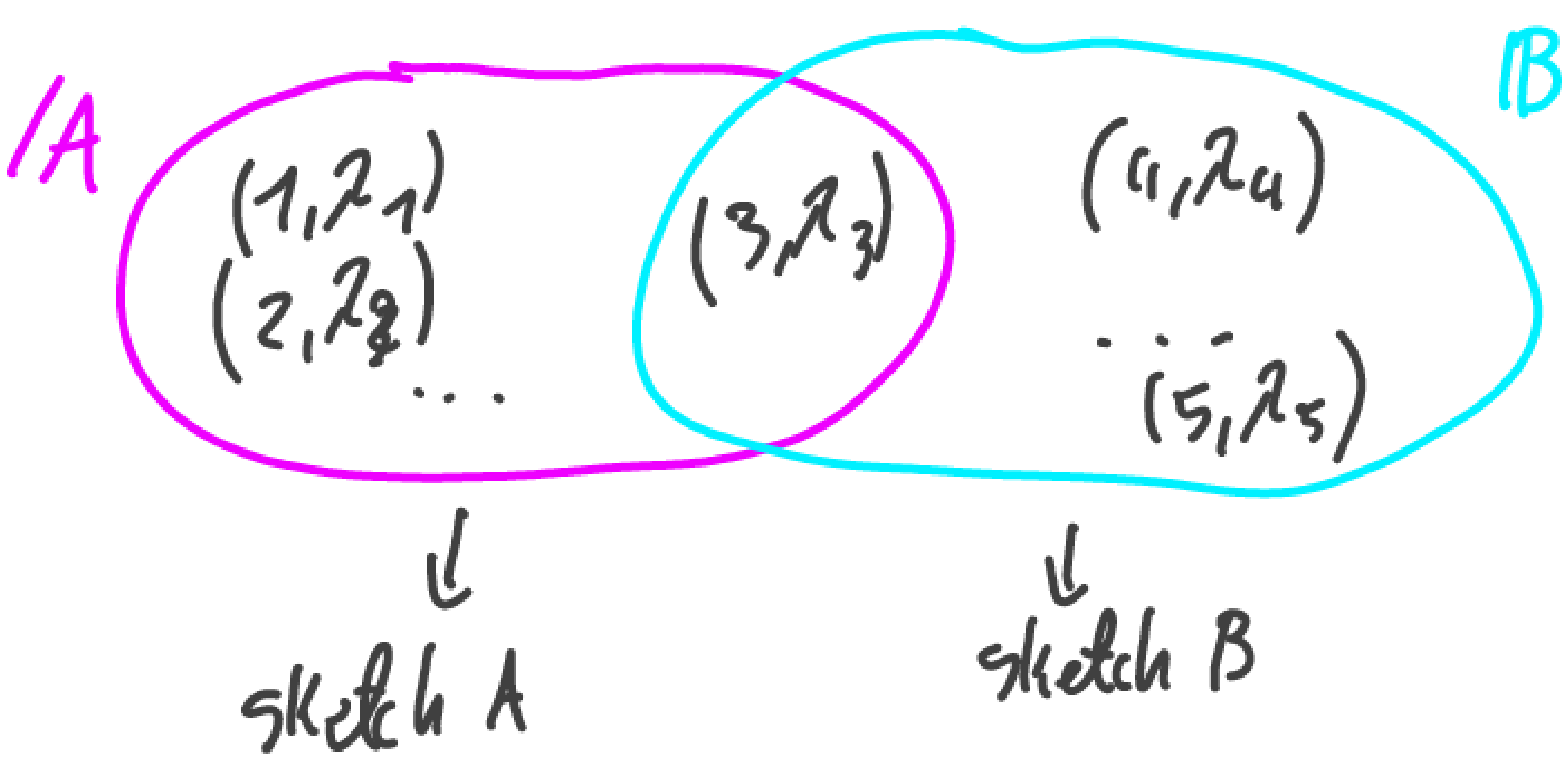
Lecture 7 : Operations on data sketches

$\mathcal{M}_1 = (\mathbb{A}, m_1)$ $\quad\quad$ $\mathcal{M}_2 = (\mathbb{B}, m_2)$

$$\begin{vmatrix} (1,\lambda_1) \\ (2,\lambda_2) \\ (3,\lambda_3) \\ (1,\lambda_1) \\ \vdots \end{vmatrix} \qquad\qquad \begin{vmatrix} (3,\lambda_3) \\ (4,\lambda_4) \\ (5,\lambda_5) \\ \vdots \end{vmatrix}$$



$|A|_w = \lambda_1 + \lambda_2 + \lambda_3$

$A = (A_1, A_2, \ldots, A_m)$,

$B = (B_1, B_2, \ldots, B_m)$,

$$A_k = \min\{S_1^{(k)}, S_2^{(k)}, S_3^{(k)}\} = \mathrm{Exp}(\lambda_1 + \lambda_2 + \lambda_3)$$

$$B_k = \min\{S_3^{(k)}, S_4^{(k)}, S_5^{(k)}\} = \mathrm{Exp}(\lambda_3 + \lambda_4 + \lambda_5)$$

$$S_i^{(k)} \sim \mathrm{Exp}(\lambda_i)$$

$$S_i^{(k)} = \frac{\ln(h(i \| k))}{-\lambda_i}$$

SUM

- $A, B \Rightarrow |A \cup B|_w = ?$, $|X|_w = \sum_{(i,\lambda_i) \in X} \lambda_i$

- Let $S_A = \{ S_i^{(k)} : (i, \lambda_i) \in A \}$,
  $S_B = \{ S_i^{(k)} : (i, \lambda_i) \in B \}$ and value of k is unimportant
  it's just the number of experiment.

- Note that $\min\{S_A \cup S_B\} = \min\{\min\{S_A\}, \min\{S_B\}\}$

thus sketch $C = (\min\{A_1, B_1\}, \dots \min\{A_m, B_m\})$ is exactly the same sketch as the sketch we would get by observing elements of $A \cup B$.

$A_1^k = \min\{S_1^1, S_2^1, S_3^1\}$, $B_1^k = \min\{S_3^1, S_4^1, S_5^1\}$

generally $\min\{S_1^{(1)}, S_2^{(1)}, S_3^{(1)}, S_4^{(1)}, S_5^{(1)}\}$

$\Downarrow$

$\min\{\min\{S_1^{(1)}, S_2^{(1)}, S_3^{(1)}\}, \min\{S_3^{(1)}, S_4^{(1)}, S_5^{(1)}\}\}$

- Since sketch C represents $A \cup B$, we can use estimator from the previous lecture:

$$\hat{\Lambda} := \frac{m-1}{\sum_{k=1}^{m} C_k} = \frac{m-1}{\sum_{k=1}^{m} \min\{A_k, B_k\}}$$

- $E[\hat{\Lambda}] = \Lambda \rightarrow E\left[\frac{m-1}{\sum \min\{A_k, B_k\}}\right] = |A \cup B|_w$

$$\rightarrow SE[\dots] = \frac{1}{\sqrt{m-2}}$$

This can be generized to any number of sets:

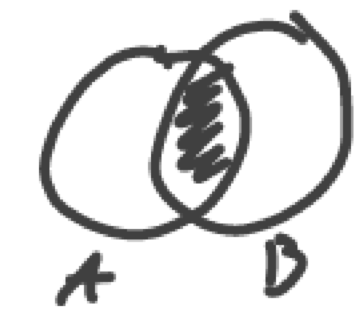
$$\min\{S_A \cup S_B \cup S_C \dots\} = \min\{\min\{S_A\}, \min\{S_B\}, \min\{S_C\} \dots\}$$

$\downarrow$

$$E\left[\frac{m-1}{\sum \min\{A_k, B_k, C_k \dots\}}\right] = |A \cup B \cup C \cup \dots|_w$$

INTERSECTION

- Jaccard similarity $J(A,B) = \dfrac{|A \cap B|}{|A \cup B|}$



nice!

- Weighted Jaccard similarity $J_w(A,B) = \dfrac{|A \cap B|_w}{|A \cup B|_w}$
- Note that $|A \cap B|_w = |A \cup B|_w \cdot J_w(A,B)$

Lemma 6

$A \rightsquigarrow A = (A_1, A_2, \dots A_m)$, $\quad B \rightsquigarrow B = (B_1, B_2, \dots B_m)$

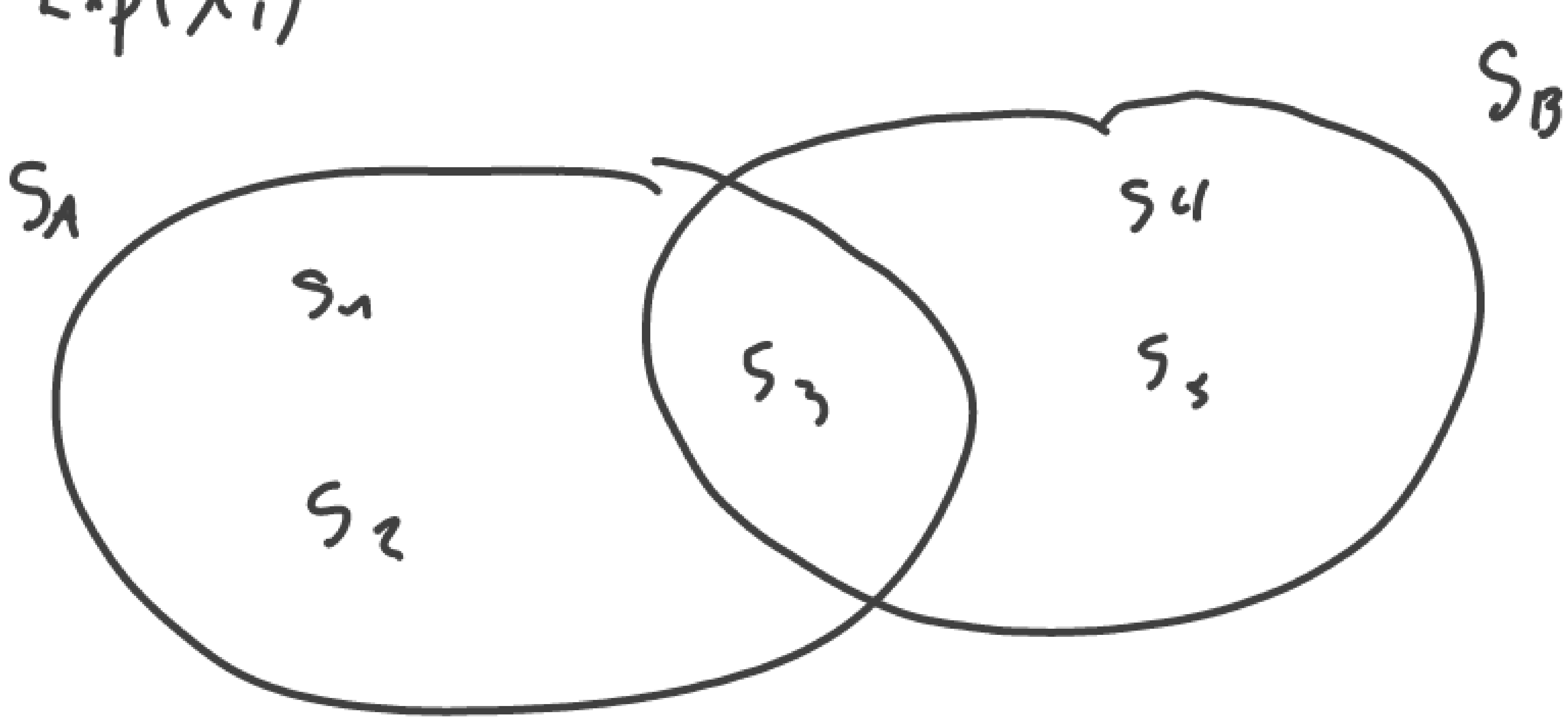
$$\left(\forall_{k \in \{1,2,3,\dots m\}}\right)\left(\Pr[A_k = B_k] = J_w(A,B)\right)$$

the probability that A and B are equal on some position is like computing J_w.

proof:

Fix k and let $S_A = \{s_i^{(k)} : (i, \lambda_i) \in A\}$, $S_B = \{s_i^{(k)} : (i, \lambda_i) \in B\}$

$s_i^{(k)} \sim Exp(\lambda_i)$



$A_k = \min\{S_A\}$ $\qquad$ $B_k = \min\{S_B\}$

to have $A_k = B_k$ the minimum must be in the intersection

$$\Pr[A_K = B_K] = \Pr[\min\{S_A \cap S_B\} < \min\{(S_A \cup S_B) \setminus (S_A \cap S_B)\}] \overset{*}{=}$$

From Lemma 2 $\circledast$:

- $\min\{S_A \cap S_B\} \sim Exp(\underbrace{|A \cap B|_w}_{= \lambda_3})$
- $\min\{(S_A \cup S_B) \setminus (S_A \cap S_B)\} \sim Exp(|(A \cup B) \setminus (A \cap B)|_w)$

also we know that

$$X \sim Exp(x) \quad , Y \sim Exp(y) \quad \rightarrow \Pr[X < Y] \overset{**}{=} \frac{x}{x+y} \quad (\text{ex.})$$

$$= \frac{|A \cap B|_w}{|A \cap B|_w + |(A \cup B) \setminus (A \cap B)|_w} = \frac{|A \cap B|_w}{|A \cup B|_w} = J_w(A, B)$$

Conclusion

- To estimate $J_w(A, B)$ it is enough to estimate $\Pr[A_K = B_K]$
- Let $\bar{J}_w(A, B) := \frac{\sum_{K=1}^{m} \mathbb{1}_{A_K = B_K}}{m}$, $\mathbb{1}_{A_K = B_K} = \begin{cases} 1 & \text{if } A_K = B_K \\ 0 & \text{if } A_K \neq B_K \end{cases}$
- $E[\bar{J}_w(A, B)] = \frac{\sum_{K=1}^{m} E[\mathbb{1}_{A_K = B_K}]}{m} = \frac{\sum_{K=1}^{m} \Pr[A_K = B_K]}{m} = \frac{m \cdot \Pr(A_K = B_K]}{m}$

$$= J_w(A, B)$$

- $\bar{I}(A, B) := \frac{m-1}{\sum_{K=1}^{m} \min\{A_K, B_K\}} \cdot \frac{\sum_{K=1}^{m} \mathbb{1}_{A_K = B_K}}{m}$

are independent

thus we get: $E[\bar{I}(A, B)] = |A \cap B|_w$

generaly:

$$\cdot\ \underbrace{\bar{I}(A,B,C,\dots)}_{|A\cap B\cap C\cap\dots|_w} := \underbrace{\frac{m-1}{\sum_{k=1}^{m}\min\{A_k,B_k,C_k,\dots\}}}_{|A\cup B\cup C\dots|_w} \cdot \underbrace{\frac{\sum_{k=1}^{m} \mathbb{1}_{A_k=B_k=C_k=\dots}}{m}}_{\boxed{J_w(A,B,C,\dots) = \frac{|A\cap B\cap C\cap\dots|_w}{|A\cup B\cup C\cup\dots|_w}}} =$$

$$\cdot\ E[\bar{I}(A,B,C\dots)] = |A\cap B\cap C\dots|_w \leftarrow \text{unbiased}$$

COMPLEMENT

$$\cdot\ \bar{D}(A,B) := \underbrace{\frac{m-1}{\sum_{k=1}^{m}\min\{A_k,B_k\}}}_{|A\cup B|_w} \cdot \underbrace{\frac{\sum_{k=1}^{m}\mathbb{1}_{A_k<B_k}}{m}}_{\frac{|A\setminus B|_w}{|A\cup B|_w}}$$

$$\cdot\ E[\bar{D}(A,B)] = |A\setminus B|_w$$

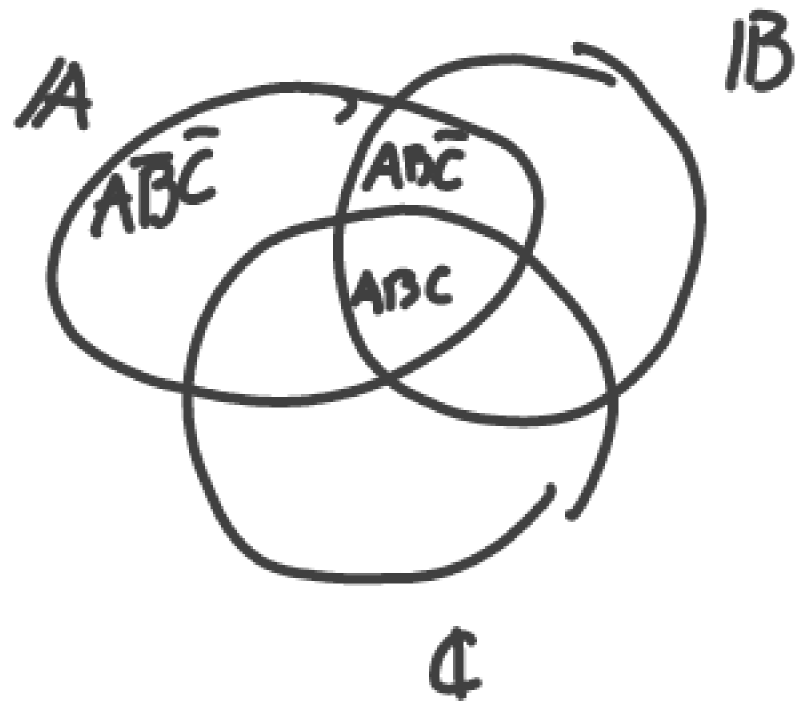
Any sequence of operations

$\mathbb{A}, \mathbb{B}, \mathbb{C}, \ldots$ can be simulated on sketches $A, B, C, \ldots$.

1. Find disjunctive normal form (DNF),

   e.g. $(\mathbb{A} \setminus \mathbb{B}) \cup (\mathbb{A} \cap \mathbb{B} \cap \mathbb{C}) = \mathbb{A} \cap \bar{\mathbb{B}} \cap \bar{\mathbb{C}} \vee \mathbb{A} \cap \mathbb{B} \cap \bar{C} \vee \mathbb{A} \cap \mathbb{B} \cap \mathbb{C}$

2. Estimate ach conjuction seperetely and sum up all estimates.



3. Based on the results we have we can estimate each conjuction

   e.g. $|\mathbb{A}\mathbb{B}\mathbb{C}|_w \rightarrow$ we use $\bar{I}(A, B, C)$

   $|\mathbb{A}\bar{\mathbb{B}}\bar{\mathbb{C}}|_w = |\mathbb{A} \setminus (\mathbb{B} \cup \mathbb{C})|$

we need to count positions in the sketch such that

$$m' = |\{k : A_k < \min\{B_k, C_k\}\}|$$

$$\cdot\, E\left[\frac{m-1}{\sum \min\{A_k, B_k, C_k\}} \cdot \frac{m'}{m}\right] = |\mathbb{A}\bar{\mathbb{B}}\bar{\mathbb{C}}|_w$$

another example

$$|AB\bar{C}D\bar{E}\bar{F}| = |ABD\bar{C}\bar{E}\bar{F}| = |ABD \setminus (C \cup E \cup F)|$$

$$m' = |\{k : A_k = B_k = D_k < \min(C_k, E_k, F_k)\}|$$

in general:

$$E\left[\frac{m-1}{\sum_{k=1}^{m} \min(A_k, B_k, C_k, D_k, E_k, F_k)} \cdot \frac{m'}{m}\right] = |AB\bar{C}D\bar{E}\bar{F}|_w$$