NASA DAG ML REPORT

Jasper Doan
Donald Bren School of Information and Computer Sciences
University of California, Irvine
Irvine, CA 92697
jasperd1@uci.edu

September 4, 2023

ABSTRACT

Develop an approach to validate expert-drawn graphs by building probabilistic graph model DAGs, through populating Machine-Learning with empirical biological data from the NASA Open Science Data Repository. Tasked with researching IAMB, Fast-IAMB, Inter-IAMB, and IAMB-FDR algorithms for medical risk assessment to let NASA HSRB formalize a shared causal flow of risk model among Risk Board stakeholders.

1 Introduction

The goal of our project is to validate expert-drawn Directed Acyclic Graphs (DAGs) for Human Spaceflight Risks, for tracking and researching risks that astronaut crews face during spaceflight. This paper goes over the process of using Bnlearn's IAMB, Fast-IAMB, Inter-IAMB, and IAMB-FDR algorithm DAG generation to formalize a shared causal flow of risk model.

Directed Acyclic Graphs (DAG) DAGs are network maps which have unidirectional arrows (directed) and do not allow feedback loops (acyclic).

In the context of the The Human System Risk Board (HSRB), DAGs are used to represent the chain of events that lead from spaceflight exposures to negative mission-level outcomes. This enables two immediate uses as well as sets the stage for further evolution of the causal networks as tools of inference.

Why DAG? Enable mathematical analysis of the relationships between factors and can potentially assess the strength of influence if quantitative values are assigned to nodes and edges.

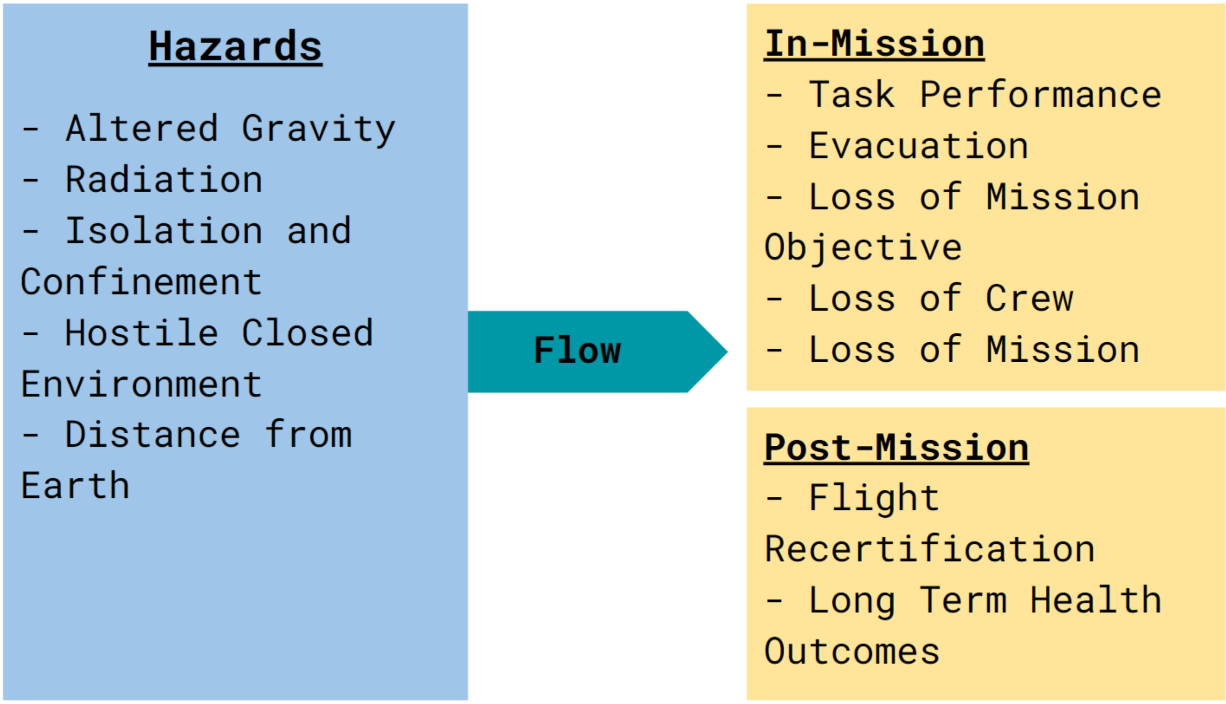
Subject to challenge and improvement based on evolving evidence. If new evidence suggests a lack of causal connection, the corresponding connection can be removed. It can aid in conveying high-level and aggregated concepts that link key components of causal flow to downstream effects in the risk domain. They facilitate communication and the development of shared mental models during board or stakeholder meetings.

DAGs for communication of complex human spaceflight risks Limits the provision of in-mission support capabilities and resources, while simultaneously increasing the need for them

Limits on these capabilities and resources stem from constraints on mass, volume, power, and data bandwidth allocations available to the vehicle's systems/habitats used by astronauts; the further a mission takes astronauts from Earth, the greater these constraints and thus the less support capability they will have. The need for capabilities and resources is increased because the further a mission goes from Earth, the longer astronauts are exposed to degradation by the spaceflight environment.

DAGs aid in prioritizing research and development Evaluation of Human System Risks is necessary to prioritize the allocation of limited research, surveillance, and technology development resources.

The previous scoring system (Red, Yellow, Green) did not consider the complex interactions and synergies between risks, which can amplify risks in other body systems or at a later time. Directed acyclic graphs (DAGs) help analyze the structure of risks and identify important factors in the causal network. Nodes in the DAG represent factors that have many effects, bridge or join risks together, or exist in the middle of the action. DAG analysis provides insights into the interdependencies and cumulative effects of risks faced by astronauts during missions.



2 Theory

Bayesian networks You can represent these relationships between variables by building a Bayesian networks. Where it shows the representation of how nodes/variables interact with each other (direct dependencies between variables) → Allows the user to affirm and make accurate predictions based on observed data.

For instance, if we observe that the weather is rainy, we can use the Bayesian network to estimate the probability of carrying an umbrella and the probability of the ground being wet. We can also do the reverse: if we know the ground is wet, we can estimate the probability that it is raining. By using probabilities and the relationships encoded in the graph, Bayesian networks allow us to reason and make inferences about the variables even when we have incomplete or uncertain information.

Constraint-based methods (What IAMB Variations do) Discover the dependencies and relationships between nodes based on data by imposing certain constraints. Identify statistical dependencies between nodes → infer the underlying structure of the Bayesian network

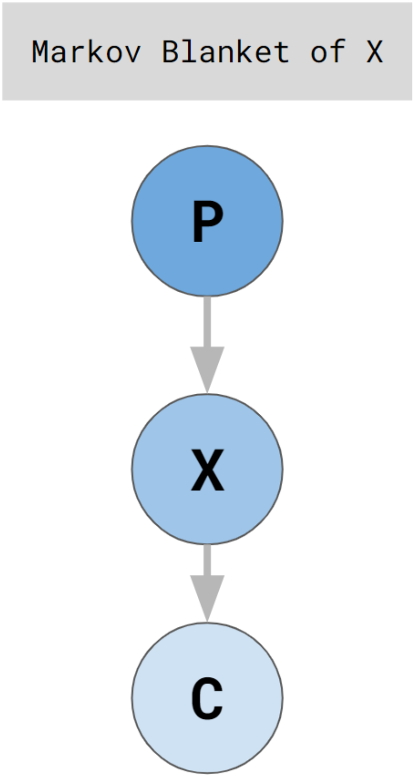
- Independence Testing: Determines the statistical independence or dependence between pairs of variables in the data. Test whether the variables are conditionally independent given other variables. If two variables are found to be independent, it suggests that there is no direct edge between them in the Bayesian network.
- Skeleton Discovery: Constructs a skeleton or an initial structure for the Bayesian network. Represents the presence or absence of edges between variables. Edges are added to the graph for variables that are found to be dependent, indicating a potential causal relationship.
- Orientation of Edges: Determine the orientation of the edges in the network. Establish the direction of causality between variables. Examining conditional dependencies and using additional tests or heuristics to infer the most likely direction of causal influence.

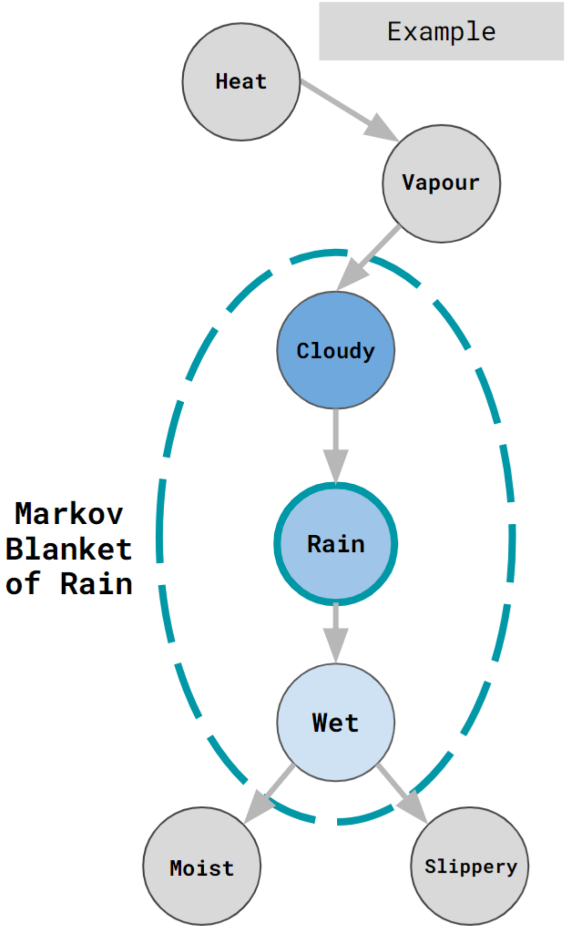
Markov Blanket The Markov blanket of a variable in a Bayesian network: Minimal set of variables that contains all the information necessary to predict the variable's value, given the values of other variables in the network.

Formally, the Markov blanket of a variable X in a Bayesian network consists of three sets of variables:

- Parents of X: These are the variables that directly influence the value of X. If A is influenced by B, then B would be a parent of A.
- Children of X: These are the variables that are directly influenced by the value of X. If C is influenced by A, then C would be a child of A.
- Parents of X's children: These are the variables that are parents of X's children. If C is influenced by B, then B would be a parent of C's child (C) and would be included in the Markov blanket of A.

Simpler terms: Imagine you have three variables: Cloudy, Rain, Wet Grass. From what you know (observed data), they do influence each other's decisions. Rain's Markov blanket consists of the variables that directly influence it. It will form a special group that has all the information needed to understand how rain will happen.





So, who would be in Rain's Markov blanket? Firstly, it would include Cloudy because, from 21 years of living, you know if its cloudy -> High-chance it will rain! Secondly, it would include any nodes that are directly influenced by rain, like wet grass, wet floor, or cold.

If you think "Rain" as a person, Rain's Markov blanket is like a small group of people who are most important to her decision-making. If you know what these people are doing, you can make a pretty good guess about what Rain is likely to do.

In a Bayesian network, Markov blanket (MB) of a variable is the group of variables that have a direct influence on it or are directly influenced by it. It's the minimal set of variables that you need to pay attention to in order to understand and predict the behavior of that variable, without worrying about all the other variables in the network.

The Markov blanket concept helps us simplify things by focusing on a smaller, important group rather than considering the entire network. It allows us to make predictions or perform calculations about a specific variable by looking only at the variables in its Markov blanket.

Generally the process of finding the Markov blanket of a variable is called Markov blanket discovery. It is a key step in many algorithms for learning the structure of a Bayesian network from data. And its Pseudo-code is as follows:

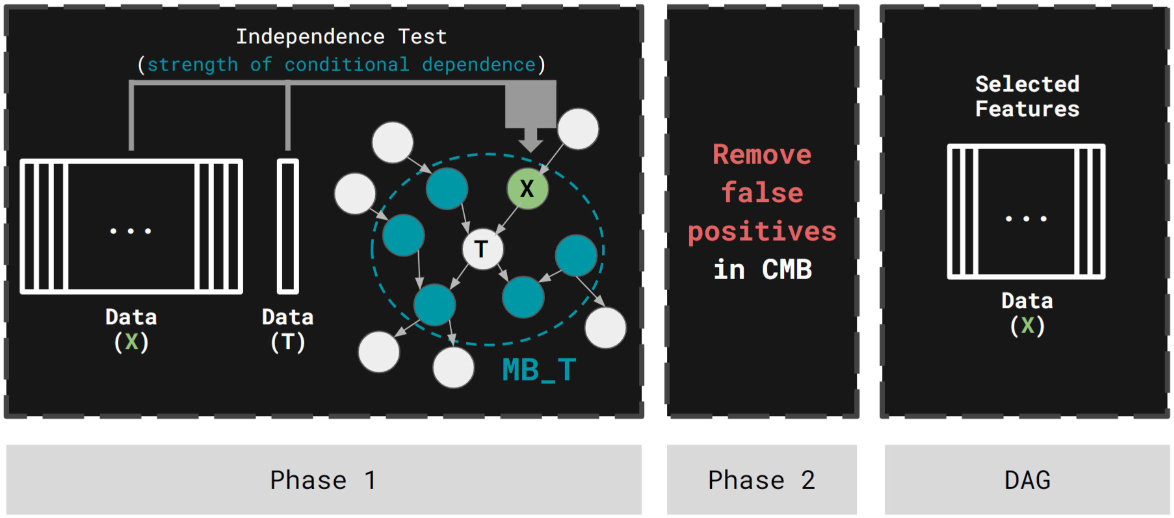
```
# 1. Preprocess the data: Handle data cleaning, handling missing values
data = preprocess_data(data)

# 2. Define the set of variables and their relationships
variables = get_variables_from_data(data)
causal_relationships = define_causal_relationships()

# 3. Initialize the Markov Blanket for each variable
markov_blanket = {}
for variable in variables:
    markov_blanket[variable] = find_markov_blanket(variable, data)

# 4. Learn causal relationships from the Markov Blanket
bayesian_network = create_empty_bayesian_network()
for variable in variables:
    parents = find_parents(variable, markov_blanket[variable])
    for parent in parents:
        bayesian_network.add_edge(parent, variable)

# 5. Output the Bayesian Network
bayesian_network.to_graph()

```

3 Algorithm Description

3.1 IAMB | Incremental Association Markov Blanket

Consists of two phases, a forward and a backward one. An estimate or copy of the Markov Blanket of a variable of interest T is kept in the MB. In the forward phase all variables that belong in $MB(T)$, including false positives enter CMB (Copy of MB) while in the backward phase the false positives are identified and removed so that $CMB = MB(T)$ in the end.

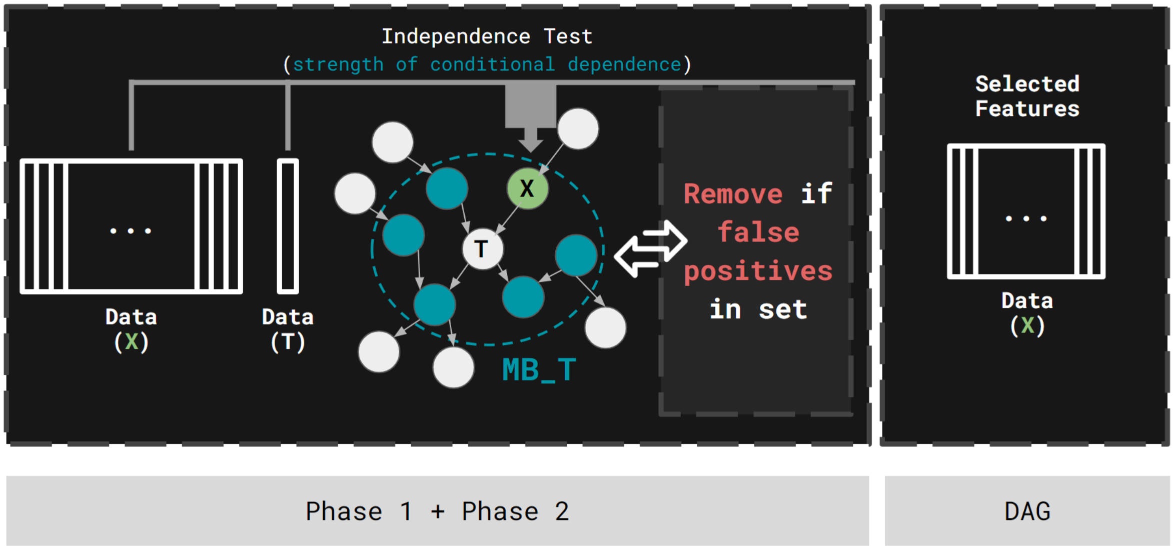
Phase I is the following: start with an empty candidate set for the CMB and admit into it (in the next iteration) the variable that maximizes a heuristic function $f(X;T|CMB)$. This is a measure of association between X and T given CMB. In backward conditioning (Phase II) we remove one-by-one the features that do not belong to the $MB(T)$ by testing whether a feature X from CMB is independent of T given the remaining CMB.



3.2 Fast-IAMB | Fast-Incremental Association Markov Blanket

Considered to be an improvement over the IAMB algorithm that aims to reduce the computational complexity. It aims to reduce the computational and time complexity of the algorithm by pruning unnecessary tests. Maintaining an order of the variables based on their likelihood of being in the Markov blanket. Uses a combination of forward and backward steps to efficiently determine the Markov blanket for each variable.

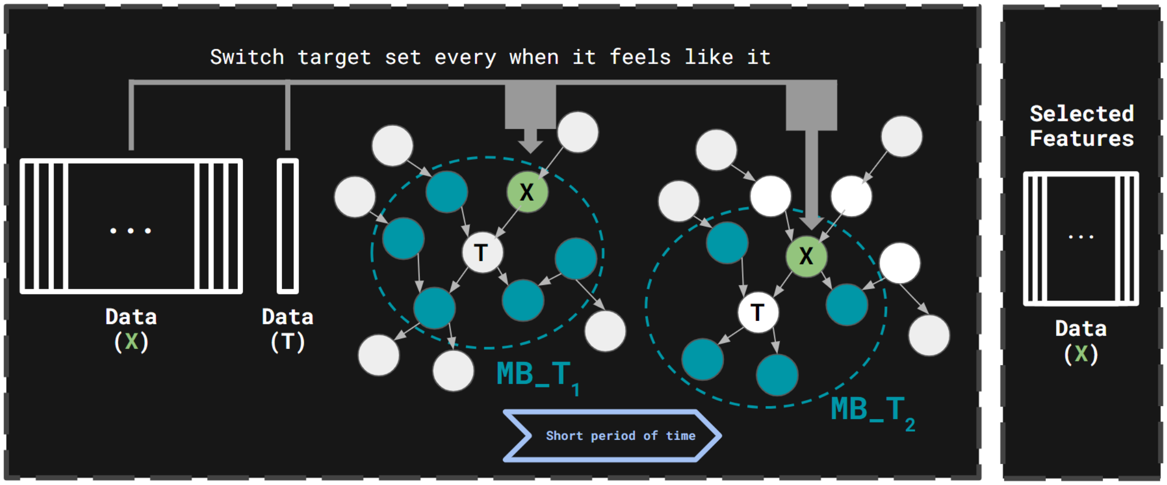
Instead of relying on traditional statistical tests like chi-squared (kai-squared). It uses a combination of forward and backward steps to determine the MB for each variable. Computationally this is just $O(n)$ while IAMB is $O(2n)$ where it needs go forward and backward, here it does it in 1 loop.



3.3 Inter-IAMB | Interleaved Incremental Association Markov Blanket

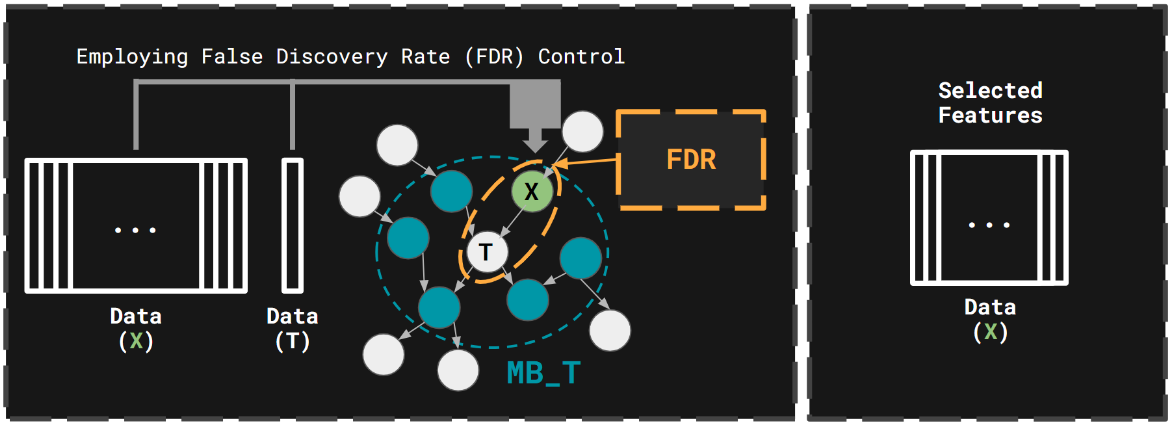
Extension of the IAMB algorithm that introduces an interleaved strategy for constructing the Markov blanket. Instead of iterating and processing one target variable at a time, it selects multiple target variables from the dataset and performs interleaved passes over the variables, gradually refining the Markov blankets. Allows for more efficient identification of variable dependencies and can lead to improved results in certain scenarios.

For each selected target variable T, it just perform the regular IAMB algorithm based on independence tests. But after a short duration of processing, it switch to another target variable and continue the process. The switching between target variables is the "interleaving" part of the algorithm. By interleaving the runs, the algorithm can potentially reduce the total number of independence tests required and, thus, improve computational efficiency.



3.4 IAMB-FDR | Incremental Association Markov Blanket with False Discovery Rate Control

Variant of the IAMB algorithm that incorporates false discovery rate (FDR) control to address the issue of multiple hypothesis testing. In structure learning, multiple statistical tests are performed, and without proper control, this can lead to an increased chance of false discoveries. IAMB-FDR applies FDR control techniques to adjust the statistical significance thresholds used in independence tests, thereby mitigating the risk of false discoveries and improving the reliability of the learned structure.



4 Data Construction

5 Examples of citations, figures, tables, references

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem. Sed neque lectus, consectetuer at, consectetuer sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus. Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui. [1, 2] and see [3].

The documentation for `natbib` may be found at

`http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf`

Of note is the command `\citet`, which produces citations appropriate for use in inline text. For example,

```
   \citet{hasselmo} investigated\dots
```

produces

> Hasselmo, et al. (1995) investigated…

`https://www.ctan.org/pkg/booktabs`

5.1 Figures

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi. [1] Sed feugiat. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Ut pellentesque augue sed urna. Vestibulum diam eros, fringilla et, consectetuer eu, nonummy id, sapien. Nullam at lectus. In sagittis ultrices mauris. Curabitur malesuada erat sit amet massa. Fusce blandit. Aliquam erat volutpat. Aliquam euismod. Aenean vel lectus. Nunc imperdiet justo nec dolor.

[1] Sample of the first footnote.

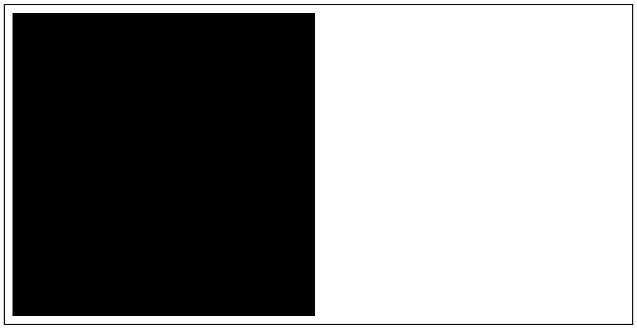


Figure 1: Sample figure caption.

Table 1: Sample table title

Part		
Name	Description	Size (μm)
Dendrite	Input terminal	$\sim$100
Axon	Output terminal	$\sim$10
Soma	Cell body	up to 10^6

$$\xi_{ij}(t) = P(x_t = i, x_{t+1} = j | y, v, w; \theta) = \frac{\alpha_i(t) a_{ij}^{w_t} \beta_j(t+1) b_j^{v_{t+1}}(y_{t+1})}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i(t) a_{ij}^{w_t} \beta_j(t+1) b_j^{v_{t+1}}(y_{t+1})} \quad (1)$$

5.2 Tables

Etiam euismod. Fusce facilisis lacinia dui. Suspendisse potenti. In mi erat, cursus id, nonummy sed, ullamcorper eget, sapien. Praesent pretium, magna in eleifend egestas, pede pede pretium lorem, quis consectetuer tortor sapien facilisis magna. Mauris quis magna varius nulla scelerisque imperdiet. Aliquam non quam. Aliquam porttitor quam a lacus. Praesent vel arcu ut tortor cursus volutpat. In vitae pede quis diam bibendum placerat. Fusce elementum convallis neque. Sed dolor orci, scelerisque ac, dapibus nec, ultricies ut, mi. Duis nec dui quis leo sagittis commodo. See awesome Table 1.

5.3 Lists

- Lorem ipsum dolor sit amet
- consectetur adipiscing elit.
- Aliquam dignissim blandit est, in dictum tortor gravida eget. In ac rutrum magna.

References

[1] George Kour and Raid Saabne. Real-time segmentation of on-line handwritten arabic script. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pages 417–422. IEEE, 2014.

[2] George Kour and Raid Saabne. Fast classification of handwritten on-line arabic characters. In *Soft Computing and Pattern Recognition (SoCPaR), 2014 6th International Conference of*, pages 312–318. IEEE, 2014.

[3] Guy Hadash, Einat Kermany, Boaz Carmeli, Ofer Lavi, George Kour, and Alon Jacovi. Estimate and replace: A novel approach to integrating deep neural networks with existing applications. *arXiv preprint arXiv:1804.09028*, 2018.