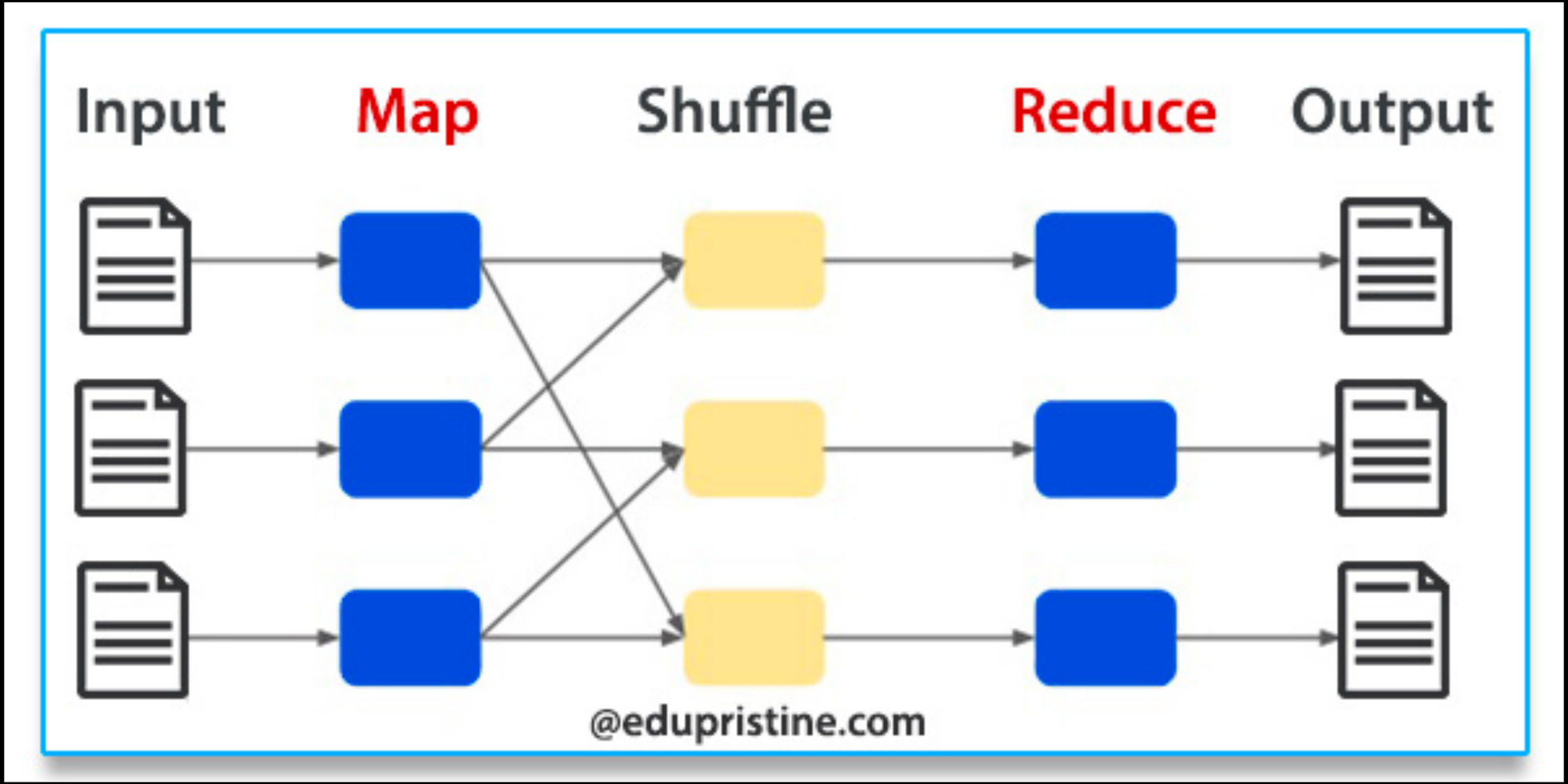
STEFANIE MUROYA LEI Y JUANPABLO HEREDIA PARILLO

TOWARDS QUANTUM FEATURE SELECTION APPLIED TO MASSIVE DATASETS

Desarrollo de un método que pueda sere escalable y útil para datasets con las dimensiones usadas en Big Data.

MARCO TEÓRICO

BIG DATA: MAPREDUCE



TABU SEARCH

- Método meta-heurístico de búsqueda.
- Métodos de búsqueda local son usados para optimización matemática.

```
BEGIN
    t ← 0;
    INITIALIZE TABU SEARCH;
    WHILE (t < t_max) DO:
        t ← t + 1;
        SEARCH NEIGHBORHOOD;
        EVALUATE CANDIDATE SOLUTIONS;
        UPDATE TABU LIST;
    END
END
```

TEORÍA DE LA INFORMACIÓN

- Teoría matemática que define los límites y posibilidades de la comunicación.
- Se basa de teoría de probabilidades y en estadística.
- Una medida clave en teoría de la información es la **entropía**.

$$H(X) = -\sum_{x} p(x) log_2 p(x)$$

ENTROPÍA

$$H(X, Y) = -\sum_{x,y} p(x, y) log_2 p(x, y)$$

ENTROPÍA CONJUNTA

$$H(X \mid Y) = -\sum_{x,y} p(x,y) log_2 p(x \mid y)$$

ENTROPÍA CONDICIONAL

INFORMACIÓN MUTUA (IM)

Mide la cantidad de información que se puede obtener sobre una variable aleatoria observando a otra.

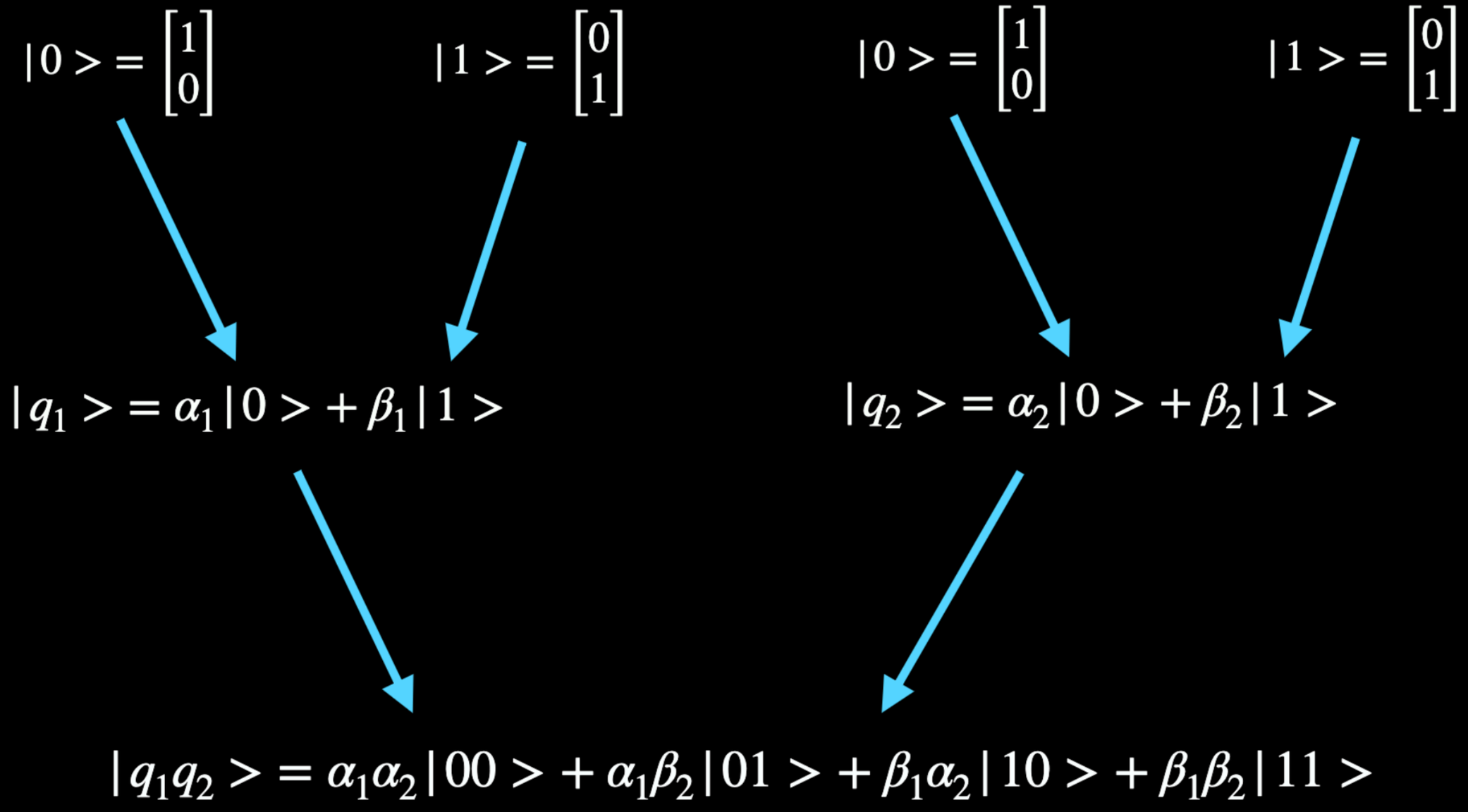
$$I(X;Y) = H(X) - H(X \mid Y)$$

INFORMACIÓN MUTUA CONDICIONAL (IMC)

Es el valor esperado de la información mutua entre 2 variables aleatorias dada una tercera.

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$$

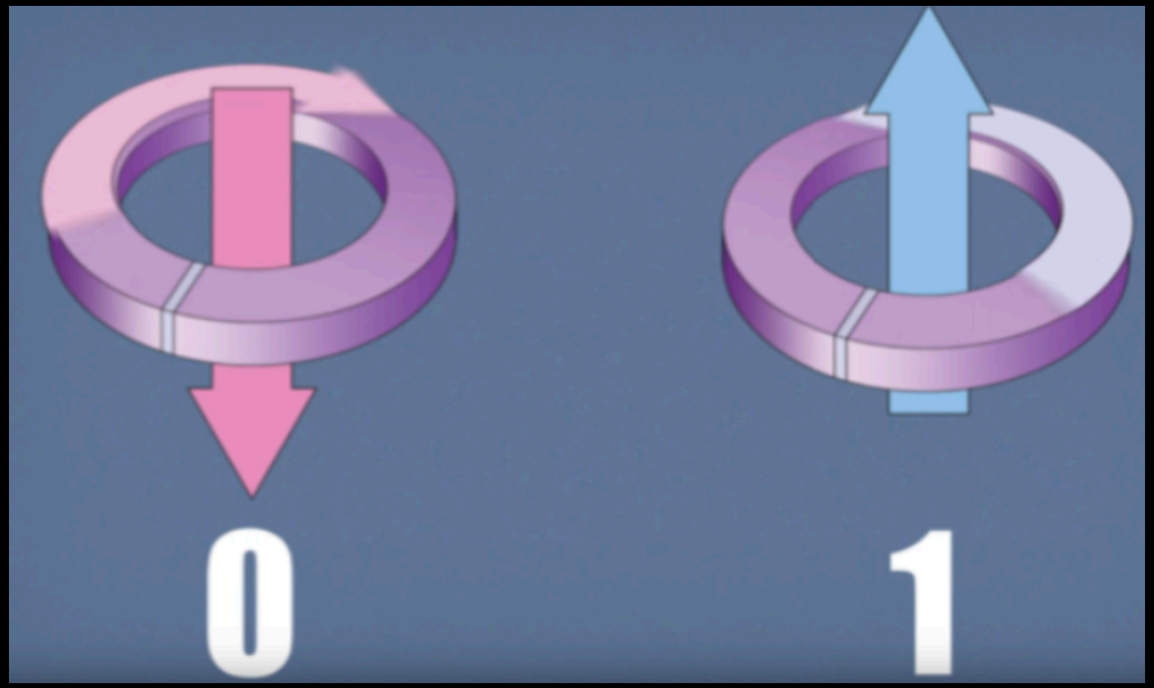
MECÁNICA CUÁNTICA

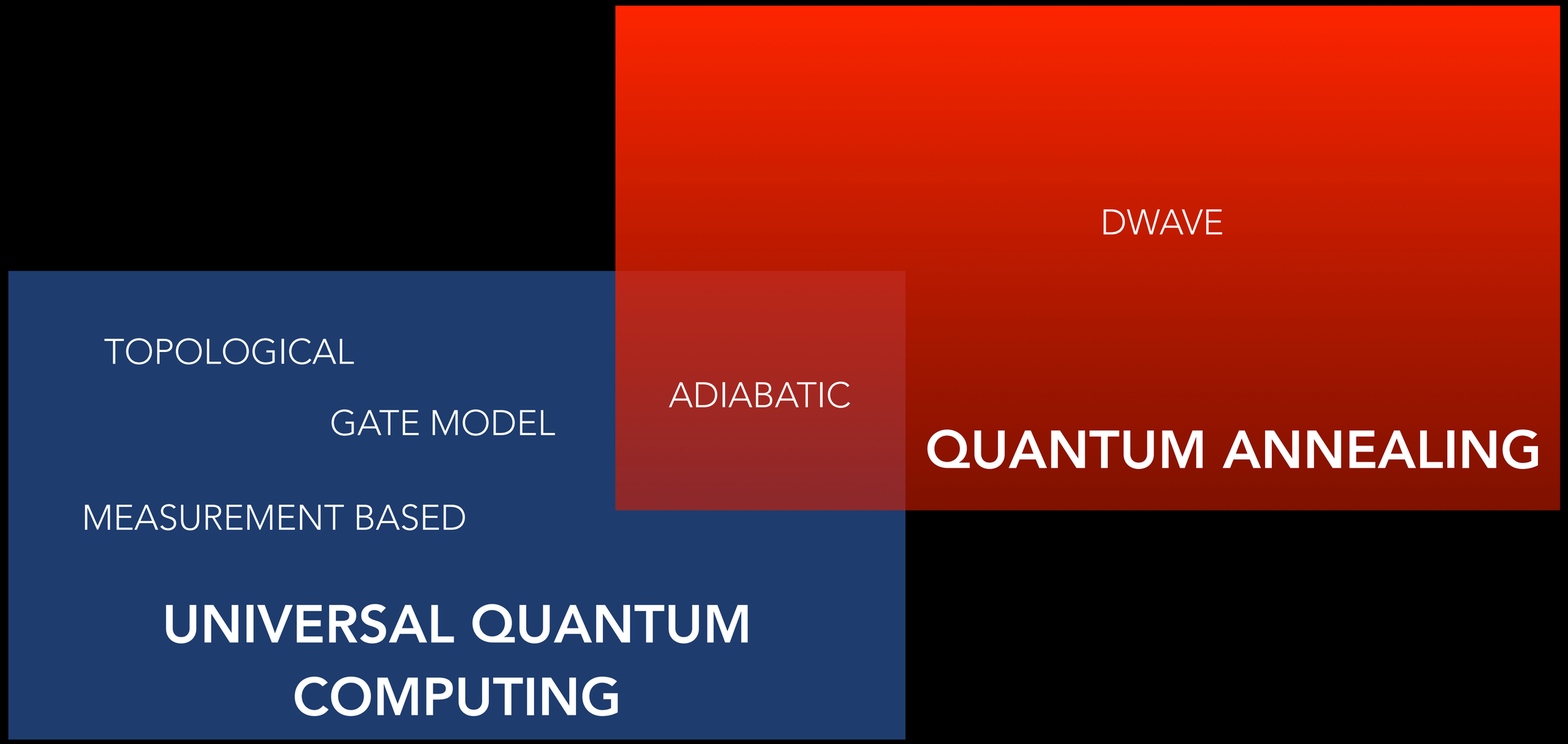

$|0> = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$
$|1> = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$
$|0> = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$
$|1> = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$
$|q_1> = \alpha_1|0> + \beta_1|1>$
$|q_2> = \alpha_2|0> + \beta_2|1>$
$|q_1q_2> = \alpha_1\alpha_2|00> + \alpha_1\beta_2|01> + \beta_1\alpha_2|10> + \beta_1\beta_2|11>$

“UN SISTEMA CUÁNTICO ES UN PRODUCTO DE ESTADOS SI ES QUE EXISTE UNA FORMA DE ESCRIBIRLO COMO UN PRODUCTO DE TENSORES DE 1-QUBIT, EN CASO CONTRARIO, ES UN SISTEMA ENTRELAZADO.”

ENTRELAZAMIENTO DE QUBITS

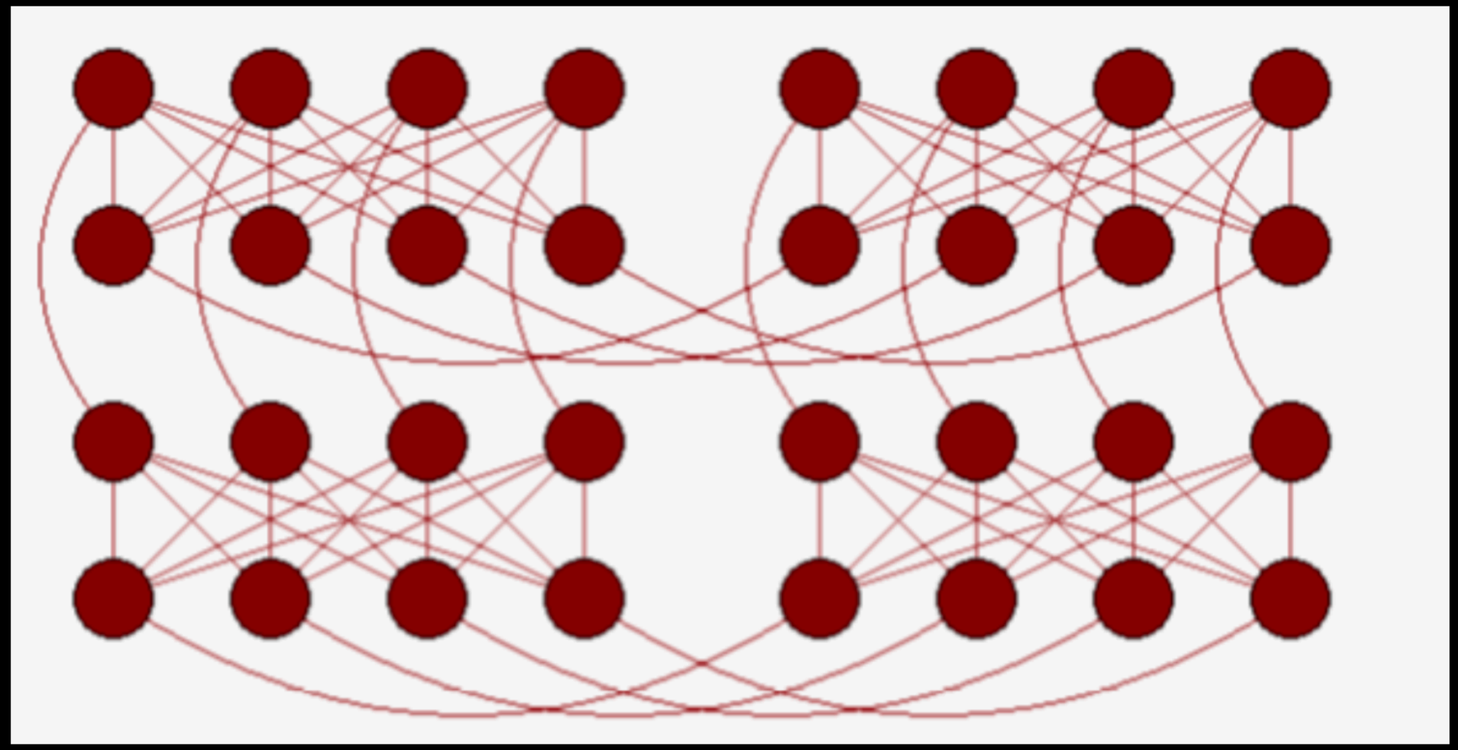
QUANTUM ANNEALING


0
1

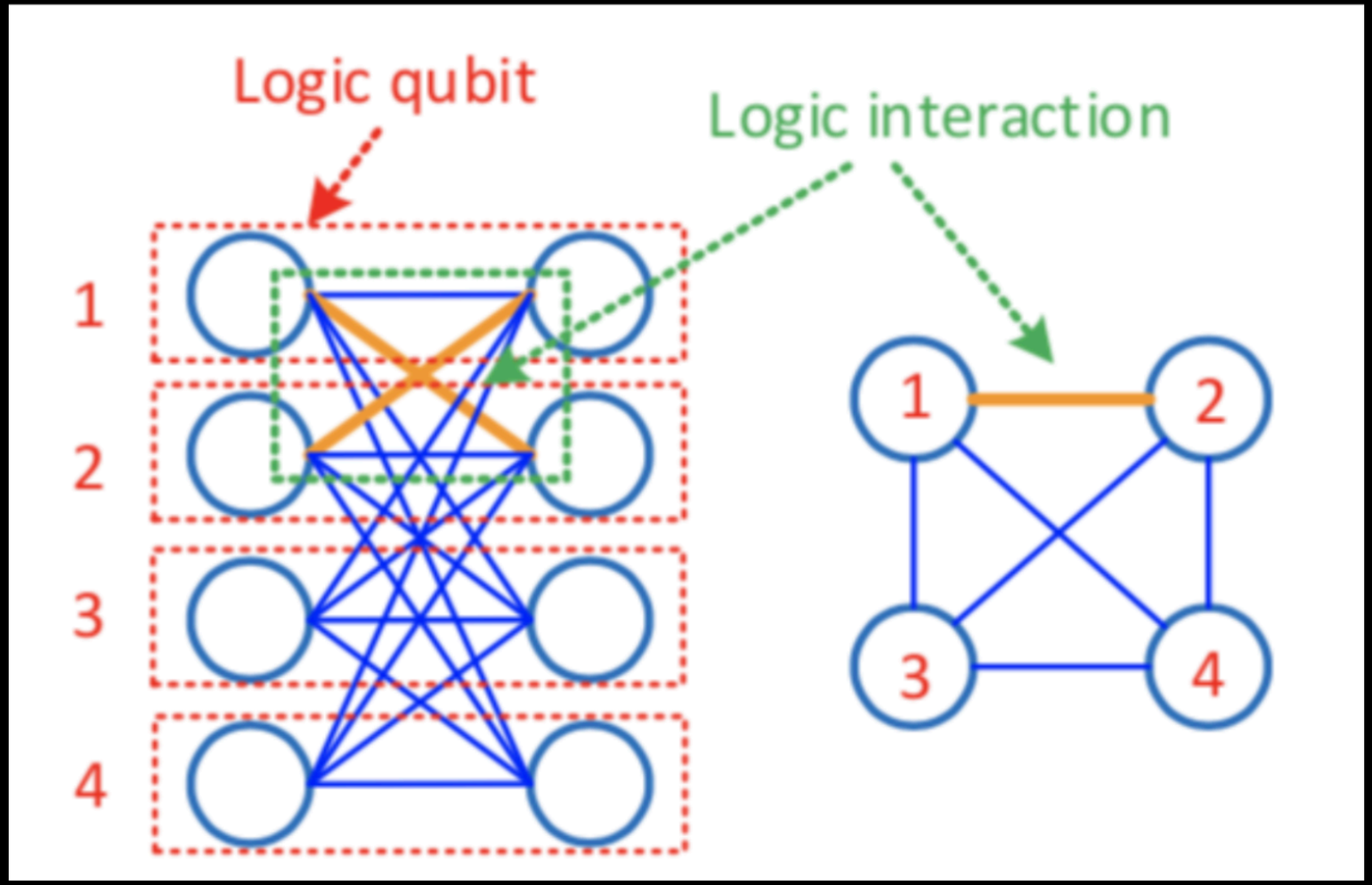

DWAVE
TOPOLOGICAL
GATE MODEL
ADIABATIC
MEASUREMENT BASED
QUANTUM ANNEALING
UNIVERSAL QUANTUM COMPUTING

ARQUITECTURA Y DISPOSICIÓN DE QUBITS

$$I = \sum_{i}^{N} h_i s_i + \sum_{i=1}^{N} \sum_{j=i+1}^{N} J_{i,j} s_i s_j$$

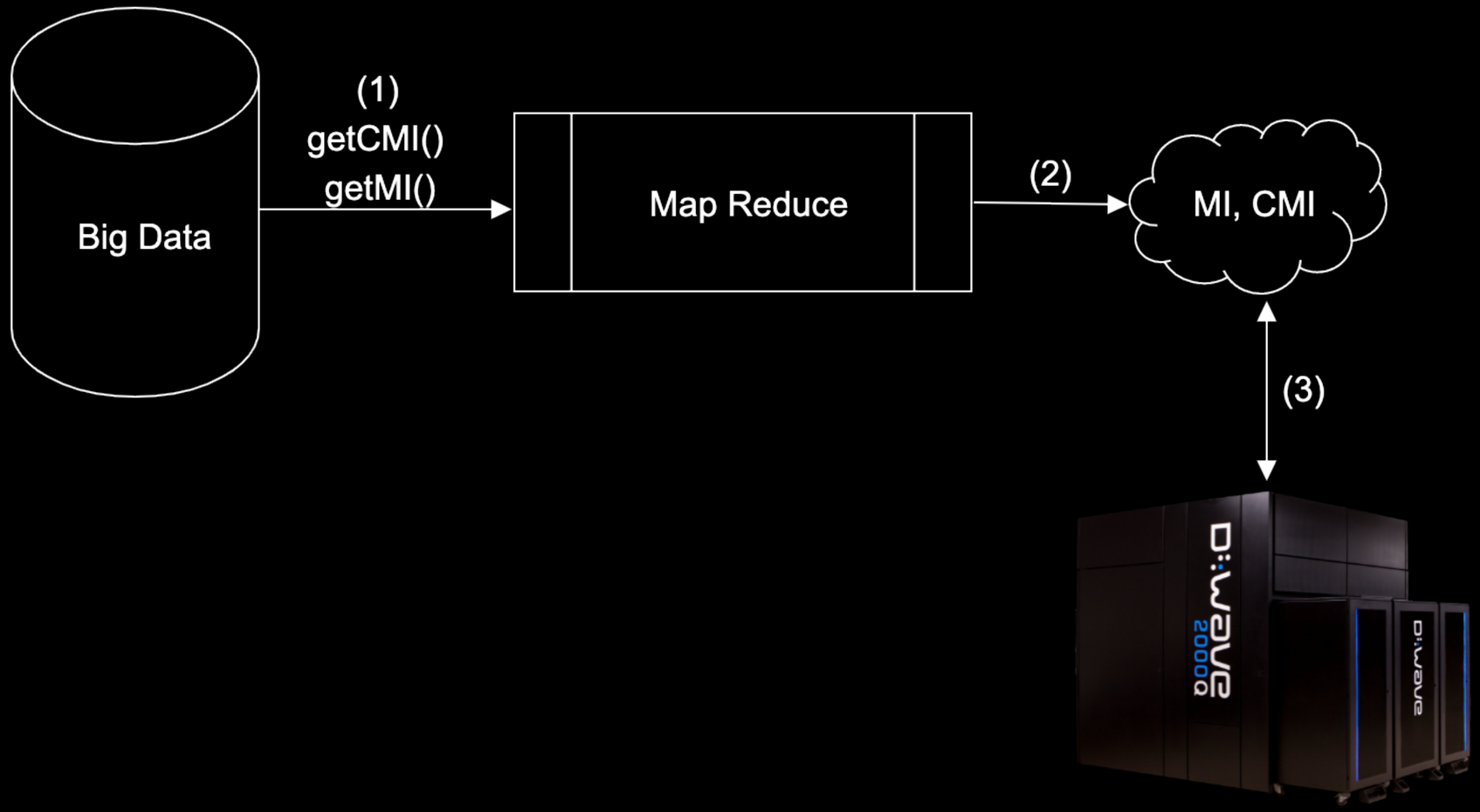


QUBITS LOGICOS VS. QUBITS FÍSICOS



PROPUESTA

ESTRUCTURA PRINCIPAL



JUSTIFICACIÓN

$$\mathcal{O}(IM\ \&\ IMC) + \mathcal{O}(k\_feat\ .\ \_selec) = \mathcal{O}(F^2) + \mathcal{O}(F!)$$

$$QUBO = \sum_{i}^{n} IM(X_i, Y)x_i + \sum_{i} \sum_{j} ICM(X_j; Y \mid X_i)x_i x_j$$

PENALIZANDO RESPUESTAS INVÁLIDAS

$$\alpha \sum_{i}^{n} (x_i - k)^2$$

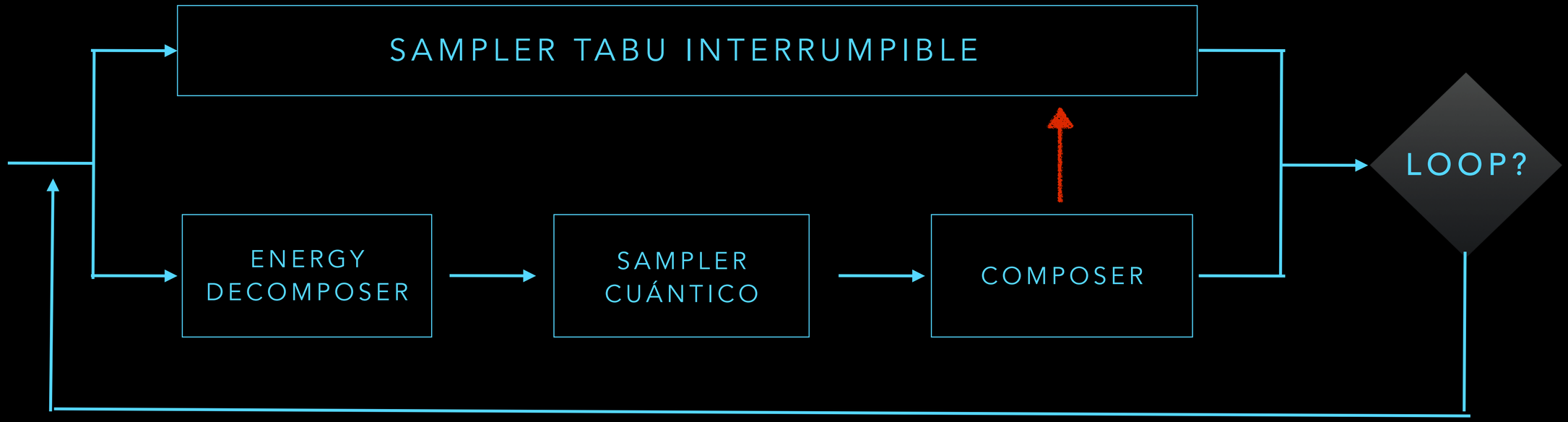
BIG DATA: HYBRID QUANTUM SOLVER

RAMAS

INTERRUPTABLE_TABU_SAMPLER()

DECOMPOSER | SAMPLER | COMPOSER

WORKFLOW PRINCIPAL



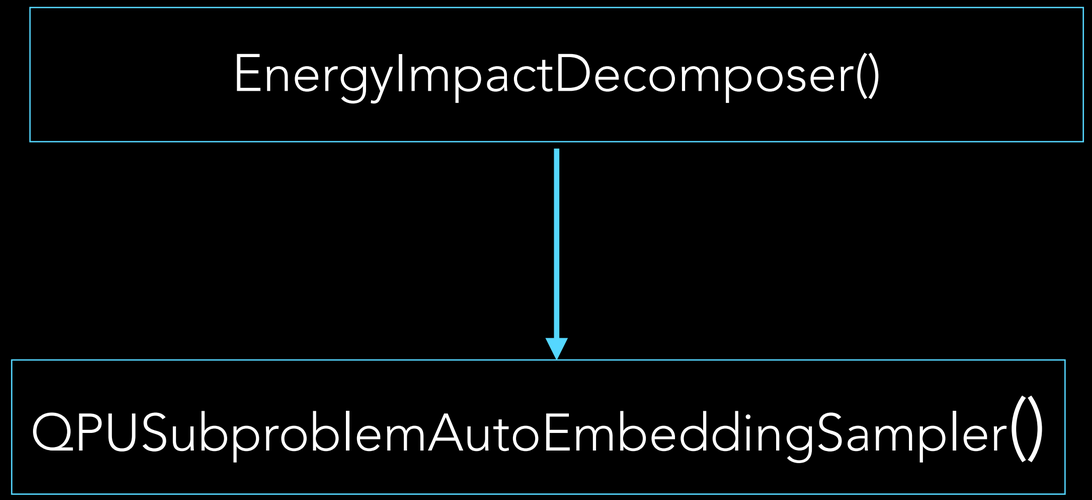
ENERGY IMPACT DECOMPOSER

- Crea subproblemas de un determinado tamaño (batches).

EnergyImpactDecomposer()

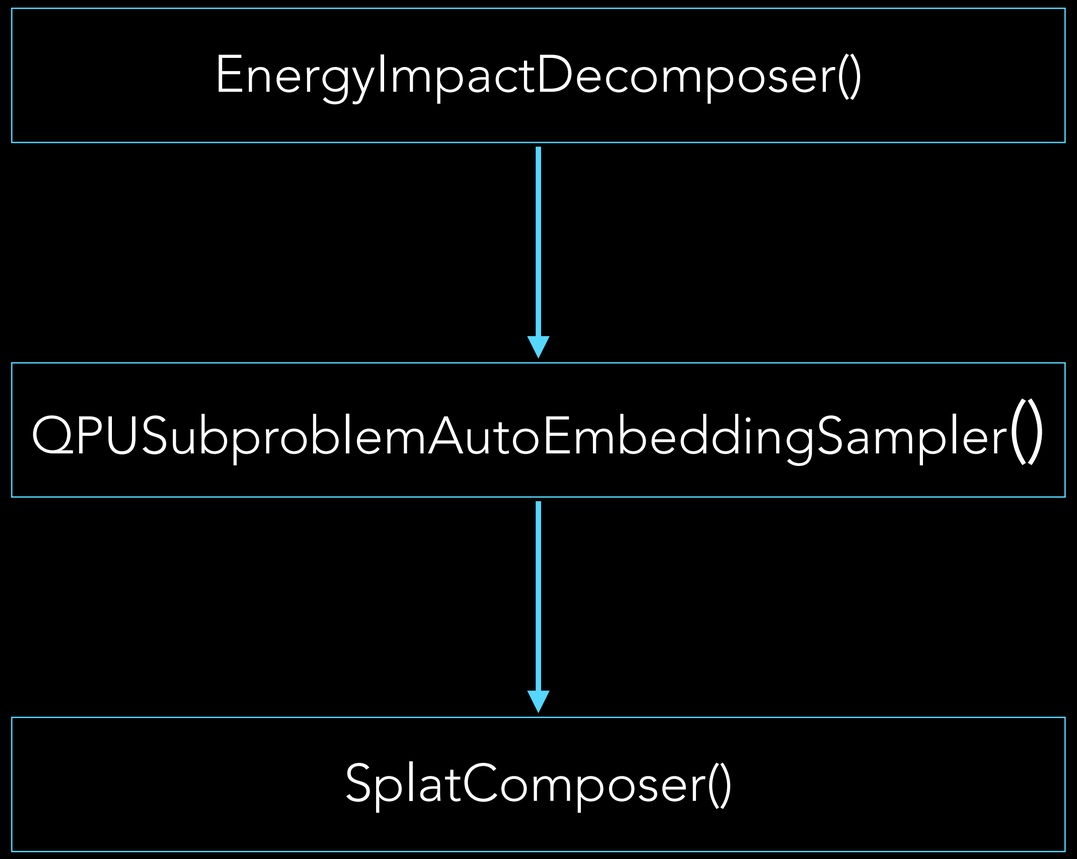
QUANTUM SAMPLER

- Ejecuta un sub-problema en la computadora cuántica



COMPOSER

- Reemplaza y actualiza las soluciones tabu.



EXPERIMENTOS

HARDWARE

CLÁSICO

- 2.3 GHz Dual-Core Intel Core i5
- 8 GB LPDDR3
- 4 núcleos físicos
- 8 núcleos lógicos

CUÁNTICO (DW_2000Q_6)

- 2041 qubits
- 1-10s tiempo de espera.
- **Total post-procesamiento:** 560 μs
- **Tiempo de iteración (anneal):** 20 μs
- **Tiempo programación qpu:** 10719 μs
- **Tiempo de lectura de qubits (por iteración):** 198 μs

SOFTWARE

CLÁSICO

- Python 3.7.6

CUÁNTICO

- **Dwave-system** 0.9.6
- **Dimod** 0.9.4
- **dwave-hybrid** 0.5.0

DATASET 1: TITANIC

¿Cuáles son los mejores atributos para predecir si un pasajero sobrevivió?

1045 registros

14 atributos

batch_size = 2

$\alpha : 10$

Qubits por Variable: 5

RESULTADO PARA K=3

0.578 segundos en QPU

¿Género?

¿ Mr.?

¿ Famoso?

$$\sum IM + \sum IMC = 1.46$$

RESPUESTA

Género

Mr.

Miss.

$$\sum IM + \sum IMC = 2.48$$

DATASET 2: HOUSING

¿Cuáles son los mejores atributos para predecir el precio de una casa?

1460 registros

batch_size: 10

80 atributos

convergencia: 3

PARA $K = 20$, $\alpha = 100$

~6 SEG

1. MSSubClass
2. LotFrontage
3. LotArea
4. Utilities
5. Neighborhood
6. YearBuilt
7. YearRemodAdd
8. Exterior1st
9. Exterior2nd
10. BsmtFinType1
11. BsmtFinSF1
12. BsmtUnfSF
13. TotalBsmtSF
14. 1stFlrSF
15. 2ndFlrSF
16. GrLivArea
17. GarageYrBlt
18. GarageArea
19. OpenPorchSF
20. MoSold
21. YrSold

PARA K = 20, $\alpha = 100000$

~6 SEG

1. MSSubClass
2. LotFrontage
3. LotArea
4. Utilities
5. Neighborhood
6. YearBuilt
7. YearRemodAdd
8. Exterior1st
9. Exterior2nd
10. BsmtFinType1
11. BsmtFinSF1
12. BsmtUnfSF
13. TotalBsmtSF
14. 1stFlrSF
15. 2ndFlrSF
16. GrLivArea
17. GarageYrBlt
18. GarageArea
19. OpenPorchSF
20. YrSold

CONCLUSIONES

- Se desarrolló un método escalable y fácil de implementar.
- Recalentamiento cuántico tiene un futuro prometedor.
- Se vieron limitaciones de hardware producto de una tecnología que es relativamente nueva en el ámbito empírico.
- Hemos demostrado supremacía cuántica.

TRABAJOS FUTUROS

- Mejorar la técnica para espectros poco amplios.
- Desarrollar componentes de workflow que mejoren el desempeño.

Bibliografía

- https://cloud.dwavesys.com/leap/
- https://docs.ocean.dwavesys.com/en/stable/docs_hybrid/intro/using.html
- https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e#:~:text=Feature%20selection%20is%20the%20process,learn%20based%20on%20irrelevant%20features.
- Brown, G., Pocock, A., Zhao, M. J., & Luján, M. (2012). "Conditional likelihood maximisation: a unifying framework for information theoretic feature selection." The journal of machine learning research, 13(1), 27-66.
- X. V. Nguyen, J. Chan, S. Romano, and J. Bailey, “Effective global approaches for mutual information based feature selection”. A. Montanaro, “Quantum algorithms: an overview,” npj Quantum Information, vol. 2, p. 15023, Jan 2016.
- A. Leonard Susskind, Quantum Mechanics.Basic Books, 2014.