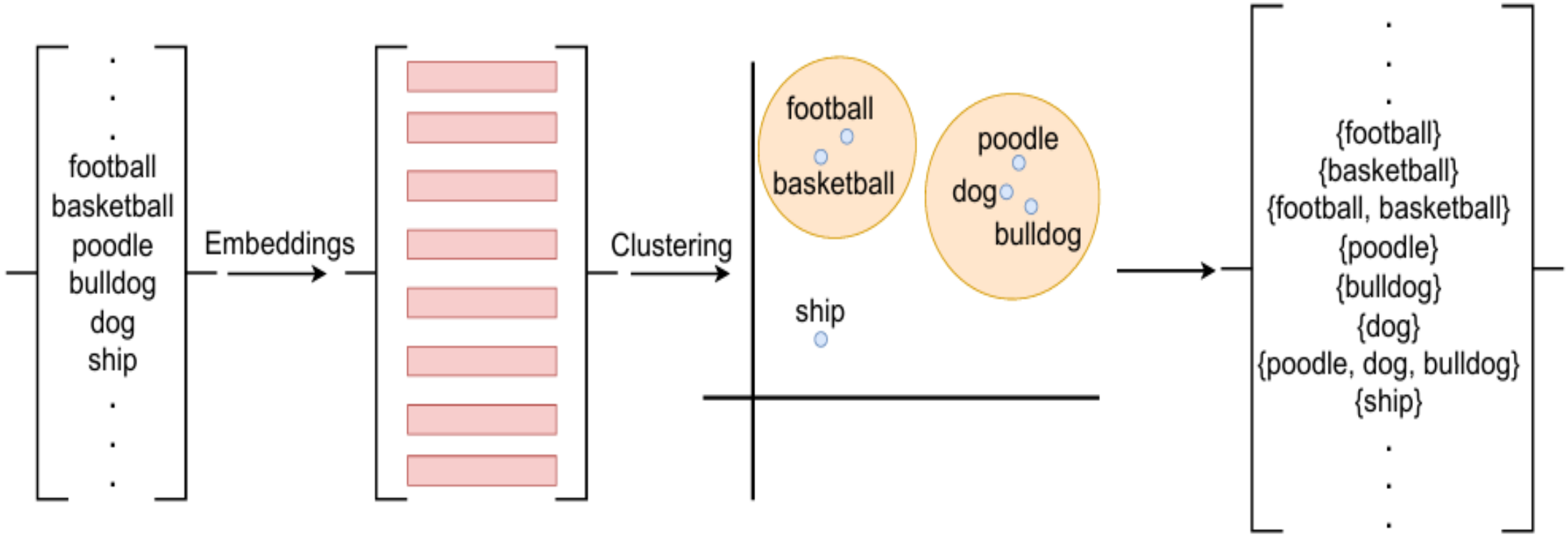
Random-Set Large Language Models

- Unlike traditional LLMs that output softmax probabilities over individual tokens, RS-LLM predicts a **belief function over sets of tokens**.
- Instead of outputs over T tokens, a vanilla RS-LLM has 2^T outputs — one for each possible subset (focal set) of tokens. In practice, a limited set of focal sets is used for efficiency.
- Ground truth is not a one-hot vector but **a belief vector**: for each focal set A, $Bel(A) = 1$ if true that the next token is in A, and zero otherwise.

Focal Set Budgeting

- Due to the massive size of the token vocabulary (~32K), using all possible token subsets 2^{32K} is infeasible. To address this, RS-LLM uses a budgeted set of focal sets.
  1. Token embeddings are extracted (e.g., from a pretrained LLM).
  2. Hierarchical clustering groups semantically similar tokens.
  3. The final budget O includes:
     - K clusters of similar tokens
     - T singleton sets (each token alone) $\rightarrow$ Total: $K + T$ focal sets

Random-Set Large Language Models



Random-Set Large Language Models

Muhammad Mubashar [1] **Shireen Kudukkil Manchingal** [1] **Fabio Cuzzolin** [1]

Abstract

Large Language Models (LLMs) are known to produce very high-quality tests and responses to our queries. But how much can we trust this generated text? In this paper, we study the problem of uncertainty quantification in LLMs. We propose a novel Random-Set Large Language Model (RS-LLM) approach which predicts finite random sets (belief functions) over the token space, rather than probability vectors as in classical LLMs. In order to allow so efficiently, we also present a methodology based on hierarchical clustering to extract and use a budget of "focal" subsets of tokens upon which the belief prediction is defined, rather than using all possible collections of tokens, making the method scalable yet effective. RS-LLMs encode the epistemic uncertainty induced in their predict the next token in an unsupervised fashion. To make these models usable, they are further fine-tuned on a particular application (Cobbe et al., 2021) and aligned to make sure the model behaves in an ethical manner and according to the human preferences (Bai et al., 2022).

However, LLMs still have limitations in their capacity to understand information and often produce false statements or "hallucinations" (Maynez et al., 2020). This makes them less trustworthy and causes hindrance to deploying LLMs in high-stake decision-making applications where the consequences of incorrect decisions are severe. Therefore, there needs to be a mechanism to make LLMs more truthful (calibration) and to associate uncertainty estimation with the model's generation. Furthermore, they need to distinguish the source of this uncertainty (Cuzzolin, 2024), as the latter can be of either aleatoric (relating to chance) or epistemic (relating to knowledge) nature (Kendall & Gal, 2017;