The Cost of Sharing:
From Topological Multitasking Limits to Semantic Horizons

G. Petri

NSIA Workshop, Netsci 2026

1/6/2026

Network Science Institute
at Northeastern University

NPL RESEARCH

Beyond TheEdge

MSCA

RUNES

erc
European Research Council
Established by the European Commission

The Cost of Sharing:
From Topological Multitasking Limits to Semantic Horizons

G. Petri

NSIA Workshop, Netsci 2026

1/6/2026

Network Science Institute
at Northeastern University

NPL RESEARCH

Beyond TheEdge

MSCA

RUNES

erc
European Research Council
Established by the European Commission

Experiment #1

Name the color of the following stimulus
and, *at the same time*, point to where it is...

Experiment #1

Name the color of the following stimulus and, *at the same time*, point to where it is...

BROWN

Experiment #1

Name the color of the following stimulus
and, *at the same time*, point to where it is...

BLUE

YELLOW

Experiment #2

point **left** if the written word is

⬅ **RED**

point **right** if the written word is

GREEN ➡

Experiment #2

point **left** if the written word is

⬅ **RED**

point **right** if the written word is

GREEN ➡

RED

Experiment #2

point **left** if the written word is

⬅ **RED**

point **right** if the written word is

GREEN ➡

GREEN

RED

Experiment #3

Name the color of the following stimulus and, *at the same time*:

point **left** if the written word is

⬅ **RED**

point **right** if the written word is

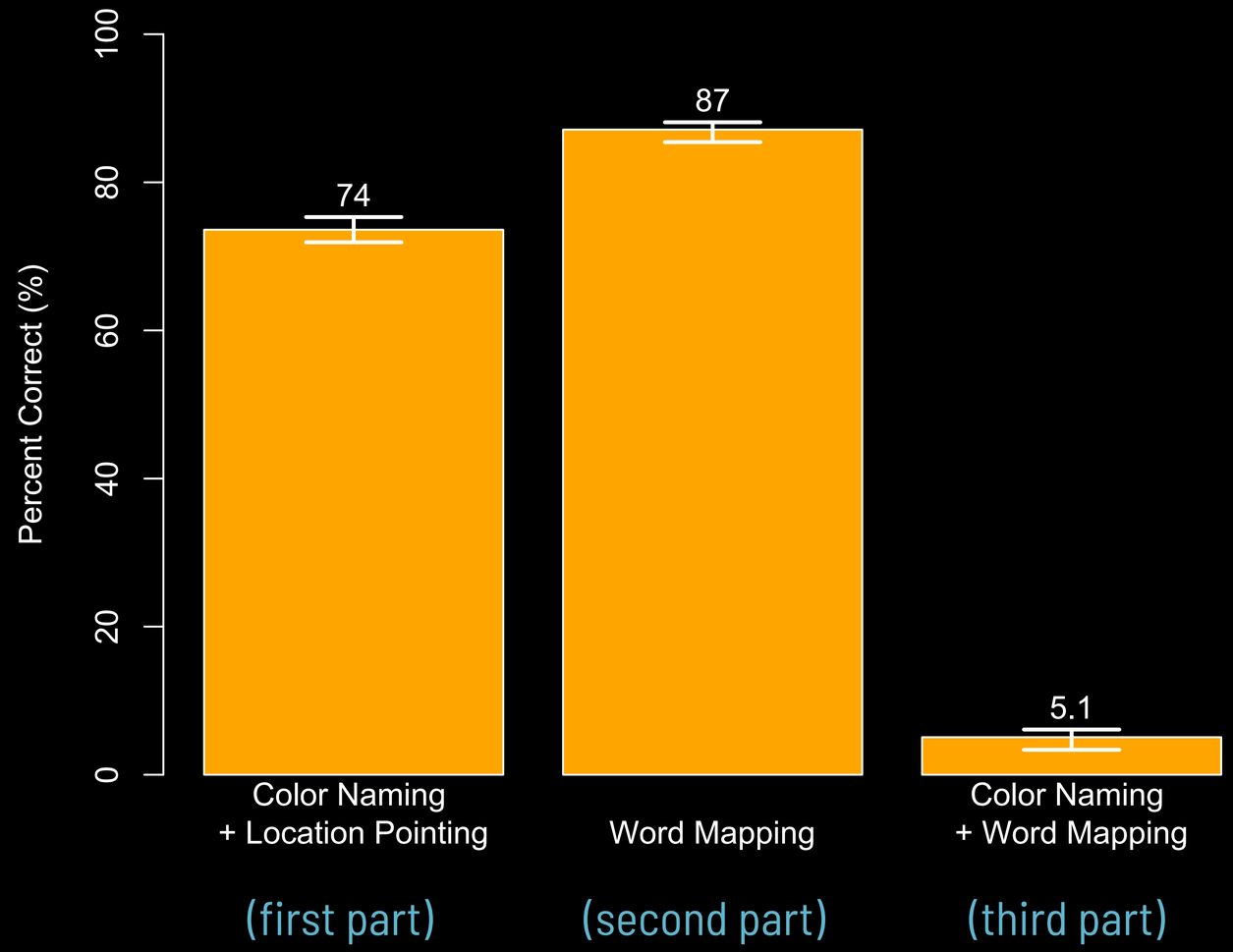
GREEN ➡

GREEN

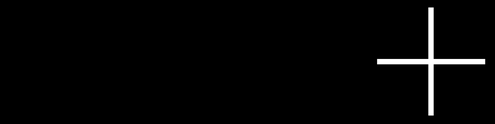
RED

Experiment

Accuracy Results



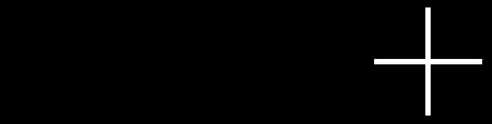
Under which conditions can we multitask?



Under which conditions can we multitask?

Understanding general computation

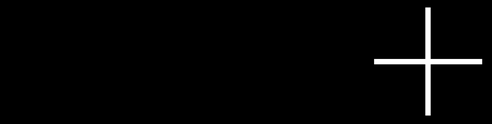
Under which conditions can we multitask?



Understanding general computation

Practical application in design of sensitive systems

Under which conditions can we multitask?

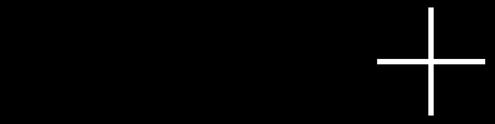


Understanding general computation

Practical application in design of sensitive systems

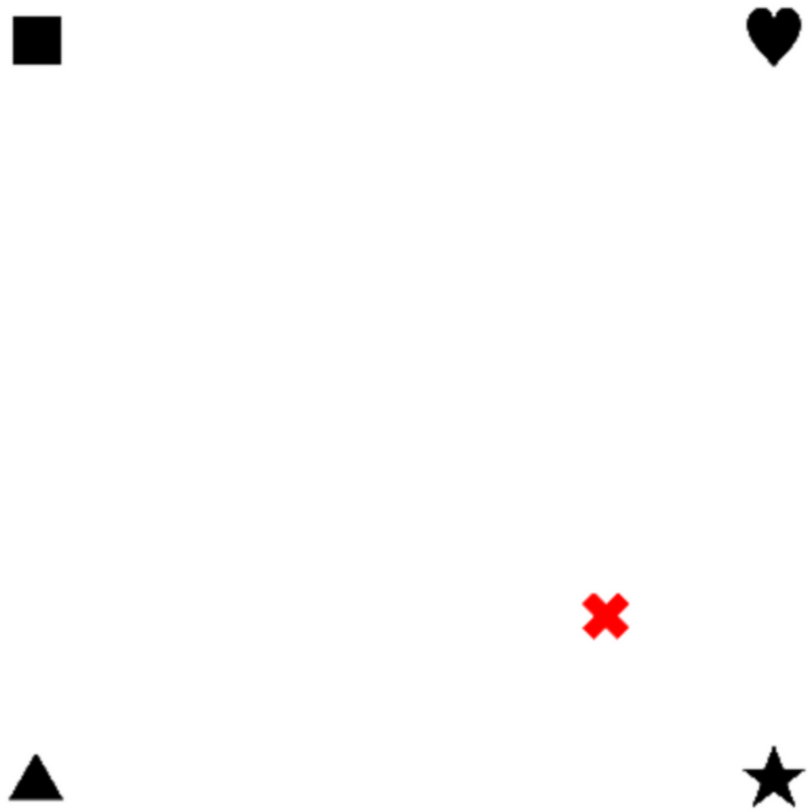
Automated AI agents

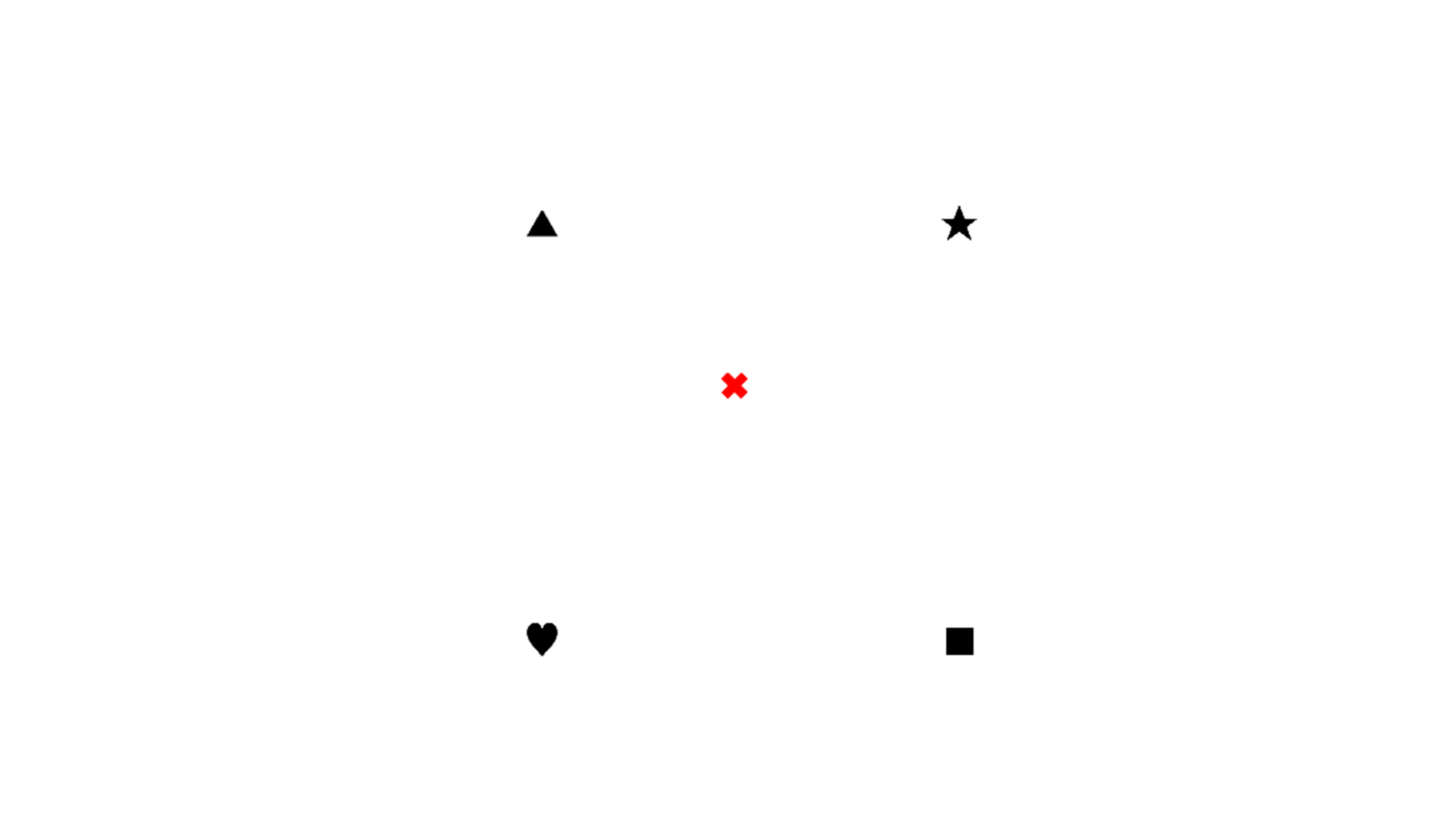
Under which conditions can we multitask?



What is the closest object to the red cross?

What is the closest object to the red cross?









These two tasks share a fundamental similarity

These two tasks share a fundamental similarity

Limits emerging from information encoding

Ok... Why?

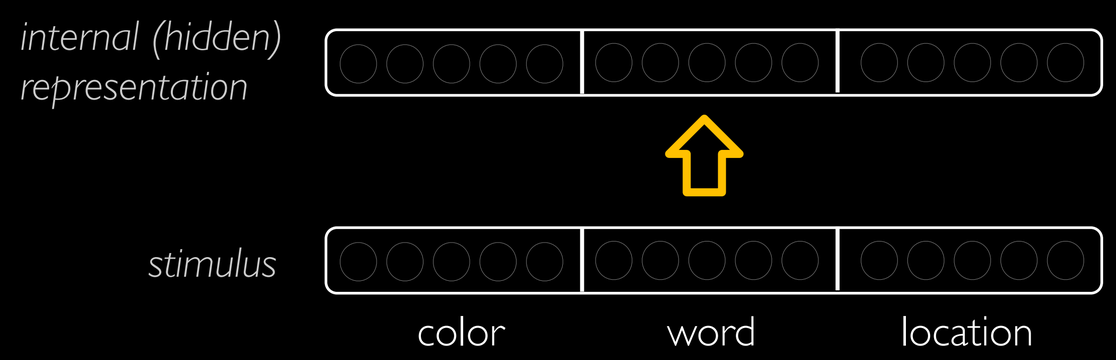
Neural Network Model of Controlled Processing

(Cohen et al., 1990 ; Feng et al., 2014; Musslick et al., 2016)

Neural Network Model of Controlled Processing

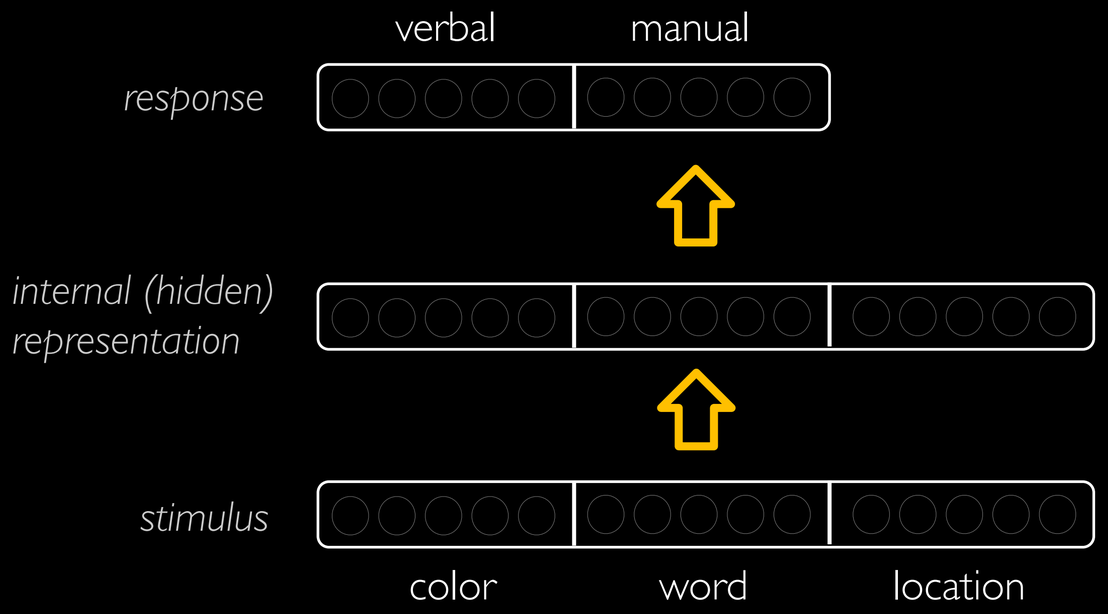
(Cohen et al., 1990 ; Feng et al., 2014; Musslick et al., 2016)

Neural Network Model of Controlled Processing



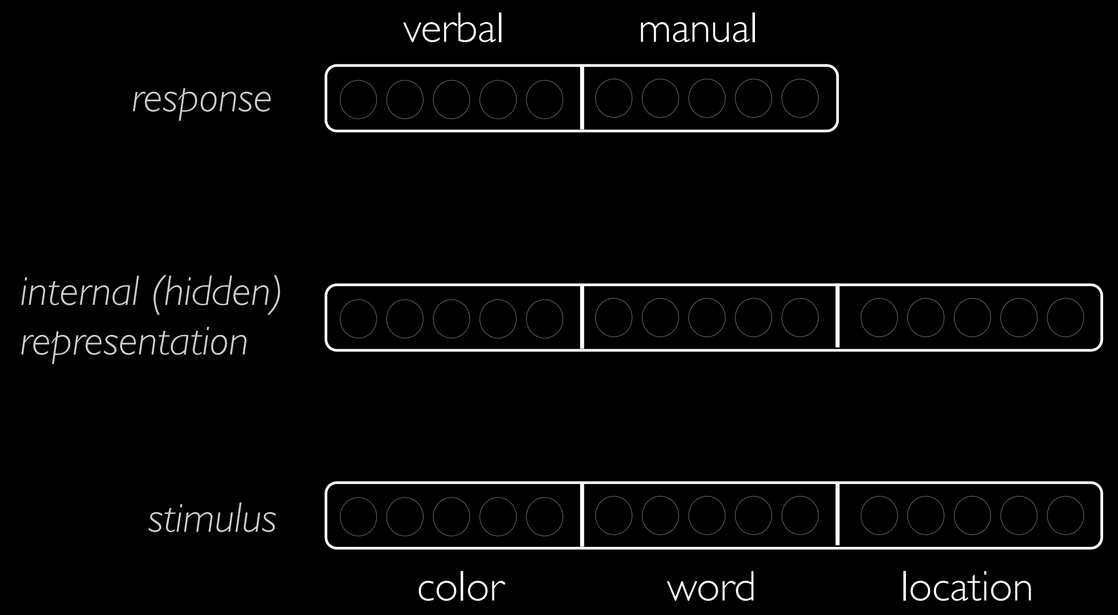
(Cohen et al., 1990 ; Feng et al., 2014; Musslick et al., 2016)

Neural Network Model of Controlled Processing



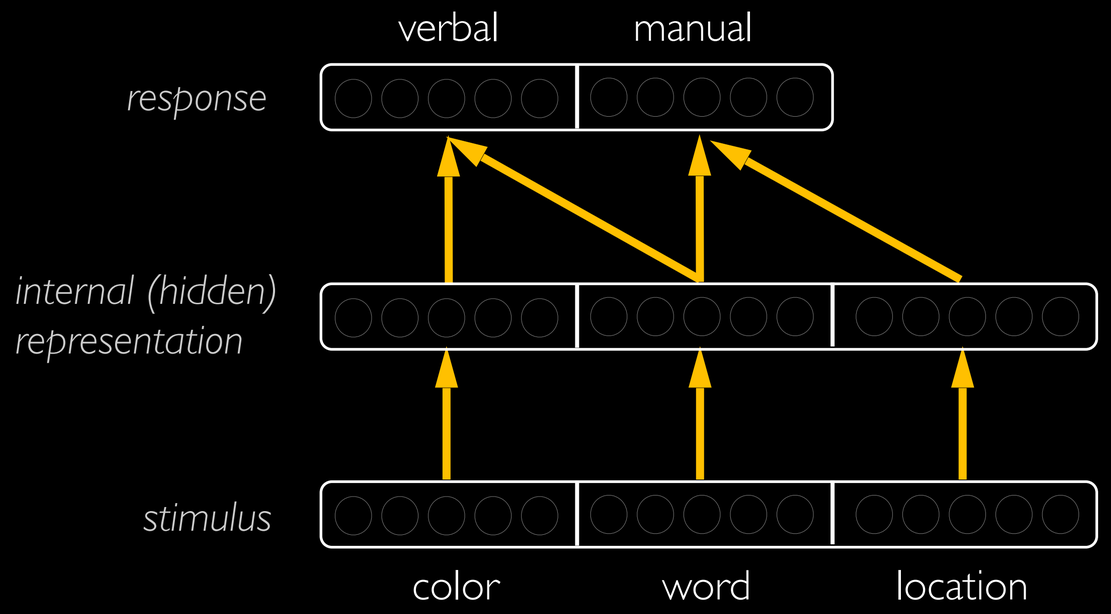
(Cohen et al., 1990 ; Feng et al., 2014; Musslick et al., 2016)

Neural Network Model of Controlled Processing



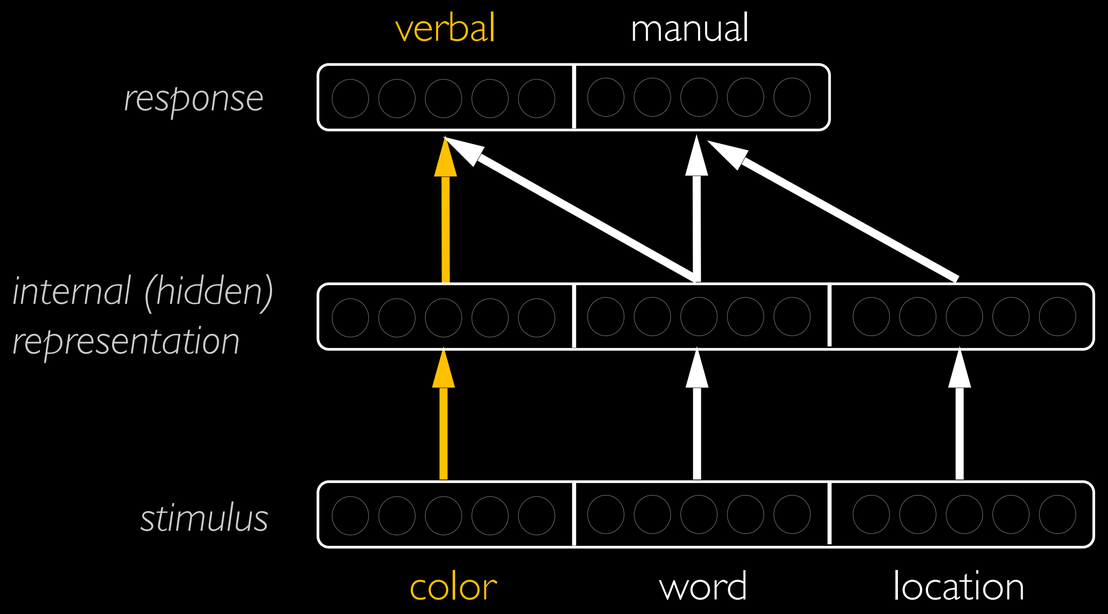
(Cohen et al., 1990 ; Feng et al., 2014; Musslick et al., 2016)

Neural Network Model of Controlled Processing



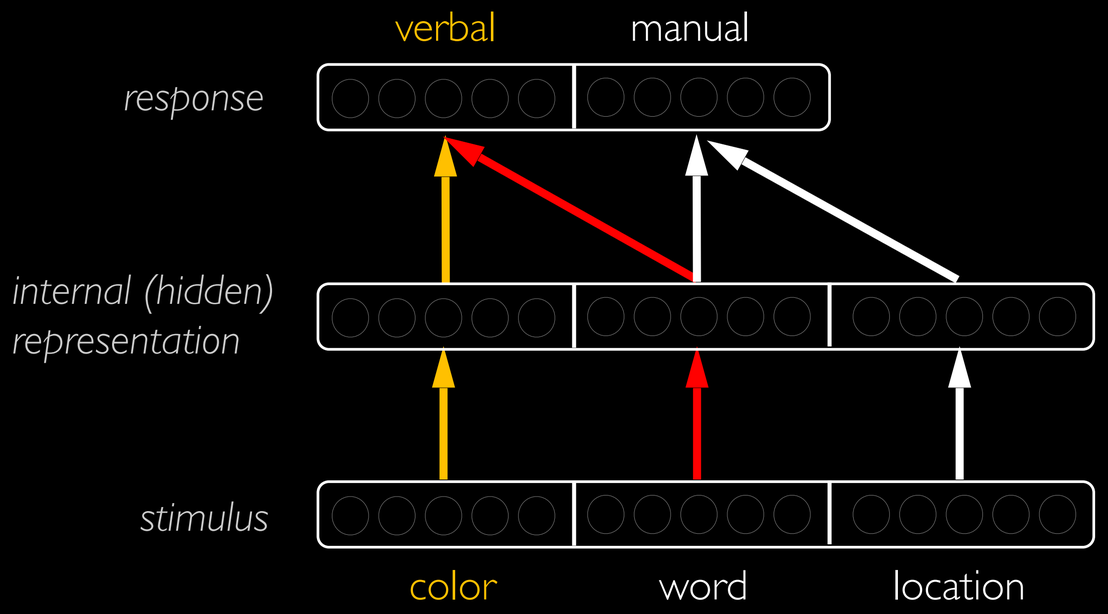
(Cohen et al., 1990 ; Feng et al., 2014; Musslick et al., 2016)

Neural Network Model of Controlled Processing



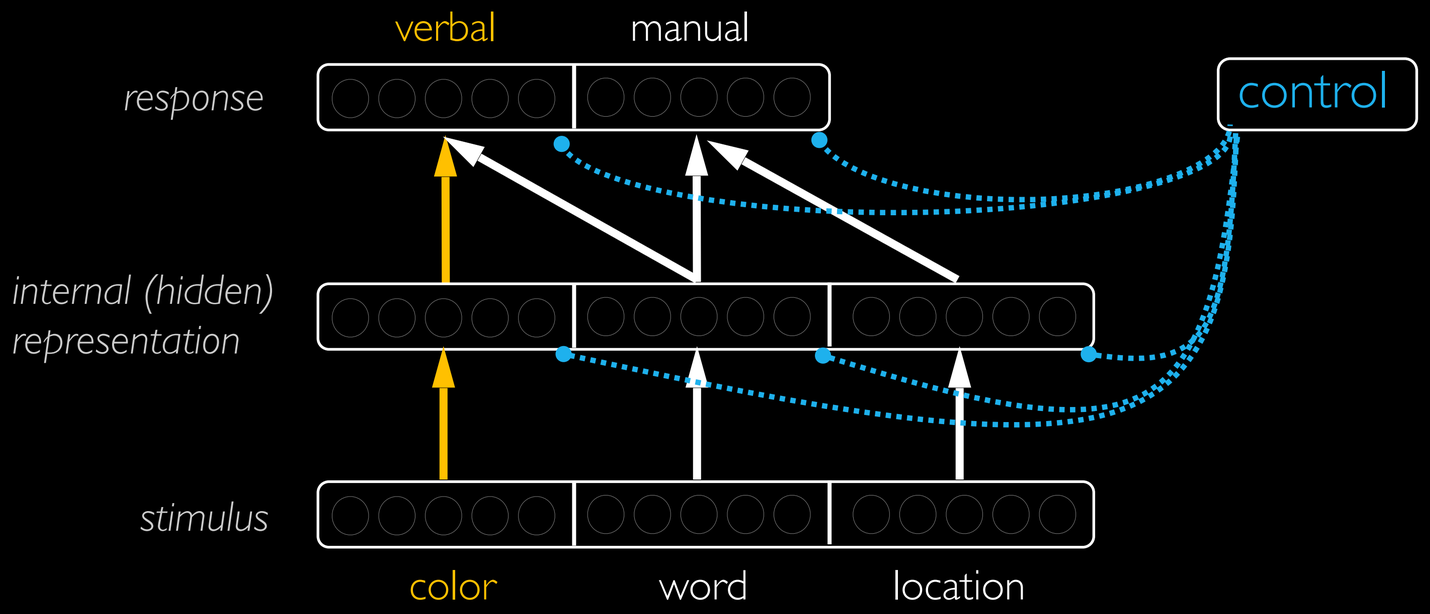
(Cohen et al., 1990 ; Feng et al., 2014; Musslick et al., 2016)

Neural Network Model of Controlled Processing



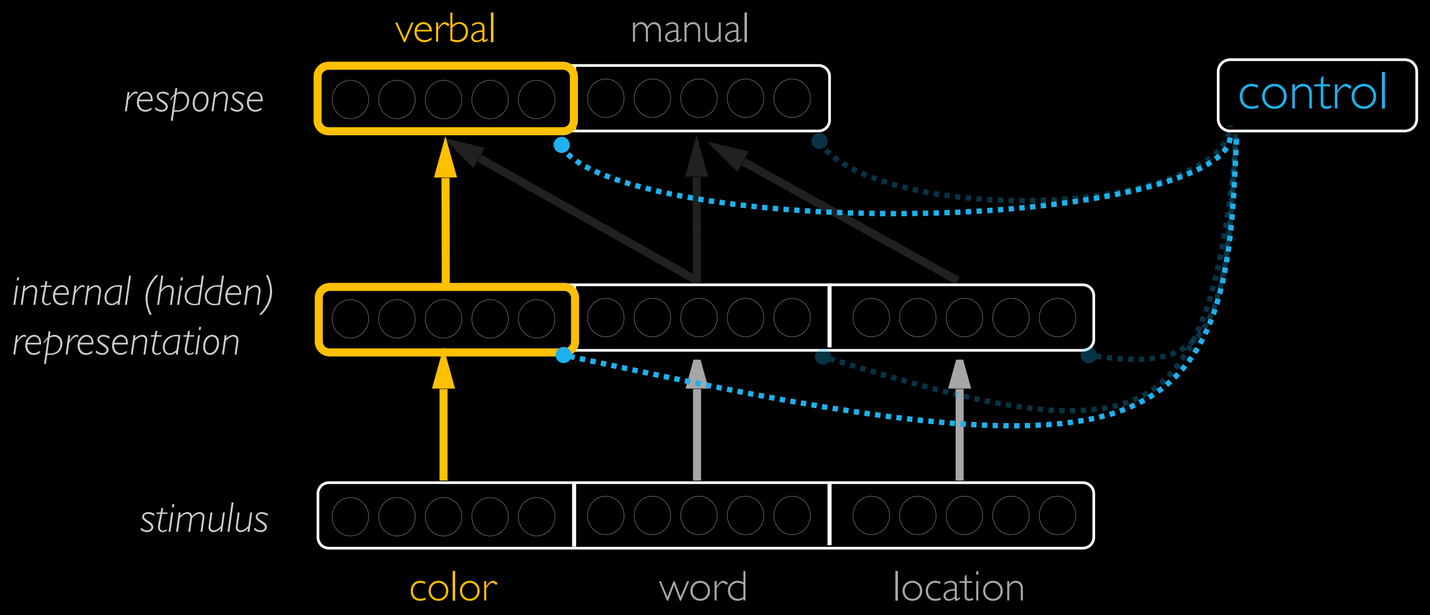
(Cohen et al., 1990 ; Feng et al., 2014; Musslick et al., 2016)

Neural Network Model of Controlled Processing



(Cohen et al., 1990 ; Feng et al., 2014; Musslick et al., 2016)

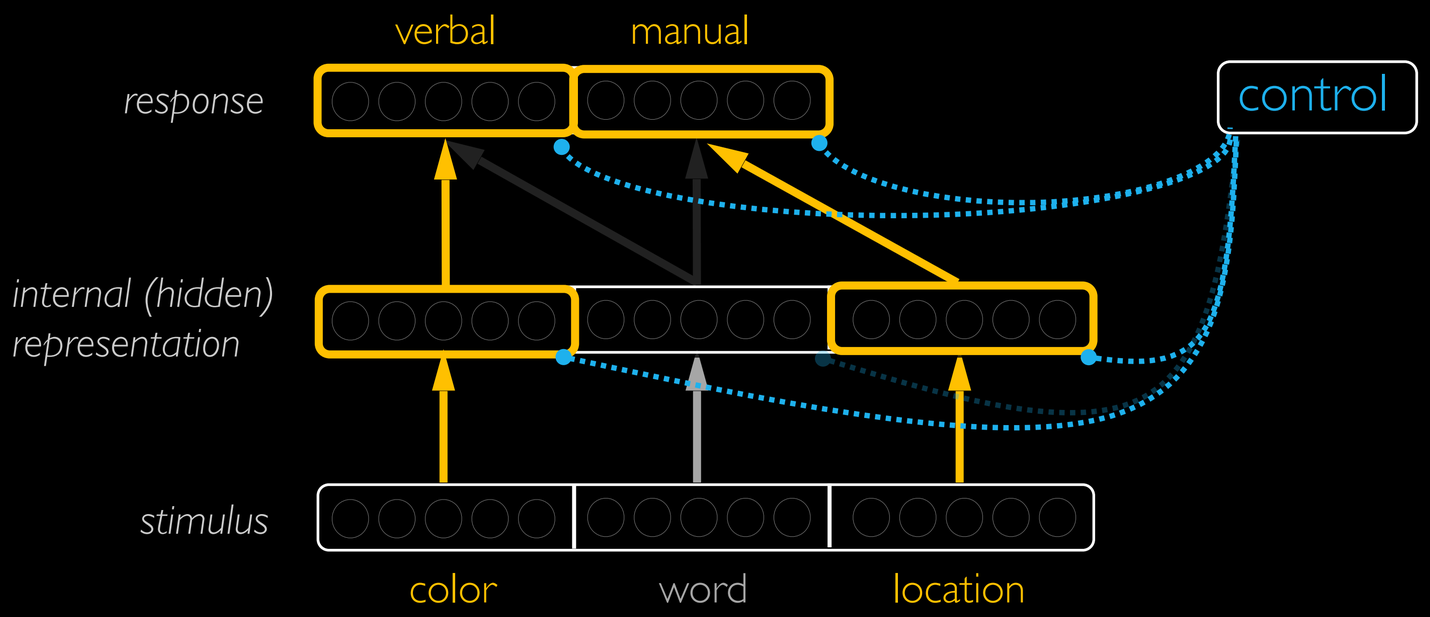
Neural Network Model of Controlled Processing



(Cohen et al., 1990 ; Feng et al., 2014; Musslick et al., 2016)

Neural Network Model of Controlled Processing

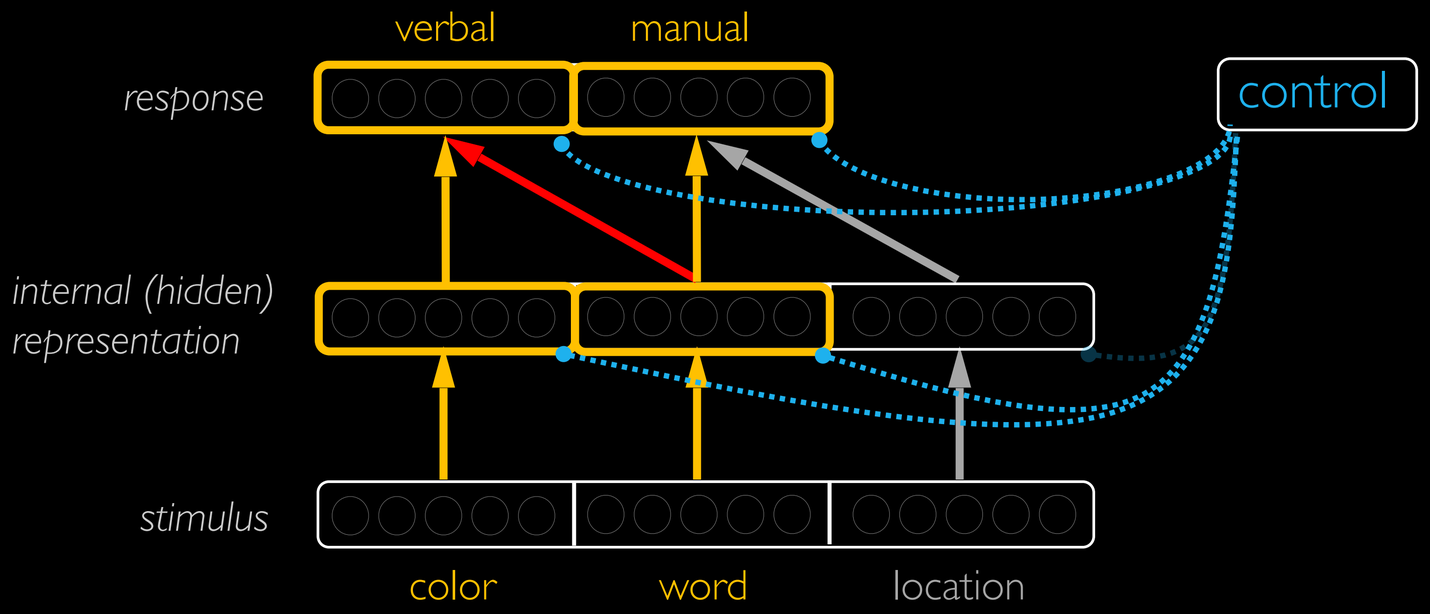
multitasking is possible



(Cohen et al., 1990 ; Feng et al., 2014; Musslick et al., 2016)

Neural Network Model of Controlled Processing

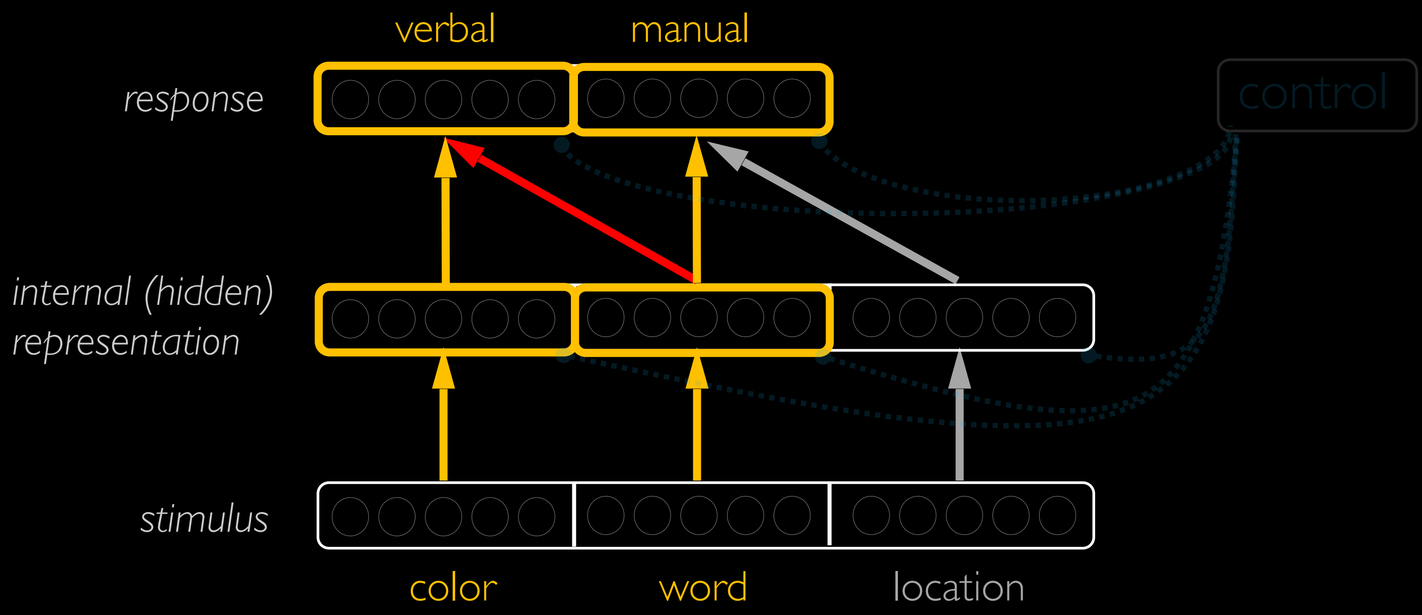
multitasking is not possible



(Cohen et al., 1990 ; Feng et al., 2014; Musslick et al., 2016)

Neural Network Model of Controlled Processing

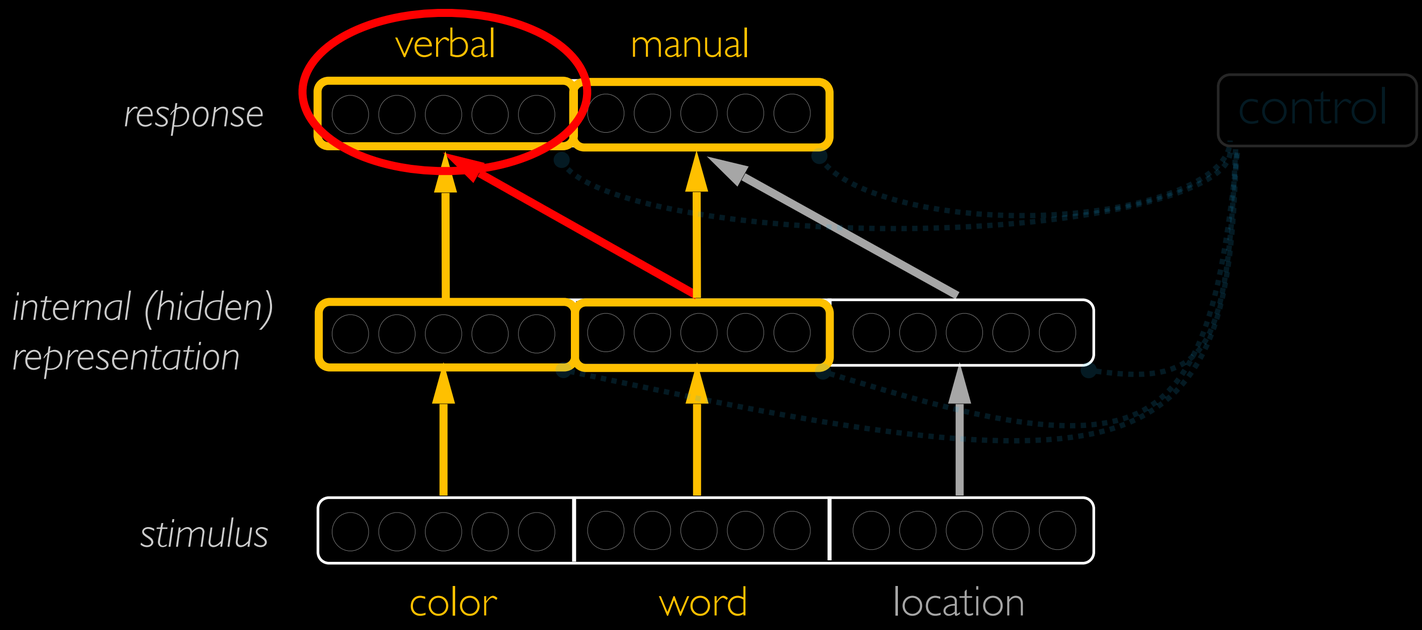
multitasking is not possible



(Cohen et al., 1990 ; Feng et al., 2014; Musslick et al., 2016)

Neural Network Model of Controlled Processing

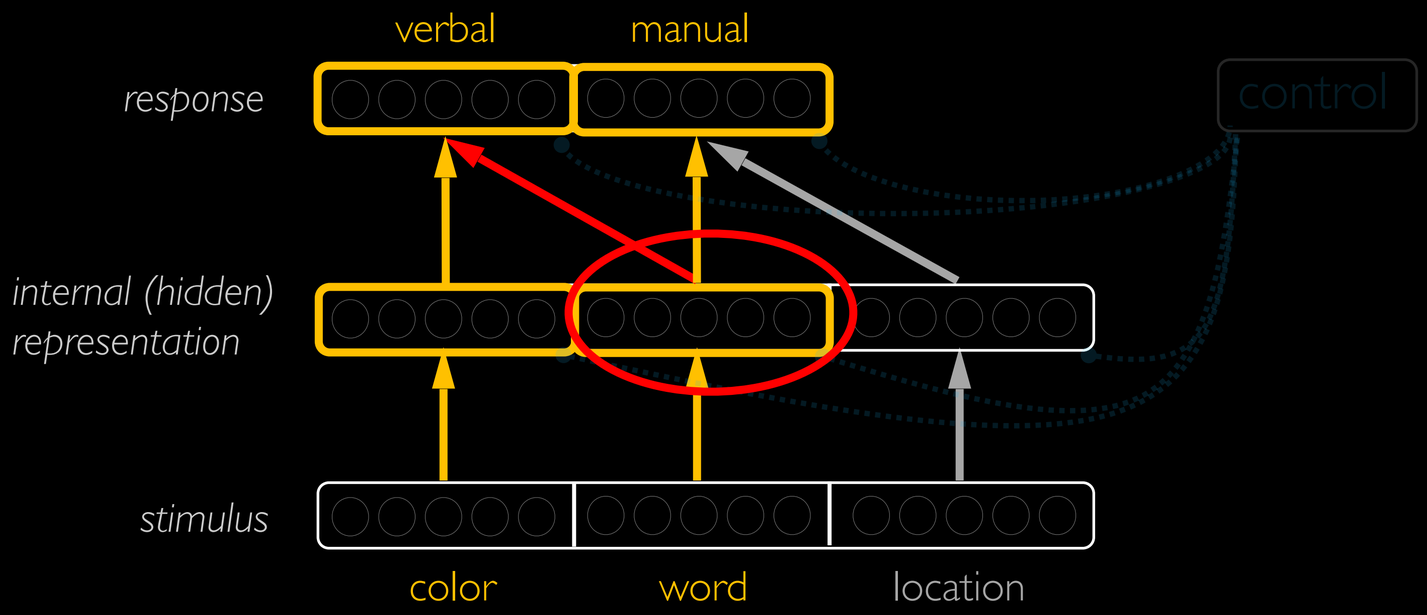
multitasking is not possible



(Cohen et al., 1990 ; Feng et al., 2014; Musslick et al., 2016)

Neural Network Model of Controlled Processing

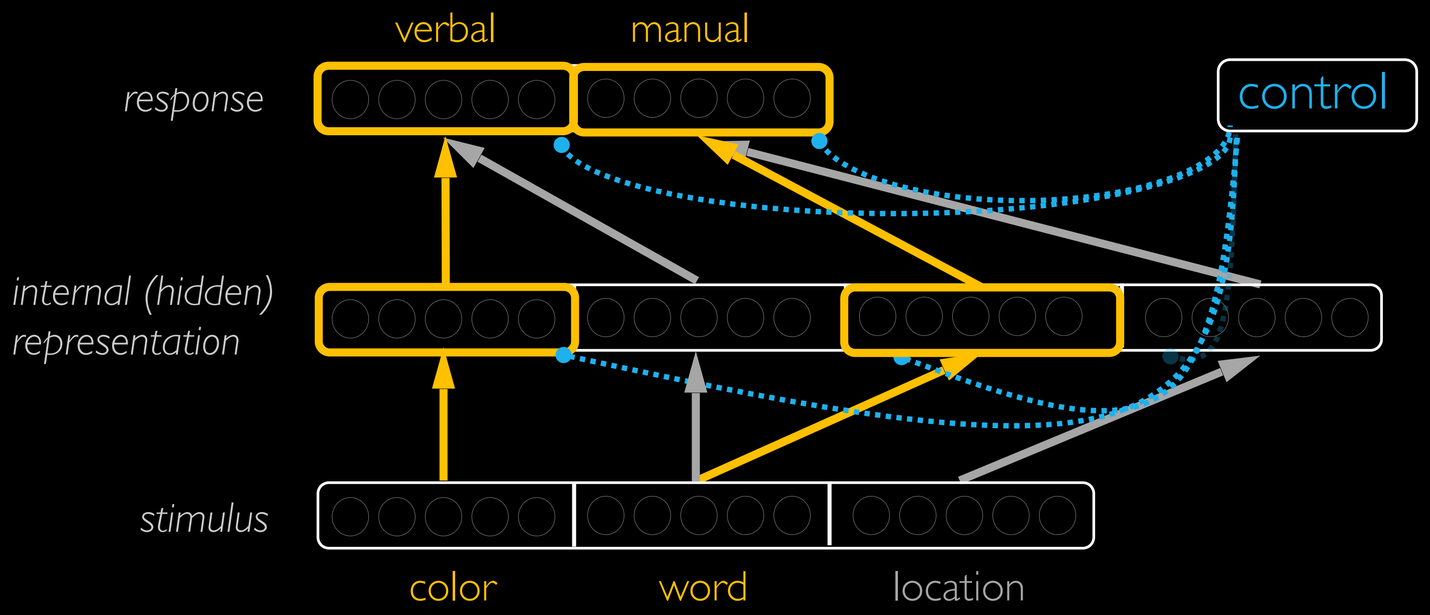
multitasking is not possible



(Cohen et al., 1990 ; Feng et al., 2014; Musslick et al., 2016)

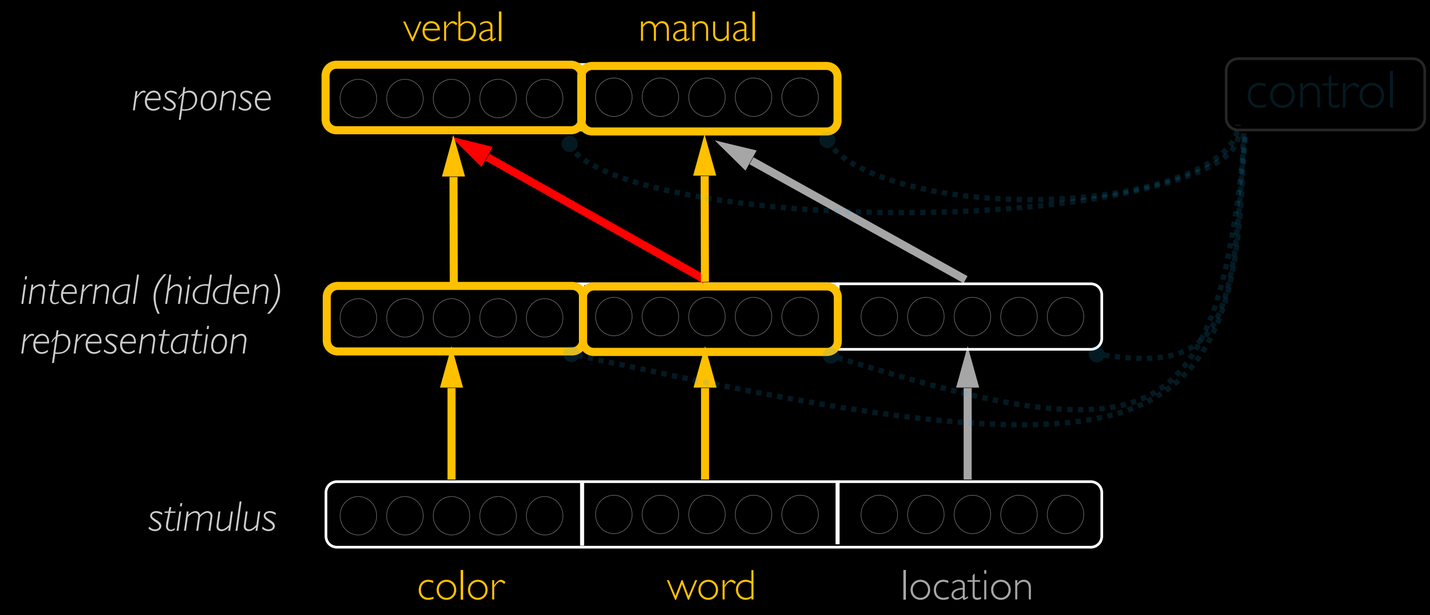
Neural Network Model of Controlled Processing

multitasking is possible



(Cohen et al., 1990 ; Feng et al., 2014; Musslick et al., 2016)

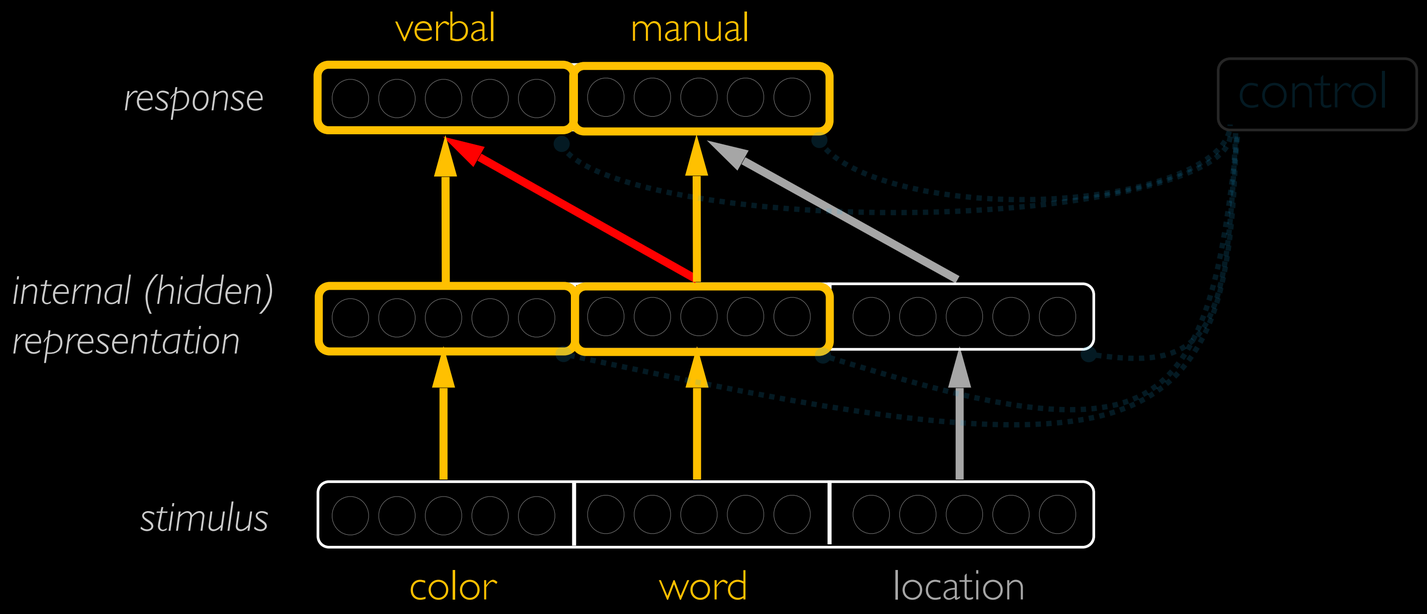
Neural Network Model of Controlled Processing

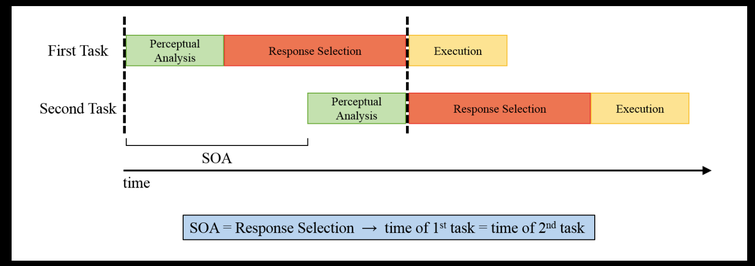


Multiple-Resource Theory

(Allport, 1972; Allport, 1980; Meyer & Kieras, 1997; Navon & Gopher, 1979; Wickens, 1984; Salvucci & Taatgen, 2008)

Neural Network Model of Controlled Processing





Harold 1994

Questions

Questions

1. Do the limitations imposed by shared representations prevail in a system as large as the brain?

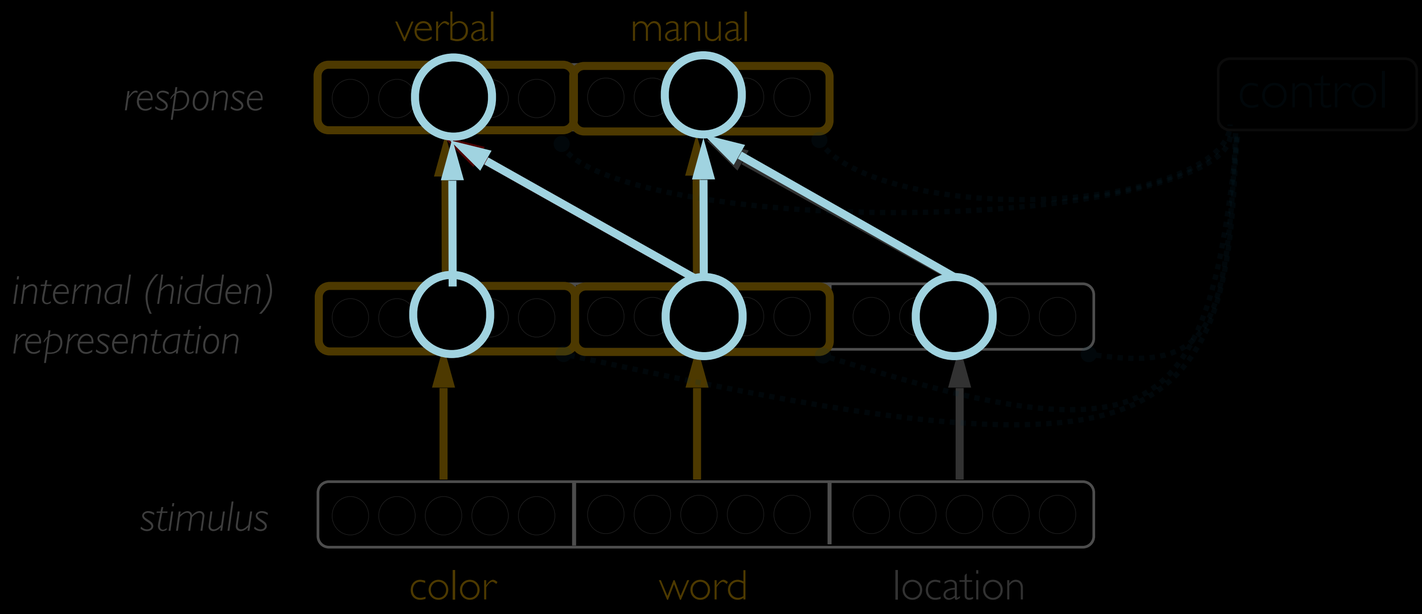
2. Assuming that shared representation cause a lot of trouble, why do we use them in the first place? (Experiment is evidence we do)

Questions

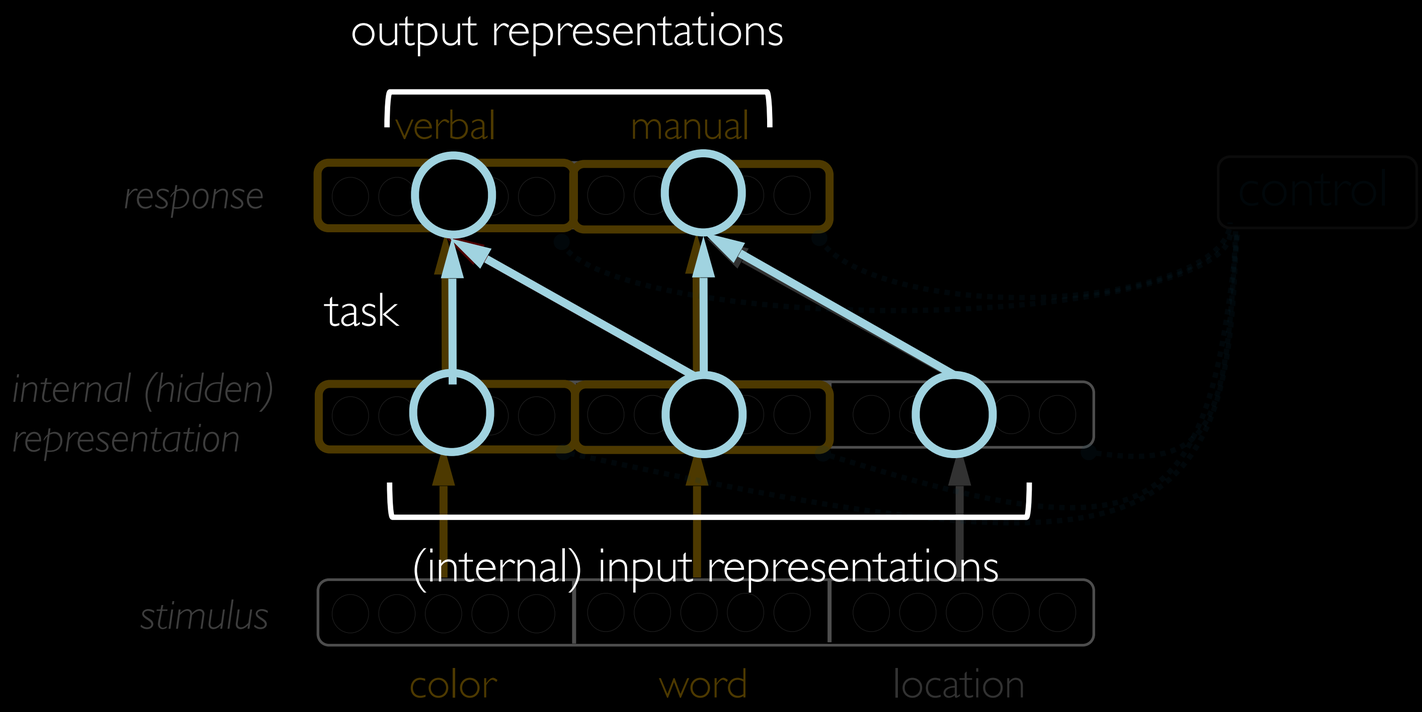
1. Do the limitations imposed by shared representations prevail in a system as large as the brain?

2. Assuming that shared representation cause a lot of trouble, why do we use them in the first place? (Experiment is evidence we do)

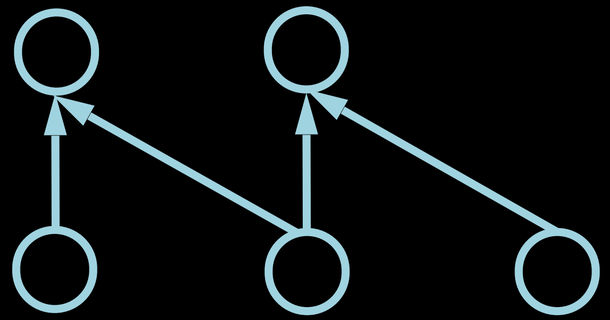
Neural Network Model of Controlled Processing



Neural Network Model of Controlled Processing

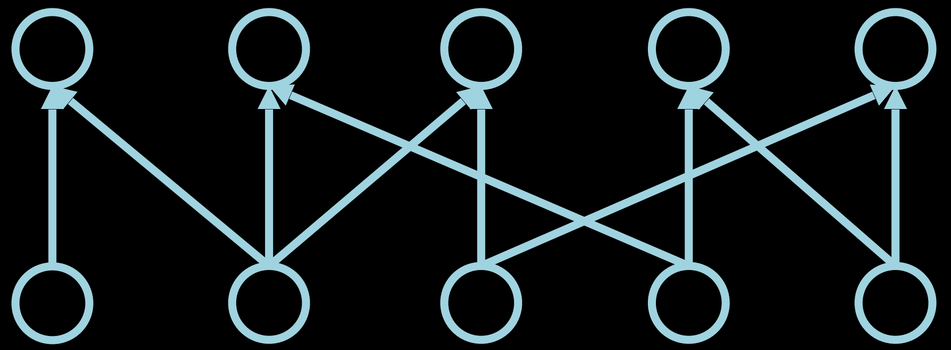


Graph-Theoretic Analysis of Multitasking Capacity



What is the maximum number of tasks
that the network can perform in parallel without interference?

Graph-Theoretic Analysis of Multitasking Capacity

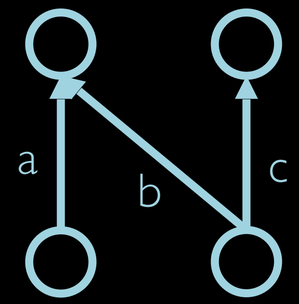


What is the maximum number of tasks
that the network can perform in parallel without interference?

Graph-Theoretic Approach

Task Dependencies

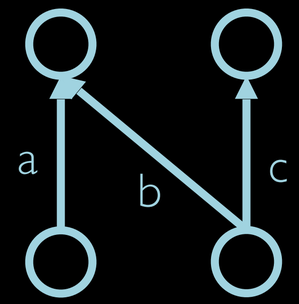
bipartite
task graph



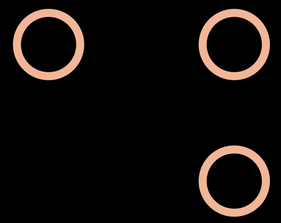
Graph-Theoretic Approach

Task Dependencies

bipartite task graph



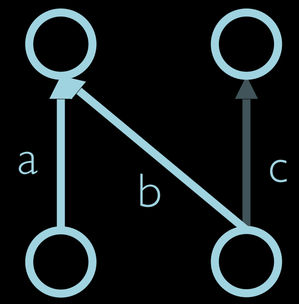
dependency graph



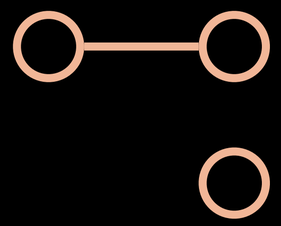
Graph-Theoretic Approach

Task Dependencies

bipartite task graph



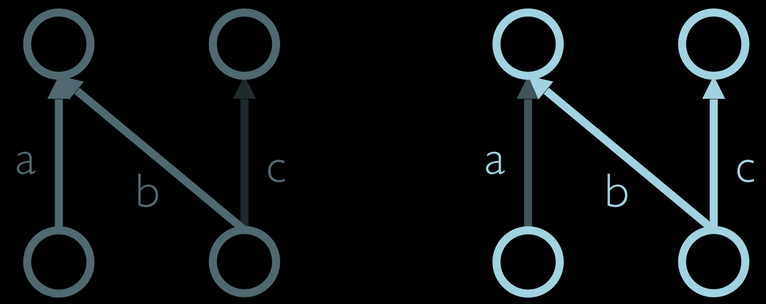
dependency graph



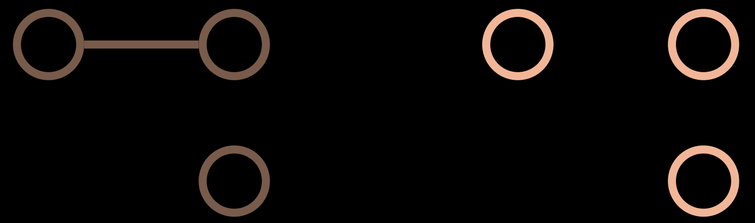
Graph-Theoretic Approach

Task Dependencies

bipartite task graph



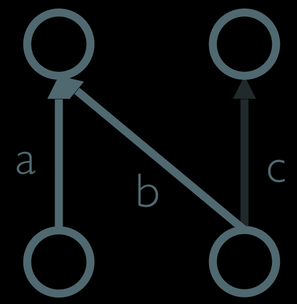
dependency graph

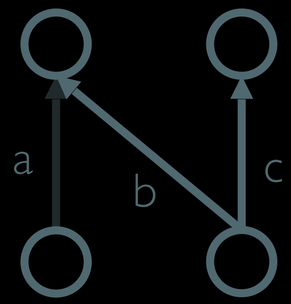


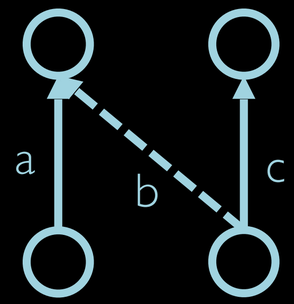
Graph-Theoretic Approach

Task Dependencies

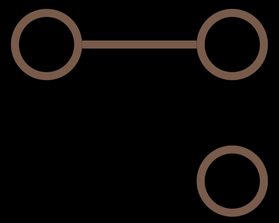
bipartite task graph

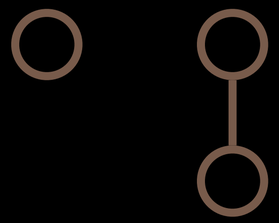


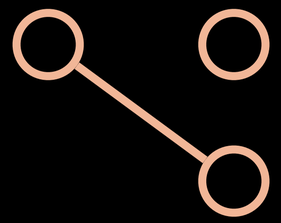




dependency graph



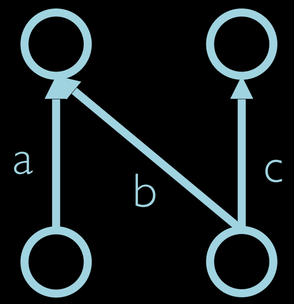




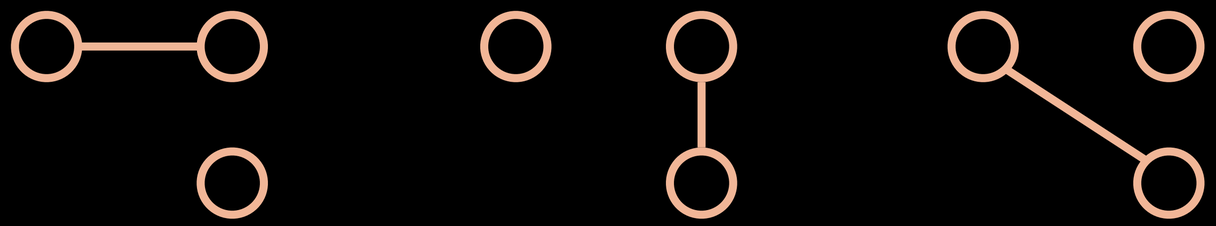
Graph-Theoretic Approach

Task Dependencies

bipartite task graph



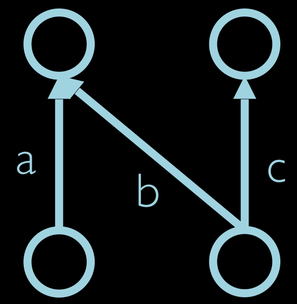
dependency graph



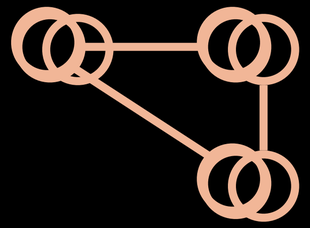
Graph-Theoretic Approach

Task Dependencies

bipartite
task graph



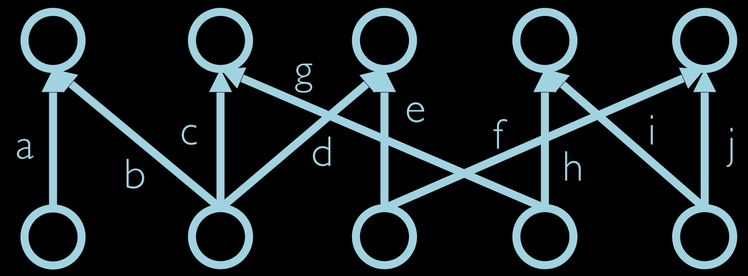
dependency
graph



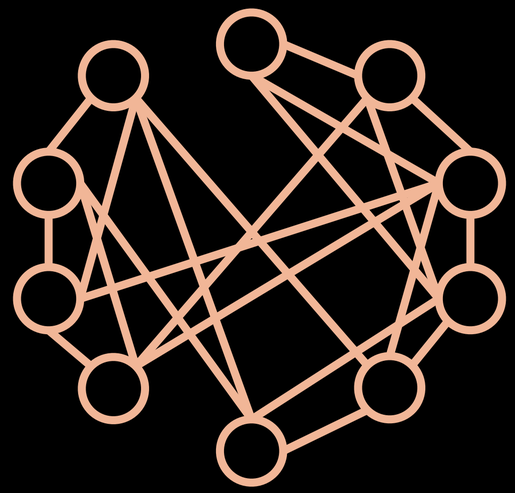
Graph-Theoretic Approach

Parallel Processing Capability

bipartite task graph



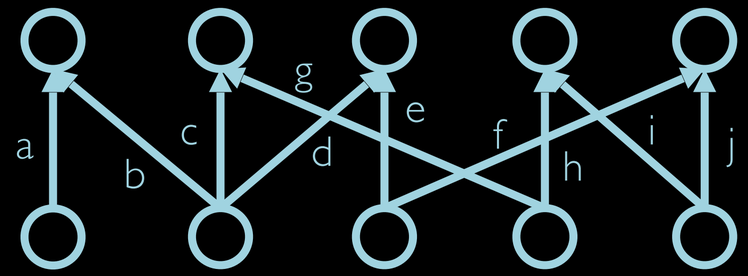
dependency graph



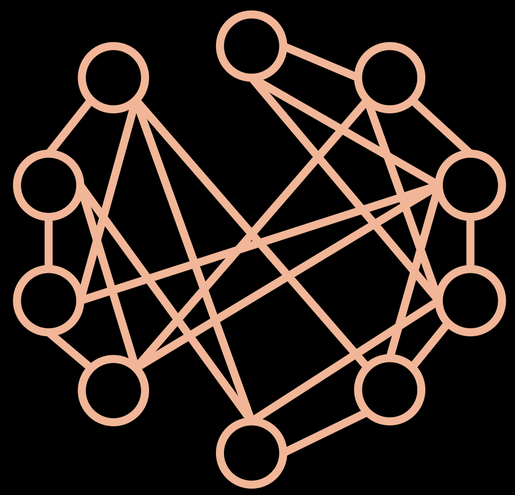
Graph-Theoretic Approach

Parallel Processing Capability

bipartite task graph



dependency graph

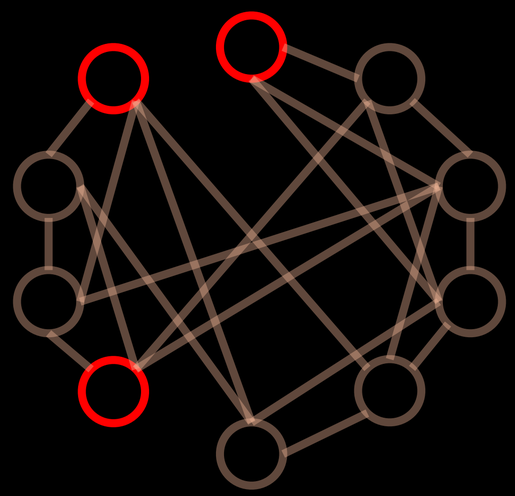


Maximum amount of tasks that can be performed in independently?

Graph-Theoretic Approach

Parallel Processing Capability

dependency graph



Maximum amount of tasks that can be performed in independently?

An independent vertex set of a graph G is a subset of the vertices such that no two vertices in the subset are connected by an edge of G.

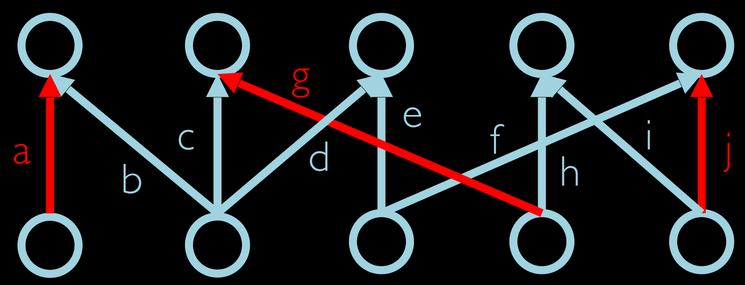
A maximum independent vertex set is an independent vertex set containing the largest possible number of vertices for a given graph.

(Musslick, Özcimder, Dey, Patwary, Willke & Cohen, CogSci, 2016)

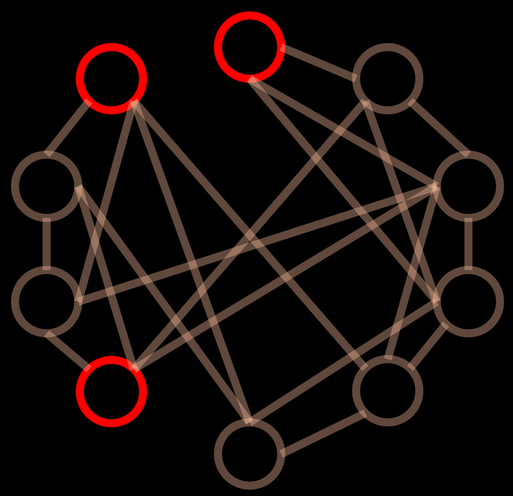
Graph-Theoretic Approach

Parallel Processing Capability

bipartite task graph



dependency graph



Maximum amount of tasks that can be performed in independently?

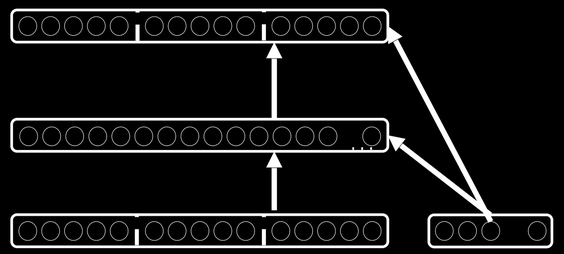
An independent vertex set of a graph G is a subset of the vertices such that no two vertices in the subset are connected by an edge of G.

A maximum independent vertex set is an independent vertex set containing the largest possible number of vertices for a given graph.

(Musslick, Özcimder, Dey, Patwary, Willke & Cohen, CogSci, 2016)

Application to Neural Systems

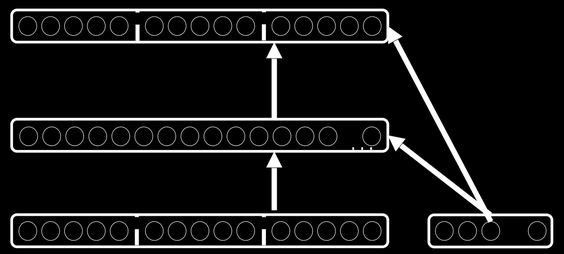
trained neural network



(Musslick, Özcimder, Dey, Patwary, Willke & Cohen, CogSci, 2016)

Application to Neural Systems

trained neural network

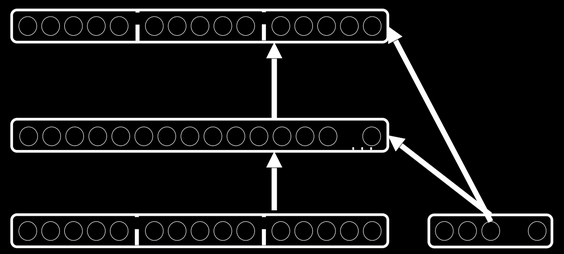


extract based on
single task representations

(Musslick, Özcimder, Dey, Patwary, Willke & Cohen, CogSci, 2016)

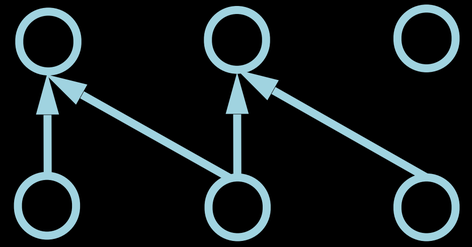
Application to Neural Systems

trained neural network



extract based on
single task representations

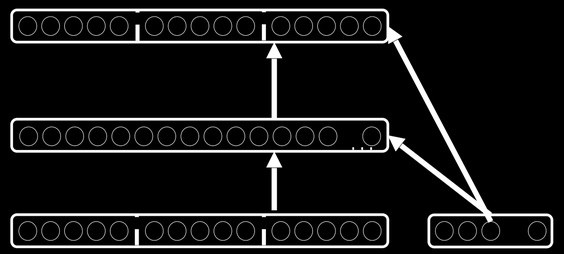
task graph



(Musslick, Özcimder, Dey, Patwary, Willke & Cohen, CogSci, 2016)

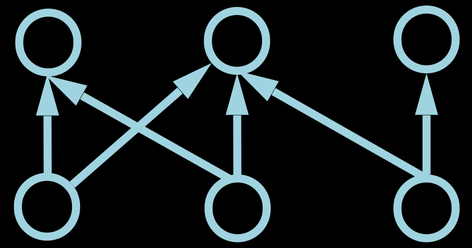
Application to Neural Systems

trained neural network

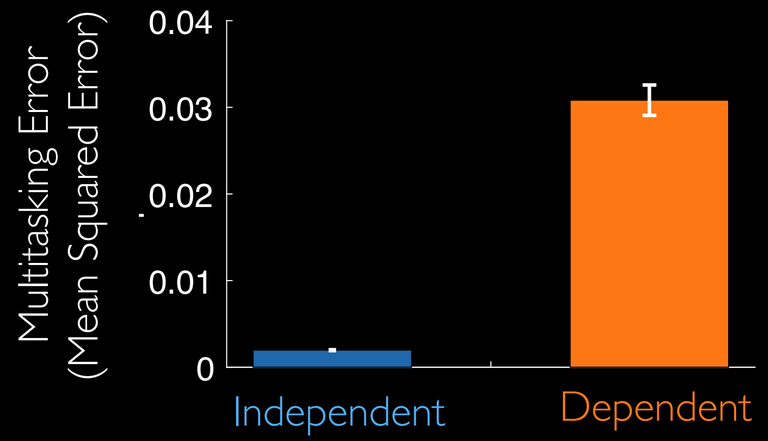


extract based on single task representations

task graph



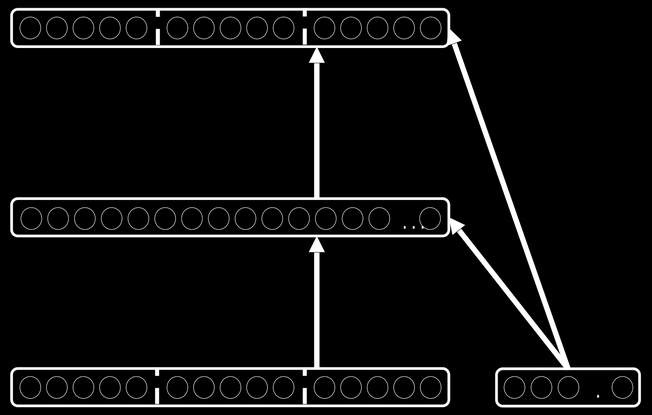
predict performance for all possible multitasking combinations



(Musslick, Özcimder, Dey, Patwary, Willke & Cohen, CogSci, 2016)

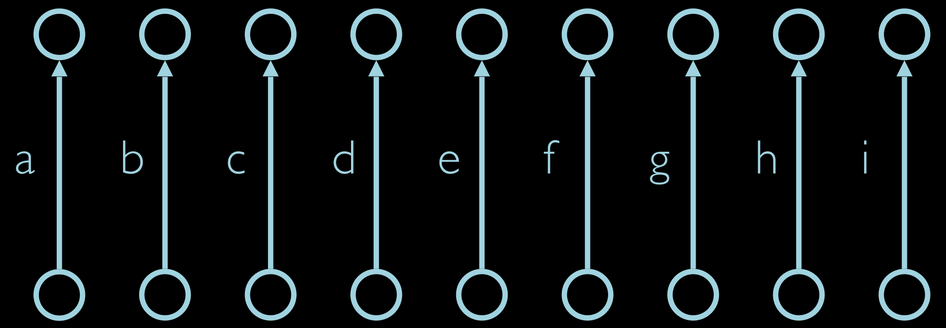
Application to Neural Systems

Extract Dependency Graph From Trained Neural Network



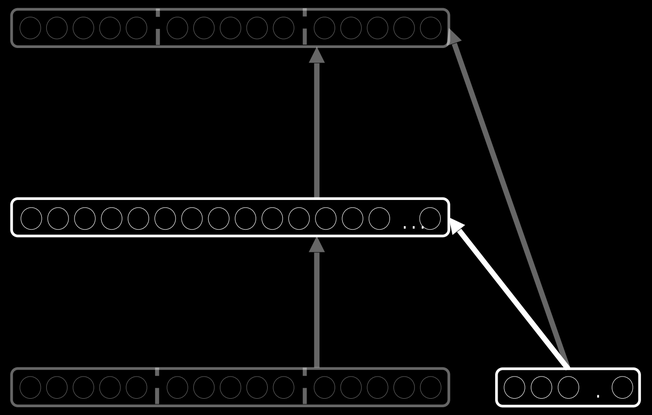
trained neural network

bipartite graph



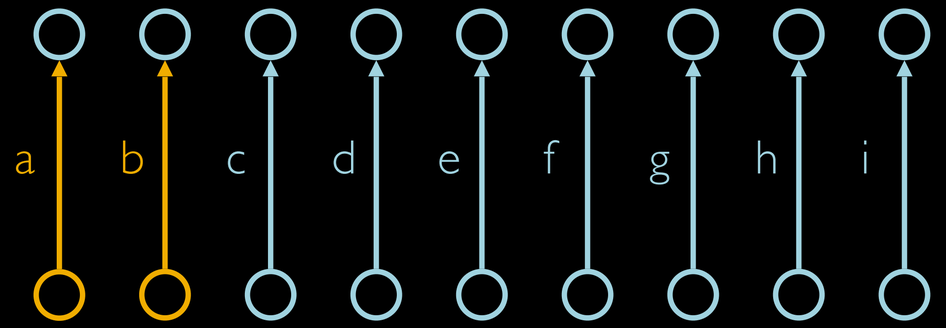
Application to Neural Systems

Extract Dependency Graph From Trained Neural Network



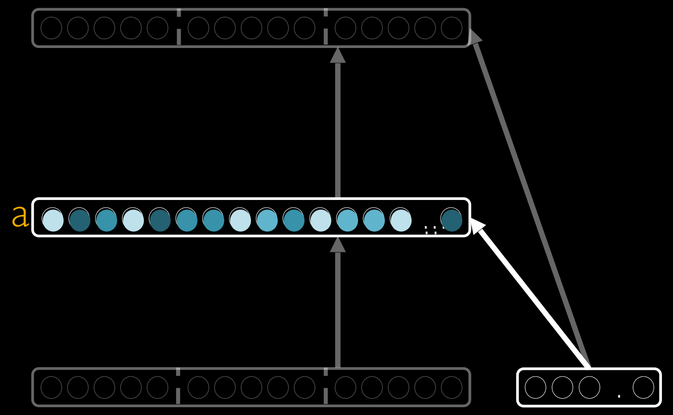
trained neural network

bipartite graph



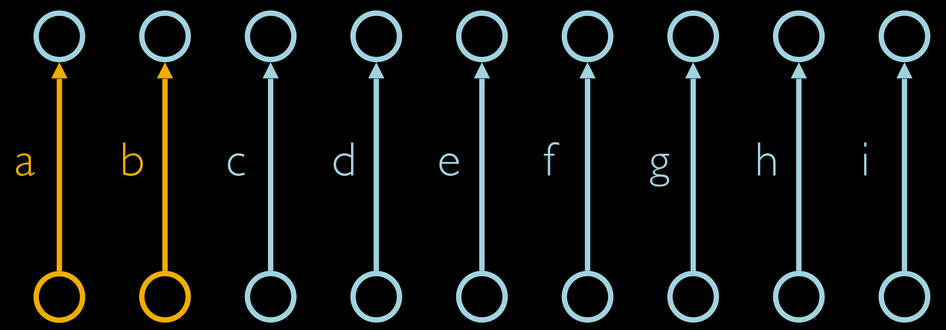
Application to Neural Systems

Extract Dependency Graph From Trained Neural Network



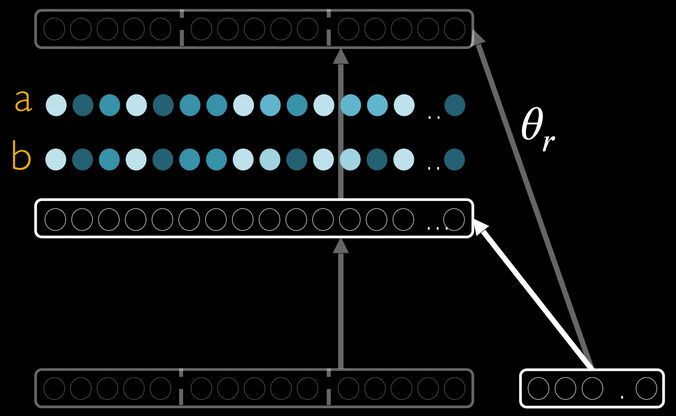
trained neural network

bipartite graph

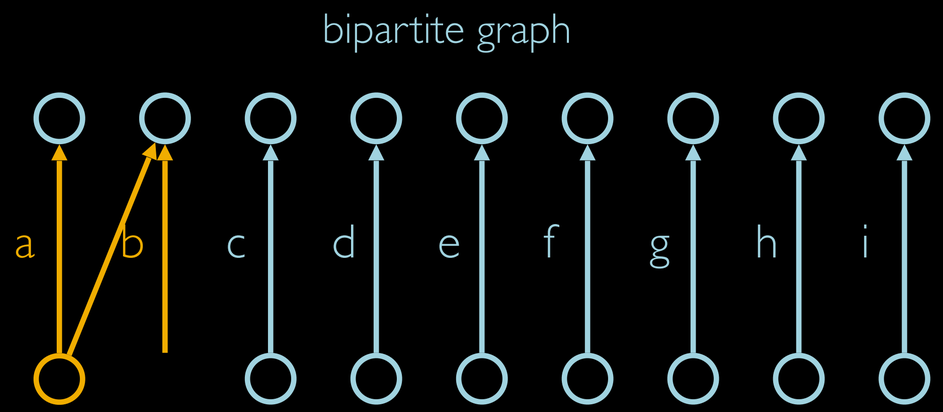


Application to Neural Systems

Extract Dependency Graph From Trained Neural Network

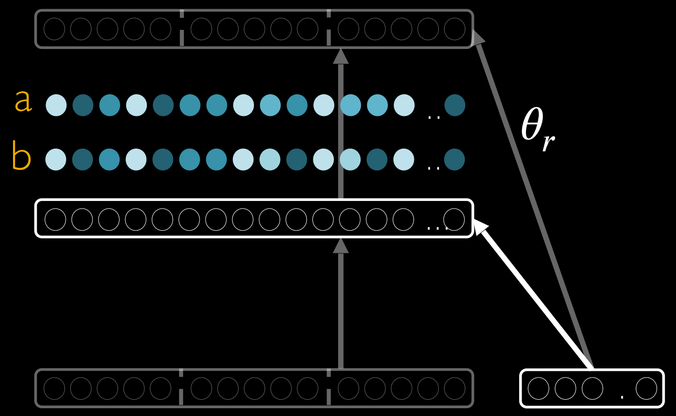


trained neural network



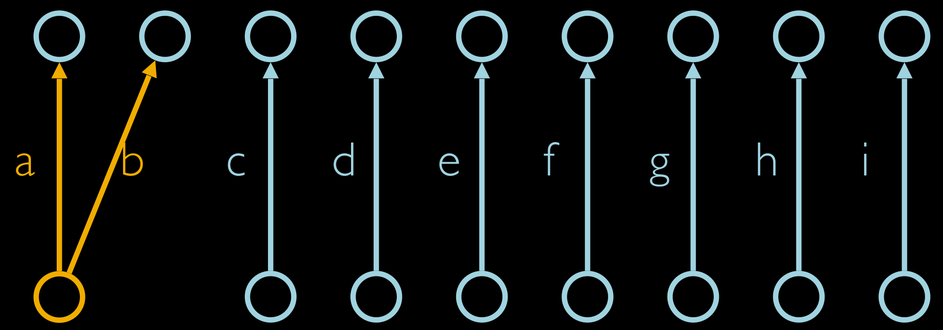
Application to Neural Systems

Extract Dependency Graph From Trained Neural Network



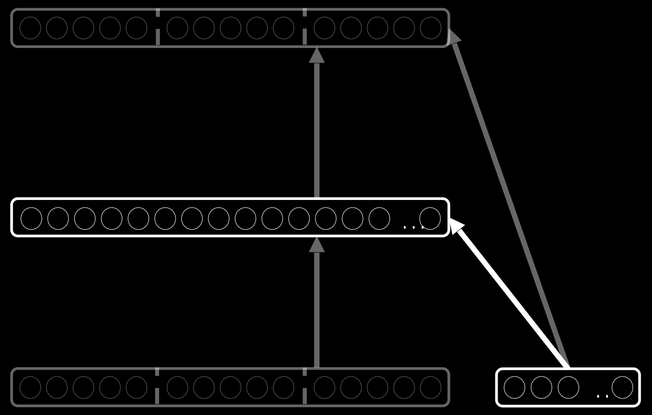
trained neural network

bipartite graph



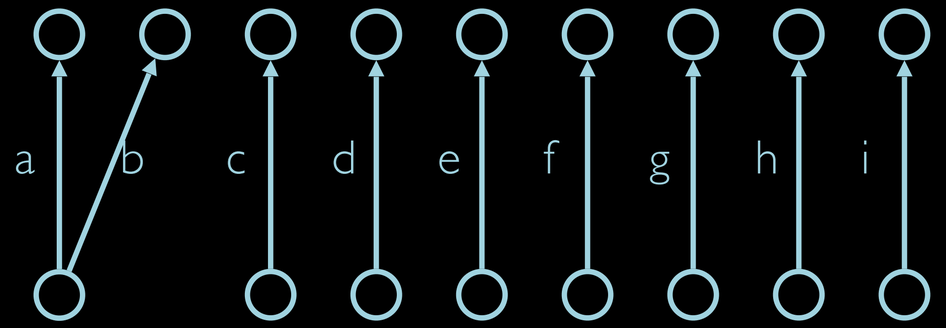
Application to Neural Systems

Extract Dependency Graph From Trained Neural Network



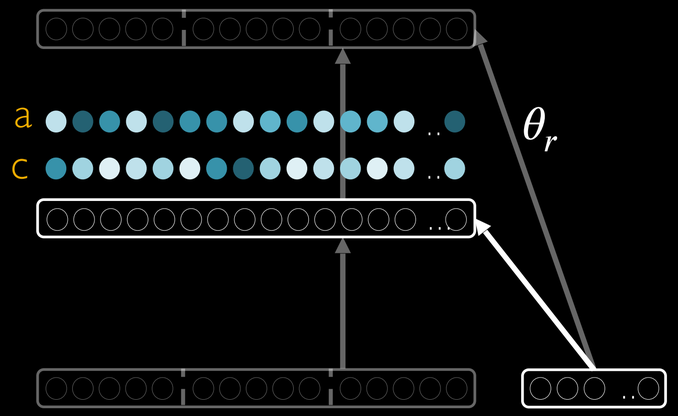
trained neural network

bipartite graph

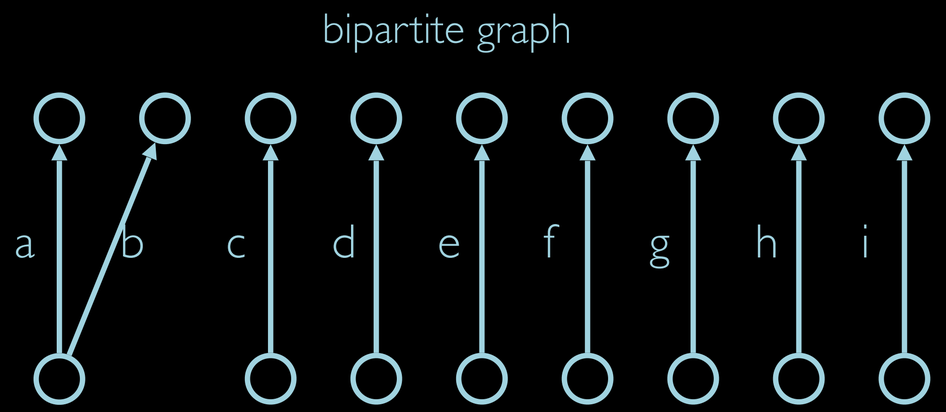


Application to Neural Systems

Extract Dependency Graph From Trained Neural Network



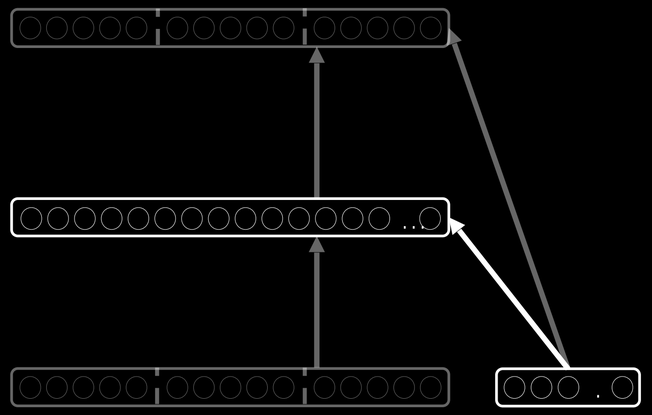
trained neural network



Application to Neural Systems

Extract Dependency Graph From Trained Neural Network

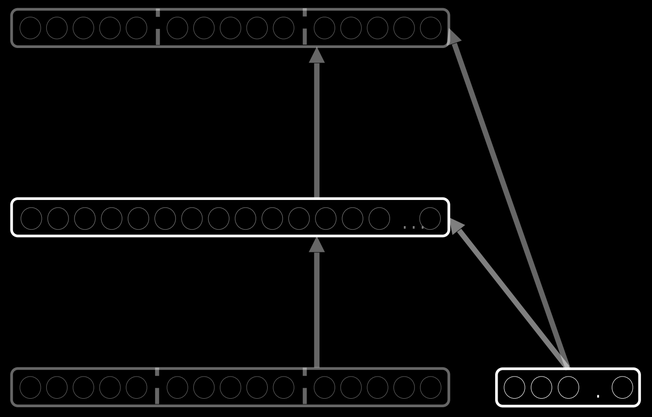
bipartite graph



trained neural network

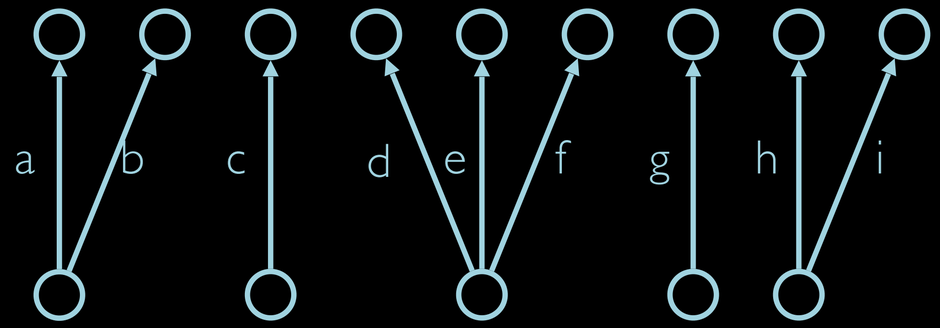
Application to Neural Systems

Extract Dependency Graph From Trained Neural Network



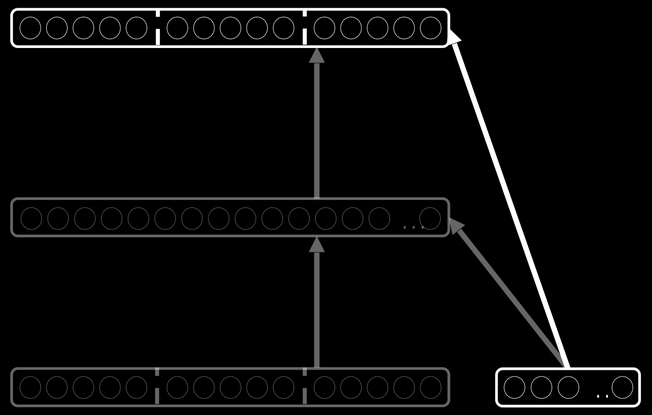
trained neural network

bipartite graph



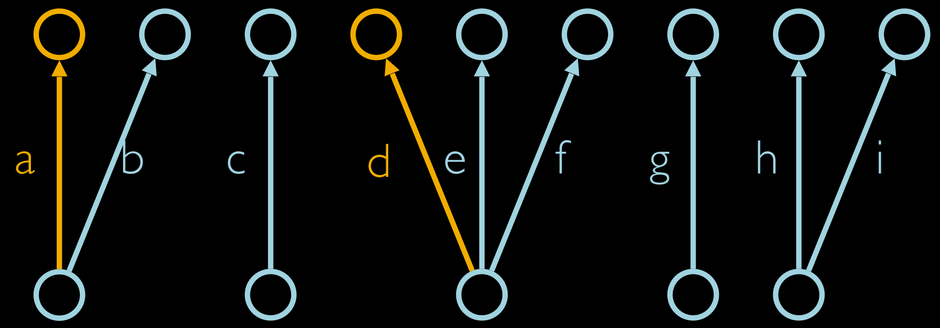
Application to Neural Systems

Extract Dependency Graph From Trained Neural Network



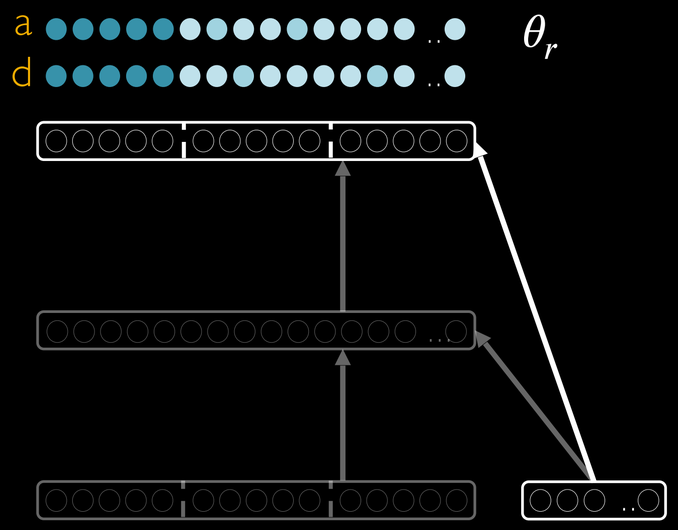
trained neural network

bipartite graph

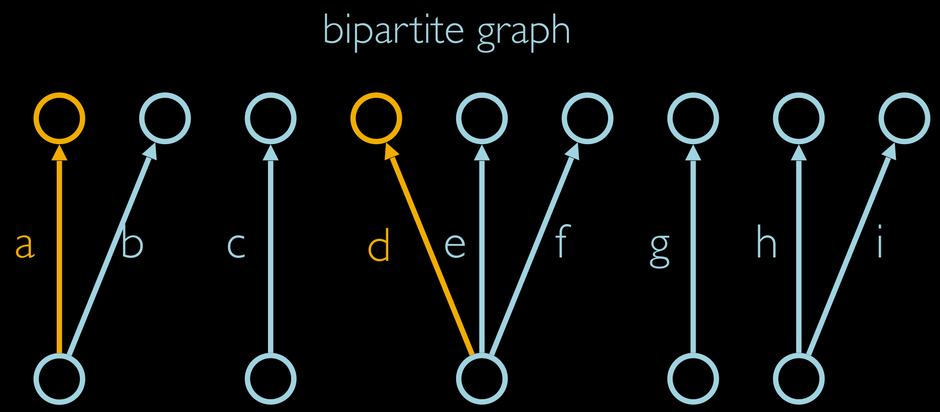


Application to Neural Systems

Extract Dependency Graph From Trained Neural Network

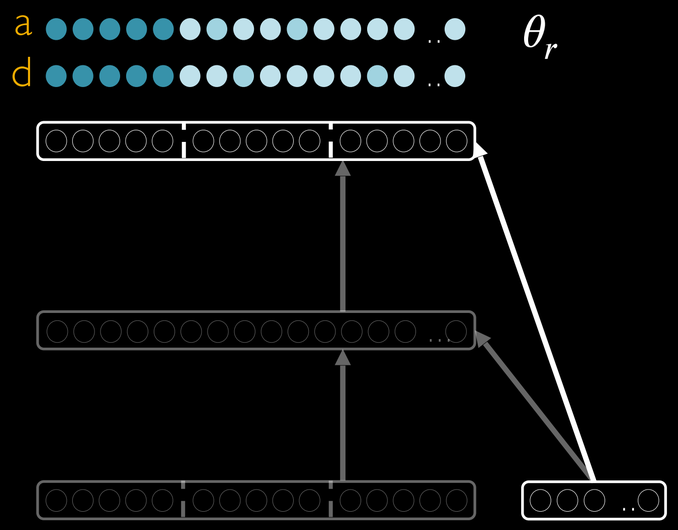


trained neural network

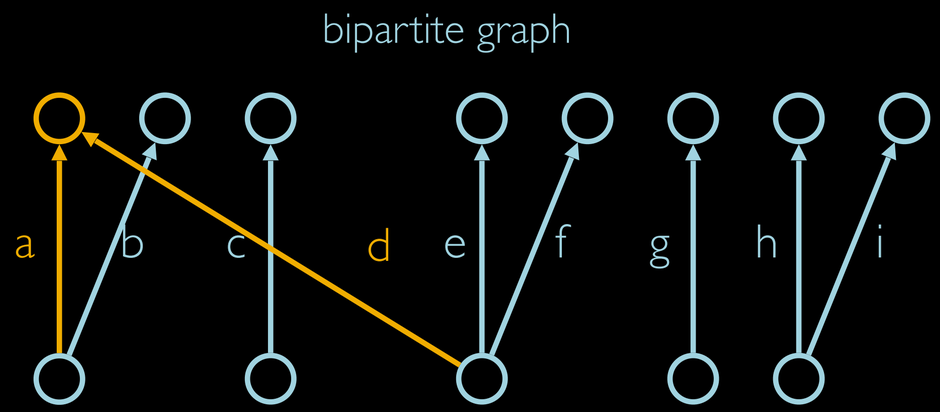


Application to Neural Systems

Extract Dependency Graph From Trained Neural Network

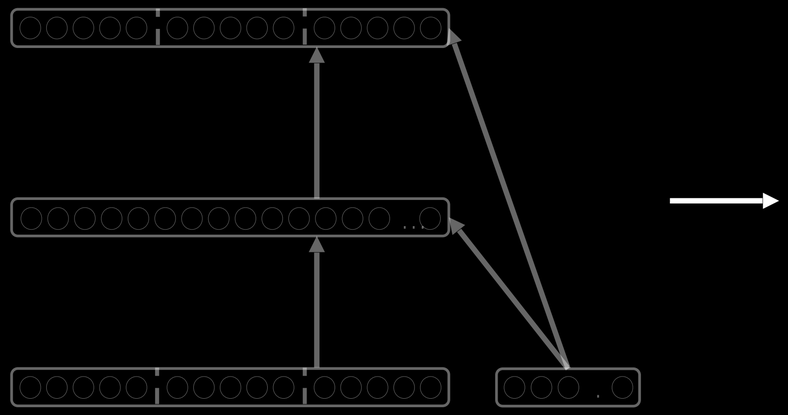


trained neural network



Application to Neural Systems

Predict Parallel Processing Performance Based On Dependency Graph

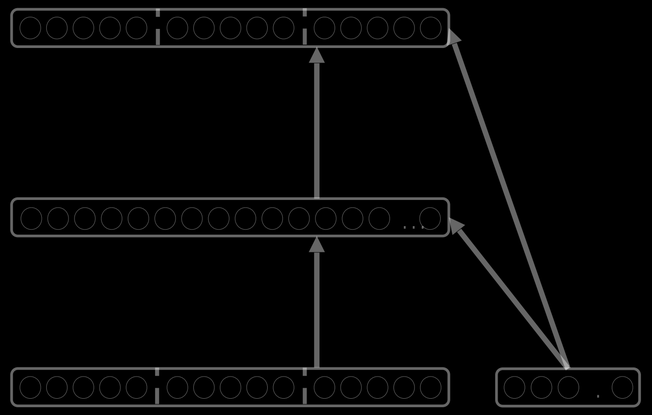


trained neural network

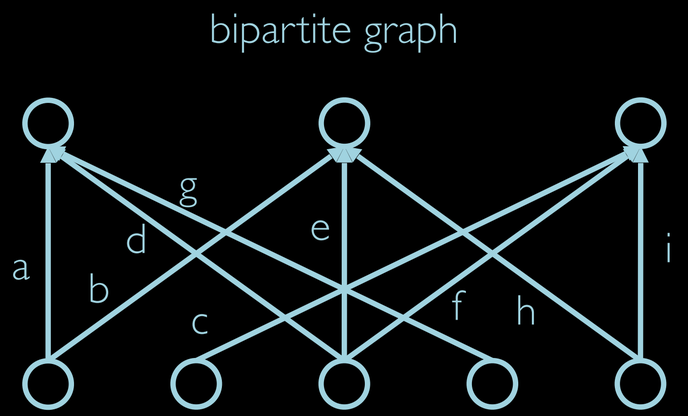
bipartite graph

Application to Neural Systems

Predict Parallel Processing Performance Based On Dependency Graph

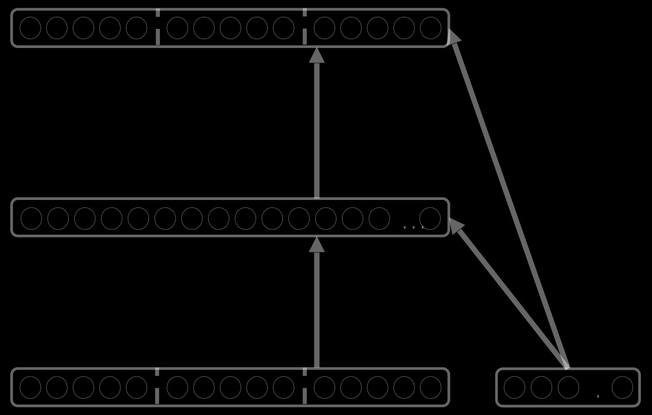


trained neural network



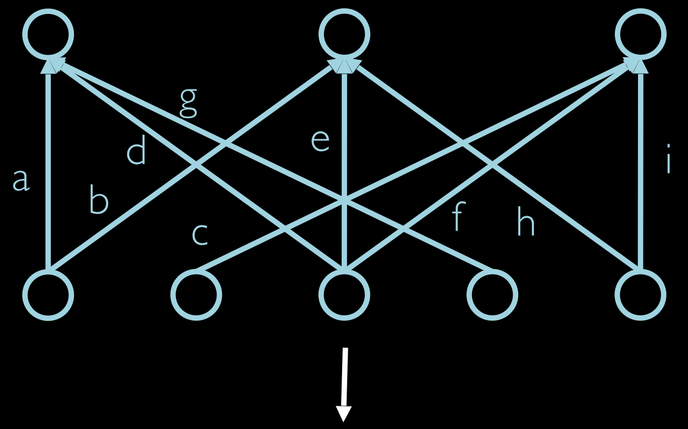
Application to Neural Systems

Predict Parallel Processing Performance Based On Dependency Graph

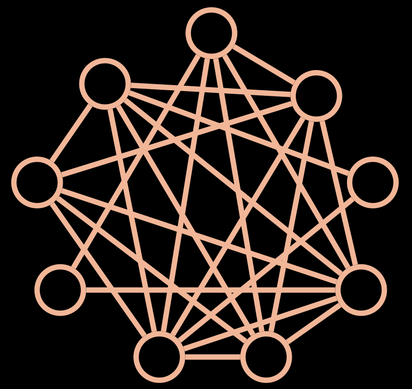


trained neural network

bipartite graph

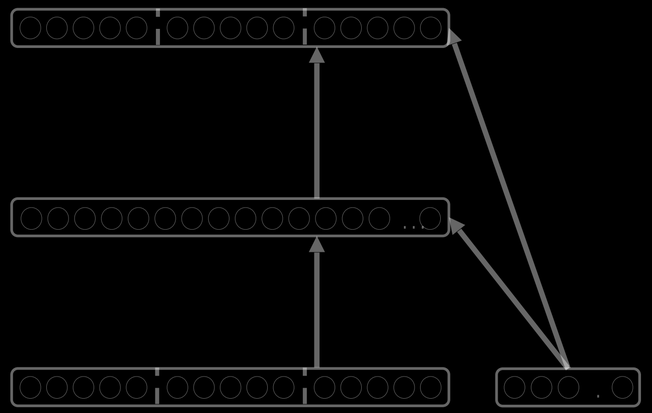


dependency graph



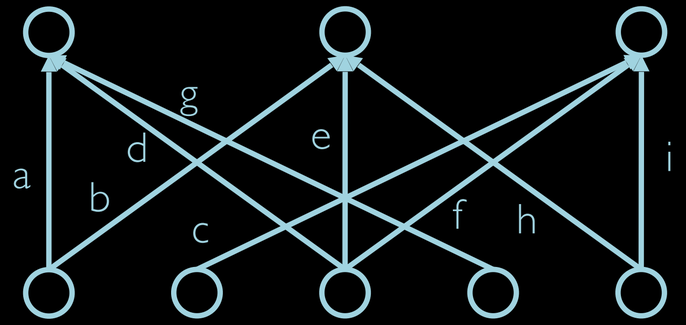
Application to Neural Systems

Predict Parallel Processing Performance Based On Dependency Graph

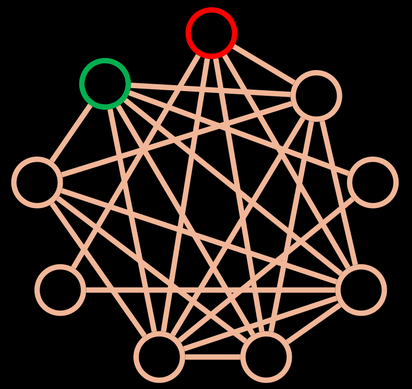


trained neural network

bipartite graph

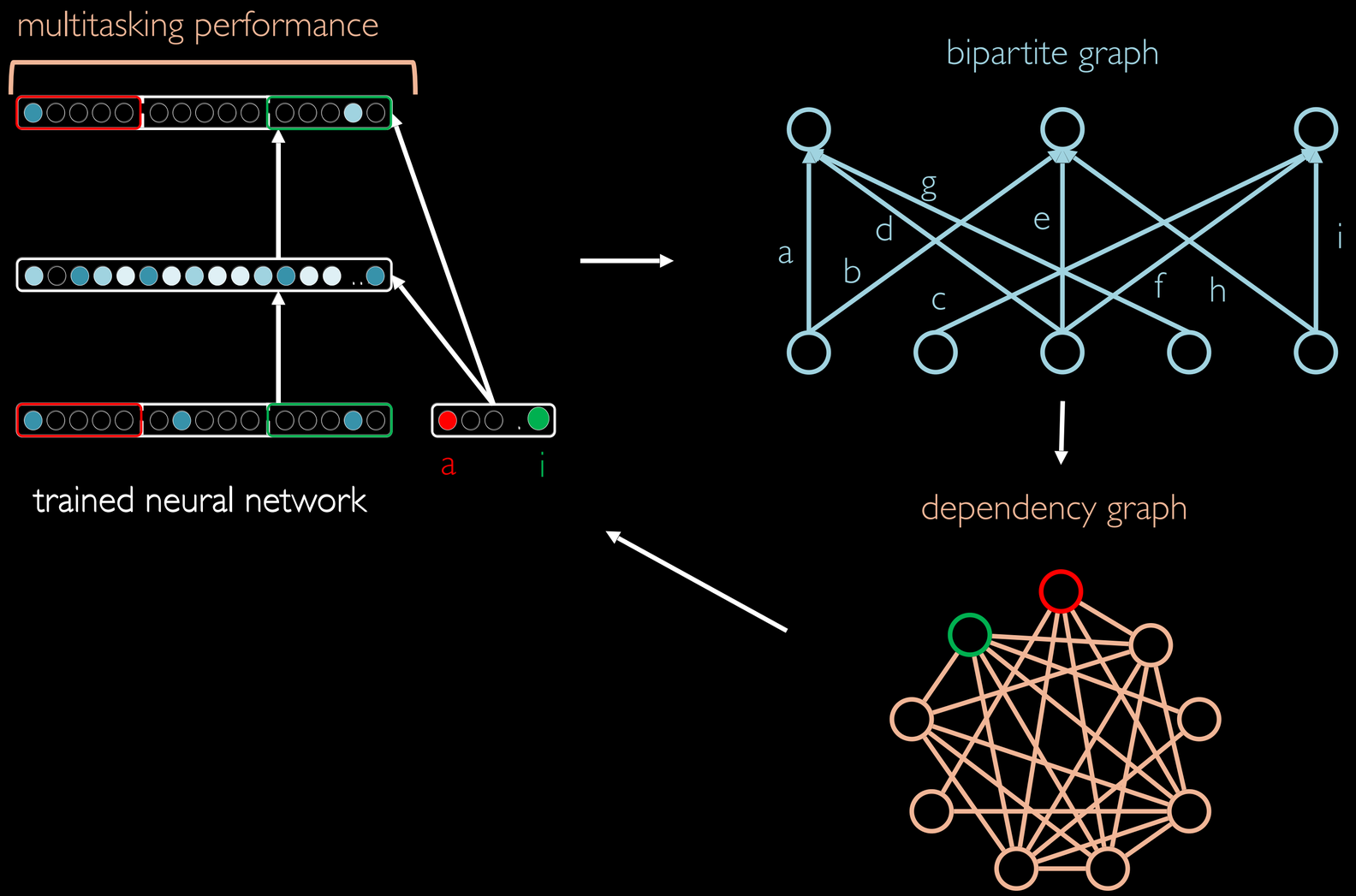


dependency graph



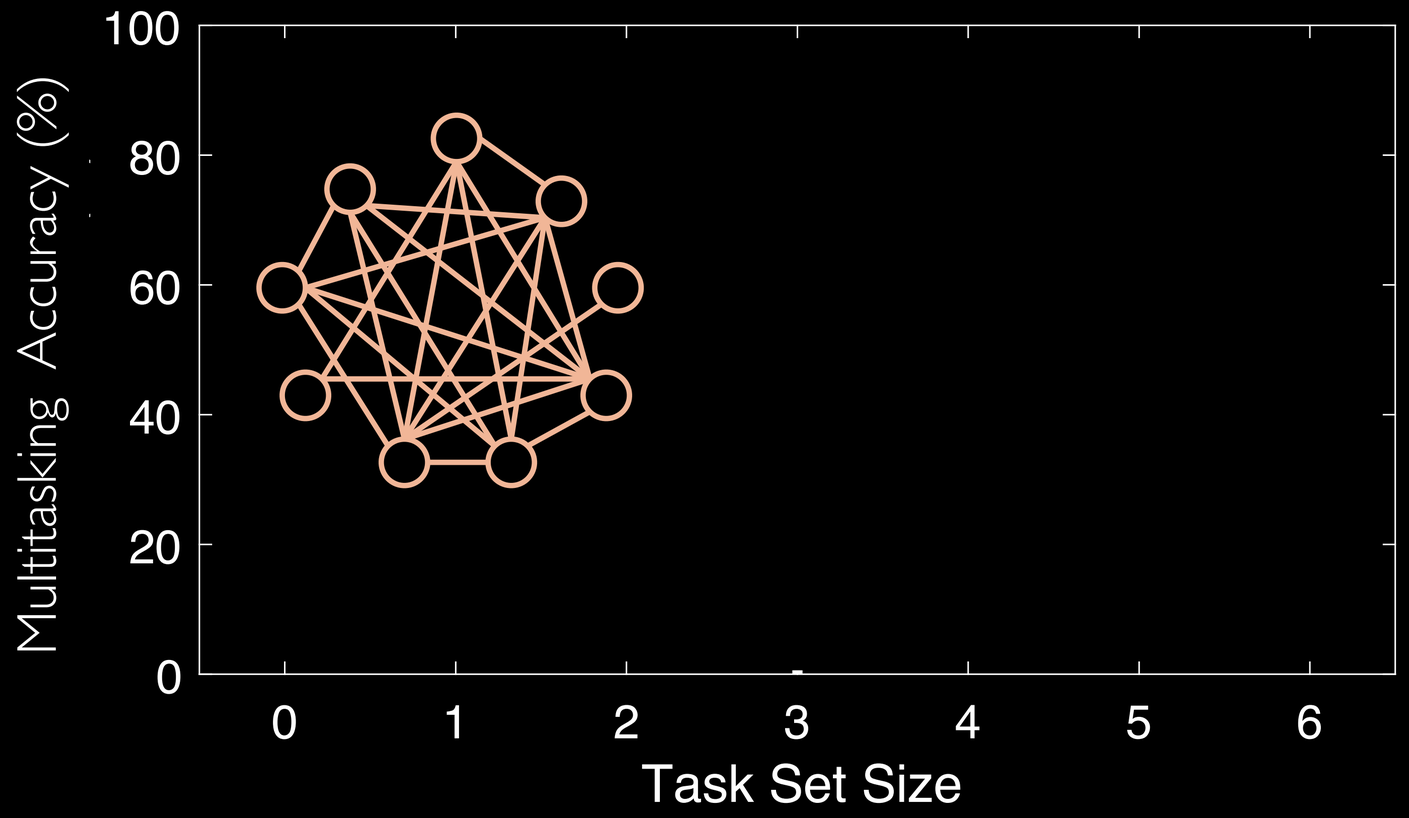
Application to Neural Systems

Predict Parallel Processing Performance Based On Dependency Graph



Application to Neural Systems

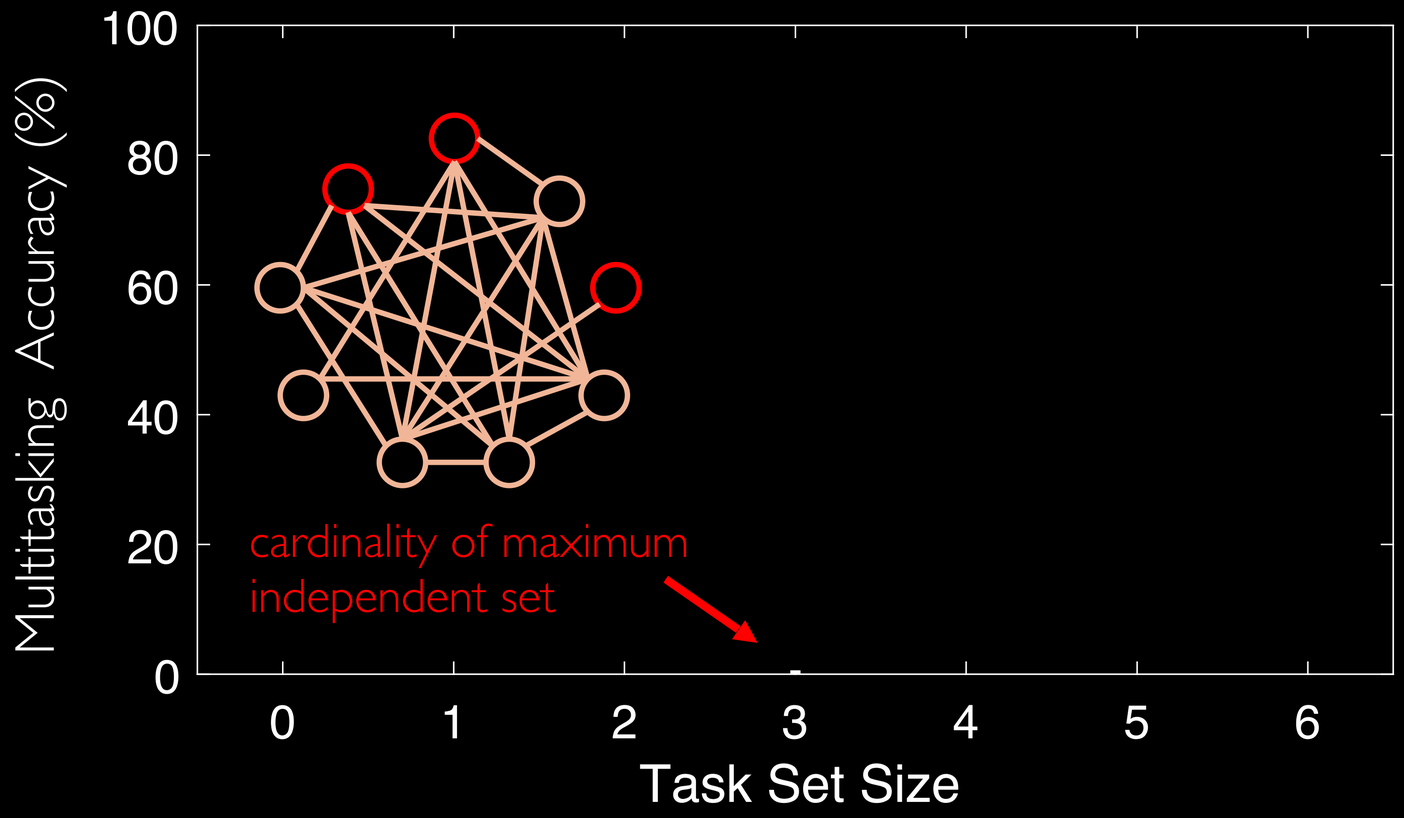
Predict Parallel Processing Capacity Based on MIS



Petri et al. Nature Physics (2021): 1-6.

Application to Neural Systems

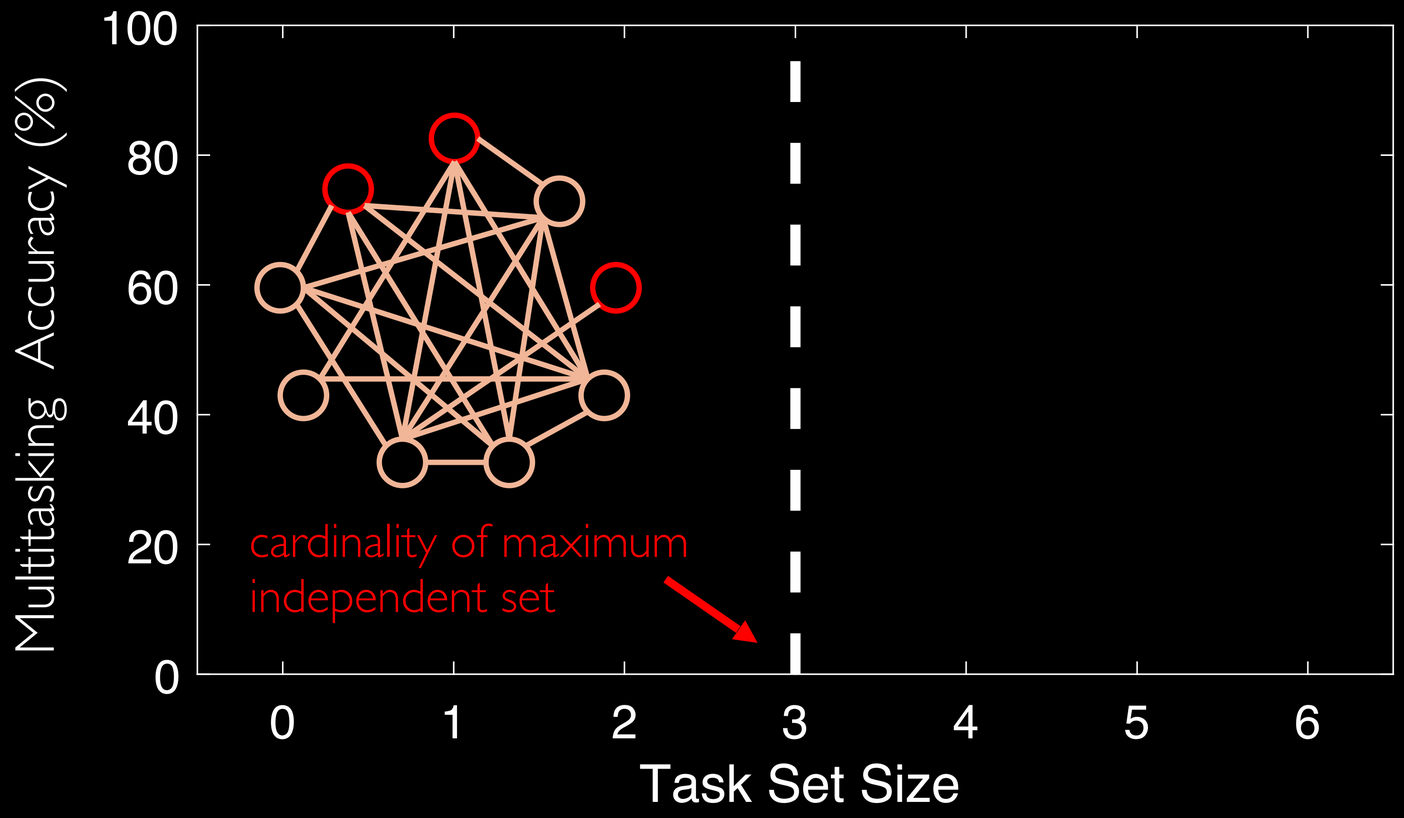
Predict Parallel Processing Capacity Based on MIS



Petri et al. Nature Physics (2021): 1-6.

Application to Neural Systems

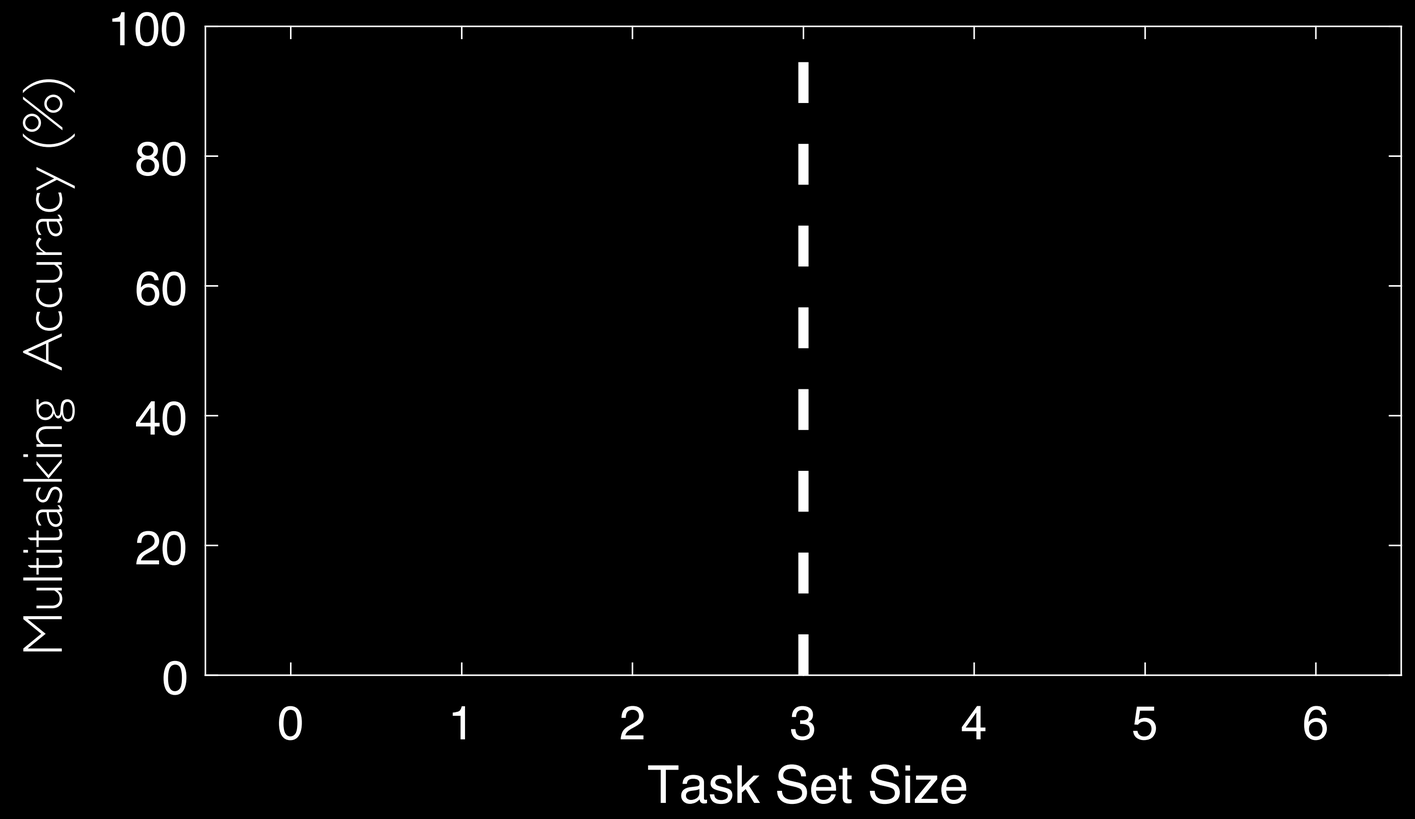
Predict Parallel Processing Capacity Based on MIS



Petri et al. Nature Physics (2021): 1-6.

Application to Neural Systems

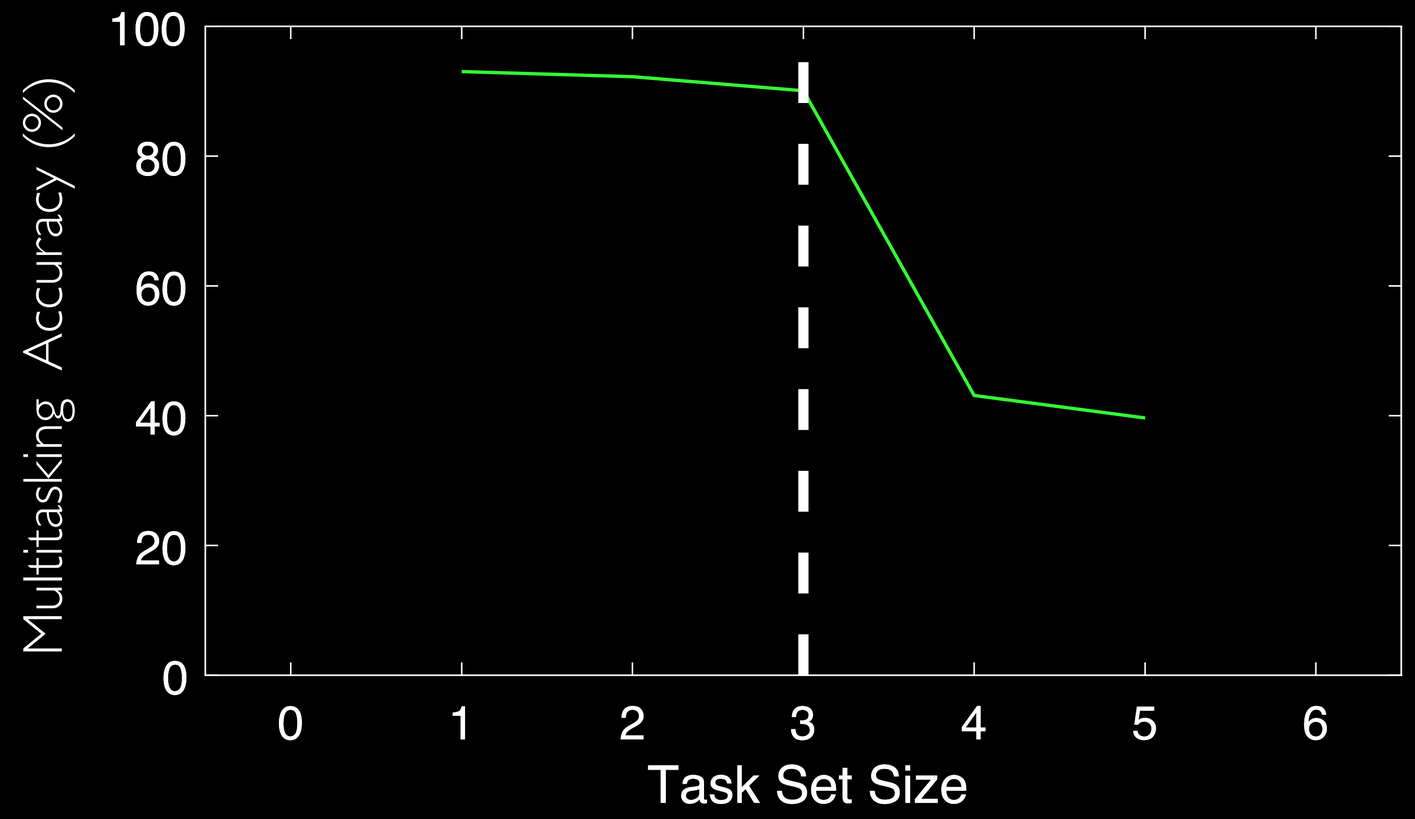
Predict Parallel Processing Capacity Based on MIS



Petri et al. Nature Physics (2021): 1-6.

Application to Neural Systems

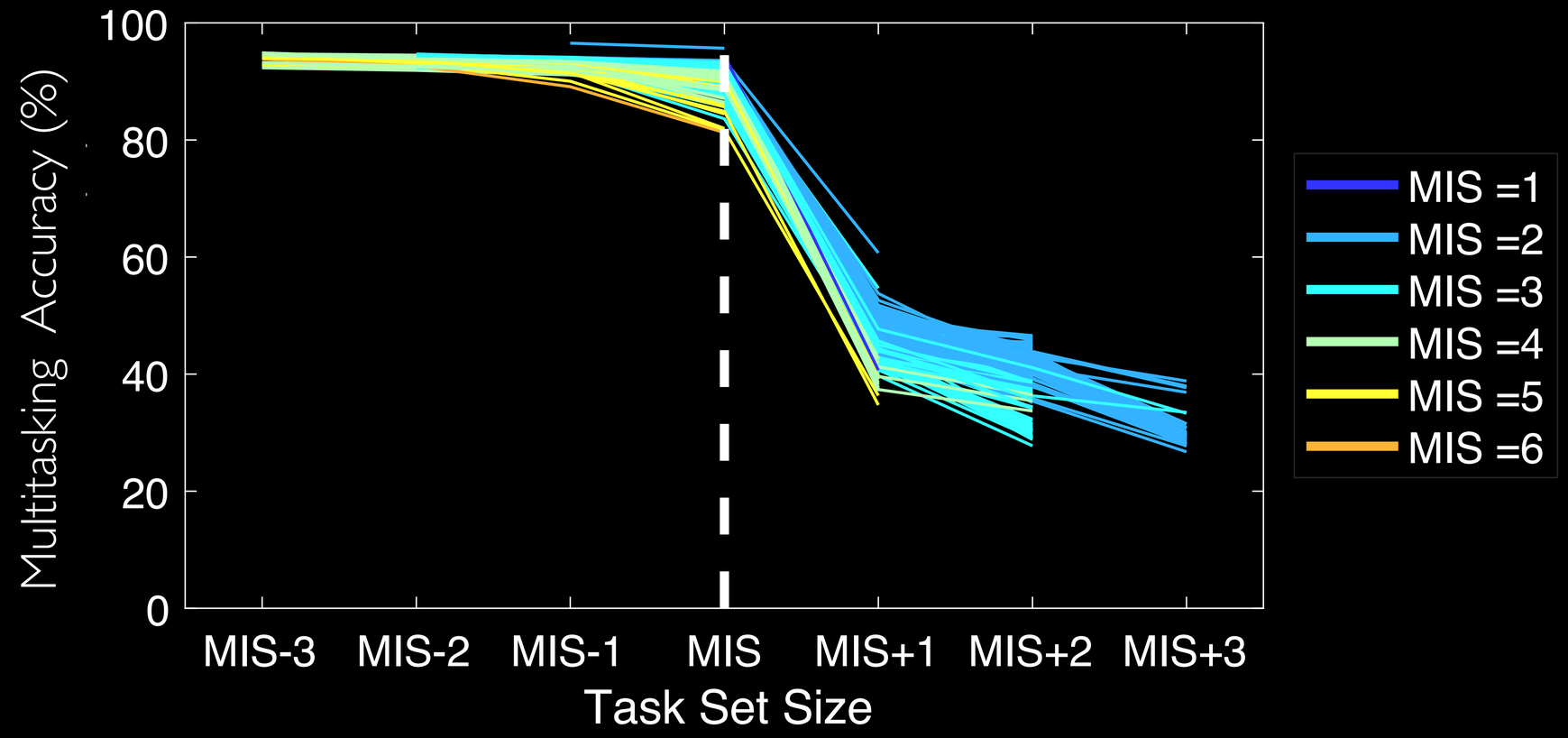
Predict Parallel Processing Capacity Based on MIS



Petri et al. Nature Physics (2021): 1-6.

Application to Neural Systems

Predict Parallel Processing Capacity Based on MIS



Petri et al. Nature Physics (2021): 1-6.

Good.
Can we scale this up?

Stat. mech. approach to MIS estimation

Max Set Packing in a nutshell

The solution comes out like this....

$$\rho_\alpha = \frac{\langle k \rangle}{\langle c \rangle}\left(1 - p_*^{c/(c-1)}\right) + (M_k(t) - M'(t))$$

$$p_* = \mathbb{E}_{\tilde{c}}\left(1 - \frac{1}{\langle k \rangle p_*} M'(t)\right)^{\tilde{c}}$$

Stat. mech. approach to MIS estimation

Max Set Packing in a nutshell

The solution comes out like this....

$$\rho_\alpha = \frac{\langle k \rangle}{\langle c \rangle}\left(1 - p_*^{c/(c-1)}\right) + (M_k(t) - M'(t))$$

$$p_* = \mathbb{E}_{\tilde{c}}\left(1 - \frac{1}{\langle k \rangle p_*} M'(t)\right)^{\tilde{c}}$$

Degree-generating function

Stat. mech. approach to MIS estimation

Max Set Packing in a nutshell

The solution comes out like this....

$$\rho_\alpha = \frac{\langle k \rangle}{\langle c \rangle}\left(1 - p_*^{c/(c-1)}\right) + (M_k(t) - M'(t))$$

$$p_* = \mathbb{E}_{\tilde{c}}\left(1 - \frac{1}{\langle k \rangle p_*} M'(t)\right)^{\tilde{c}}$$

Degree-generating function

Specifying to Gaussian degree distributions, it becomes...

$$M_k(t) = \mathrm{e}^{\langle k \rangle t + \sigma^2 t^2/2}$$

Stat. mech. approach to MIS estimation

Max Set Packing in a nutshell

The solution comes out like this....

$$\rho_\alpha = \frac{\langle k \rangle}{\langle c \rangle}\left(1 - p_*^{c/(c-1)}\right) + (M_k(t) - M'(t))$$

$$p_* = \mathbb{E}_{\tilde{c}}\left(1 - \frac{1}{\langle k \rangle p_*} M'(t)\right)^{\tilde{c}}$$

Degree-generating function

Specifying to Gaussian degree distributions, it becomes...

$$M_k(t) = \mathrm{e}^{\langle k \rangle t + \sigma^2 t^2/2}$$

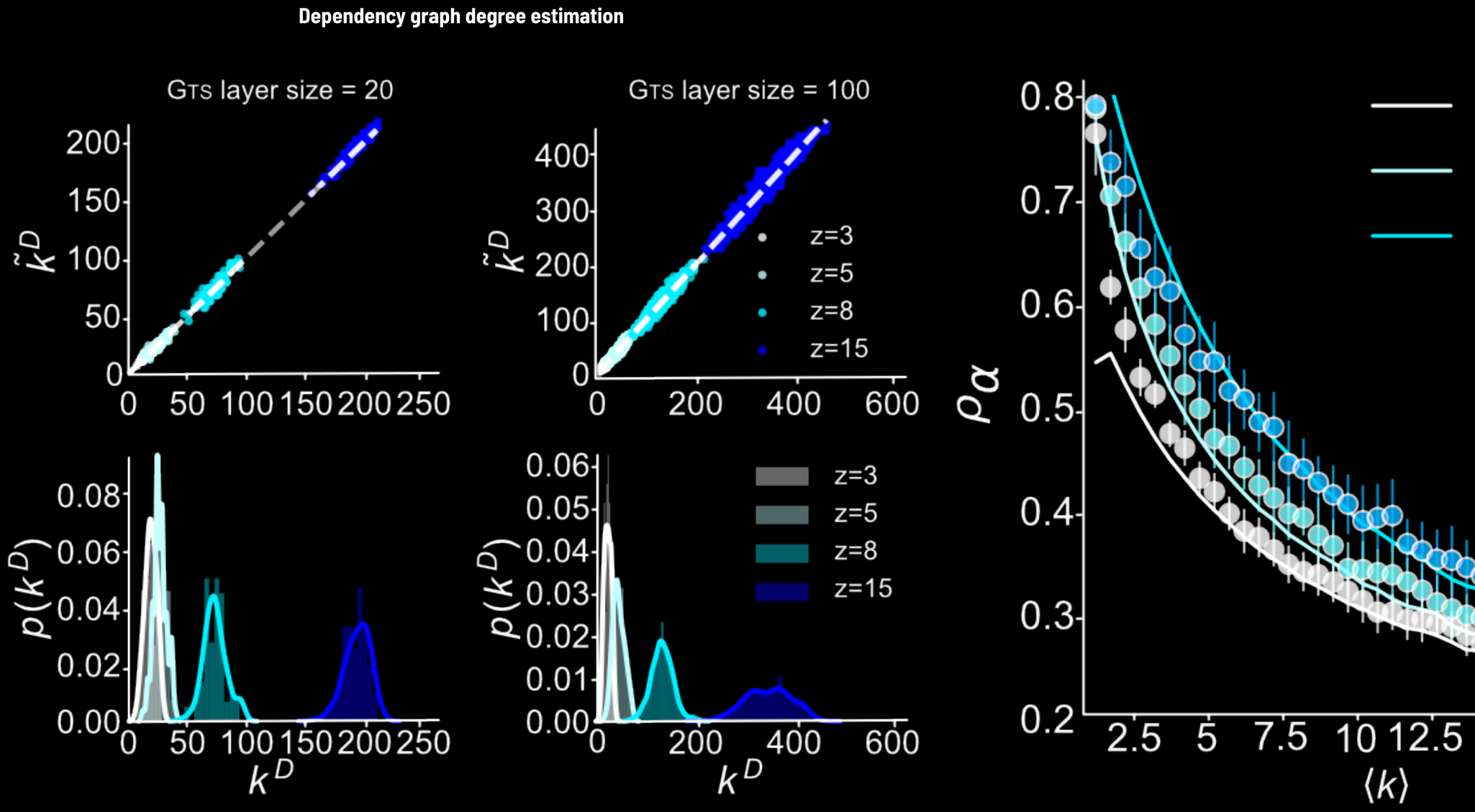
$$\rho_\alpha = \frac{\langle k \rangle}{\langle c \rangle}\left(1 - p_*^{c/(c-1)}\right) + M_k(\ln p_*)(1 - \langle k \rangle - \sigma^2 \ln p_*)$$

Stat. mech. approach to MIS estimation

MIS results

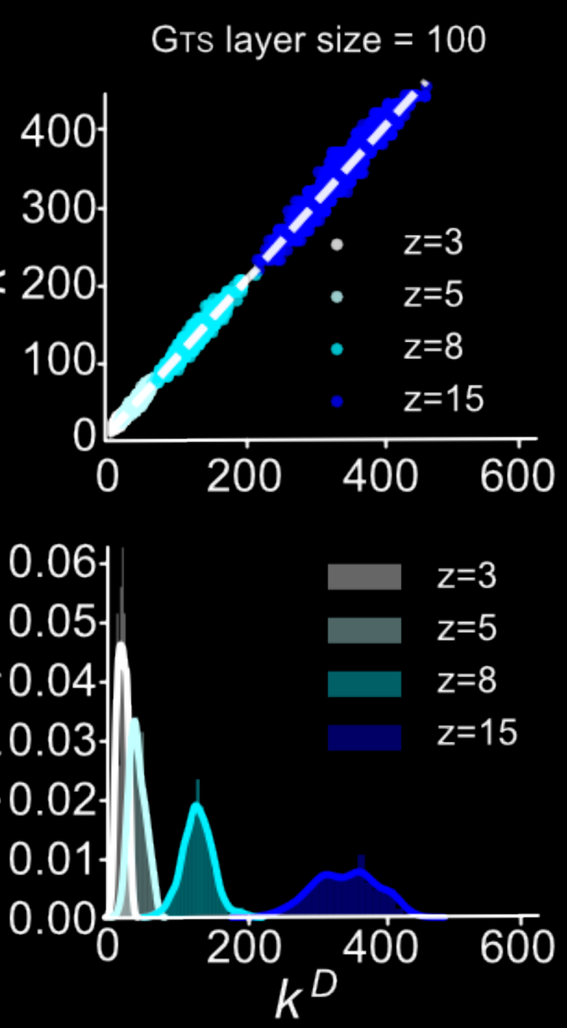
Stat. mech. approach to MIS estimation

MIS results

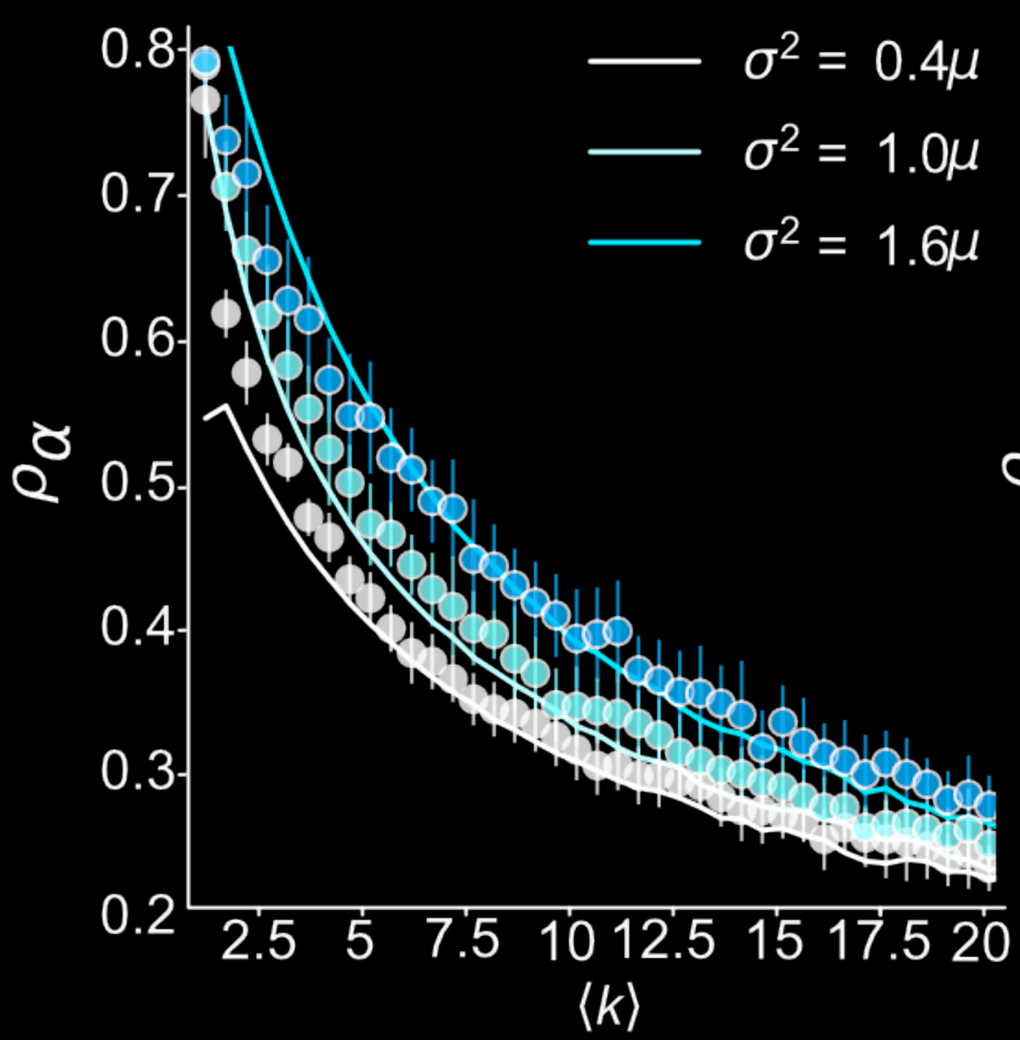


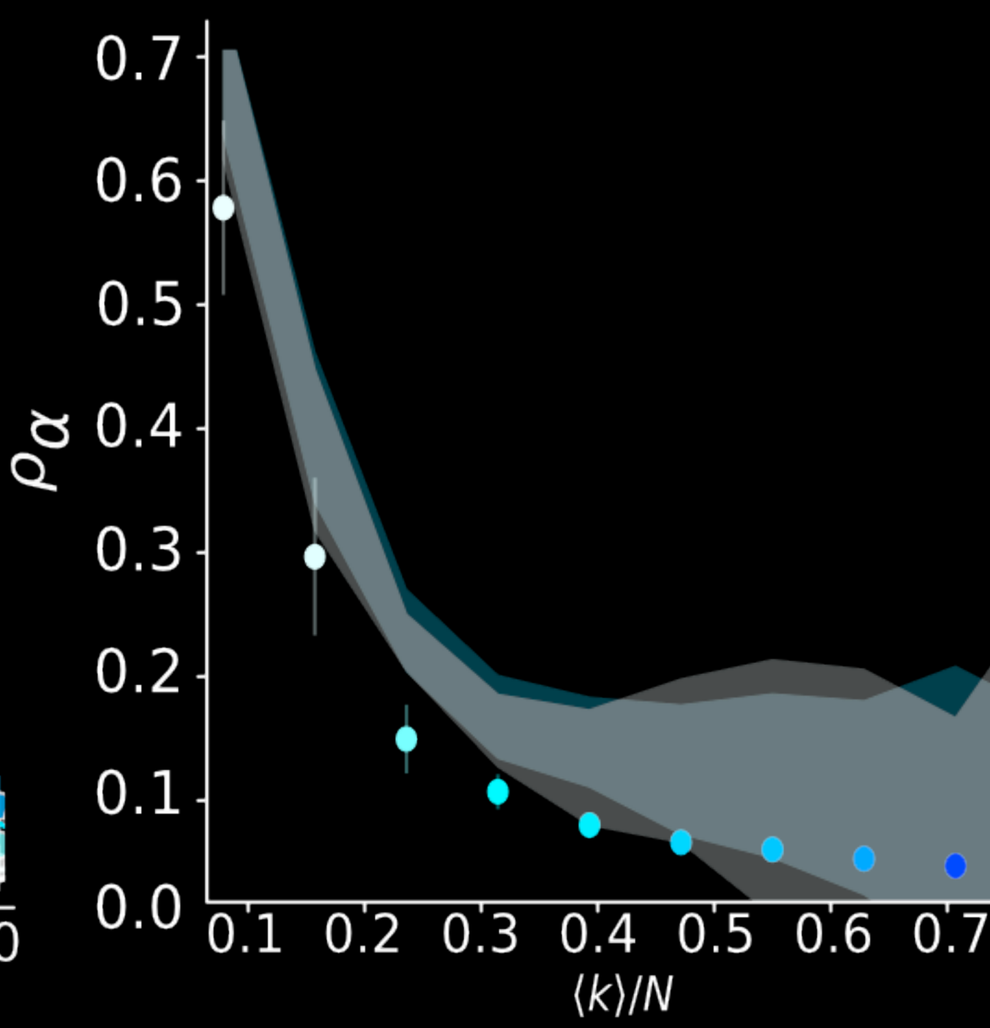
Stat. mech. approach to MIS estimation

MIS results



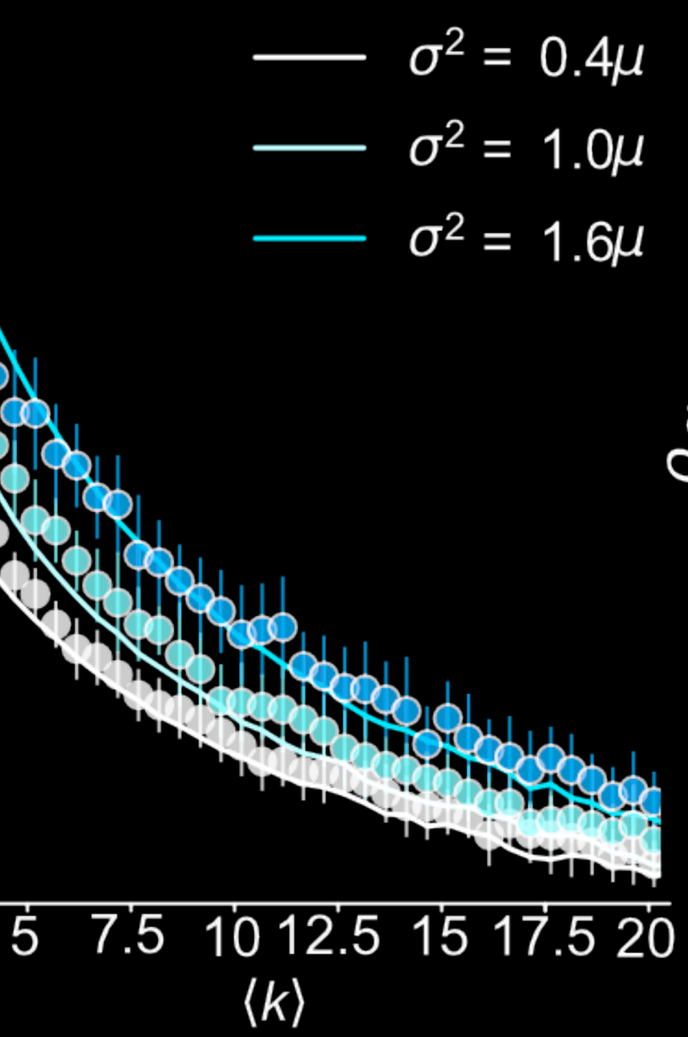
MIS prediction as function of density and heterogeneity



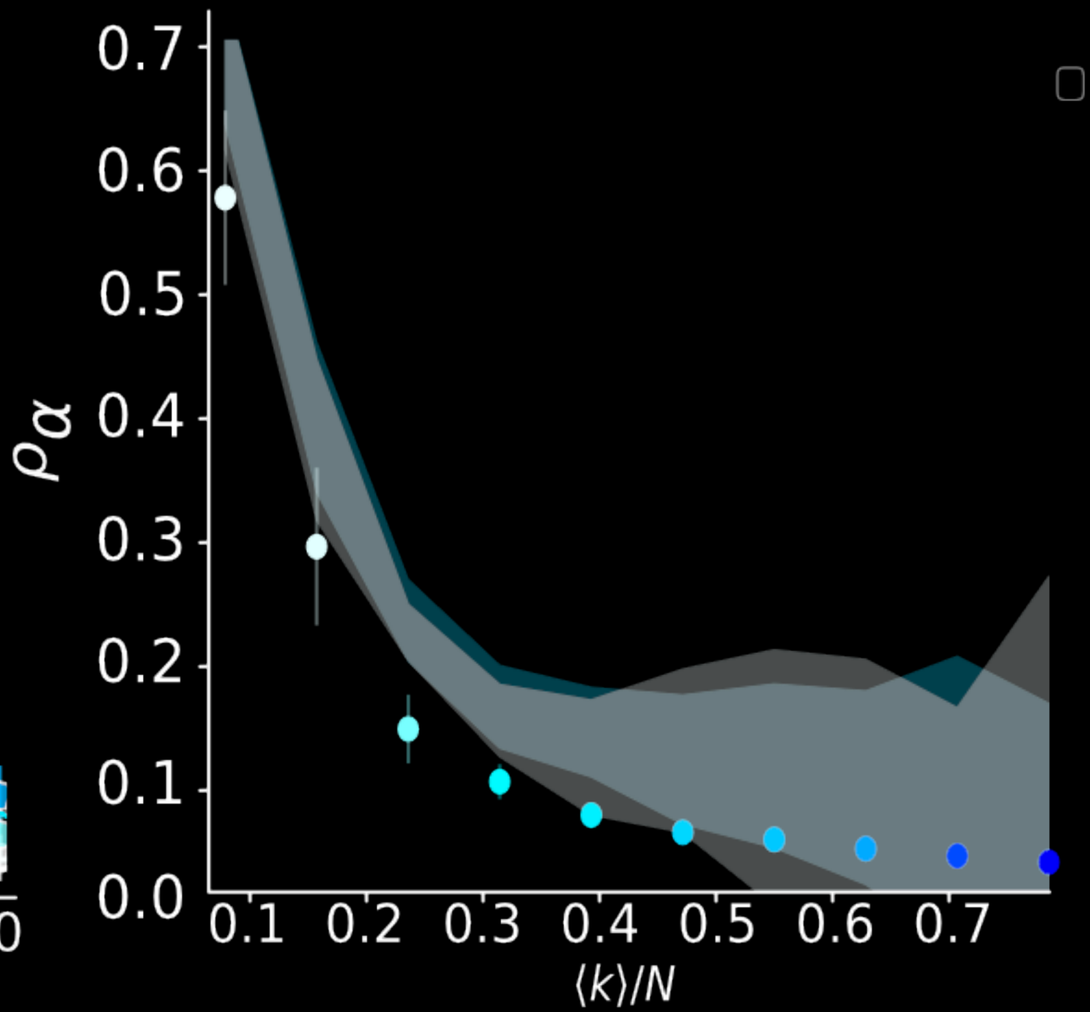


Stat. mech. approach to MIS estimation

MIS results



All together!



This holds for the MIS.
That is, a specific set.

This holds for the MIS.
That is, a specific set.

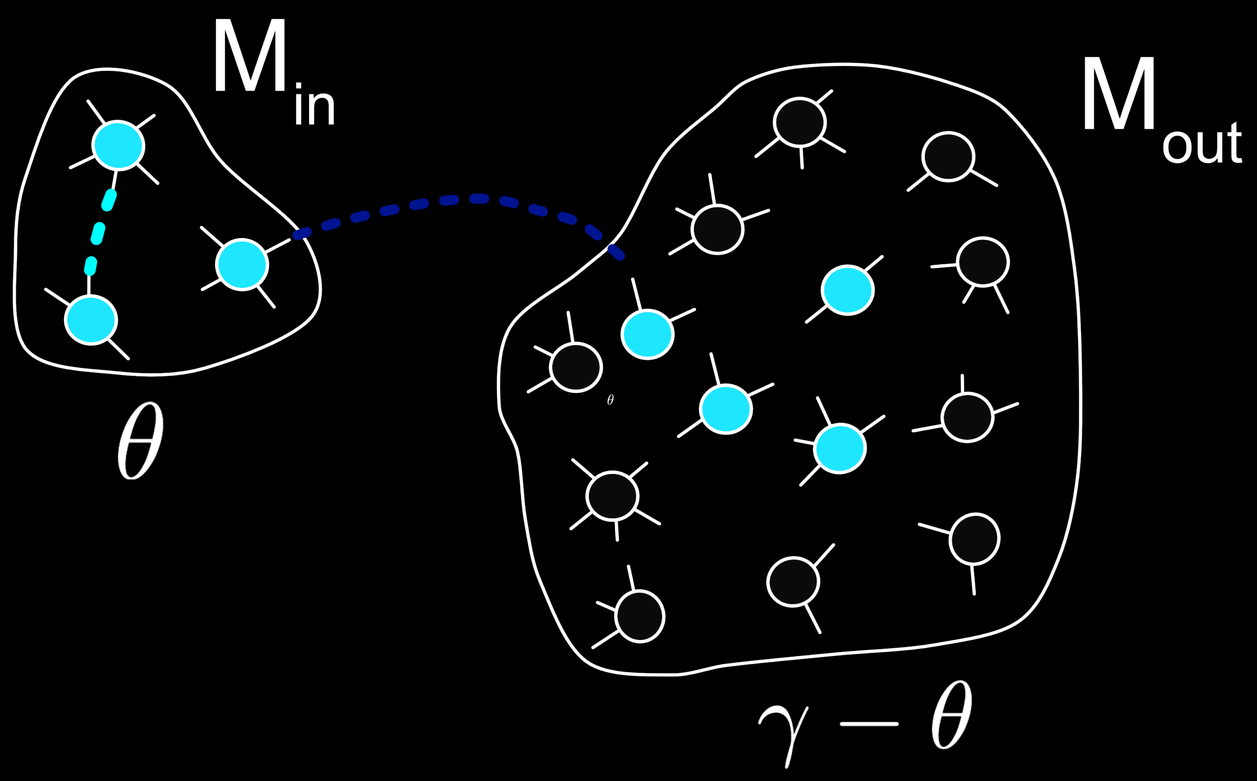
What happens if I give you
γ randomly chosen tasks?

This holds for the MIS.
That is, a specific set.

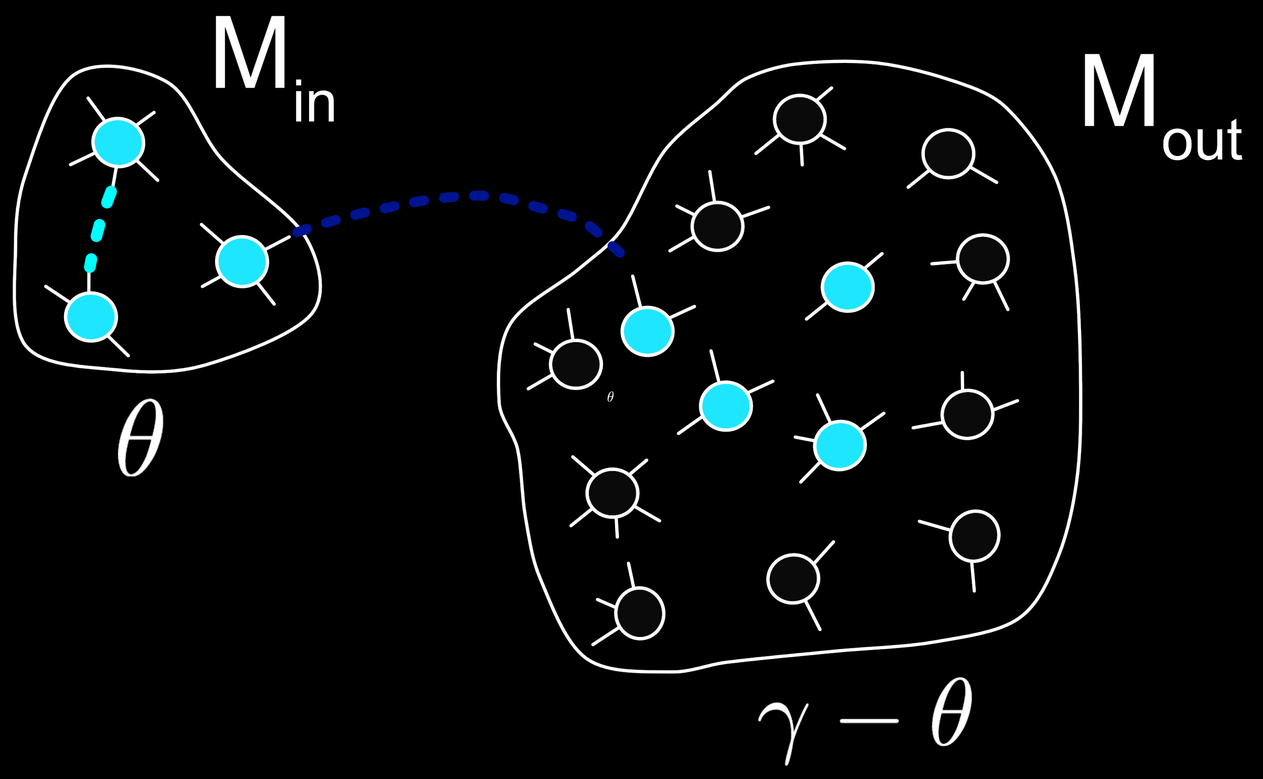
What happens if I give you
γ randomly chosen tasks?

**expected capacity vs
maximum capacity**

Estimation of expected parallel capacity



Estimation of expected parallel capacity

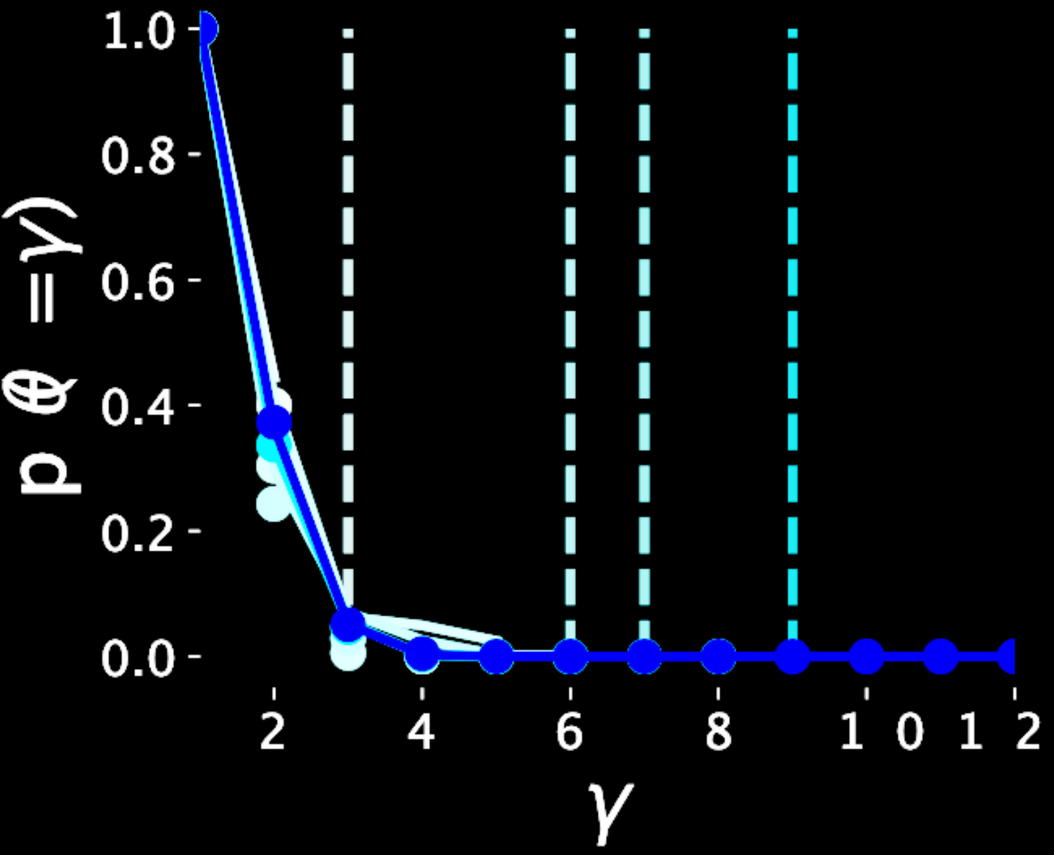


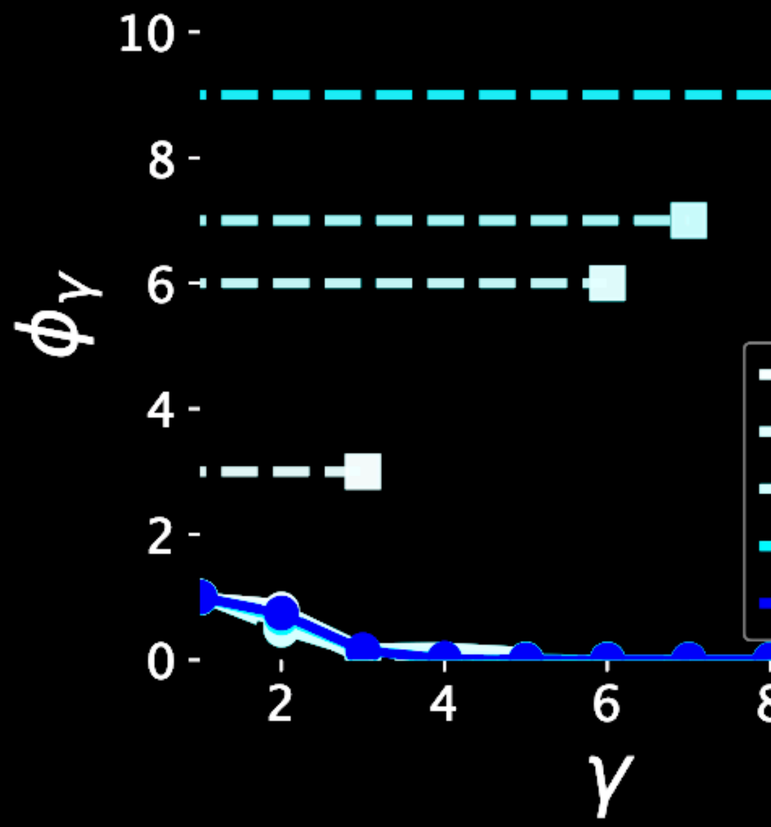
$$P(\theta;\gamma,\mathcal{G}_{\mathrm{D}}) \simeq \left(1-\frac{\langle k^2\rangle}{2M_{\mathrm{D}}}\right)^{\binom{\theta}{2}} \left(\frac{\theta\langle k\rangle^2}{2M_{\mathrm{D}}}\right)^{\gamma-\theta}$$

Estimation of expected parallel capacity

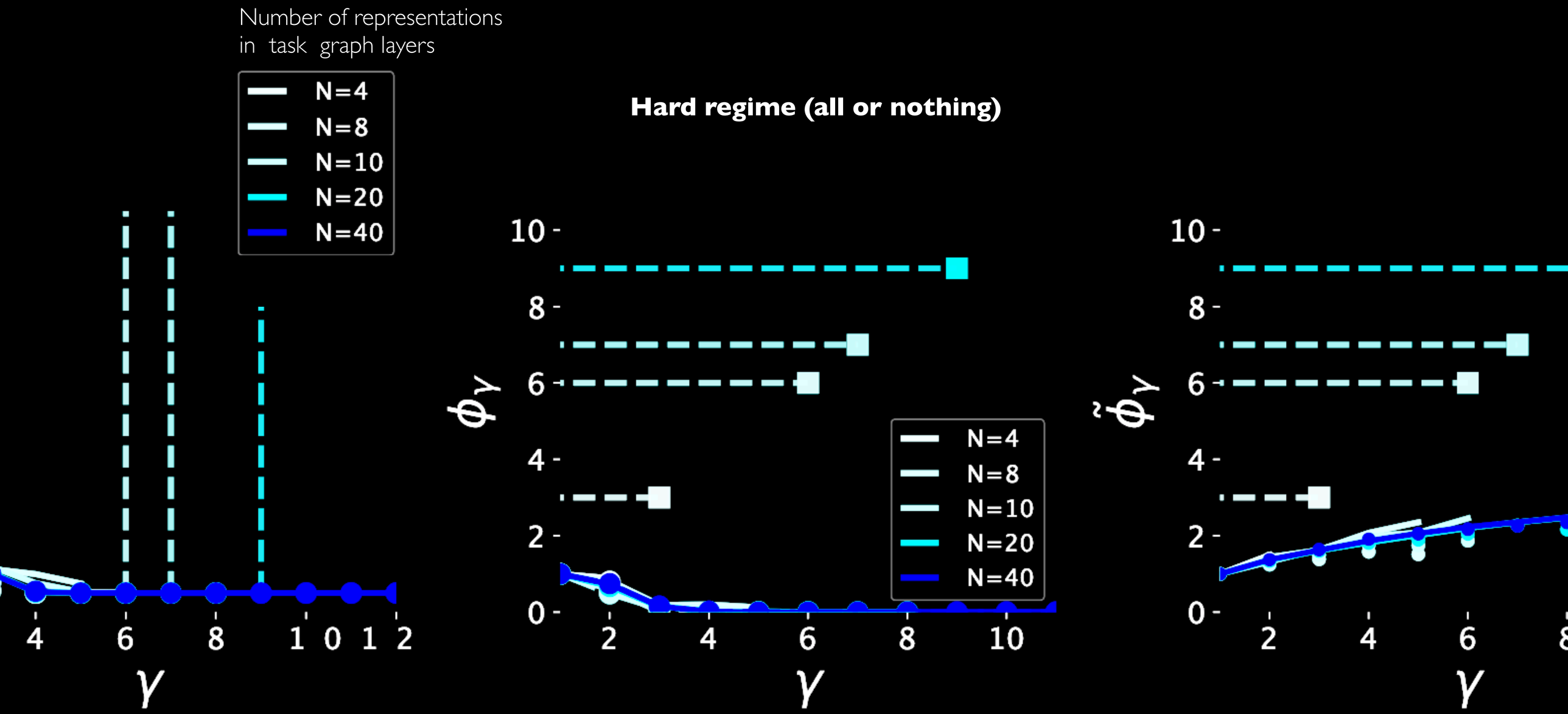
Number of representations
in task graph layers

Probability of successful execution



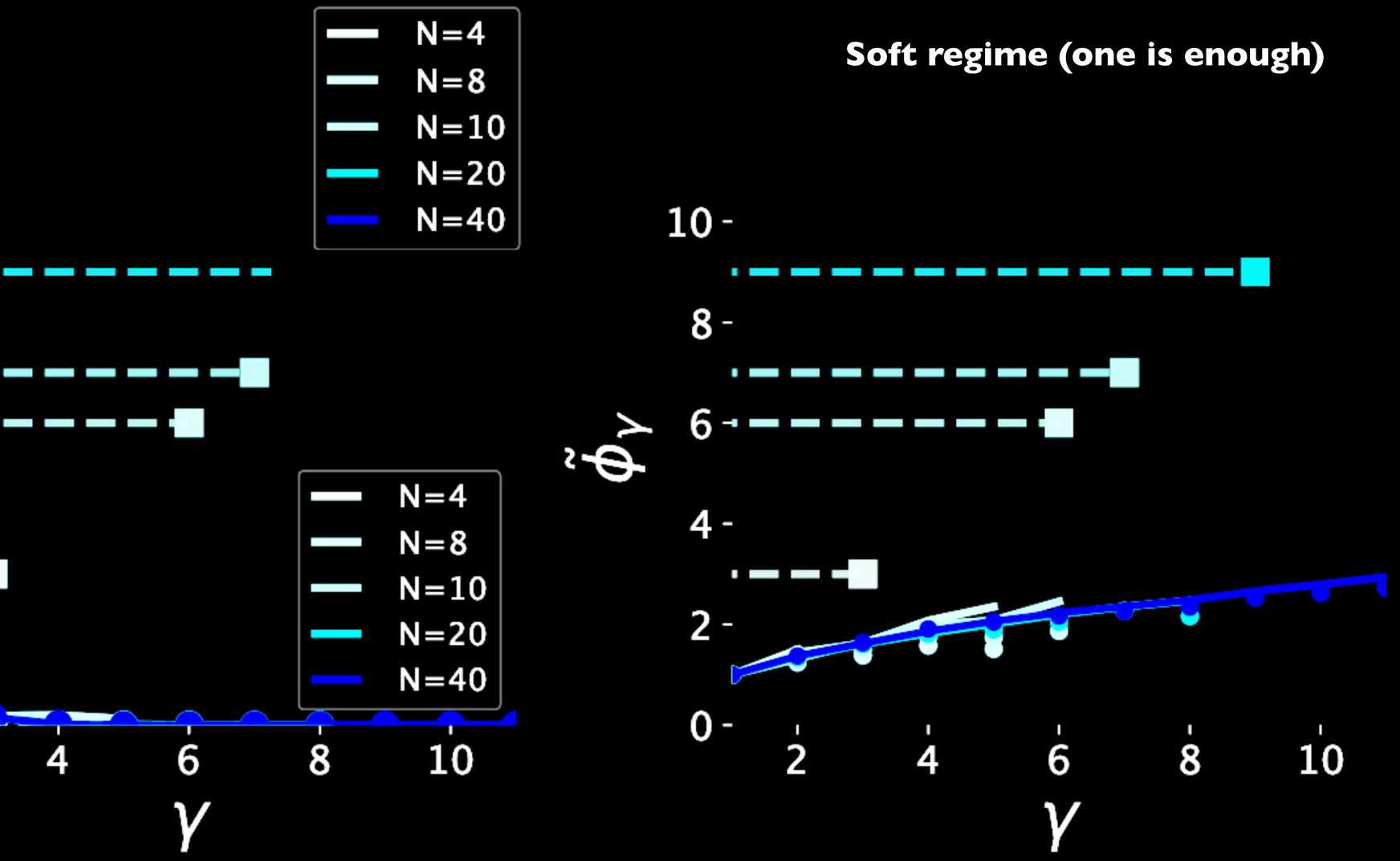


Estimation of expected parallel capacity



Estimation of expected parallel capacity

Number of representations
in task graph layers



Soft regime (one is enough)

Take-home #1

Take-home #1

Results:

- we can extract dependency patterns from data
- severe limits to max. capacity under modest repr. sharing
- expected capacity is even worse

Take-home #1

Results:

- we can extract dependency patterns from data
- severe limits to max. capacity under modest repr. sharing
- expected capacity is even worse

A little too rigid...

No weights, simple networks,
no control!!

Can we extend this?

Neuroscience

An Information-Theoretic Approach to Reward Rate Optimization in the Tradeoff Between Controlled and Automatic Processing in Neural Network Architectures

Giovanni Petri, Sebastian Musslick, Jonathan D. Cohen

Questions

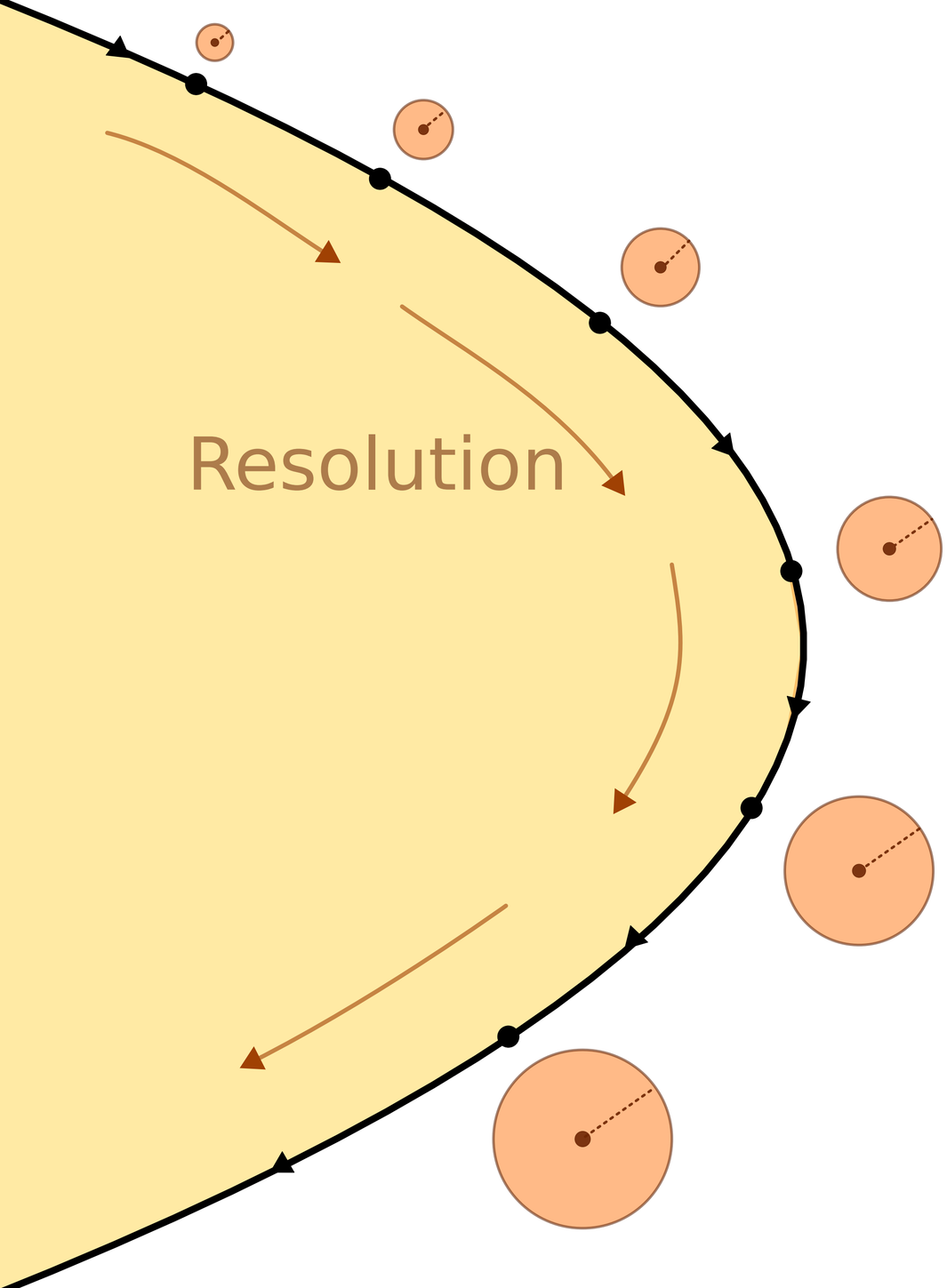
1. Do the limitations imposed by shared representations prevail in a system as large as the brain?

2. Assuming that shared representation cause a lot of trouble, why do we use them in the first place? (Experiment is evidence we do)

Questions

1. Do the limitations imposed by shared representations prevail in a system as large as the brain?

2. Assuming that shared representation cause a lot of trouble, why do we use them in the first place? (Experiment is evidence we do)



Miller's Law

Generalization $\longleftrightarrow$ Parallel Capacity

Toward a Universal Law of Generalization for Psychological Science

ROGER N. SHEPARD

Generalization ⟷ Parallel Capacity

Toward a Universal Law of Generalization for Psychological Science

ROGER N. SHEPARD

Generalization ⟷ Parallel Capacity

THE PSYCHOLOGICAL REVIEW

THE MAGICAL NUMBER SEVEN, PLUS OR MINUS TWO: SOME LIMITS ON OUR CAPACITY FOR PROCESSING INFORMATION [1]

GEORGE A. MILLER

Harvard University

Toward a Universal Law of Generalization for Psychological Science

ROGER N. SHEPARD

Psychometrica 1957; Science, 1987

Toward a Universal Law of Generalization for Psychological Science

ROGER N. SHEPARD

$$\mathcal{G}_{x\hat{x}} \triangleq \left(\frac{p_{x\hat{x}} \cdot p_{\hat{x}x}}{p_{\hat{x}\hat{x}} \cdot p_{xx}} \right)^{\frac{1}{2}}$$

Psychometrica 1957; Science, 1987

Toward a Universal Law of Generalization for Psychological Science

ROGER N. SHEPARD

$$\mathcal{G}_{x\hat{x}} \triangleq \left(\frac{p_{x\hat{x}} \cdot p_{\hat{x}x}}{p_{\hat{x}\hat{x}} \cdot p_{xx}} \right)^{\frac{1}{2}}$$

$$D_{ik}^{R} \cong -\log \left\{ (1 + C^{R}) \left(\frac{P_{ik}^{R} \cdot P_{ki}^{R}}{P_{ii}^{R} \cdot P_{kk}^{R}} \right)^{1/2} - C^{R} \right\}.$$

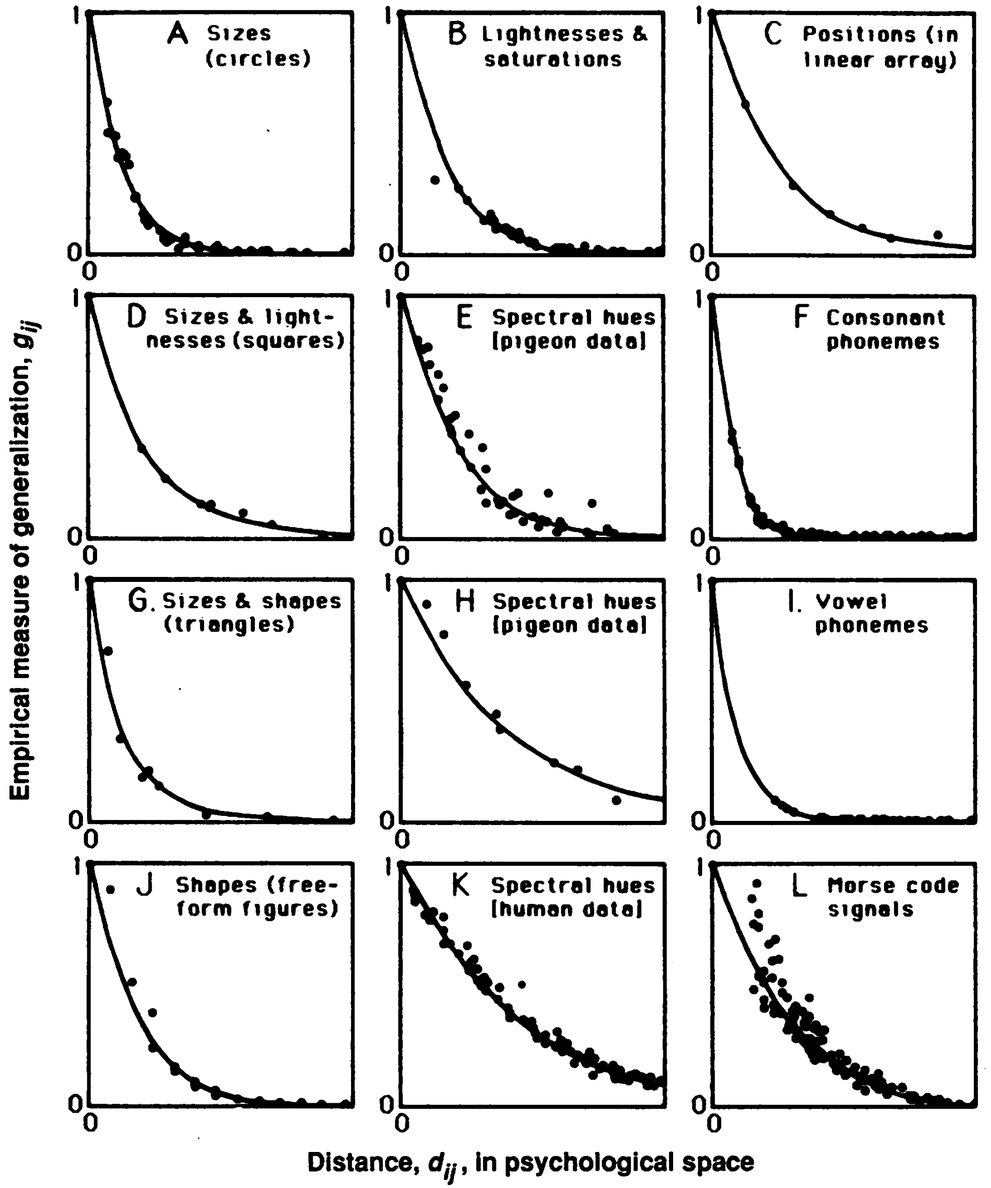
Psychometrica 1957; Science, 1987

Toward a Universal Law of Generalization for Psychological Science

ROGER N. SHEPARD

$$\mathcal{G}_{x\hat{x}} \triangleq \left(\frac{p_{x\hat{x}} \cdot p_{\hat{x}x}}{p_{\hat{x}\hat{x}} \cdot p_{xx}} \right)^{\frac{1}{2}}$$

$$D_{ik}^{R} \cong -\log \left\{ (1 + C^{R}) \left(\frac{P_{ik}^{R} \cdot P_{ki}^{R}}{P_{ii}^{R} \cdot P_{kk}^{R}} \right)^{1/2} - C^{R} \right\}.$$



Psychometrica 1957; Science, 1987

COGNITIVE PSYCHOLOGY

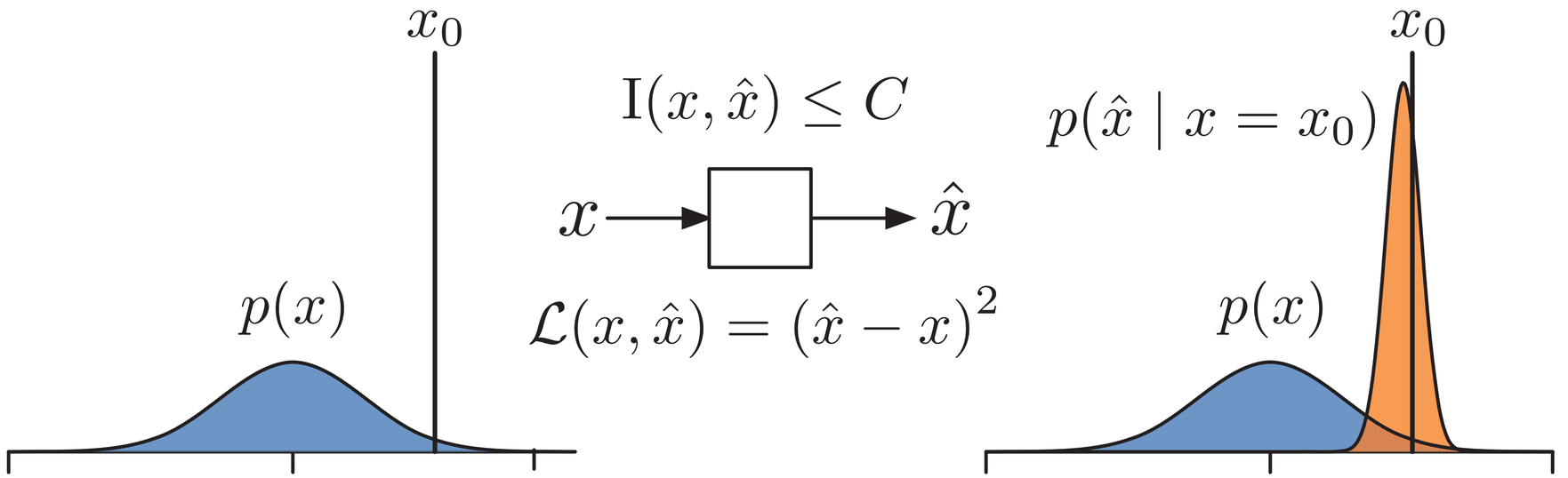
Efficient coding explains the universal law of generalization in human perception

Chris R. Sims*

COGNITIVE PSYCHOLOGY

Efficient coding explains the universal law of generalization in human perception

Chris R. Sims*



COGNITIVE PSYCHOLOGY

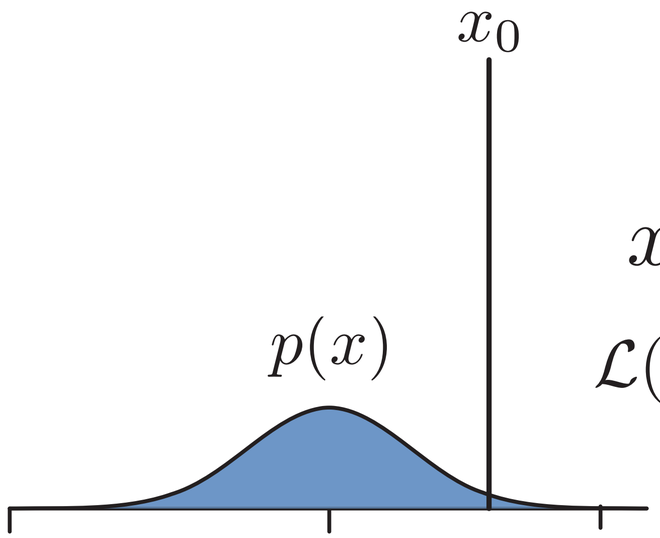
Efficient coding explains the universal law of generalization in human perception

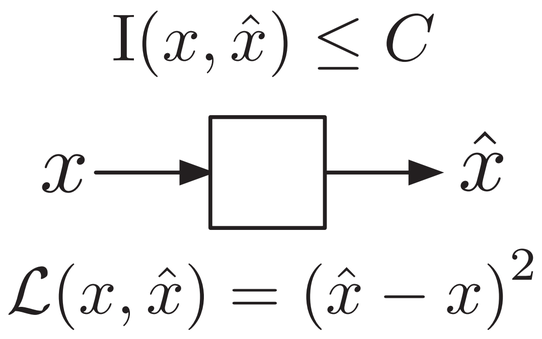
Chris R. Sims*

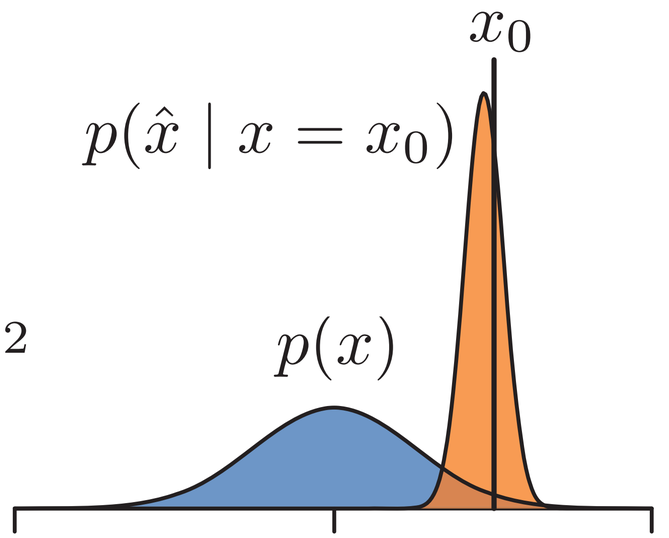
RDT

$$p(\hat{x} \mid x) = \frac{p(\hat{x}) \exp\big(s\, \mathcal{L}(x, \hat{x})\big)}{\int p(\hat{x}) \exp\big(s\, \mathcal{L}(x, \hat{x})\big)\, d\hat{x}},$$

$$p(\hat{x}) = \int p(x) p(\hat{x} \mid x) \mathrm{d}x\,.$$







COGNITIVE PSYCHOLOGY

Efficient coding explains the universal law of generalization in human perception

Chris R. Sims*

RDT

$$p(\hat{x} \mid x) = \frac{p(\hat{x}) \exp\big(s\, \mathcal{L}(x, \hat{x})\big)}{\int p(\hat{x}) \exp\big(s\, \mathcal{L}(x, \hat{x})\big)\, d\hat{x}},$$

$$p(\hat{x}) = \int p(x) p(\hat{x} \mid x) \mathrm{d}x\,.$$

COGNITIVE PSYCHOLOGY

Efficient coding explains the universal law of generalization in human perception

Chris R. Sims*

RDT

$$p(\hat{x} \mid x) = \frac{p(\hat{x}) \exp\big(s\, \mathcal{L}(x, \hat{x})\big)}{\int p(\hat{x}) \exp\big(s\, \mathcal{L}(x, \hat{x})\big)\, d\hat{x}},$$

$$\mathcal{G}_{x\hat{x}} \triangleq \left(\frac{p_{x\hat{x}} \cdot p_{\hat{x}x}}{p_{\hat{x}\hat{x}} \cdot p_{xx}} \right)^{\frac{1}{2}}$$

$$p(\hat{x}) = \int p(x) p(\hat{x} \mid x) \mathrm{d}x\,.$$

COGNITIVE PSYCHOLOGY

Efficient coding explains the universal law of generalization in human perception

Chris R. Sims*

RDT

$$p(\hat{x} \mid x) = \frac{p(\hat{x}) \exp\big(s\, \mathcal{L}(x, \hat{x})\big)}{\int p(\hat{x}) \exp\big(s\, \mathcal{L}(x, \hat{x})\big)\, d\hat{x}},$$

$$\mathcal{G}_{x\hat{x}} \triangleq \left(\frac{p_{x\hat{x}} \cdot p_{\hat{x}x}}{p_{\hat{x}\hat{x}} \cdot p_{xx}} \right)^{\frac{1}{2}}$$

$$p(\hat{x}) = \int p(x) p(\hat{x} \mid x) \mathrm{d}x\,.$$

$$\mathcal{G}_{x\hat{x}} = \exp\left[s \frac{1}{2} \Big(\mathcal{L}(x, \hat{x}) + \mathcal{L}(\hat{x}, x) - \mathcal{L}(x, x) - \mathcal{L}(\hat{x}, \hat{x}) \Big) \right]$$

COGNITIVE PSYCHOLOGY

Efficient coding explains the universal law of generalization in human perception

Chris R. Sims*

RDT

$$p(\hat{x} \mid x) = \frac{p(\hat{x}) \exp\big(s\, \mathcal{L}(x, \hat{x})\big)}{\int p(\hat{x}) \exp\big(s\, \mathcal{L}(x, \hat{x})\big)\, d\hat{x}},$$

$$\mathcal{G}_{x\hat{x}} \triangleq \left(\frac{p_{x\hat{x}} \cdot p_{\hat{x}x}}{p_{\hat{x}\hat{x}} \cdot p_{xx}}\right)^{\frac{1}{2}}$$

$$p(\hat{x}) = \int p(x) p(\hat{x} \mid x) \mathrm{d}x\,.$$

$$\mathcal{G}_{x\hat{x}} = \exp\left[s \frac{1}{2}\Big(\mathcal{L}(x, \hat{x}) + \mathcal{L}(\hat{x}, x) - \mathcal{L}(x, x) - \mathcal{L}(\hat{x}, \hat{x})\Big)\right]$$

$$\mathcal{G}_{x\hat{x}} = \exp\big[s\ \mathcal{L}(x, \hat{x})\big]$$

COGNITIVE PSYCHOLOGY

Efficient coding explains the universal law of generalization in human perception

Chris R. Sims*

RDT

$$p(\hat{x} \mid x) = \frac{p(\hat{x}) \exp\big(s\, \mathcal{L}(x, \hat{x})\big)}{\int p(\hat{x}) \exp\big(s\, \mathcal{L}(x, \hat{x})\big)\, d\hat{x}},$$

$$p(\hat{x}) = \int p(x) p(\hat{x} \mid x) \mathrm{d}x\,.$$

$$\mathcal{G}_{x\hat{x}} = \exp\left[s \frac{1}{2}\Big(\mathcal{L}(x, \hat{x}) + \mathcal{L}(\hat{x}, x) - \mathcal{L}(x, x) - \mathcal{L}(\hat{x}, \hat{x})\Big)\right]$$

$$\mathcal{G}_{x\hat{x}} \triangleq \left(\frac{p_{x\hat{x}} \cdot p_{\hat{x}x}}{p_{\hat{x}\hat{x}} \cdot p_{xx}}\right)^{\frac{1}{2}}$$

$$\mathcal{G}_{x\hat{x}} = \exp[s\ \mathcal{L}(x, \hat{x})]$$

this can be computed iteratively
(Blahut-Arimoto)

COGNITIVE PSYCHOLOGY

Efficient coding explains the universal law of generalization in human perception

Chris R. Sims*

RDT

$$p(\hat{x} \mid x) = \frac{p(\hat{x}) \exp\big(s\, \mathcal{L}(x, \hat{x})\big)}{\int p(\hat{x}) \exp\big(s\, \mathcal{L}(x, \hat{x})\big)\, d\hat{x}},$$

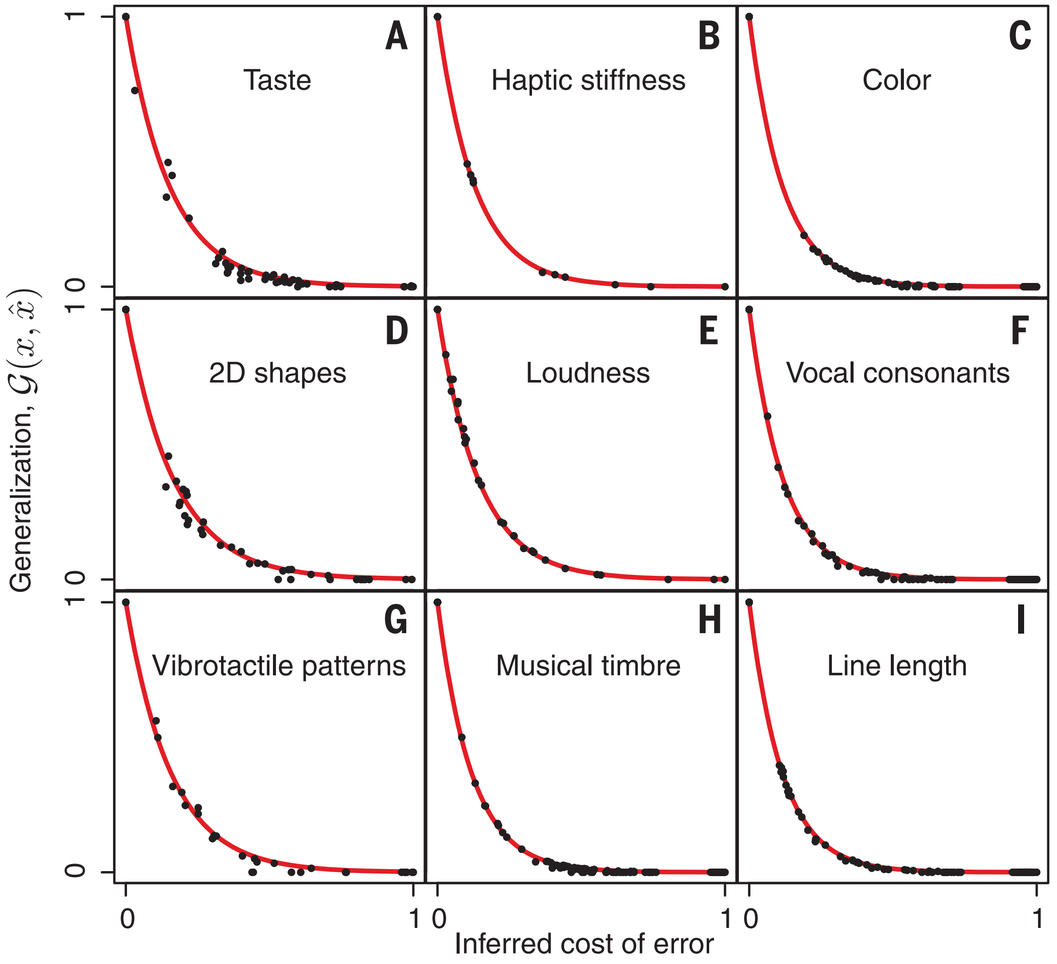
$$\mathcal{G}_{x\hat{x}} \triangleq \left(\frac{p_{x\hat{x}} \cdot p_{\hat{x}x}}{p_{\hat{x}\hat{x}} \cdot p_{xx}}\right)^{\frac{1}{2}}$$

$$p(\hat{x}) = \int p(x) p(\hat{x} \mid x) \mathrm{d}x\,.$$

$$\mathcal{G}_{x\hat{x}} = \exp\left[s \frac{1}{2}\Big(\mathcal{L}(x, \hat{x}) + \mathcal{L}(\hat{x}, x) - \mathcal{L}(x, x) - \mathcal{L}(\hat{x}, \hat{x})\Big)\right]$$

$$\mathcal{G}_{x\hat{x}} = \exp[s\ \mathcal{L}(x, \hat{x})]$$

this can be computed iteratively (Blahut-Arimoto)



THE PSYCHOLOGICAL REVIEW

THE MAGICAL NUMBER SEVEN, PLUS OR MINUS TWO: SOME LIMITS ON OUR CAPACITY FOR PROCESSING INFORMATION [1]

GEORGE A. MILLER

Harvard University

THE PSYCHOLOGICAL REVIEW

THE MAGICAL NUMBER SEVEN, PLUS OR MINUS TWO: SOME LIMITS ON OUR CAPACITY FOR PROCESSING INFORMATION [1]

GEORGE A. MILLER

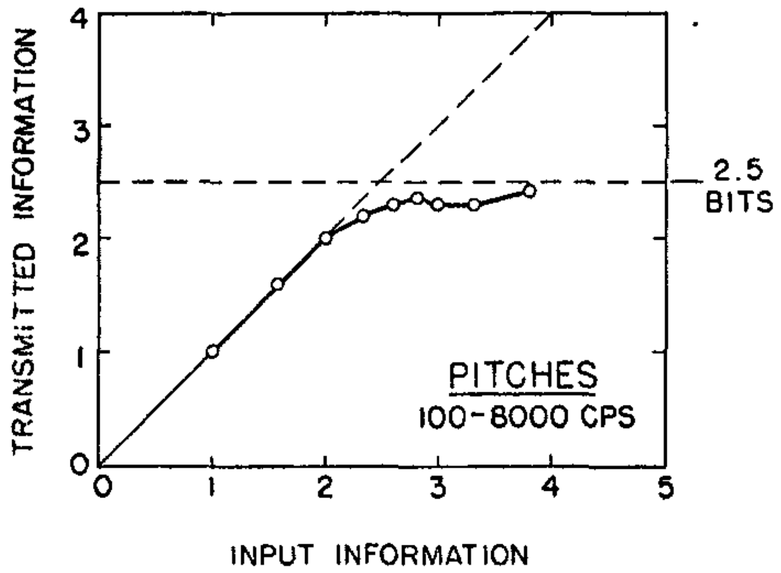
Harvard University

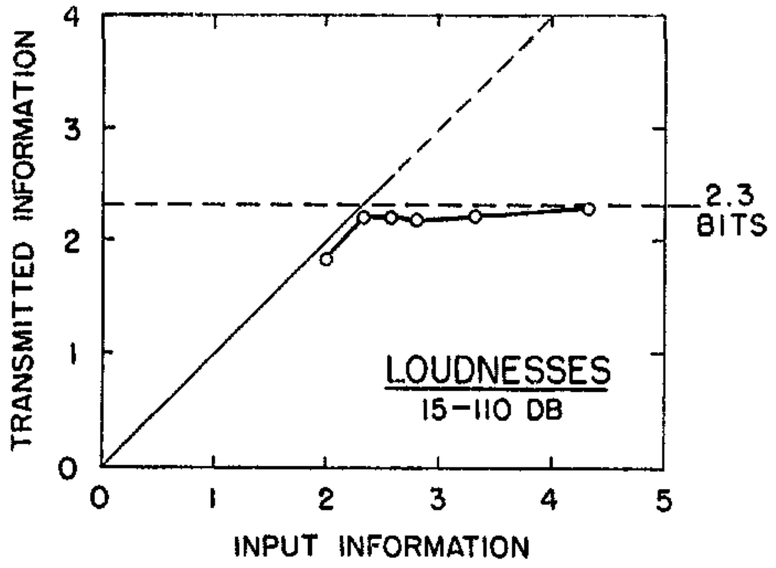
THE PSYCHOLOGICAL REVIEW

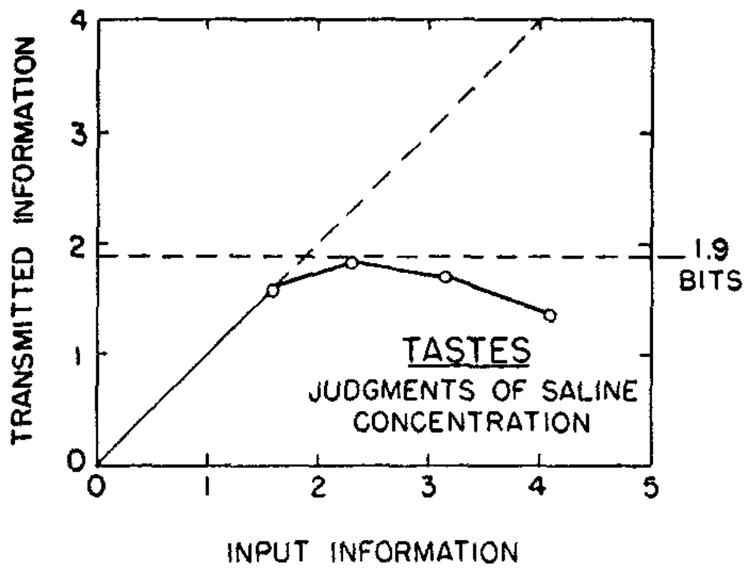
THE MAGICAL NUMBER SEVEN, PLUS OR MINUS TWO: SOME LIMITS ON OUR CAPACITY FOR PROCESSING INFORMATION [1]

GEORGE A. MILLER

Harvard University





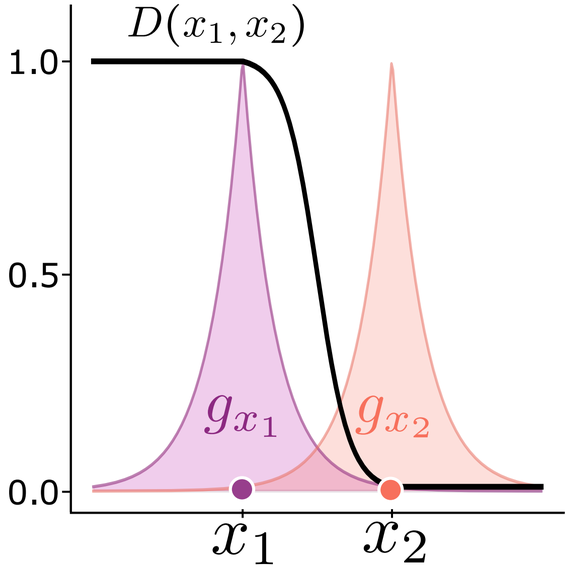


Generalization vs Processing

Bound by semanticity: universal laws governing the generalization-identification tradeoff

Marco Nurisso[1,2], Jesseba Fernando[3,4], Raj Deshpande[5], Alan Perotti[2], Raja Marjieh[6], Steven M. Frankland[7], Richard L. Lewis[8], Taylor W. Webb[9], Declan Campbell[10], Francesco Vaccarino[1], Jonathan D. Cohen[6,10], Giovanni Petri[2,5,11]

a No resolution



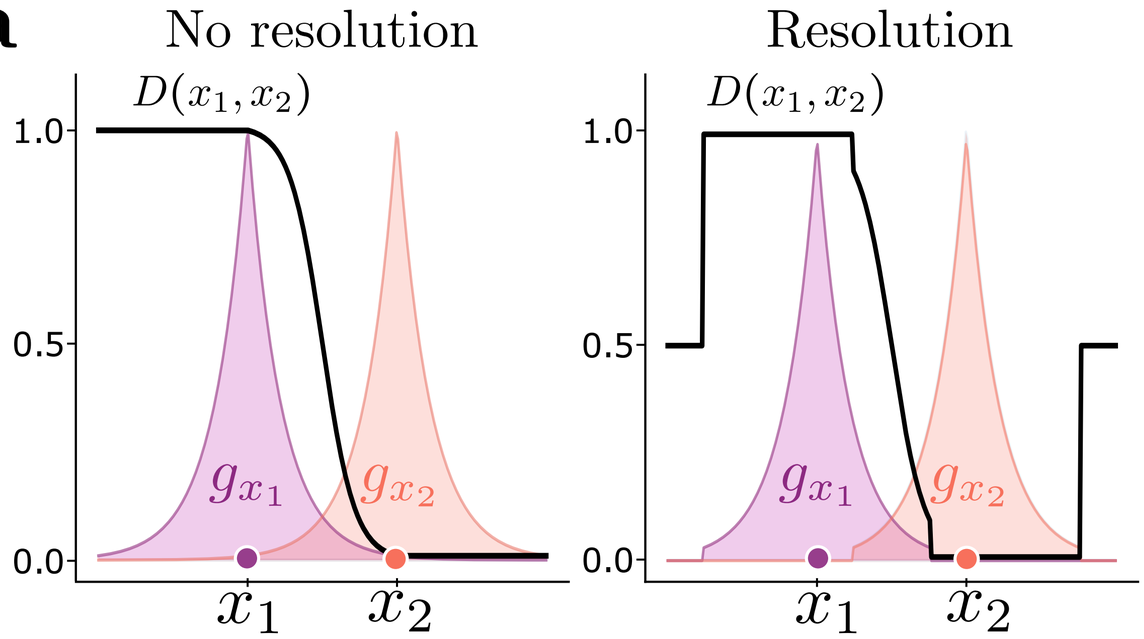
$$D(Y = x_1 \,|\, p) = \frac{G(x_1, p)}{G(x_1, p) + G(x_2, p)}$$

Generalization vs Processing

Bound by semanticity: universal laws governing the generalization-identification tradeoff

Marco Nurisso[1,2], Jesseba Fernando[3,4], Raj Deshpande[5], Alan Perotti[2], Raja Marjieh[6], Steven M. Frankland[7], Richard L. Lewis[8], Taylor W. Webb[9], Declan Campbell[10], Francesco Vaccarino[1], Jonathan D. Cohen[6,10], Giovanni Petri[2,5,11]

a

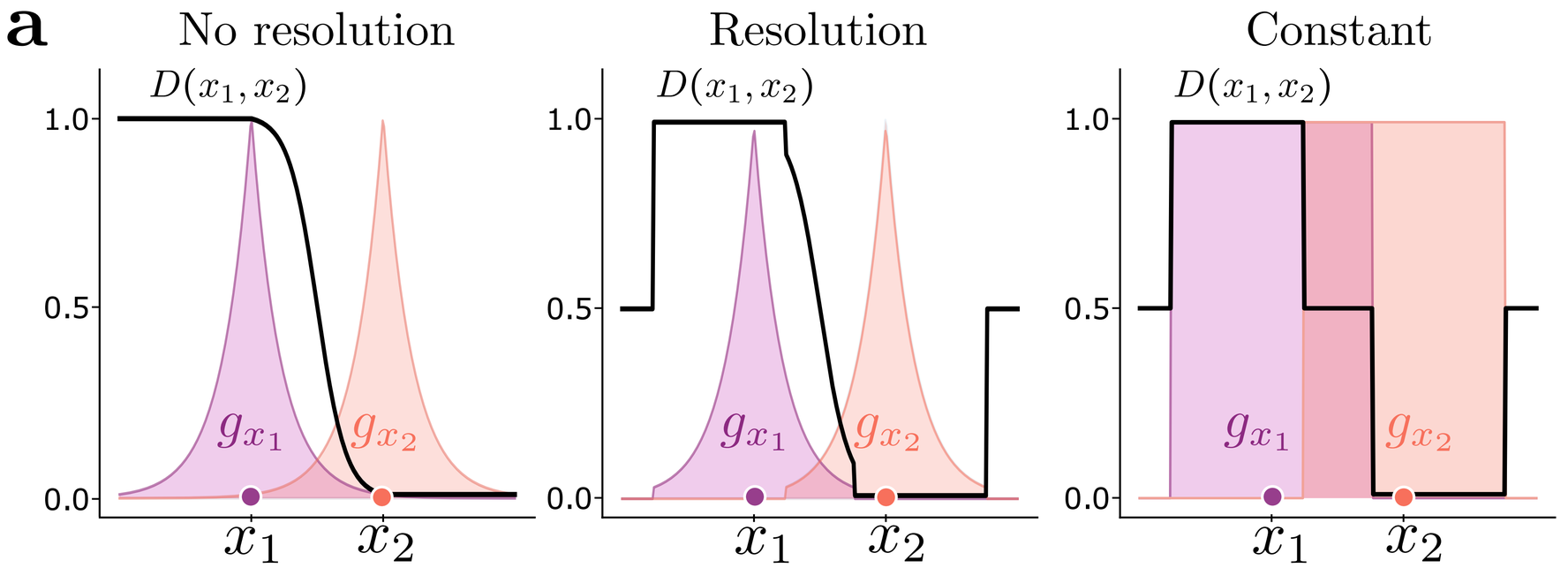


$$D(Y = x_1 \,|\, p) = \frac{G(x_1, p)}{G(x_1, p) + G(x_2, p)}$$

Generalization vs Processing

Bound by semanticity: universal laws governing the generalization-identification tradeoff

Marco Nurisso[1,2], Jesseba Fernando[3,4], Raj Deshpande[5], Alan Perotti[2], Raja Marjieh[6], Steven M. Frankland[7], Richard L. Lewis[8], Taylor W. Webb[9], Declan Campbell[10], Francesco Vaccarino[1], Jonathan D. Cohen[6,10], Giovanni Petri[2,5,11]

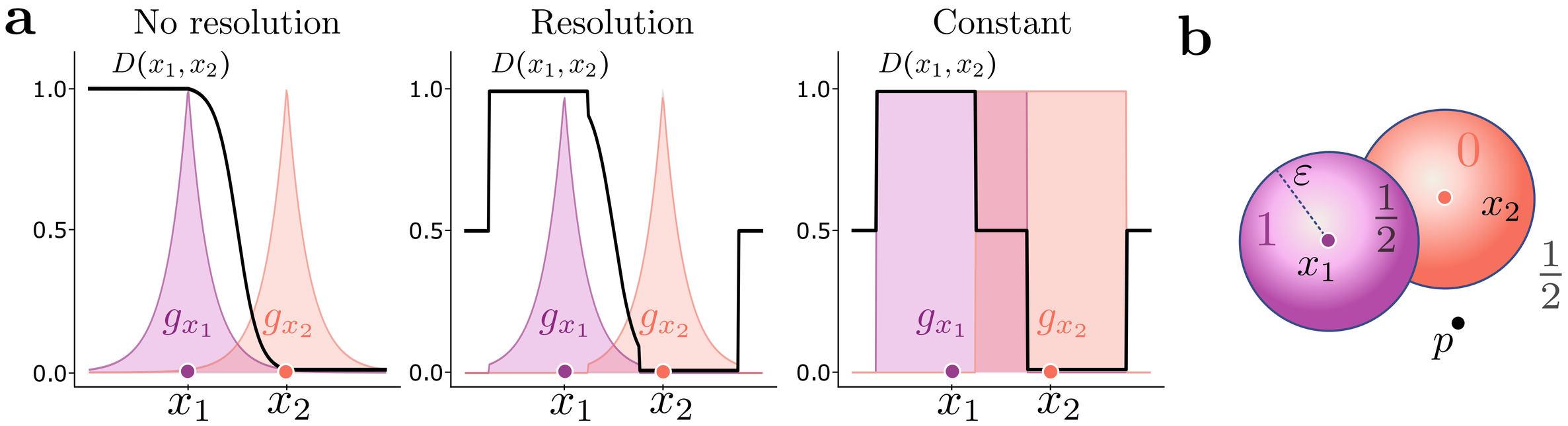


$$D(Y = x_1 \,|\, p) = \frac{G(x_1, p)}{G(x_1, p) + G(x_2, p)}$$

Generalization vs Processing

Bound by semanticity: universal laws governing the generalization-identification tradeoff

Marco Nurisso[1,2], Jesseba Fernando[3,4], Raj Deshpande[5], Alan Perotti[2], Raja Marjieh[6], Steven M. Frankland[7], Richard L. Lewis[8], Taylor W. Webb[9], Declan Campbell[10], Francesco Vaccarino[1], Jonathan D. Cohen[6,10], Giovanni Petri[2,5,11]



$$D(Y = x_1 \,|\, p) = \frac{G(x_1, p)}{G(x_1, p) + G(x_2, p)}$$

Generalization vs Processing

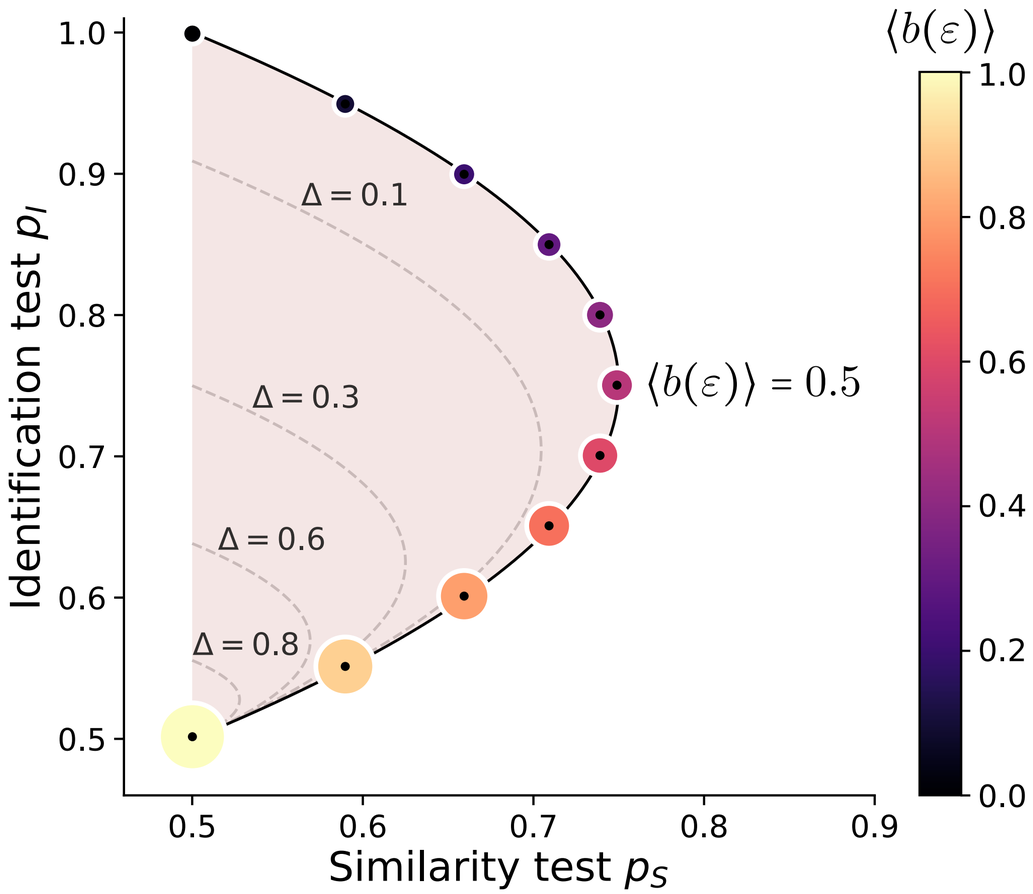
Bound by semanticity: universal laws governing the generalization-identification tradeoff

Marco Nurisso[1,2], Jesseba Fernando[3,4], Raj Deshpande[5], Alan Perotti[2], Raja Marjieh[6], Steven M. Frankland[7], Richard L. Lewis[8], Taylor W. Webb[9], Declan Campbell[10], Francesco Vaccarino[1], Jonathan D. Cohen[6,10], Giovanni Petri[2,5,11]

Generalization vs Processing

Bound by semanticity: universal laws governing the generalization-identification tradeoff

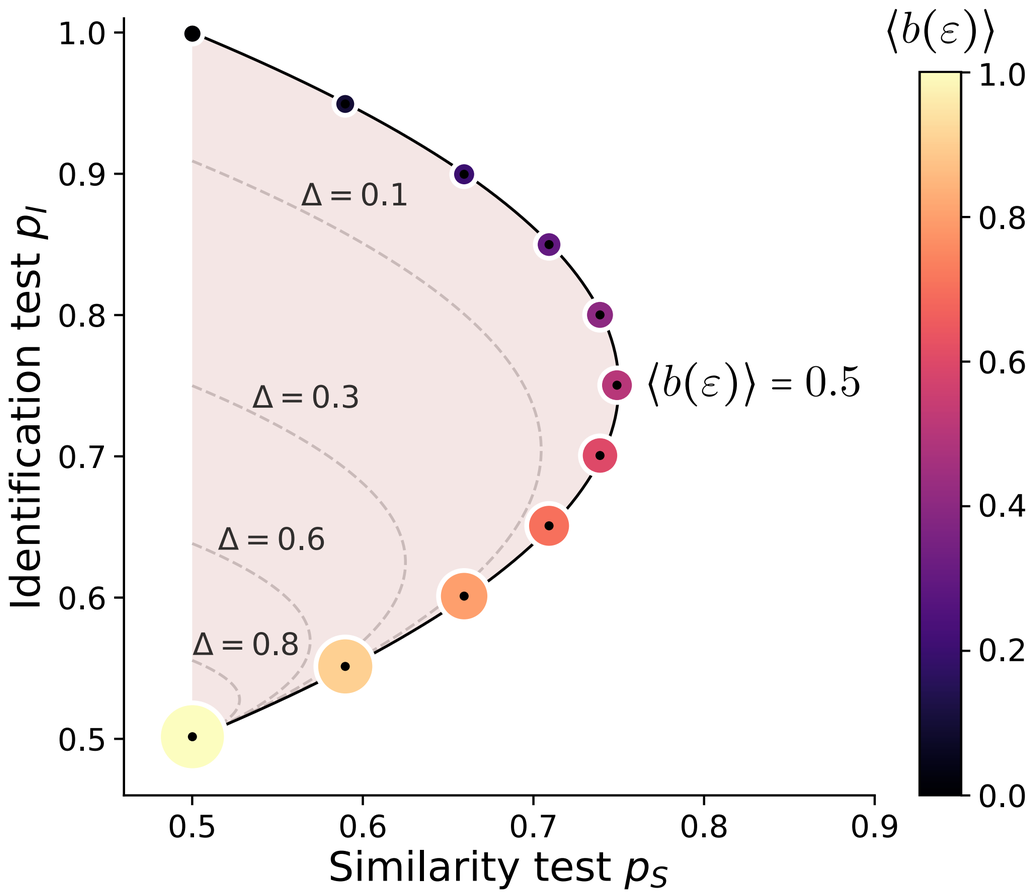
Marco Nurisso[1,2], Jesseba Fernando[3,4], Raj Deshpande[5], Alan Perotti[2], Raja Marjieh[6], Steven M. Frankland[7], Richard L. Lewis[8], Taylor W. Webb[9], Declan Campbell[10], Francesco Vaccarino[1], Jonathan D. Cohen[6,10], Giovanni Petri[2,5,11]



Generalization vs Processing

Bound by semanticity: universal laws governing the generalization-identification tradeoff

Marco Nurisso[1,2], Jesseba Fernando[3,4], Raj Deshpande[5], Alan Perotti[2], Raja Marjieh[6], Steven M. Frankland[7], Richard L. Lewis[8], Taylor W. Webb[9], Declan Campbell[10], Francesco Vaccarino[1], Jonathan D. Cohen[6,10], Giovanni Petri[2,5,11]



Theorem 1 (2-item tests)**.** *Let (M, d, Σ, ν) be a separable metric probability space. If, for every $p \in M$, b_p is absolutely continuous on every closed sub-interval of $[0, \infty)$, then, for the noise-free constant similarity function $g = g_{\varepsilon;0}$ it holds that*

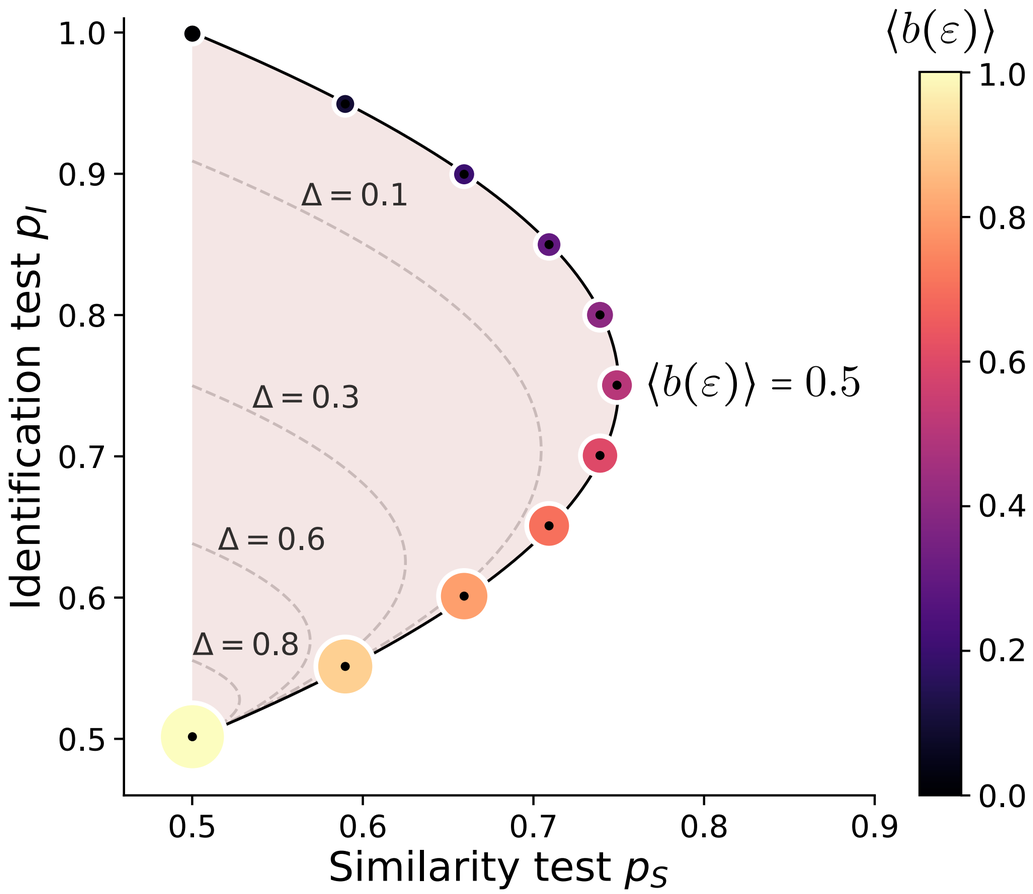
$$p_S(\varepsilon) = \frac{1}{2} + \langle b(\varepsilon)\rangle - \langle b(\varepsilon)\rangle^2 - \mathrm{Var}(b(\varepsilon)), \tag{3}$$

$$p_I(\varepsilon) = 1 - \frac{1}{2}\langle b(\varepsilon)\rangle. \tag{4}$$

Generalization vs Processing

Bound by semanticity: universal laws governing the generalization-identification tradeoff

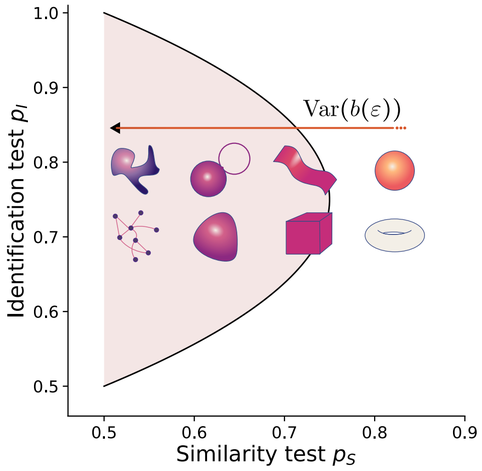
Marco Nurisso[1,2], Jesseba Fernando[3,4], Raj Deshpande[5], Alan Perotti[2], Raja Marjieh[6], Steven M. Frankland[7], Richard L. Lewis[8], Taylor W. Webb[9], Declan Campbell[10], Francesco Vaccarino[1], Jonathan D. Cohen[6,10], Giovanni Petri[2,5,11]



Theorem 1 (2-item tests). *Let (M, d, Σ, ν) be a separable metric probability space. If, for every $p \in M$, b_p is absolutely continuous on every closed sub-interval of $[0, \infty)$, then, for the noise-free constant similarity function $g = g_{\varepsilon;0}$ it holds that*

$$p_S(\varepsilon) = \frac{1}{2} + \langle b(\varepsilon)\rangle - \langle b(\varepsilon)\rangle^2 - \mathrm{Var}(b(\varepsilon)), \tag{3}$$

$$p_I(\varepsilon) = 1 - \frac{1}{2}\langle b(\varepsilon)\rangle. \tag{4}$$



Theorem 2 (Noise). *Under the same assumptions of Theorem 1, for the two-item similarity and identification tests with constant similarity functions $g = g_{\varepsilon;\Delta}$ with noise level $\Delta \geq 0$ it holds that*

$$p_S(\varepsilon, \Delta) = \frac{1}{2} + \frac{1-\Delta}{1+\Delta}(\langle b(\varepsilon)\rangle - \langle b(\varepsilon)^2\rangle), \tag{5}$$

$$p_I(\varepsilon, \Delta) = \frac{2 - (1-\Delta)\langle b(\varepsilon)\rangle}{2 + 2\Delta}. \tag{6}$$

Generalization vs Processing

Multi-item identification decays!

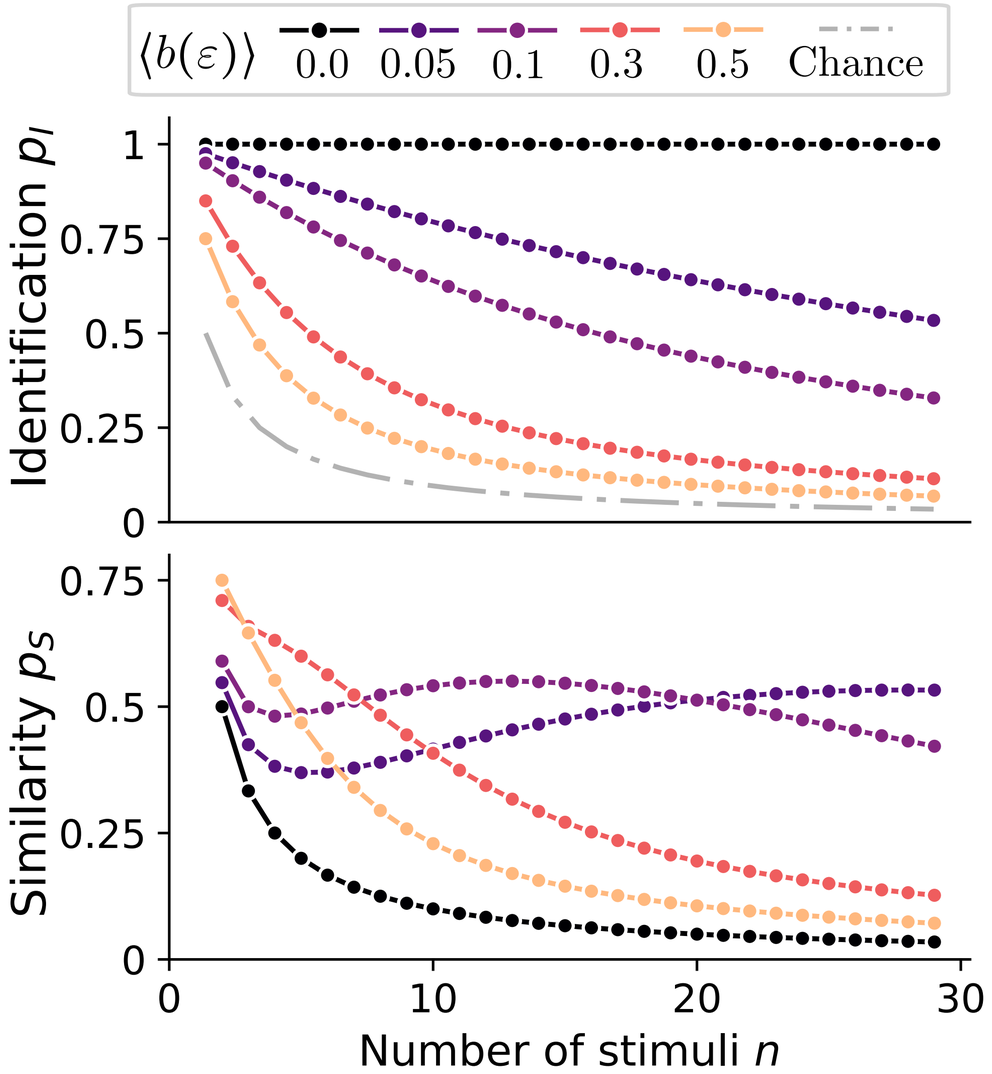
Bound by semanticity: universal laws governing the generalization-identification tradeoff

Marco Nurisso[1,2], Jesseba Fernando[3,4], Raj Deshpande[5], Alan Perotti[2], Raja Marjieh[6], Steven M. Frankland[7], Richard L. Lewis[8], Taylor W. Webb[9], Declan Campbell[10], Francesco Vaccarino[1], Jonathan D. Cohen[6,10], Giovanni Petri[2,5,11]

Theorem 3 (n-item tests)**.** *Under the same assumptions of Theorem 1, for the constant noise-free ($\Delta = 0$) similarity function $g = g_{\varepsilon;0}$ we have that*

$$p_S^n(\varepsilon) = \mathbb{E}_{p\sim\nu}\left[\frac{1}{n} + \sum_{k=1}^{n-1} \frac{(1-b_p(\varepsilon))^{n-k} - (1-b_p(\varepsilon))^n}{k}\right], \tag{7}$$

$$p_I^n(\varepsilon) = \mathbb{E}_{p\sim\nu}\left[\frac{1-(1-b_p(\varepsilon))^n}{nb_p(\varepsilon)}\right]. \tag{8}$$

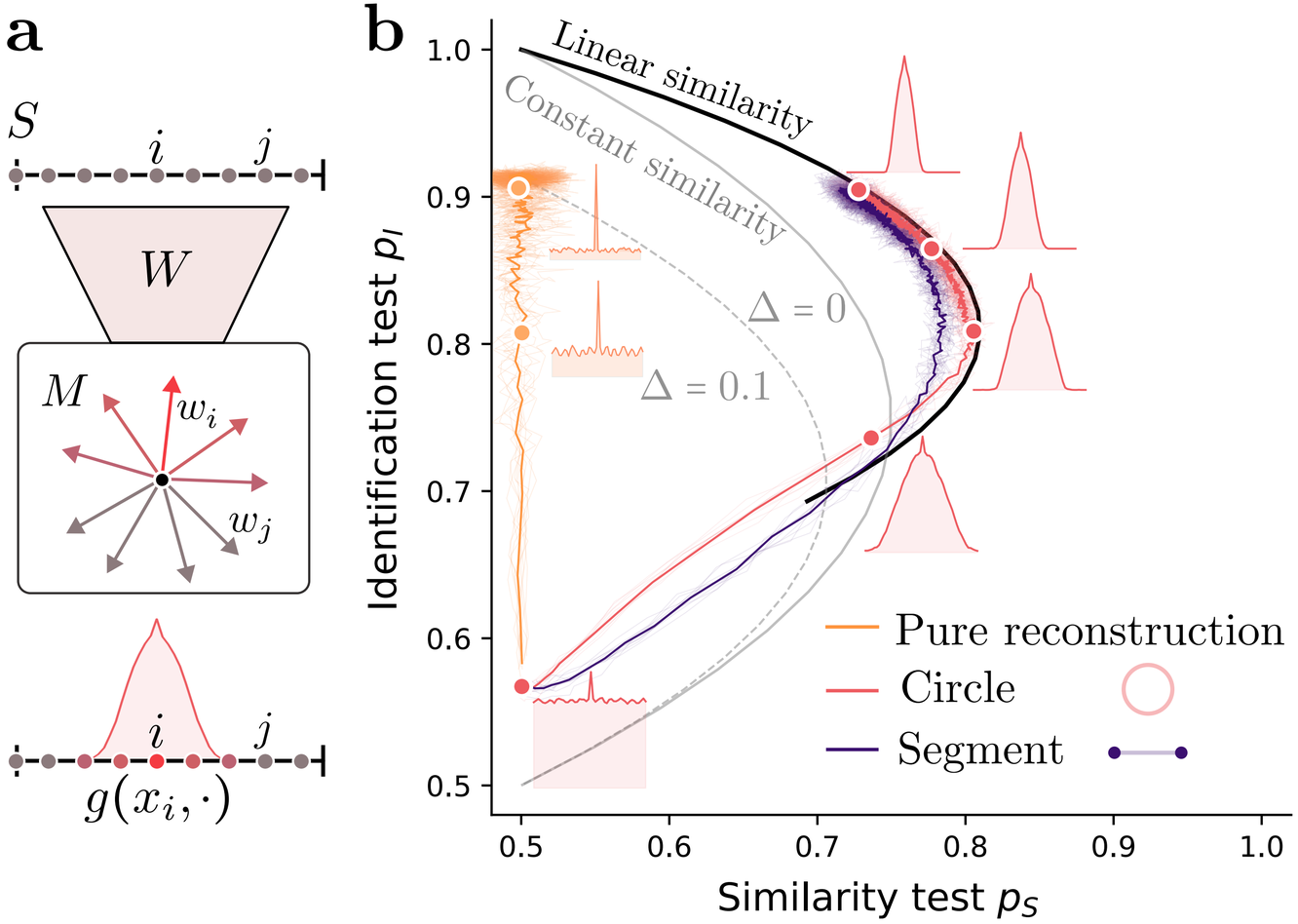


Is there an emergent resolution?

Generalization vs Processing

Bound by semanticity: universal laws governing the generalization-identification tradeoff

Marco Nurisso[1,2], Jesseba Fernando[3,4], Raj Deshpande[5], Alan Perotti[2], Raja Marjieh[6], Steven M. Frankland[7], Richard L. Lewis[8], Taylor W. Webb[9], Declan Campbell[10], Francesco Vaccarino[1], Jonathan D. Cohen[6,10], Giovanni Petri[2,5,11]



$$f(x) = \sigma(W^\top W x),$$

$$L_{\text{rec}} = \sum_{i=1}^{l} \left\| e_i - \sigma(W^\top W e_i) \right\|^2 = \sum_{i=1}^{l} \left\| e_i - \sigma(W^\top w_i) \right\|^2.$$

$$D_i = \frac{\sigma(w_i^\top w_k)}{\sigma(w_i^\top w_k) + \sigma(w_j^\top w_k)}, \; D_j = \frac{\sigma(w_j^\top w_k)}{\sigma(w_i^\top w_k) + \sigma(w_j^\top w_k)}.$$

$$L_{\text{sim}} = -\frac{1}{2} D_{\hat{i}}.$$

Generalization vs Processing

Bound by semanticity: universal laws governing the generalization-identification tradeoff

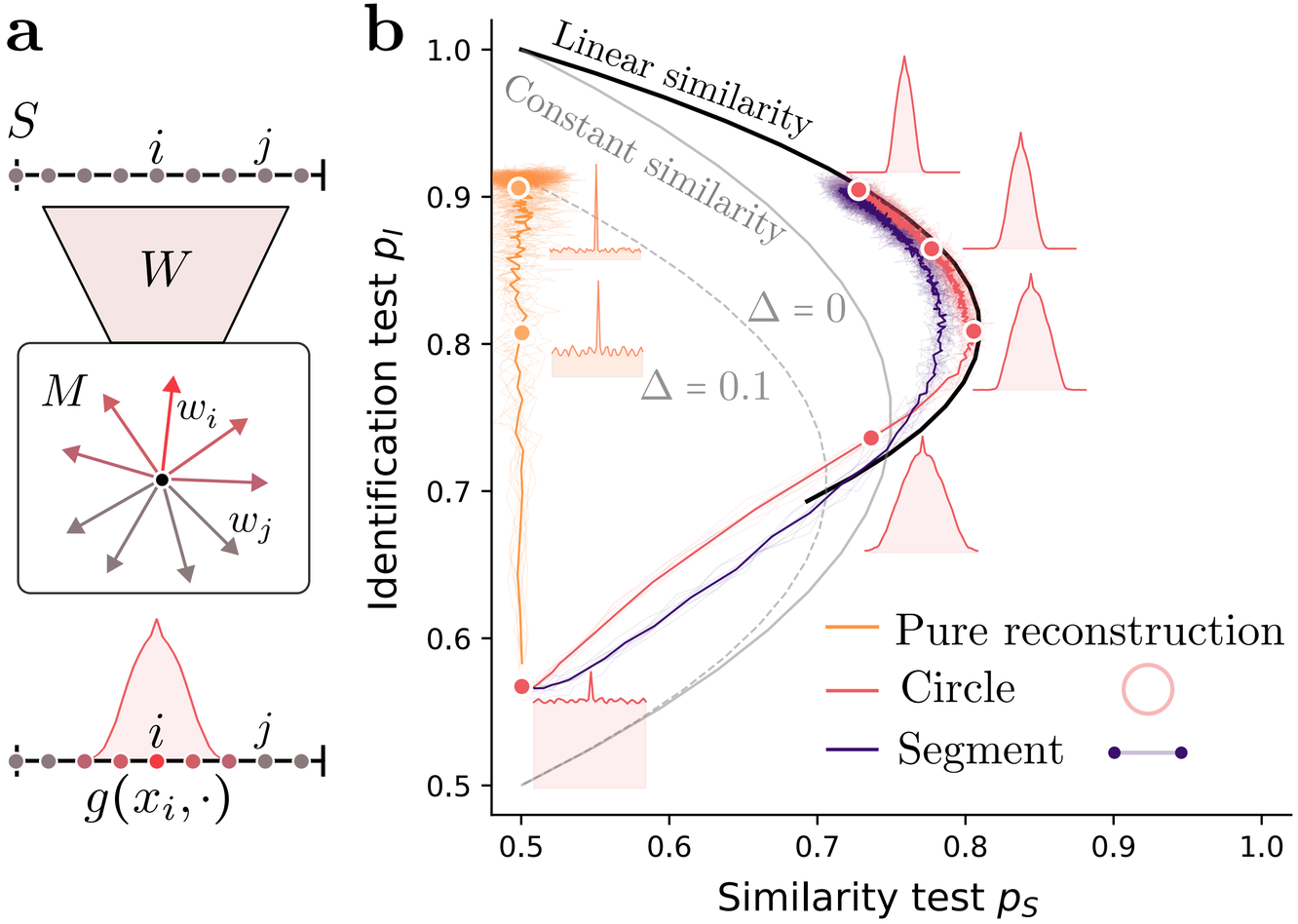
Marco Nurisso[1,2], Jesseba Fernando[3,4], Raj Deshpande[5], Alan Perotti[2], Raja Marjieh[6], Steven M. Frankland[7], Richard L. Lewis[8], Taylor W. Webb[9], Declan Campbell[10], Francesco Vaccarino[1], Jonathan D. Cohen[6,10], Giovanni Petri[2,5,11]

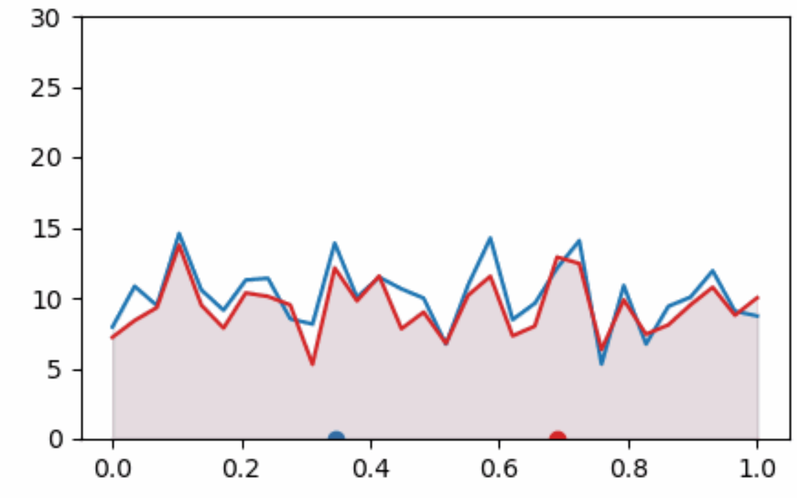
$$f(x) = \sigma(W^\top W x),$$

$$L_{\text{rec}} = \sum_{i=1}^{l} \left\| e_i - \sigma(W^\top W e_i) \right\|^2 = \sum_{i=1}^{l} \left\| e_i - \sigma(W^\top w_i) \right\|^2.$$

$$D_i = \frac{\sigma(w_i^\top w_k)}{\sigma(w_i^\top w_k) + \sigma(w_j^\top w_k)}, \; D_j = \frac{\sigma(w_j^\top w_k)}{\sigma(w_i^\top w_k) + \sigma(w_j^\top w_k)}.$$

$$L_{\text{sim}} = -\frac{1}{2} D_{\hat{\imath}}.$$

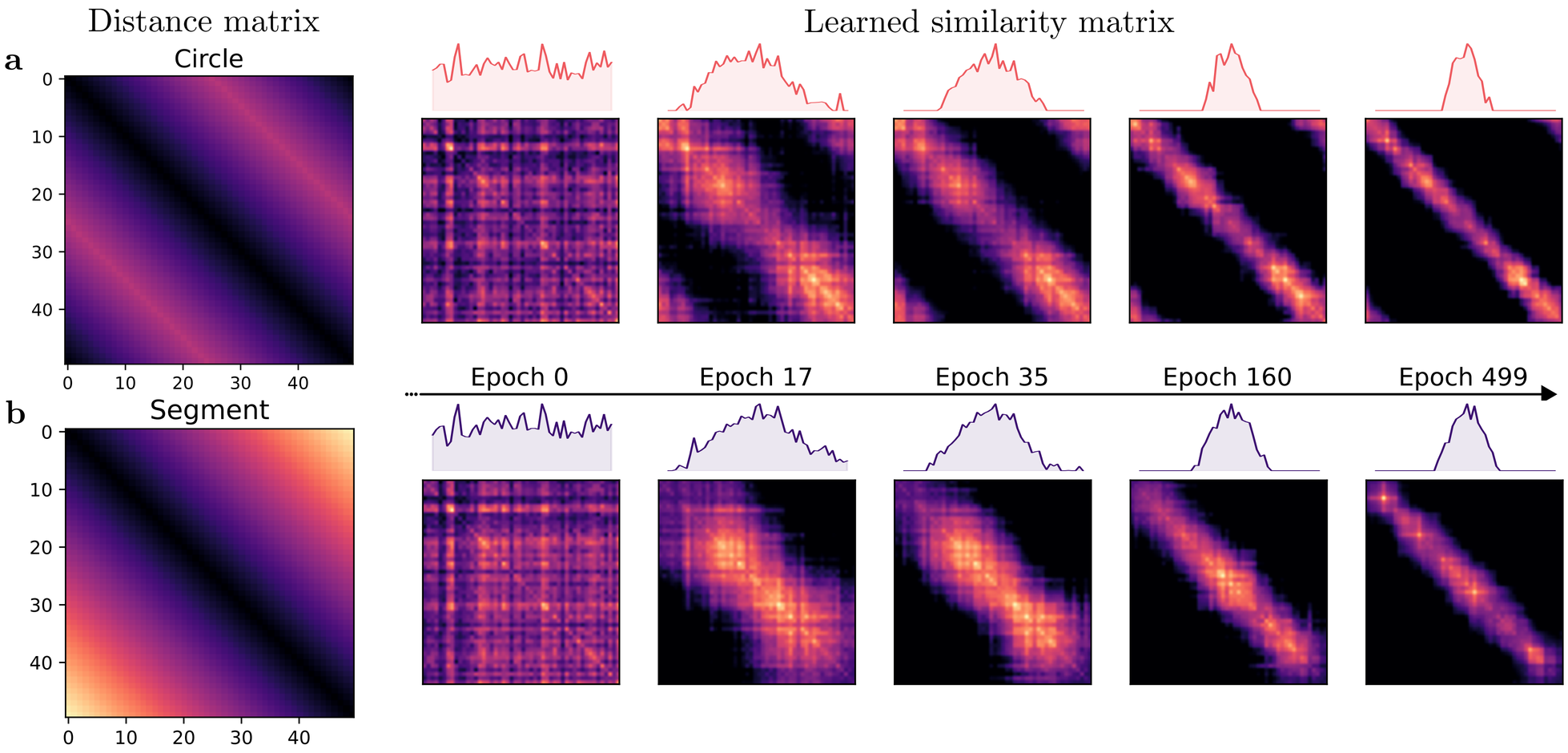




Generalization vs Processing

Bound by semanticity: universal laws governing the generalization-identification tradeoff

Marco Nurisso[1,2], Jesseba Fernando[3,4], Raj Deshpande[5], Alan Perotti[2], Raja Marjieh[6], Steven M. Frankland[7], Richard L. Lewis[8], Taylor W. Webb[9], Declan Campbell[10], Francesco Vaccarino[1], Jonathan D. Cohen[6,10], Giovanni Petri[2,5,11]



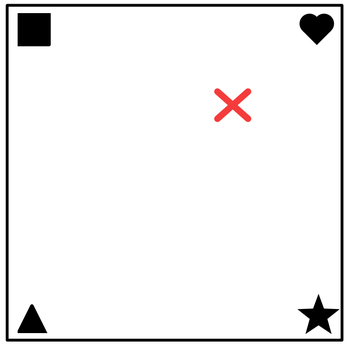
Link to percolation on random lattice models?

Generalization vs Processing

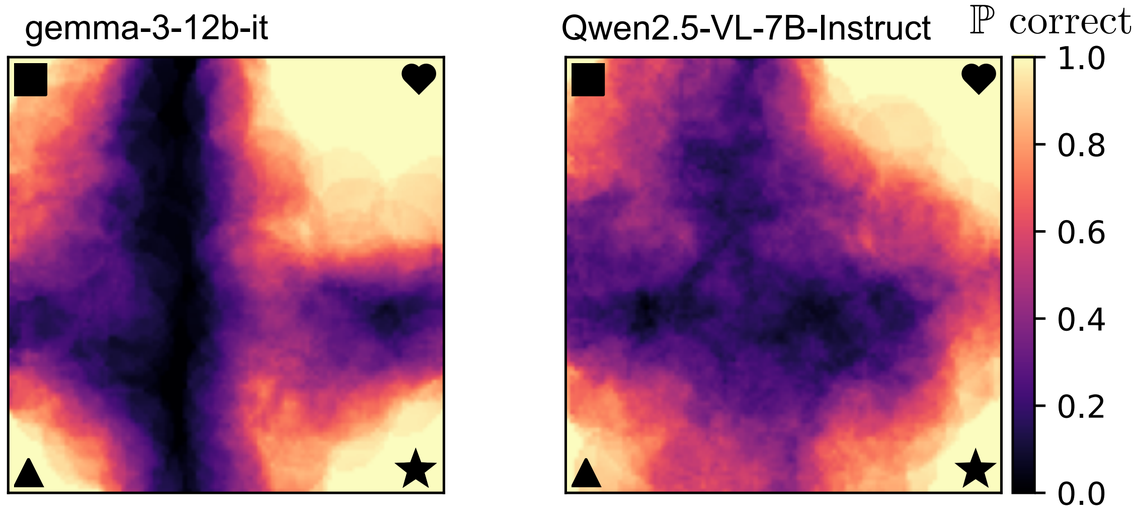
Bound by semanticity: universal laws governing the generalization-identification tradeoff

Marco Nurisso[1,2], Jesseba Fernando[3,4], Raj Deshpande[5], Alan Perotti[2], Raja Marjieh[6], Steven M. Frankland[7], Richard L. Lewis[8], Taylor W. Webb[9], Declan Campbell[10], Francesco Vaccarino[1], Jonathan D. Cohen[6,10], Giovanni Petri[2,5,11]

VLM spatial similarity task



Which **black shape** is closest to the red X?



Generalization vs Processing

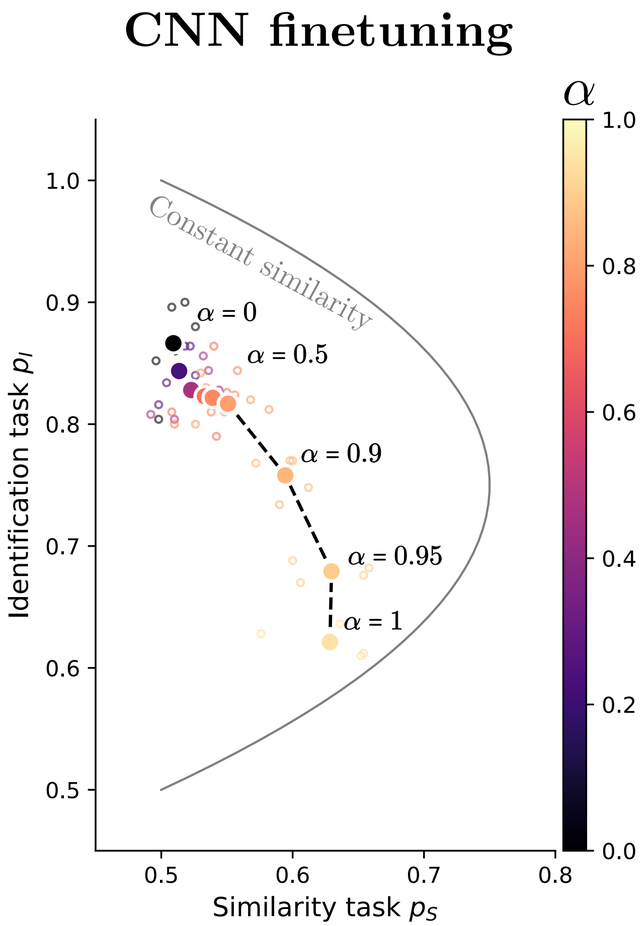
Bound by semanticity: universal laws governing the generalization-identification tradeoff

Marco Nurisso[1,2], Jesseba Fernando[3,4], Raj Deshpande[5], Alan Perotti[2], Raja Marjieh[6], Steven M. Frankland[7], Richard L. Lewis[8], Taylor W. Webb[9], Declan Campbell[10], Francesco Vaccarino[1], Jonathan D. Cohen[6,10], Giovanni Petri[2,5,11]

Generalization vs Processing

Bound by semanticity: universal laws governing the generalization-identification tradeoff

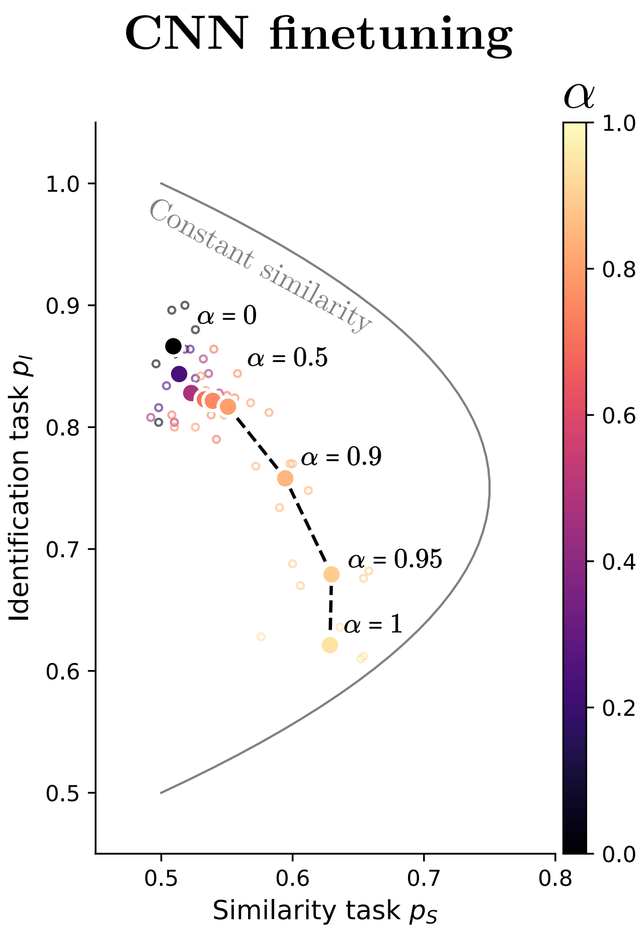
Marco Nurisso[1,2], Jesseba Fernando[3,4], Raj Deshpande[5], Alan Perotti[2], Raja Marjieh[6], Steven M. Frankland[7], Richard L. Lewis[8], Taylor W. Webb[9], Declan Campbell[10], Francesco Vaccarino[1], Jonathan D. Cohen[6,10], Giovanni Petri[2,5,11]

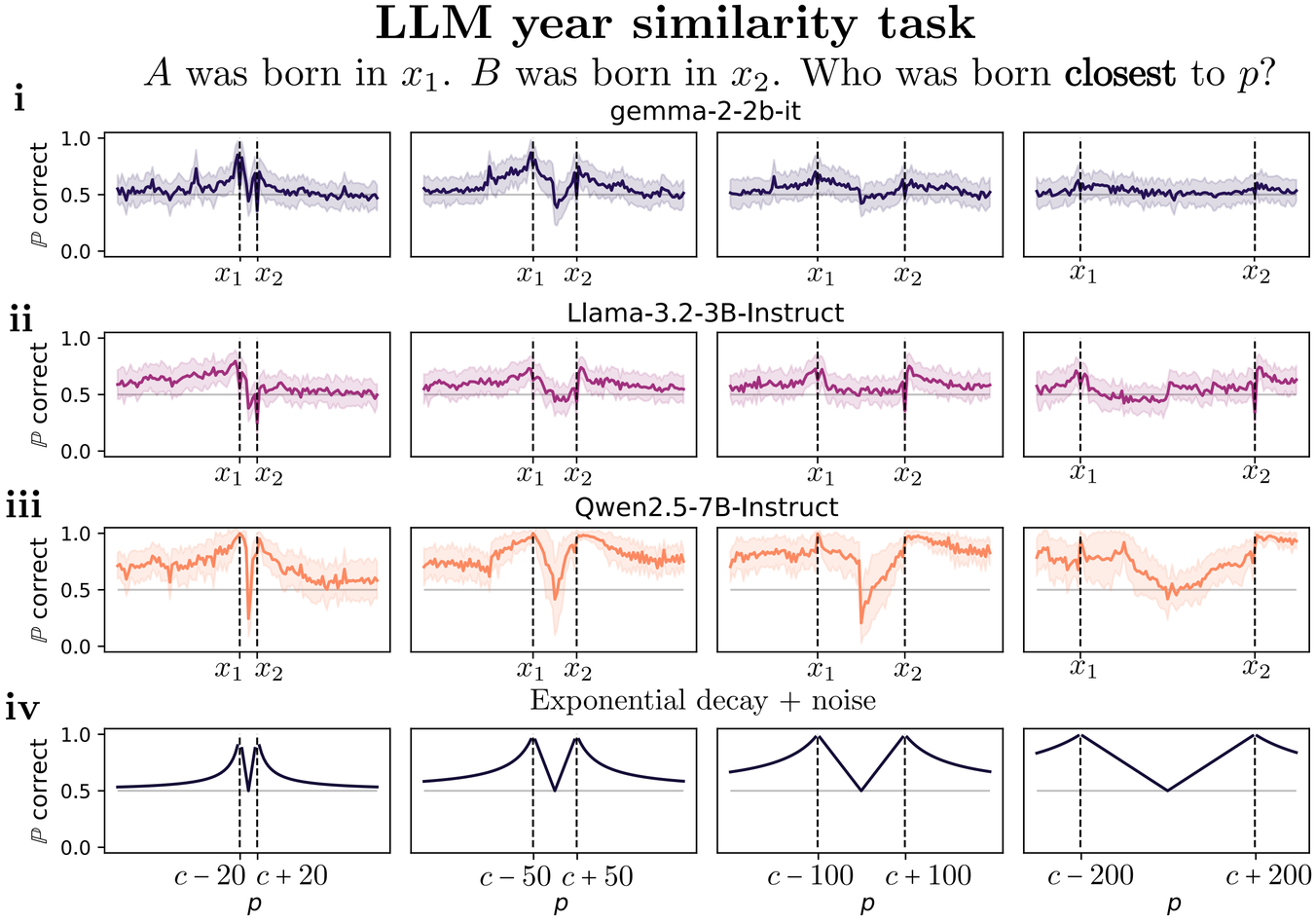


Generalization vs Processing

Bound by semanticity: universal laws governing the generalization-identification tradeoff

Marco Nurisso[1,2], Jesseba Fernando[3,4], Raj Deshpande[5], Alan Perotti[2], Raja Marjieh[6], Steven M. Frankland[7], Richard L. Lewis[8], Taylor W. Webb[9], Declan Campbell[10], Francesco Vaccarino[1], Jonathan D. Cohen[6,10], Giovanni Petri[2,5,11]

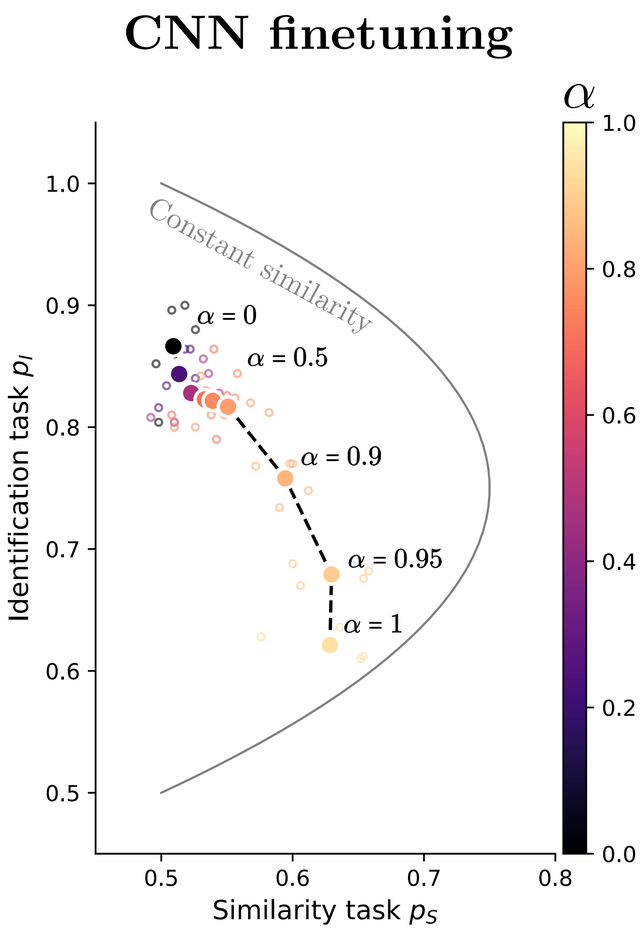


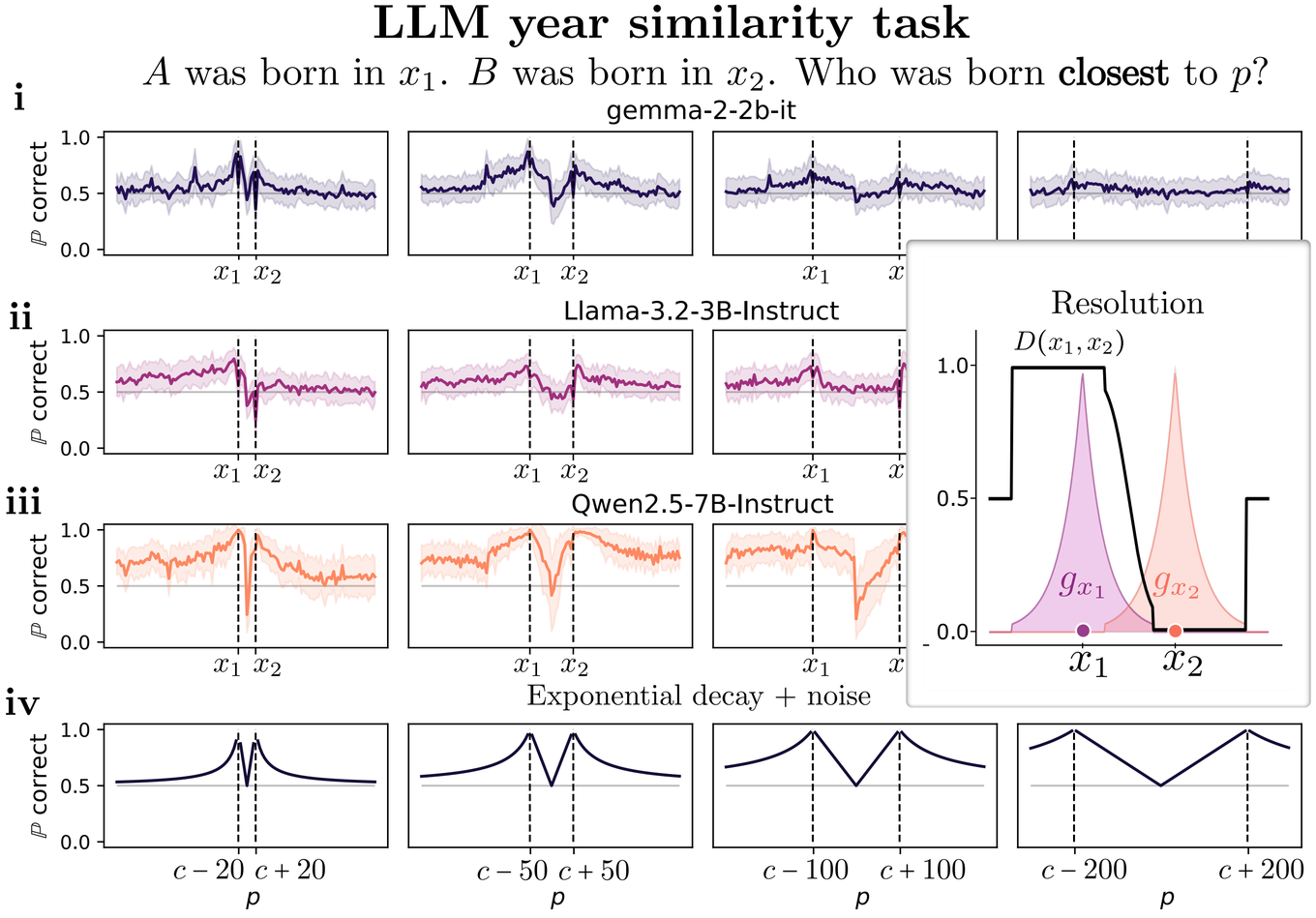


Generalization vs Processing

Bound by semanticity: universal laws governing the generalization-identification tradeoff

Marco Nurisso[1,2], Jesseba Fernando[3,4], Raj Deshpande[5], Alan Perotti[2], Raja Marjieh[6], Steven M. Frankland[7], Richard L. Lewis[8], Taylor W. Webb[9], Declan Campbell[10], Francesco Vaccarino[1], Jonathan D. Cohen[6,10], Giovanni Petri[2,5,11]





Take-home

Take-home

Results:

- Any resolution limit —> tradeoff generalization vs processing
- Resolution emerges independently!

Take-home

Results:

- Any resolution limit —> tradeoff generalization vs processing
- Resolution emerges independently!

Next:

Take-home

Results:

- Any resolution limit —> tradeoff generalization vs processing
- Resolution emerges independently!

Next:

- We can steer representations to reduce/increase performances (Savietto, ICML 2026)

Take-home

Results:

- Any resolution limit —> tradeoff generalization vs processing
- Resolution emerges independently!

Next:

- We can steer representations to reduce/increase performances (Savietto, ICML 2026)
- How does compositionality affects the tradeoff?

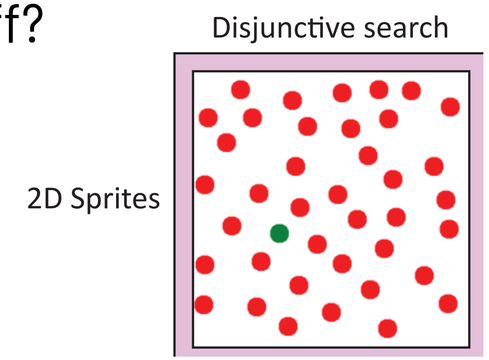
Take-home

Results:

- Any resolution limit —> tradeoff generalization vs processing
- Resolution emerges independently!

Next:

- We can steer representations to reduce/increase performances (Savietto, ICML 2026)
- How does compositionality affects the tradeoff?



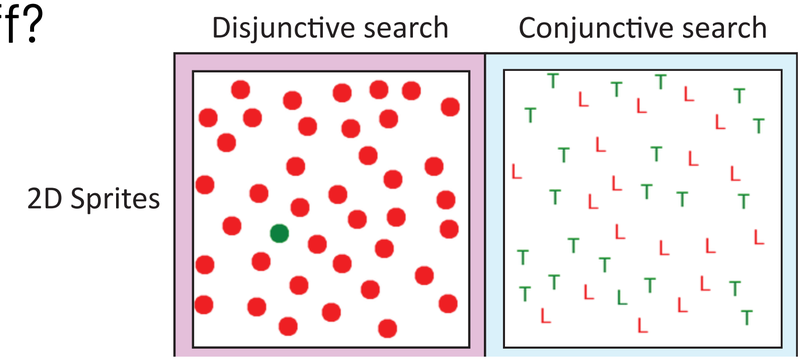
Take-home

Results:

- Any resolution limit —> tradeoff generalization vs processing
- Resolution emerges independently!

Next:

- We can steer representations to reduce/increase performances (Savietto, ICML 2026)
- How does compositionality affects the tradeoff?



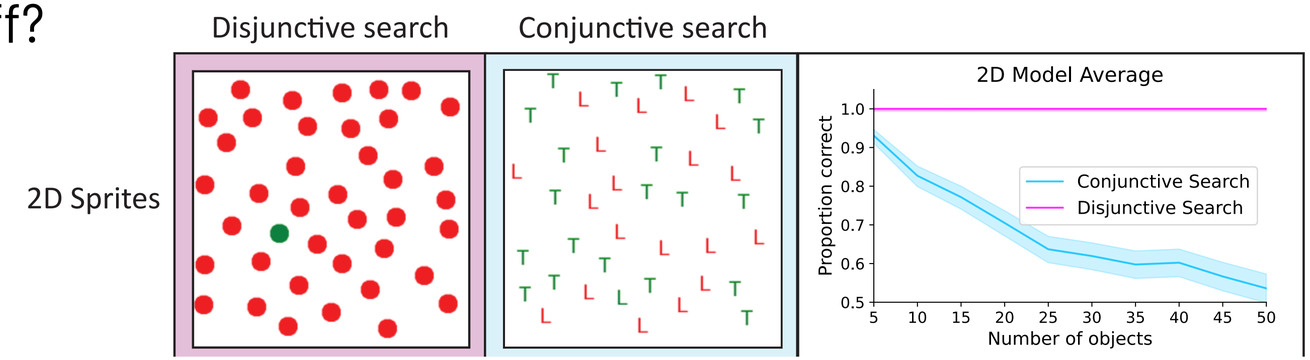
Take-home

Results:

- Any resolution limit —> tradeoff generalization vs processing
- Resolution emerges independently!

Next:

- We can steer representations to reduce/increase performances (Savietto, ICML 2026)
- How does compositionality affects the tradeoff?



Take-home

Results:

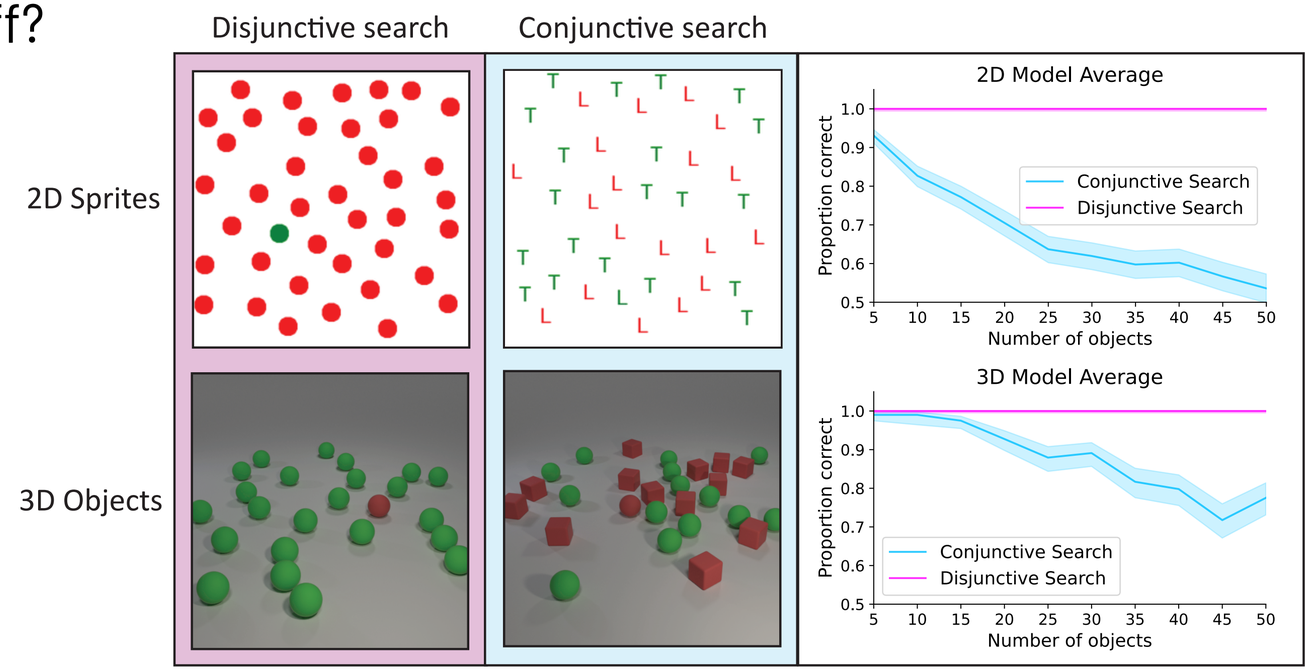
- Any resolution limit —> tradeoff generalization vs processing
- Resolution emerges independently!

Next:

- We can steer representations to reduce/increase performances (Savietto, ICML 2026)
- How does compositionality affects the tradeoff?

Understanding the Limits of Vision Language Models Through the Lens of the Binding Problem

Declan Campbell[1], Sunayana Rane[2], Tyler Giallanza[2], Nicolò De Sabbata[3], Kia Ghods[1], Amogh Joshi[1], Alexander Ku[4], Steven M. Frankland[5], Thomas L. Griffiths[2,3]*, Jonathan D. Cohen[1,2]*, and Taylor W. Webb[6]*



Take-home

Results:

- Any resolution limit —> tradeoff generalization vs processing
- Resolution emerges independently!

Next:

- We can steer representations to reduce/increase performances (Savietto, ICML 2026)
- How does compositionality affects the tradeoff?

Take-home

Results:

- Any resolution limit —> tradeoff generalization vs processing
- Resolution emerges independently!

Next:

- We can steer representations to reduce/increase performances (Savietto, ICML 2026)
- How does compositionality affects the tradeoff?

$$\textbf{TPR formulation of Miller-Shepard's law}$$

$$\Phi = \sum_{i=1}^{n} f^{(i)} \otimes r^{(i)} = FR^{\top},$$

Take-home

Results:

- Any resolution limit —> tradeoff generalization vs processing
- Resolution emerges independently!

Next:

- We can steer representations to reduce/increase performances (Savietto, ICML 2026)
- How does compositionality affects the tradeoff?

TPR formulation of Miller-Shepard's law

$$\Phi = \sum_{i=1}^{n} f^{(i)} \otimes r^{(i)} = FR^{\top},$$

filler
(red, square, etc..)

Take-home

Results:

- Any resolution limit —> tradeoff generalization vs processing
- Resolution emerges independently!

Next:

- We can steer representations to reduce/increase performances (Savietto, ICML 2026)
- How does compositionality affects the tradeoff?

TPR formulation of Miller-Shepard's law

$$\Phi = \sum_{i=1}^{n} f^{(i)} \otimes r^{(i)} = FR^{\top},$$

filler (red, square, etc..)

Role (abstract object slot)

Take-home

Results:

- Any resolution limit —> tradeoff generalization vs processing
- Resolution emerges independently!

Next:

- We can steer representations to reduce/increase performances (Savietto, ICML 2026)
- How does compositionality affects the tradeoff?

TPR formulation of Miller-Shepard's law

$$\Phi = \sum_{i=1}^{n} f^{(i)} \otimes r^{(i)} = FR^\top,$$

filler (red, square, etc..)

Role (abstract object slot)

$$\begin{aligned}\hat{f} &= \Phi r = FR^\top r^{(k)} \\ &= f^{(k)} + \sum_{j \neq k} f^{(j)} (r^{(j)})^\top r^{(k)} \\ &= f^{(k)} + \sum_{j \neq k} G^R_{jk} f^{(j)},\end{aligned}$$

Take-home

Results:

- Any resolution limit —> tradeoff generalization vs processing
- Resolution emerges independently!

Next:

- We can steer representations to reduce/increase performances (Savietto, ICML 2026)
- How does compositionality affects the tradeoff?

TPR formulation of Miller-Shepard's law

$$\Phi = \sum_{i=1}^{n} f^{(i)} \otimes r^{(i)} = FR^{\top},$$

filler (red, square, etc..)

Role (abstract object slot)

$$\begin{aligned}\hat{f} &= \Phi r = FR^{\top} r^{(k)} \\ &= f^{(k)} + \sum_{j \neq k} f^{(j)} \left(r^{(j)}\right)^{\top} r^{(k)} \\ &= f^{(k)} + \sum_{j \neq k} G^{R}_{jk} f^{(j)},\end{aligned}$$

feature

Take-home

Results:

- Any resolution limit —> tradeoff generalization vs processing
- Resolution emerges independently!

Next:

- We can steer representations to reduce/increase performances (Savietto, ICML 2026)
- How does compositionality affects the tradeoff?

TPR formulation of Miller-Shepard's law

$$\Phi = \sum_{i=1}^{n} f^{(i)} \otimes r^{(i)} = FR^\top,$$

filler (red, square, etc..)

Role (abstract object slot)

$$\begin{aligned}\hat{f} &= \Phi r = FR^\top r^{(k)} \\ &= f^{(k)} + \sum_{j \neq k} f^{(j)} (r^{(j)})^\top r^{(k)} \\ &= f^{(k)} + \sum_{j \neq k} G^R_{jk} f^{(j)},\end{aligned}$$

feature

interference

Take-home

Results:

- Any resolution limit —> tradeoff generalization vs processing
- Resolution emerges independently!

Next:

- We can steer representations to reduce/increase performances (Savietto, ICML 2026)
- How does compositionality affects the tradeoff?

Take-home

Results:

- Any resolution limit —> tradeoff generalization vs processing
- Resolution emerges independently!

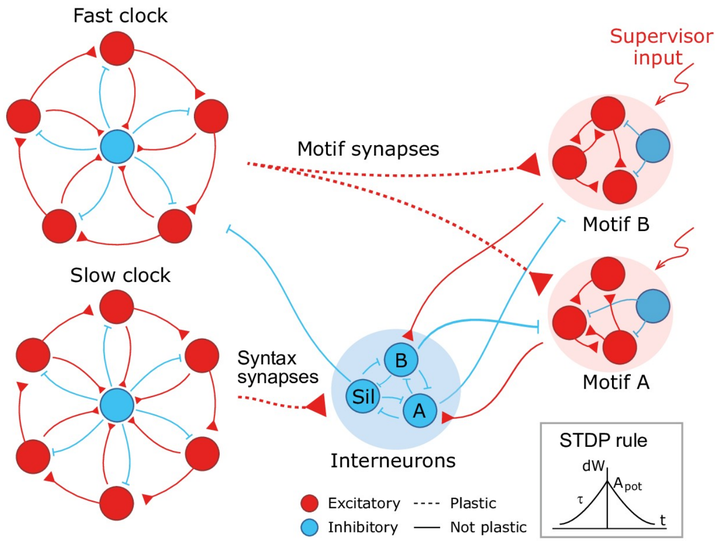
Next:

- We can steer representations to reduce/increase performances (Savietto, ICML 2026)
- How does compositionality affects the tradeoff?

RESEARCH ARTICLE

Learning compositional sequences with multiple time scales through a hierarchical network of spiking neurons

Amadeus Maes[1], Mauricio Barahona[2], Claudia Clopath[1]*



Take-home

Results:

- Any resolution limit —> tradeoff generalization vs processing
- Resolution emerges independently!

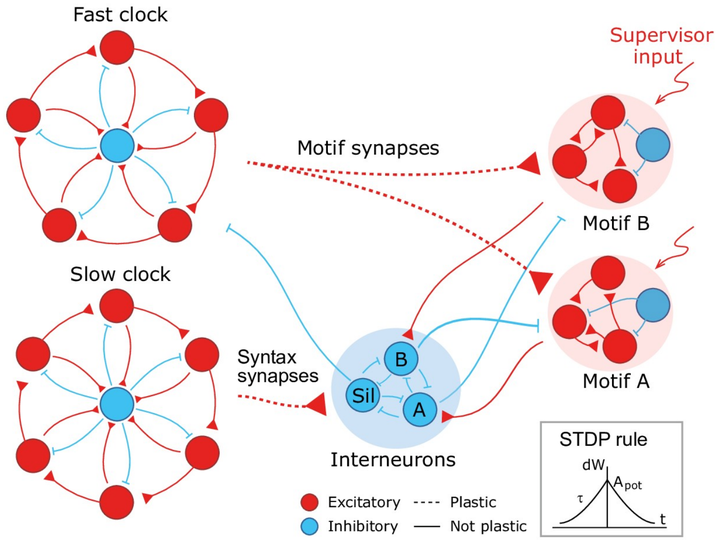
Next:

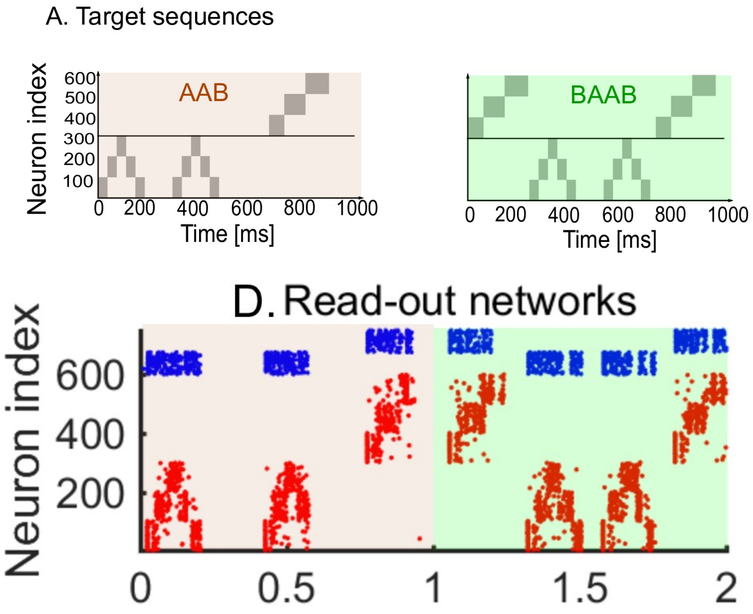
- We can steer representations to reduce/increase performances (Savietto, ICML 2026)
- How does compositionality affects the tradeoff?

RESEARCH ARTICLE

Learning compositional sequences with multiple time scales through a hierarchical network of spiking neurons

Amadeus Maes[1], Mauricio Barahona[2], Claudia Clopath[1]*





Take-home

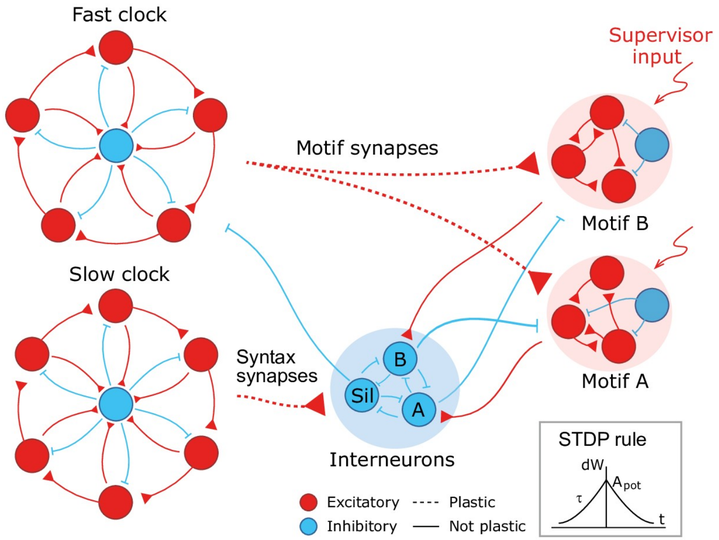
Results:

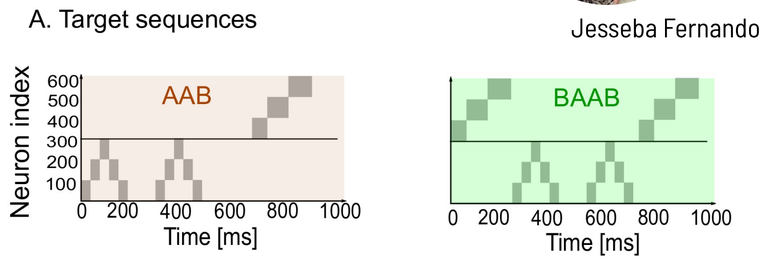
- Any resolution limit —> tradeoff generalization vs processing
- Resolution emerges independently!

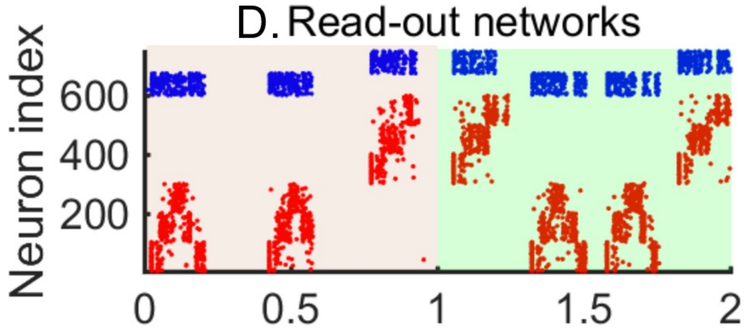
Next:

- We can steer representations to reduce/increase performances (Savietto, ICML 2026)
- How does compositionality affects the tradeoff?

Jesseba Fernando







RESEARCH ARTICLE

Learning compositional sequences with multiple time scales through a hierarchical network of spiking neurons

Amadeus Maes[1], Mauricio Barahona[2], Claudia Clopath[1]*

Take-home

Results:

- Any resolution limit —> tradeoff generalization vs processing
- Resolution emerges independently!

Next:

- We can steer representations to reduce/increase performances (Savietto, ICML 2026)
- How does compositionality affects the tradeoff?
- Does this depend on scale? (i.e. What if we renormalize representations?)

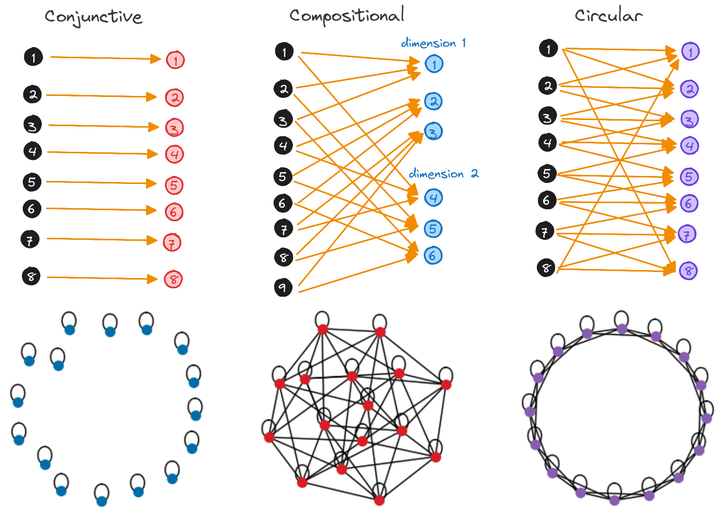
Take-home

Results:

- Any resolution limit —> tradeoff generalization vs processing
- Resolution emerges independently!

Next:

- We can steer representations to reduce/increase performances (Savietto, ICML 2026)
- How does compositionality affects the tradeoff?
- Does this depend on scale? (i.e. What if we renormalize representations?)



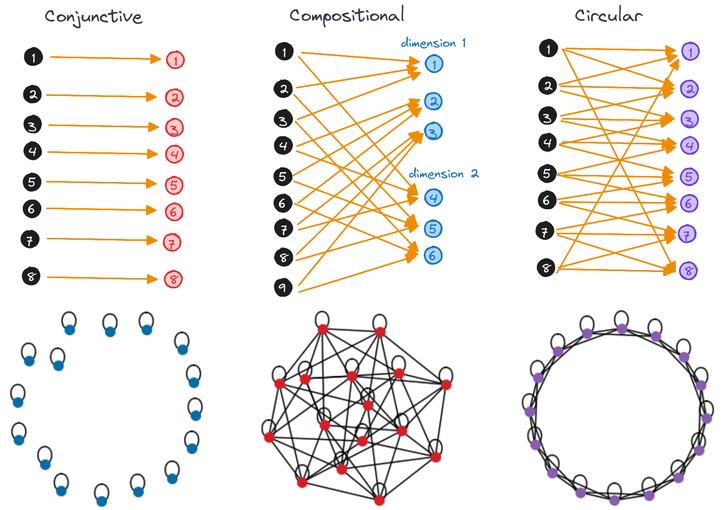
Take-home

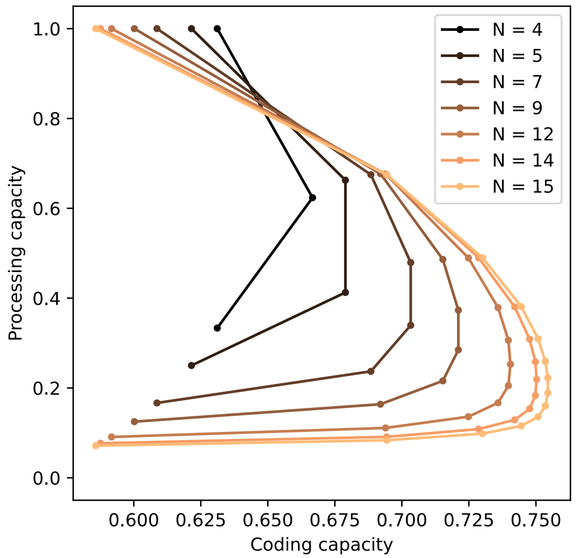
Results:

- Any resolution limit —> tradeoff generalization vs processing
- Resolution emerges independently!

Next:

- We can steer representations to reduce/increase performances (Savietto, ICML 2026)
- How does compositionality affects the tradeoff?
- Does this depend on scale? (i.e. What if we renormalize representations?)





Take-home

Results:

- Any resolution limit —> tradeoff generalization vs processing
- Resolution emerges independently!

Next:

- We can steer representations to reduce/increase performances (Savietto, ICML 2026)
- How does compositionality affects the tradeoff?
- Does this depend on scale? (i.e. What if we renormalize representations?)

Take-home

Results:

- Any resolution limit —> tradeoff generalization vs processing
- Resolution emerges independently!

Next:

- We can steer representations to reduce/increase performances (Savietto, ICML 2026)
- How does compositionality affects the tradeoff?
- Does this depend on scale? (i.e. What if we renormalize representations?)

Hot takes:

Take-home

Results:

- Any resolution limit —> tradeoff generalization vs processing
- Resolution emerges independently!

Next:

- We can steer representations to reduce/increase performances (Savietto, ICML 2026)
- How does compositionality affects the tradeoff?
- Does this depend on scale? (i.e. What if we renormalize representations?)

Hot takes:

- All living/cog systems (immune system, cells, your football team, etc..) face this tradeoff! What does it imply?

Take-home

Results:

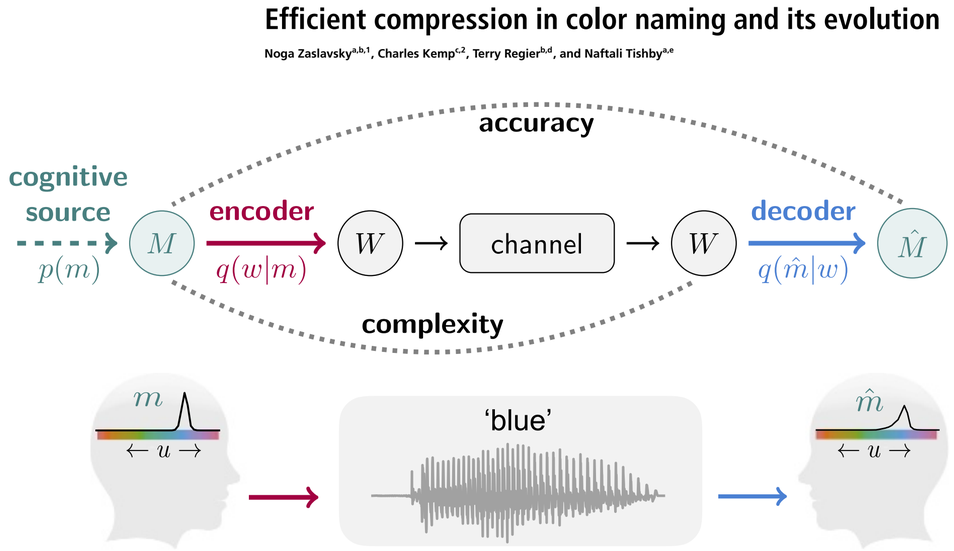
- Any resolution limit —> tradeoff generalization vs processing
- Resolution emerges independently!

Next:

- We can steer representations to reduce/increase performances (Savietto, ICML 2026)
- How does compositionality affects the tradeoff?
- Does this depend on scale? (i.e. What if we renormalize representations?)

Hot takes:

- All living/cog systems (immune system, cells, your football team, etc..) face this tradeoff! What does it imply?



Take-home

Results:

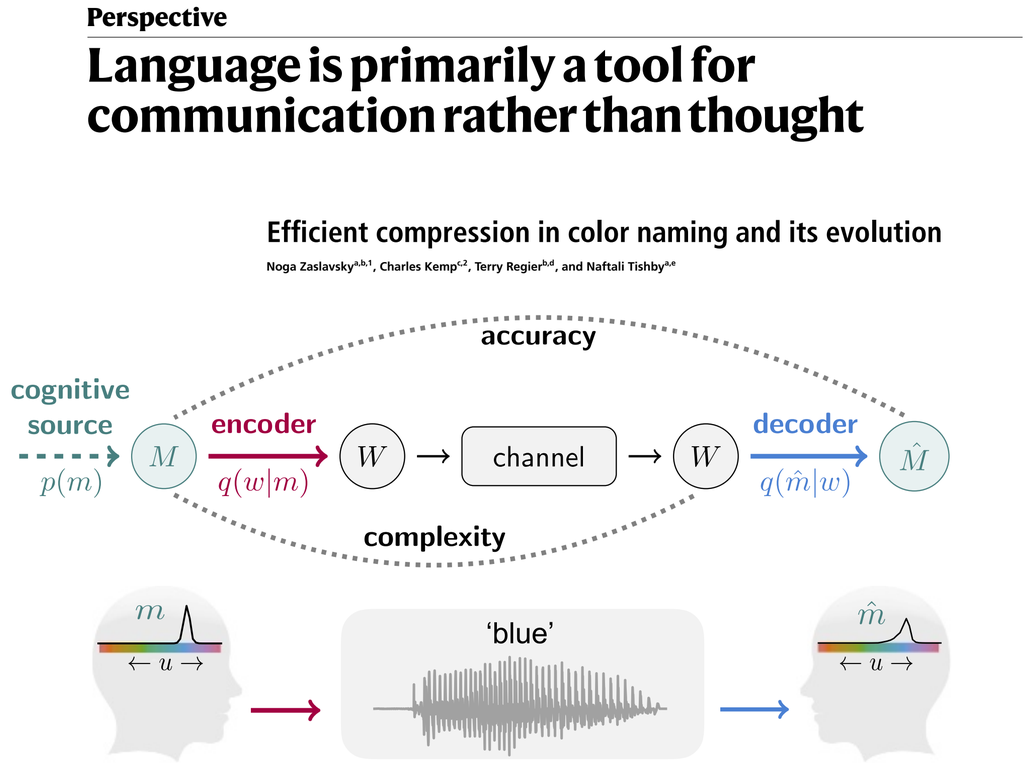
- Any resolution limit —> tradeoff generalization vs processing
- Resolution emerges independently!

Next:

- We can steer representations to reduce/increase performances (Savietto, ICML 2026)
- How does compositionality affects the tradeoff?
- Does this depend on scale? (i.e. What if we renormalize representations?)

Hot takes:

- All living/cog systems (immune system, cells, your football team, etc..) face this tradeoff! What does it imply?



Take-home

Results:

- Any resolution limit —> tradeoff generalization vs processing
- Resolution emerges independently!

Next:

- We can steer representations to reduce/increase performances (Savietto, ICML 2026)
- How does compositionality affects the tradeoff?
- Does this depend on scale? (i.e. What if we renormalize representations?)

Hot takes:

- All living/cog systems (immune system, cells, your football team, etc..) face this tradeoff! What does it imply?

Take-home

Results:

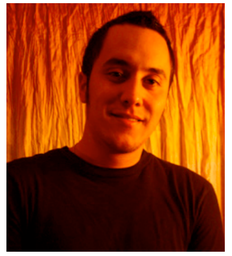
- Any resolution limit —> tradeoff generalization vs processing
- Resolution emerges independently!

Next:

- We can steer representations to reduce/increase performances (Savietto, ICML 2026)
- How does compositionality affects the tradeoff?
- Does this depend on scale? (i.e. What if we renormalize representations?)

Hot takes:

- All living/cog systems (immune system, cells, your football team, etc..) face this tradeoff! What does it imply?



Biswadip Dey

H. Kayan Oczimder

Nesreen Ahmed

Ted Willke

nature physics

ARTICLES

https://doi.org/10.1038/s41567-021-01170-x

Check for updates

Topological limits to the parallel processing capability of network architectures

eLife Home Magazine Community About

Neuroscience

An Information-Theoretic Approach to Reward Rate Optimization in the Tradeoff Between Controlled and Automatic Processing in Neural Network Architectures

Giovanni Petri, Sebastian Musslick, Jonathan D. Cohen

Sebastian Müsslick

Jonathan Cohen

Bound by semanticity: universal laws governing the generalization-identification tradeoff

Marco Nurisso

Raj Deshpande

Jesseba Fernando

Alan Perotti

Raja Marjieh, Steven M. Frankland,
Richard L. Lewis, Taylor W. Webb,
Declan Campbell, Francesco Vaccarino

The Geometry of Representational Failures in Vision Language Models

Take-home

Results:

- Any resolution limit —> tradeoff generalization vs processing
- Resolution emerges independently!

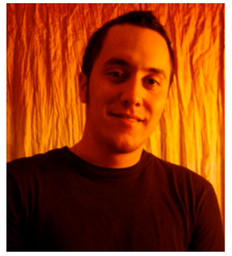
Next:

- We can steer representations to reduce/increase performances (Savietto, ICML 2026)
- How does compositionality affects the tradeoff?
- Does this depend on scale? (i.e. What if we renormalize representations?)

Hot takes:

- All living/cog systems (immune system, cells, your football team, etc..) face this tradeoff! What does it imply?

Thank you!



Biswadip Dey

H. Kayan Oczimder

Nesreen Ahmed

Ted Willke

nature physics

ARTICLES

https://doi.org/10.1038/s41567-021-01170-x

Check for updates

Topological limits to the parallel processing capability of network architectures

eLife Home Magazine Community About

Neuroscience

An Information-Theoretic Approach to Reward Rate Optimization in the Tradeoff Between Controlled and Automatic Processing in Neural Network Architectures

Giovanni Petri, Sebastian Musslick, Jonathan D. Cohen

Sebastian Müsslick

Jonathan Cohen

Bound by semanticity: universal laws governing the generalization-identification tradeoff

Marco Nurisso

Raj Deshpande

Jesseba Fernando

Alan Perotti

Raja Marjieh, Steven M. Frankland, Richard L. Lewis, Taylor W. Webb, Declan Campbell, Francesco Vaccarino

The Geometry of Representational Failures in Vision Language Models

Backup

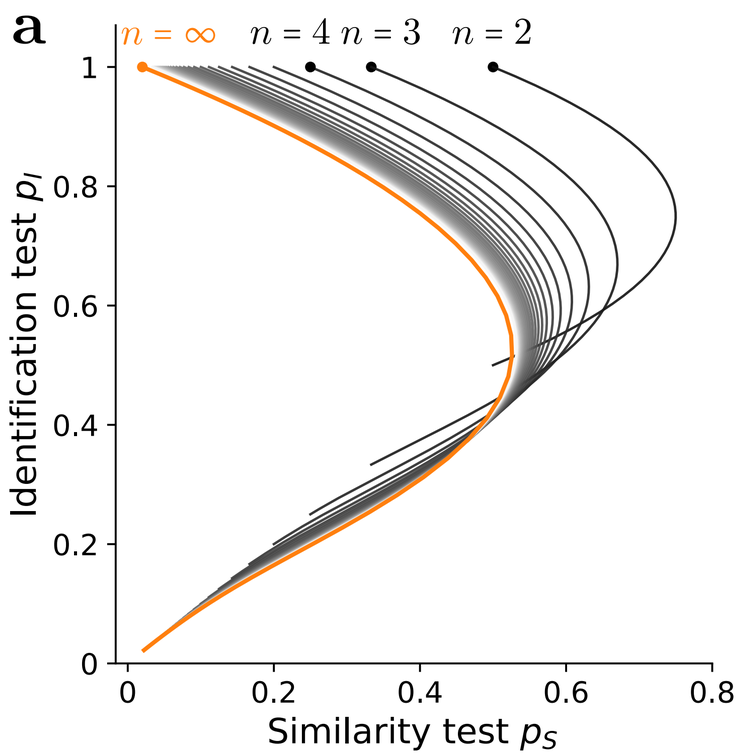
Generalization vs Processing

Multiple objects

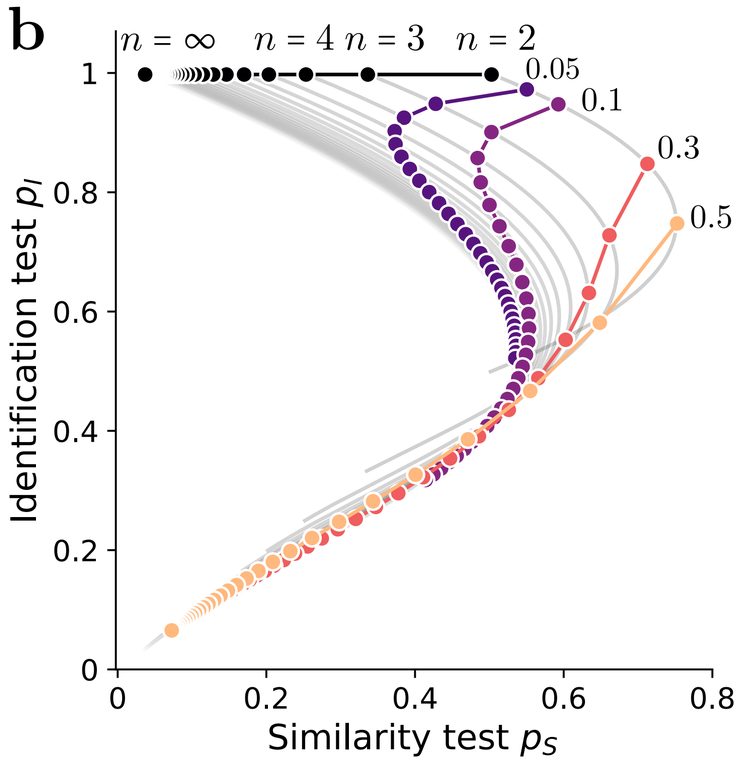
Bound by semanticity: universal laws governing the generalization-identification tradeoff

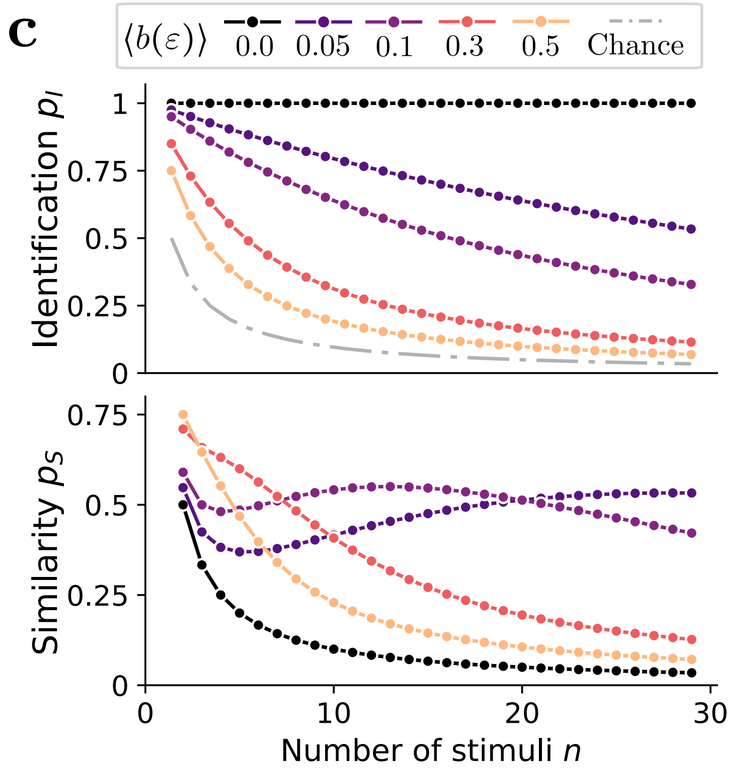
Marco Nurisso[1,2], Jesseba Fernando[3,4], Raj Deshpande[5], Alan Perotti[2], Raja Marjieh[6], Steven M. Frankland[7], Richard L. Lewis[8], Taylor W. Webb[9], Declan Campbell[10], Francesco Vaccarino[1], Jonathan D. Cohen[6,10], Giovanni Petri[2,5,11]

Multiple n - All resolutions



Multiple n - Fixed resolutions





Theorem 3 (n-item tests)**.** *Under the same assumptions of Theorem 1, for the constant noise-free ($\Delta = 0$) similarity function $g = g_{\varepsilon;0}$ we have that*

$$p_S^n(\varepsilon) = \mathbb{E}_{p\sim\nu}\left[\frac{1}{n} + \sum_{k=1}^{n-1} \frac{(1-b_p(\varepsilon))^{n-k} - (1-b_p(\varepsilon))^n}{k}\right], \tag{7}$$

$$p_I^n(\varepsilon) = \mathbb{E}_{p\sim\nu}\left[\frac{1-(1-b_p(\varepsilon))^n}{n b_p(\varepsilon)}\right]. \tag{8}$$

Binary code networks

Representational vs processing capacity

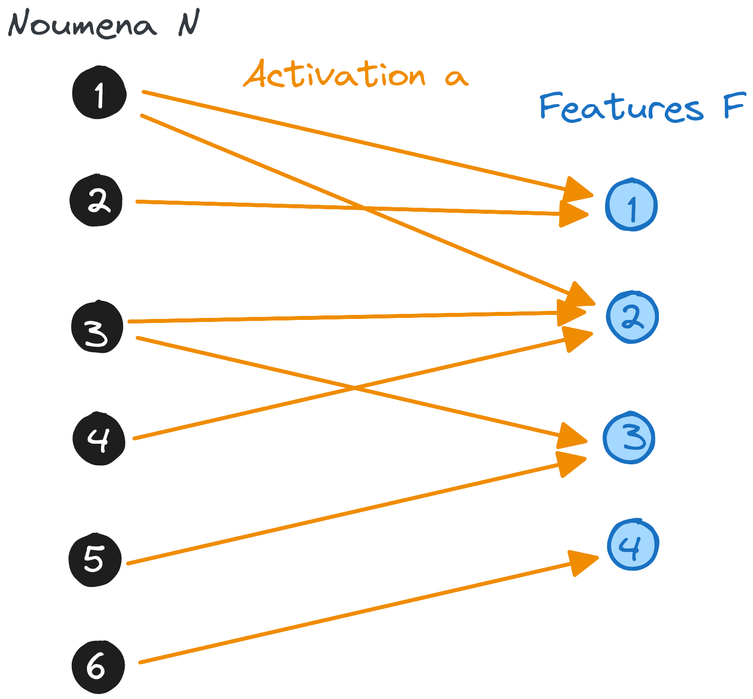
Nurisso, Deshpande, et al in prep

Representational vs processing capacity

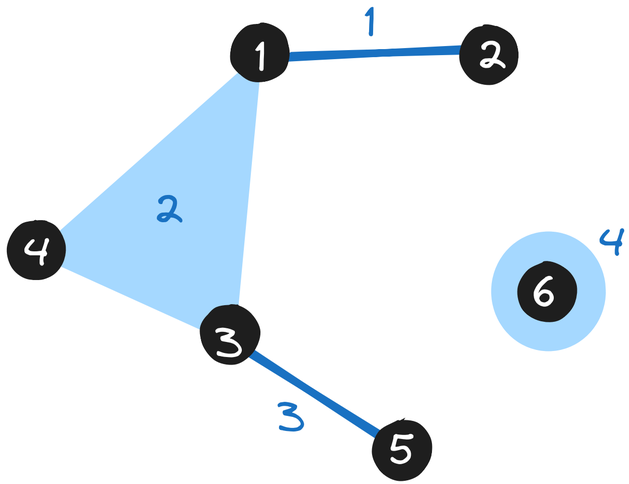
A code of N is a map

$$a : N \to 2^F$$

Bipartite graph representation



Hypergraph representation



Nurisso, Deshpande, et al in prep

Representational vs processing capacity

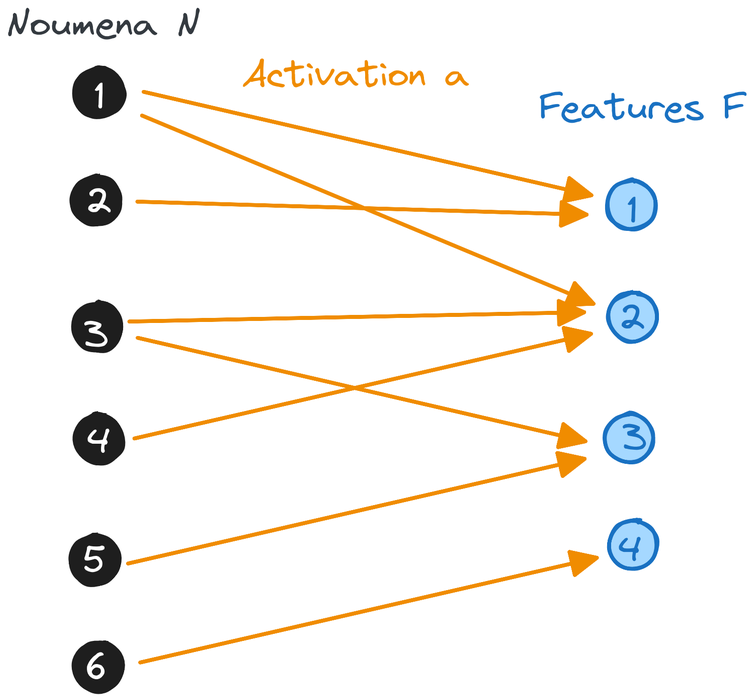
A code of N is a map

$$a : N \to 2^F$$

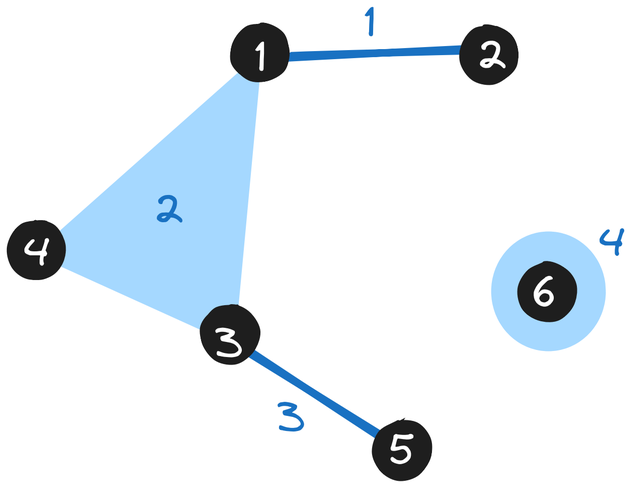
1-to-1 code (deterministically)

$a(x) \neq a(y)$ if $x \neq y$

Bipartite graph representation



Hypergraph representation



Nurisso, Deshpande, et al in prep

Representational vs processing capacity

A code of N is a map

$$a : N \to 2^F$$

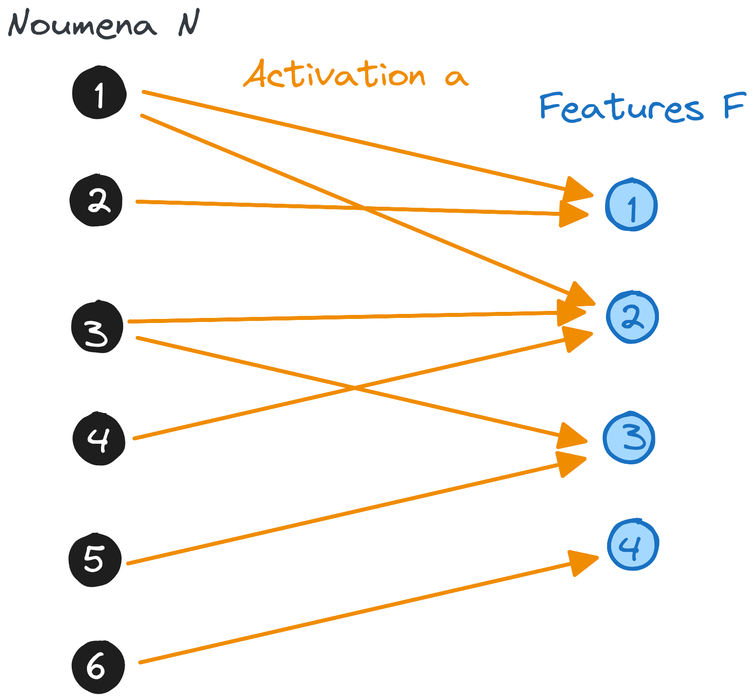
1-to-1 code (deterministically)

$$a(x) \neq a(y) \text{ if } x \neq y$$

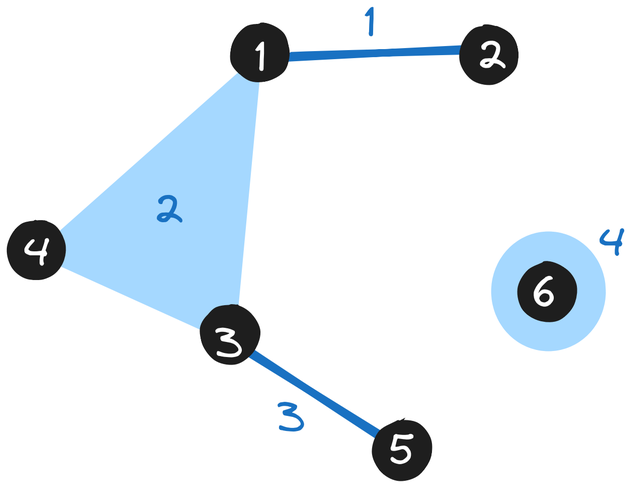
Dist $p(x)$ on N induced dist $\mathbb{P}(a(x))$

$$I(X; a(X)) = -\sum_{x \in N} p_x \log_2 \mathbb{P}(a(x)),$$

Bipartite graph representation



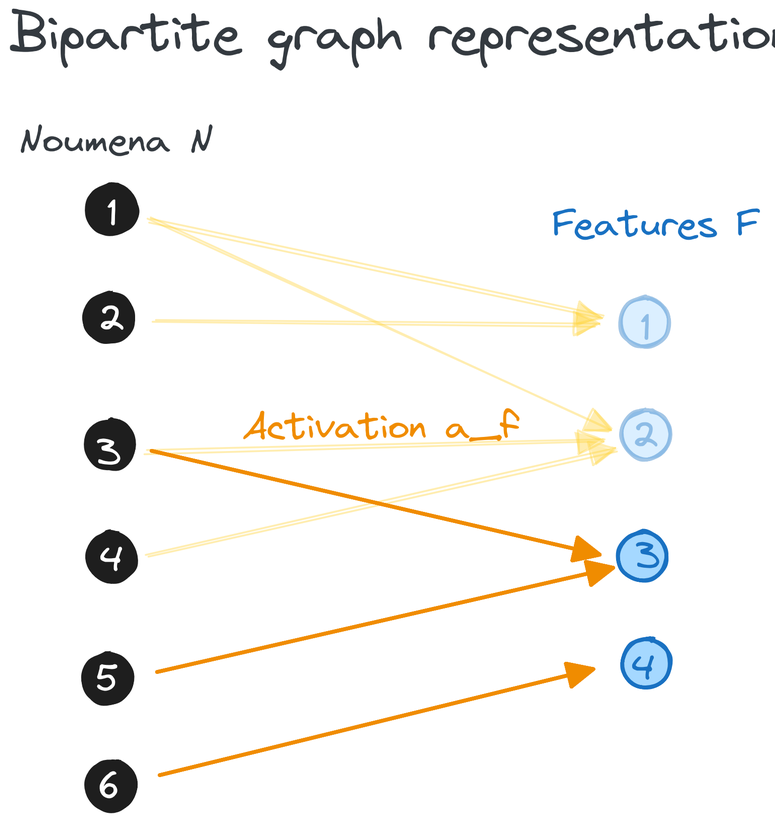
Hypergraph representation

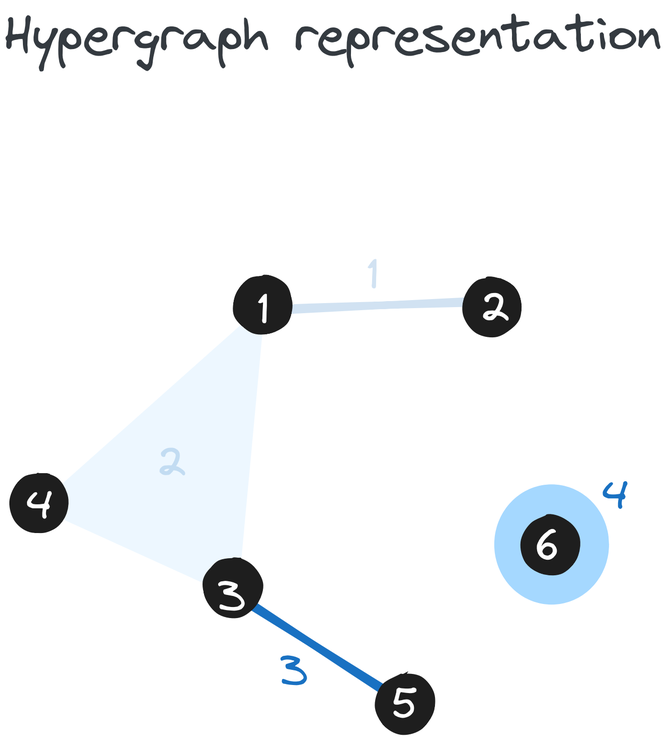


Nurisso, Deshpande, et al in prep

Representational vs processing capacity

Representation





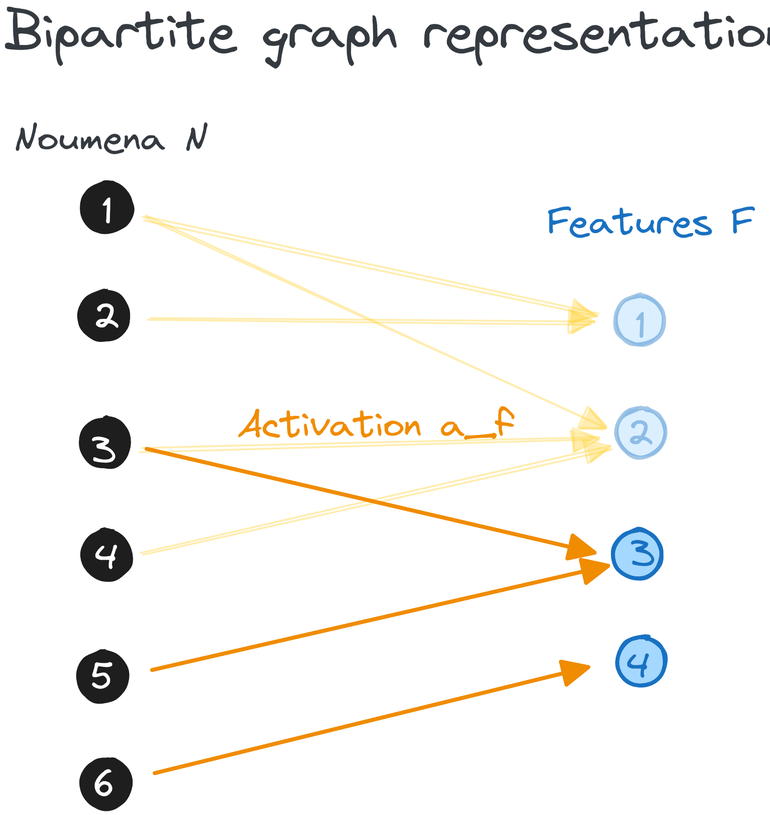
Nurisso, Deshpande, et al in prep

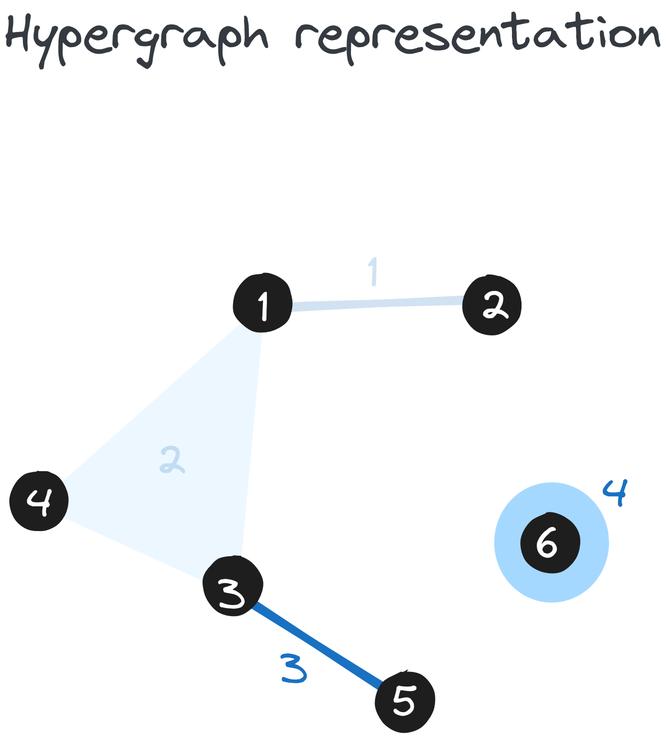
Representational vs processing capacity

Representation

Feature restriction f:

$$a_f : N \to 2^f, \quad a_f(x) = a(x) \cap f,$$





Nurisso, Deshpande, et al in prep

Representational vs processing capacity

Representation

Feature restriction f:

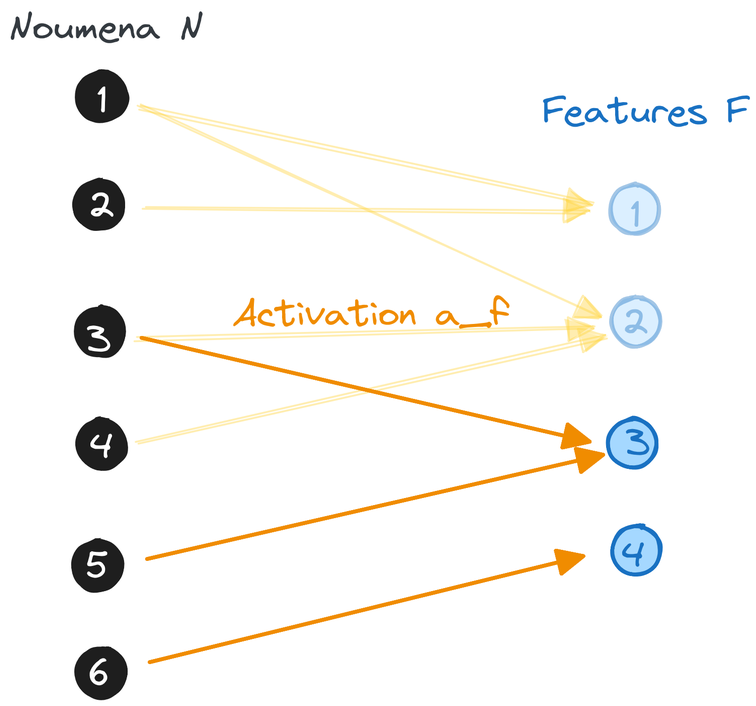
$$a_f : N \to 2^f, \;\; a_f(x) = a(x) \cap f,$$

Info coded by f:

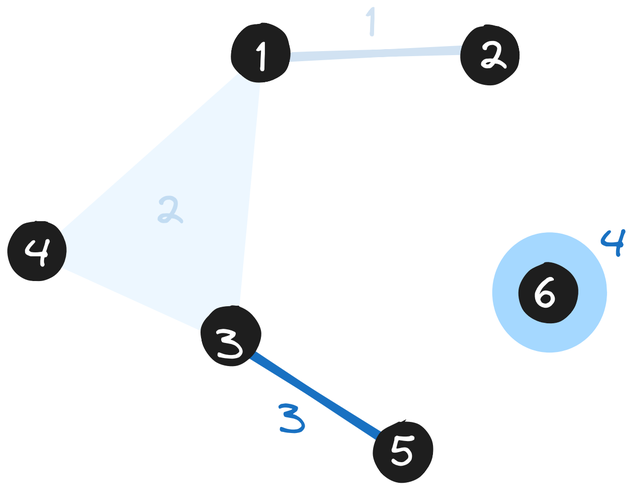
$$I_r(X; a) := \mathbb{E}_{f \in \binom{N}{r}}[I(x, a_f(X))]$$

$$= \sum_{f \in \binom{F}{r}} \mathbb{P}[f] I(X; a_f(X)),$$

Bipartite graph representation



Hypergraph representation



Nurisso, Deshpande, et al in prep

Representational vs processing capacity

Representation

Feature restriction f:

$$a_f : N \to 2^f, \;\; a_f(x) = a(x) \cap f,$$

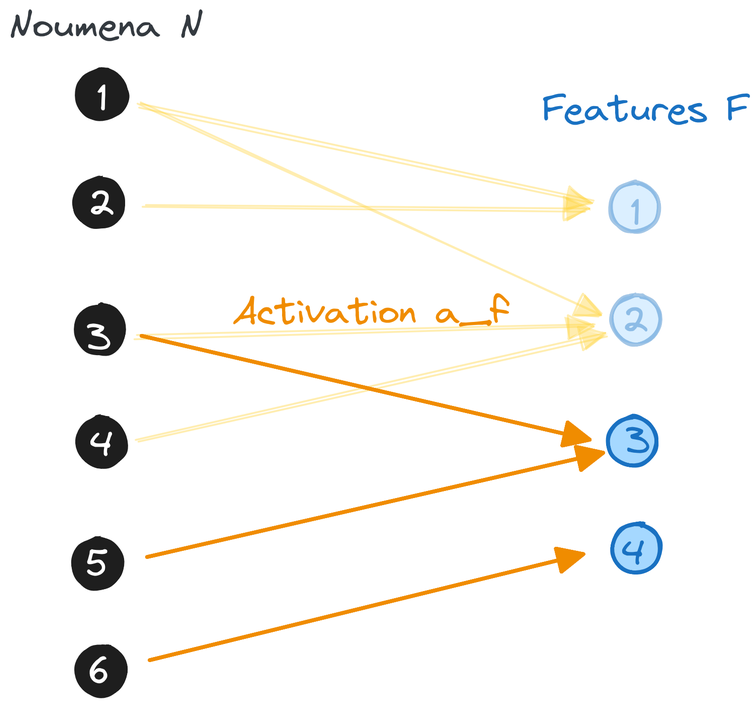
Info coded by f:

$$I_r(X; a) := \mathbb{E}_{f \in \binom{N}{r}}[I(x, a_f(X))]$$

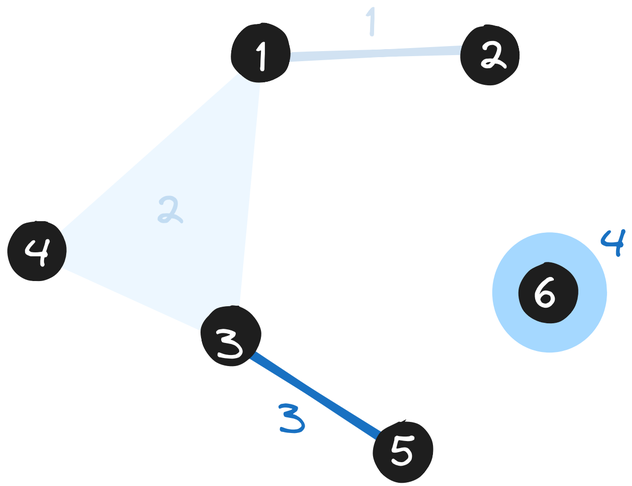
$$= \sum_{f \in \binom{F}{r}} \mathbb{P}[f] I(X; a_f(X)),$$

A generalizing code should increase $I_r(X, a)$ rapidly with r

Bipartite graph representation



Hypergraph representation



Nurisso, Deshpande, et al in prep

Representational vs processing capacity

Representation

Feature restriction f:

$$a_f : N \to 2^f, \;\; a_f(x) = a(x) \cap f,$$

Info coded by f:

$$I_r(X; a) := \mathbb{E}_{f \in \binom{N}{r}}[I(x, a_f(X))]$$

$$= \sum_{f \in \binom{F}{r}} \mathbb{P}[f] I(X; a_f(X)),$$

A generalizing code should increase $I_r(X, a)$ rapidly with r

Nurisso, Deshpande, et al in prep

Representational vs processing capacity

Representation

Feature restriction f:

$$a_f : N \to 2^f, \;\; a_f(x) = a(x) \cap f,$$

Info coded by f:

$$I_r(X; a) := \mathbb{E}_{f \in \binom{N}{r}}[I(x, a_f(X))]$$

$$= \sum_{f \in \binom{F}{r}} \mathbb{P}[f] I(X; a_f(X)),$$

A generalizing code should increase $I_r(X, a)$ rapidly with r

Processing

Nurisso, Deshpande, et al in prep

Representational vs processing capacity

Representation

Feature restriction f:

$$a_f : N \to 2^f, \;\; a_f(x) = a(x) \cap f,$$

Info coded by f:

$$I_r(X; a) := \mathbb{E}_{f \in \binom{N}{r}}[I(x, a_f(X))]$$

$$= \sum_{f \in \binom{F}{r}} \mathbb{P}[f] I(X; a_f(X)),$$

A generalizing code should increase $I_r(X, a)$ rapidly with r

Processing

Apple —> (1,0,0,0,1)

Banana —> (1,1,0,0,1)

Nurisso, Deshpande, et al in prep

Representational vs processing capacity

Representation

Feature restriction f:

$$a_f : N \to 2^f, \;\; a_f(x) = a(x) \cap f,$$

Info coded by f:

$$I_r(X; a) := \mathbb{E}_{f \in \binom{N}{r}}[I(x, a_f(X))]$$

$$= \sum_{f \in \binom{F}{r}} \mathbb{P}[f] I(X; a_f(X)),$$

A generalizing code should increase $I_r(X, a)$ rapidly with r

Processing

Apple —> (1,0,0,0,1)

Banana —> (1,1,0,0,1)

(Apple, banana) —> (1,1,0,0,1)

Nurisso, Deshpande, et al in prep

Representational vs processing capacity

Representation

Feature restriction f:

$$a_f : N \to 2^f, \;\; a_f(x) = a(x) \cap f,$$

Info coded by f:

$$I_r(X; a) := \mathbb{E}_{f \in \binom{N}{r}}[I(x, a_f(X))]$$

$$= \sum_{f \in \binom{F}{r}} \mathbb{P}[f] I(X; a_f(X)),$$

A generalizing code should increase $I_r(X, a)$ rapidly with r

Processing

Apple –> (1,0,0,0,1)

Banana –> (1,1,0,0,1)

(Apple, banana) –> (1,1,0,0,1)

$$a^k : \binom{N}{k} \to 2^F, \quad a^k((x_1, \ldots, x_k)) = \bigcup_{i=1}^{k} a(x_i)$$

Nurisso, Deshpande, et al in prep

Representational vs processing capacity

Representation

Feature restriction f:

$$a_f : N \to 2^f, \;\; a_f(x) = a(x) \cap f,$$

Info coded by f:

$$I_r(X;a) := \mathbb{E}_{f \in \binom{N}{r}}[I(x, a_f(X))]$$

$$= \sum_{f \in \binom{F}{r}} \mathbb{P}[f] I(X; a_f(X)),$$

A generalizing code should increase $I_r(X, a)$ rapidly with r

Processing

Apple –> (1,0,0,0,1)

Banana –> (1,1,0,0,1)

(Apple, banana) –> (1,1,0,0,1)

$$a^k : \binom{N}{k} \to 2^F, \quad a^k((x_1, \ldots, x_k)) = \bigcup_{i=1}^{k} a(x_i)$$

$$I^k(X;a) = \sum_{S \in \binom{N}{k}} \mathbb{P}[S] \log_2 \mathbb{P}(a^k(S)).$$

Nurisso, Deshpande, et al in prep

Representational vs processing capacity

Representation

Feature restriction f:

$$a_f : N \to 2^f, \;\; a_f(x) = a(x) \cap f,$$

Info coded by f:

$$I_r(X;a) := \mathbb{E}_{f \in \binom{N}{r}}[I(x, a_f(X))]$$

$$= \sum_{f \in \binom{F}{r}} \mathbb{P}[f] I(X; a_f(X)),$$

A generalizing code should increase $I_r(X, a)$ rapidly with r

Processing

Apple —> (1,0,0,0,1)

Banana —> (1,1,0,0,1)

(Apple, banana) —> (1,1,0,0,1)

$$a^k : \binom{N}{k} \to 2^F, \quad a^k((x_1, \ldots, x_k)) = \bigcup_{i=1}^{k} a(x_i)$$

$$I^k(X;a) = \sum_{S \in \binom{N}{k}} \mathbb{P}[S] \log_2 \mathbb{P}(a^k(S)).$$

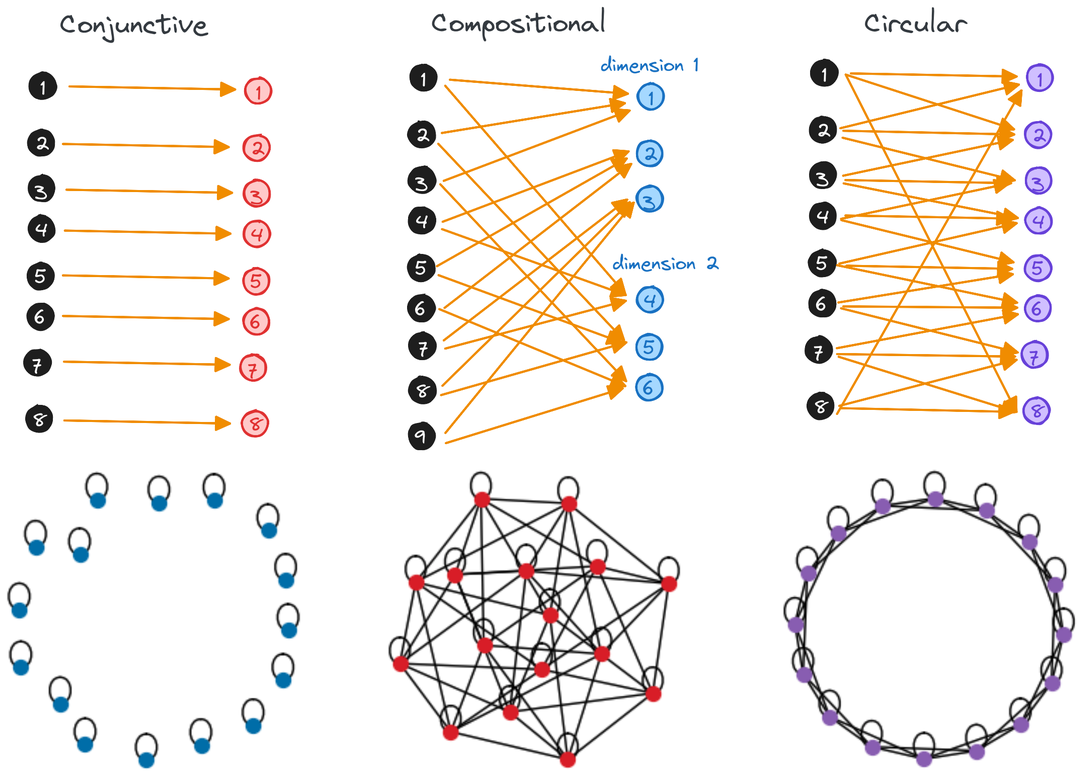
A good processing code should not decrease $I^k(X, a)$ with k

Nurisso, Deshpande, et al in prep

Representational vs processing capacity

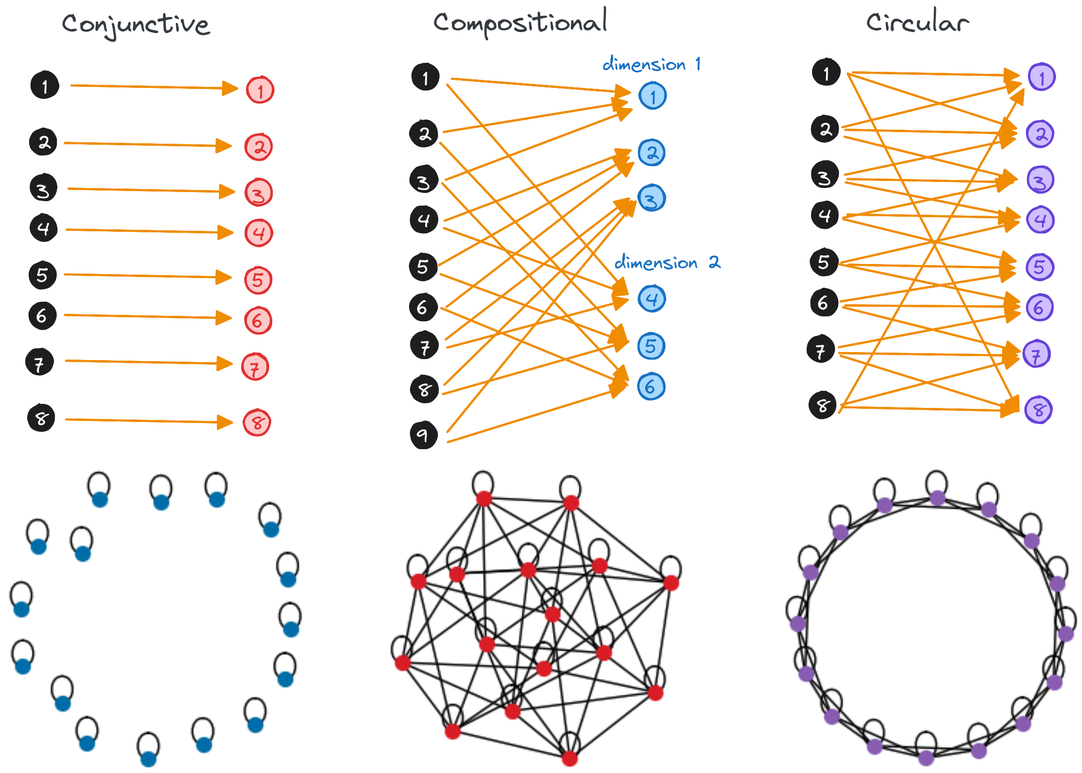
Nurisso, Deshpande, et al in prep

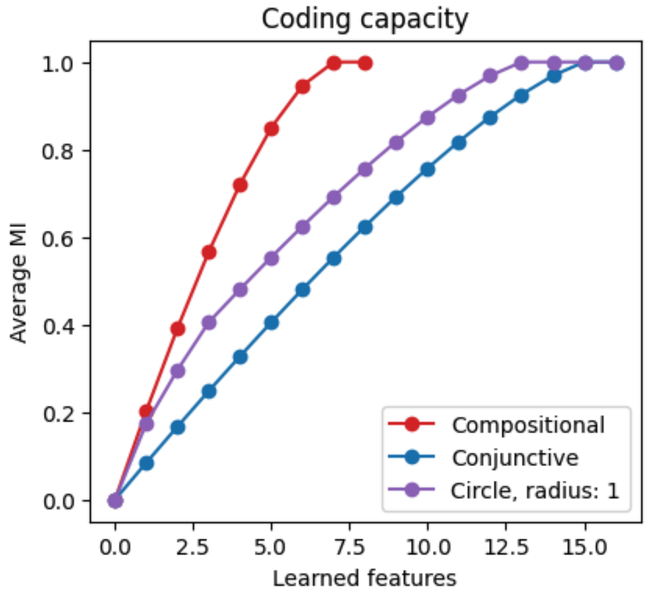
Representational vs processing capacity

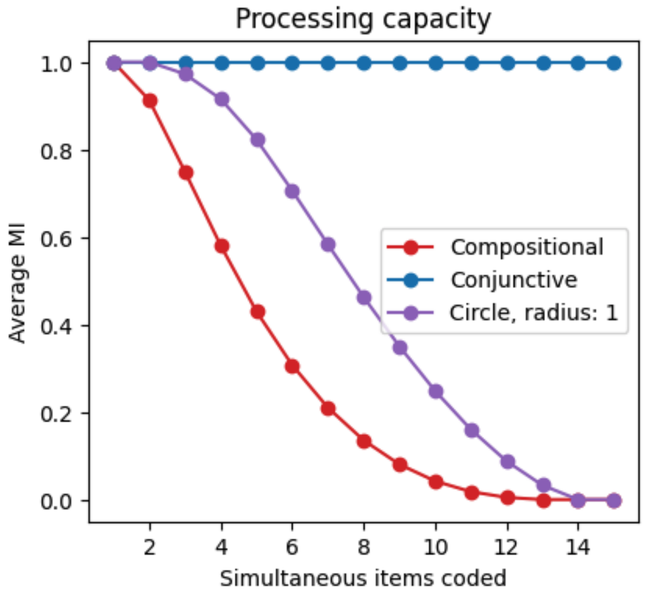


Nurisso, Deshpande, et al in prep

Representational vs processing capacity

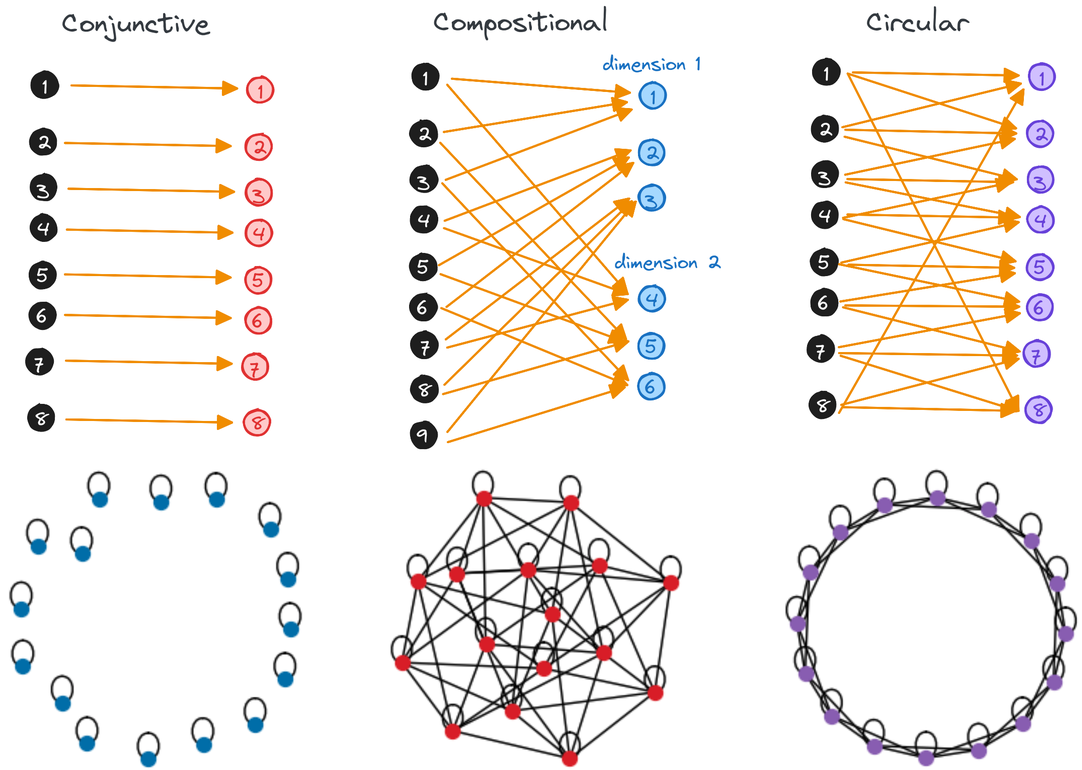


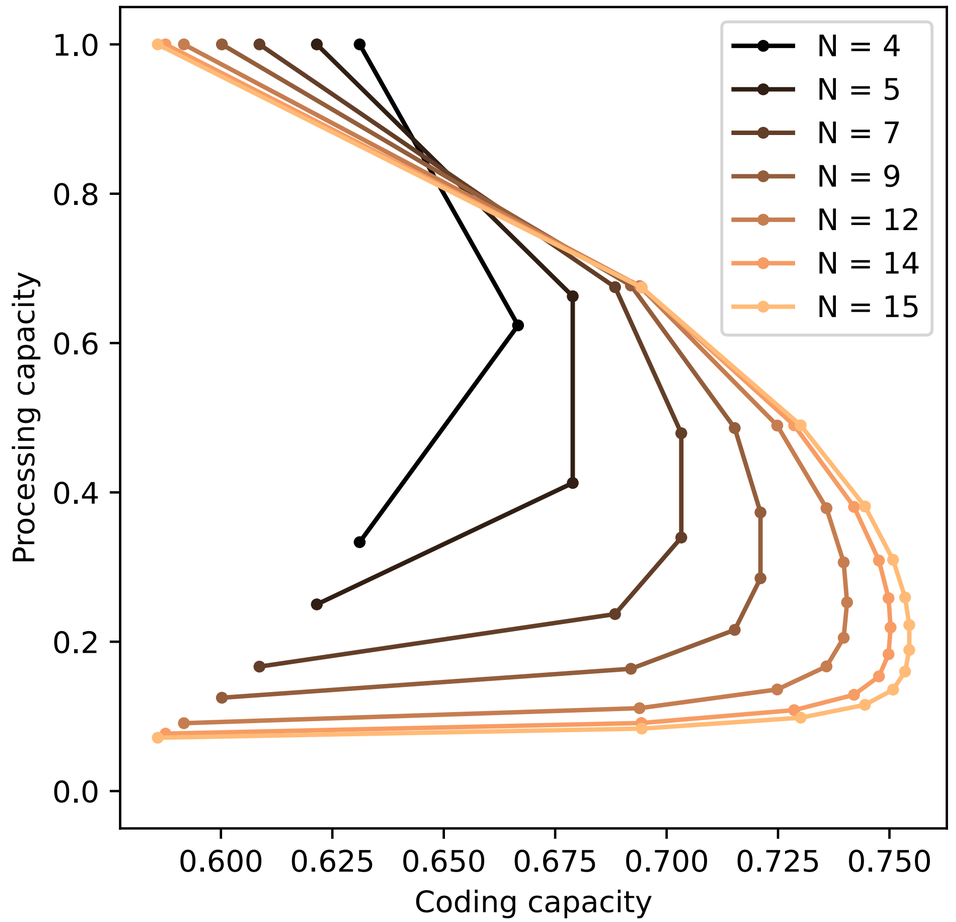




Nurisso, Deshpande, et al in prep

Representational vs processing capacity





Nurisso, Fernando, Deshpande, et al in prep

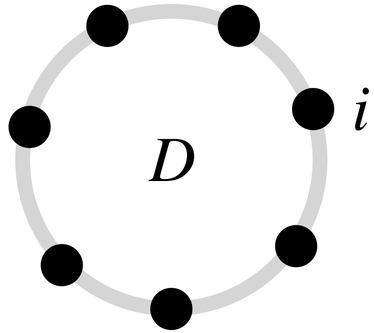
Emergence of resolution in simple neural networks

Implementation

Consider the *toy model of superposition* architecture $f(x) = \sigma(W^\top W x)$.

The data is composed of n items (encoded as one-hots) + a distance matrix D (circle)

i-th embedding $z_i = We_i = W_i$

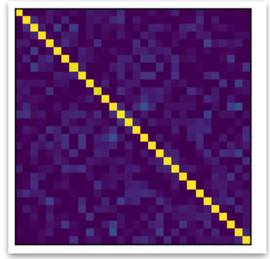


I train it using two losses:

1. **Reconstruction loss** $L_{rec} = \frac{1}{n}\sum_{i=1}^{n} \|e_i - f(e_i)\|_2^2$

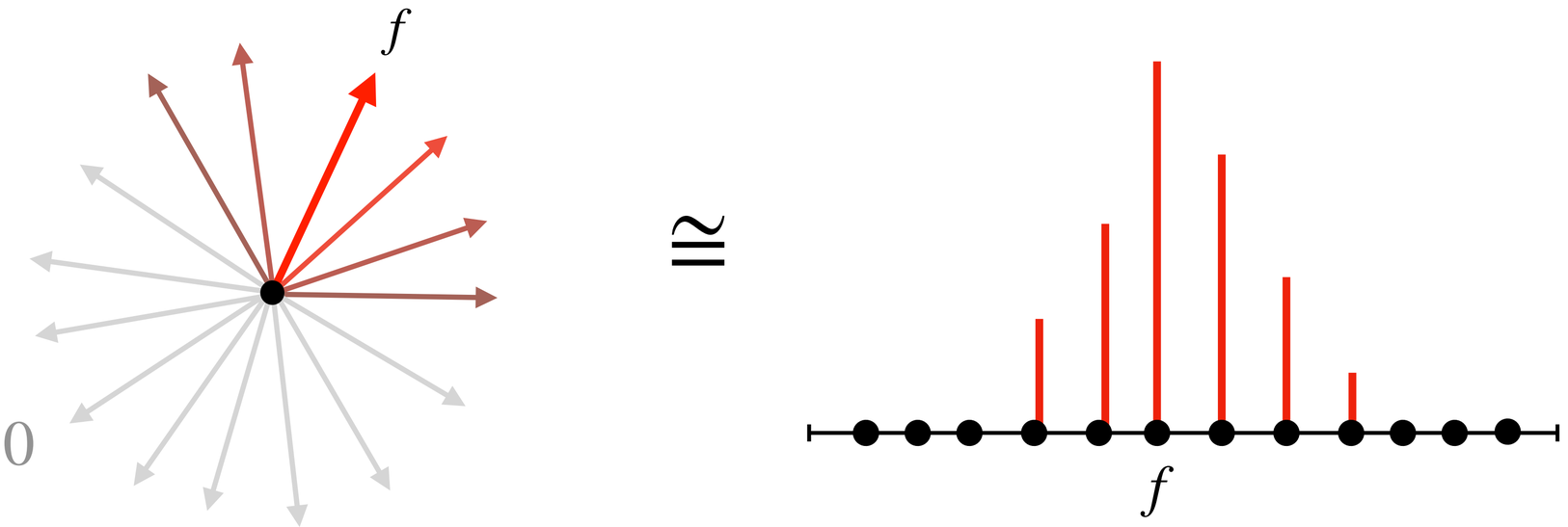
2. **Similarity test loss** = pick a triplet (i, j, k) and compute the probability $p_{ik} = \frac{\sigma(z_i^\top z_k)}{\sigma(z_i^\top z_k) + \sigma(z_j^\top z_k)}$ and then compute Cross Entropy loss L_{sim} against the true index of the closest one w.r.t. D

Reconstruction ≅ Identification pushes for the embeddings to be "ReLU-orthogonal"



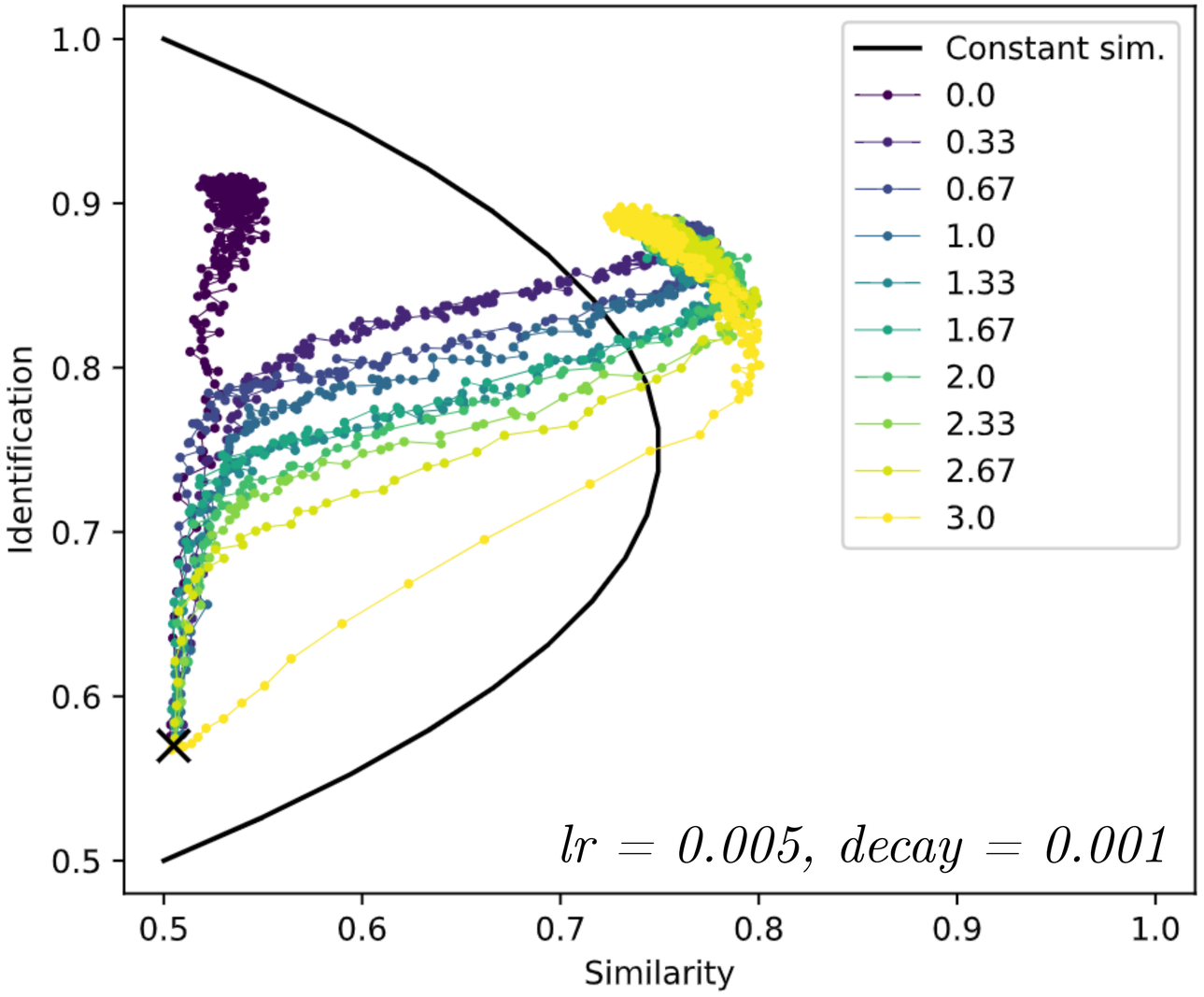
Similarity pushes for the embeddings to adapt to the metric space structure

Reconstruction drives the model to put many features in a quasi-orthogonal state
$\Longrightarrow$ Similarity encourages a "metric" resolution $\Longrightarrow$ Miller's law (?)



Total loss $= \lambda_{\text{rec}} L_{\text{rec}} + \lambda_{\text{sim}} L_{\text{sim}}$

We can look at the training profiles when we change the values of λ_{rec}, λ_{sim}

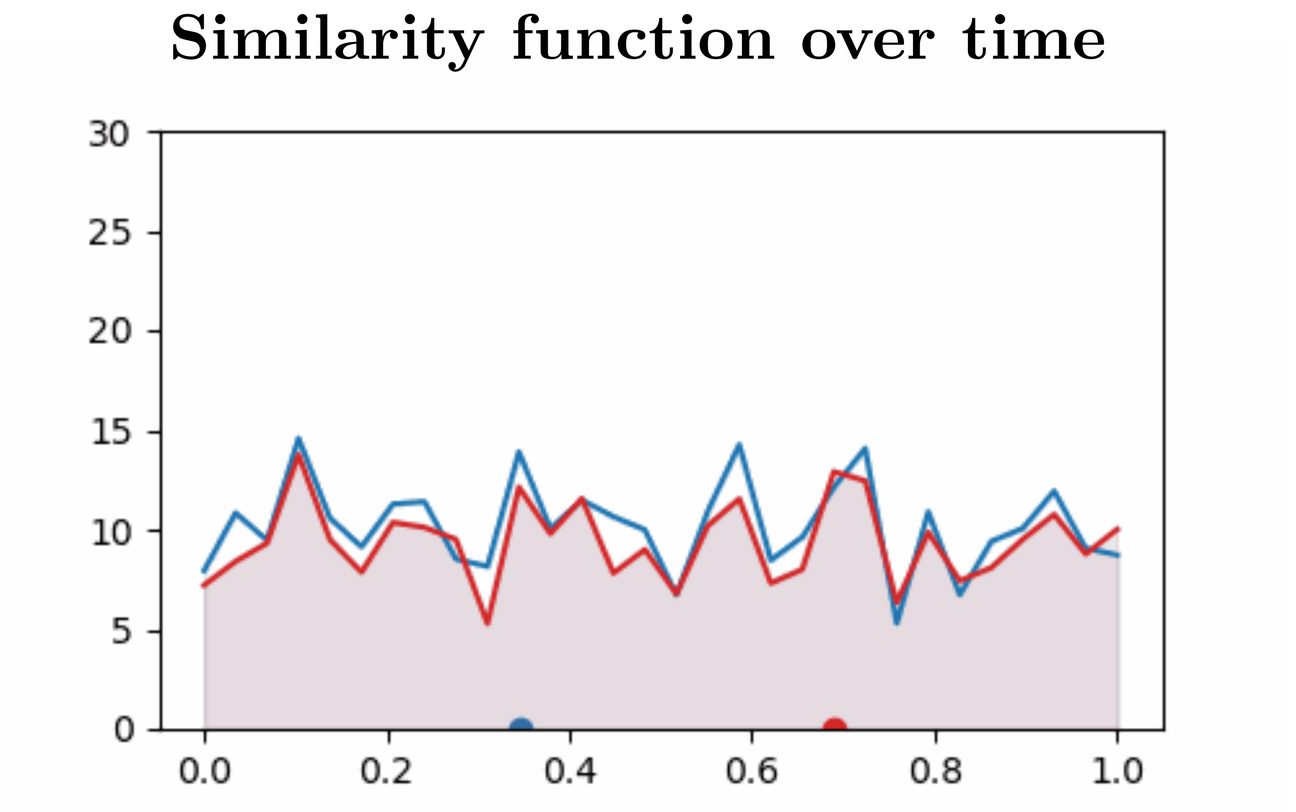


It seems that, even for similarity-dominated trainings, there is a force leading the model to better I at the expense of some G (*opt. issue?*).

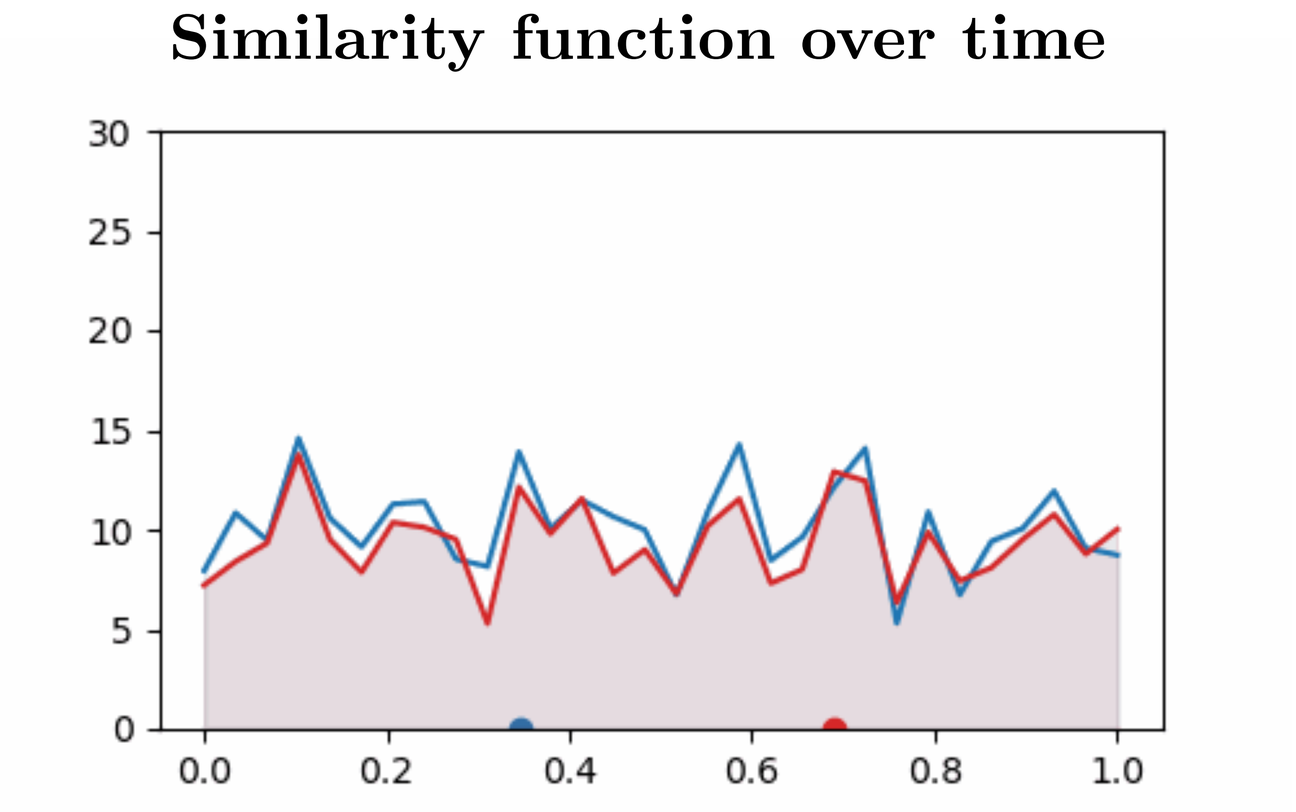
This is an optimization issue but it reveals the law's curve

The max identification is due to the fact that we're squeezing 30 features in 10 dimensions.

Train a single model only on similarity for *500 epochs* with *0.001 lr and 0 decay*



Train a single model only on similarity for *500 epochs* with *0.001 lr and 0 decay*



The shape of the similarity function suggests that the constant similarity of the theory may not be a good fit.

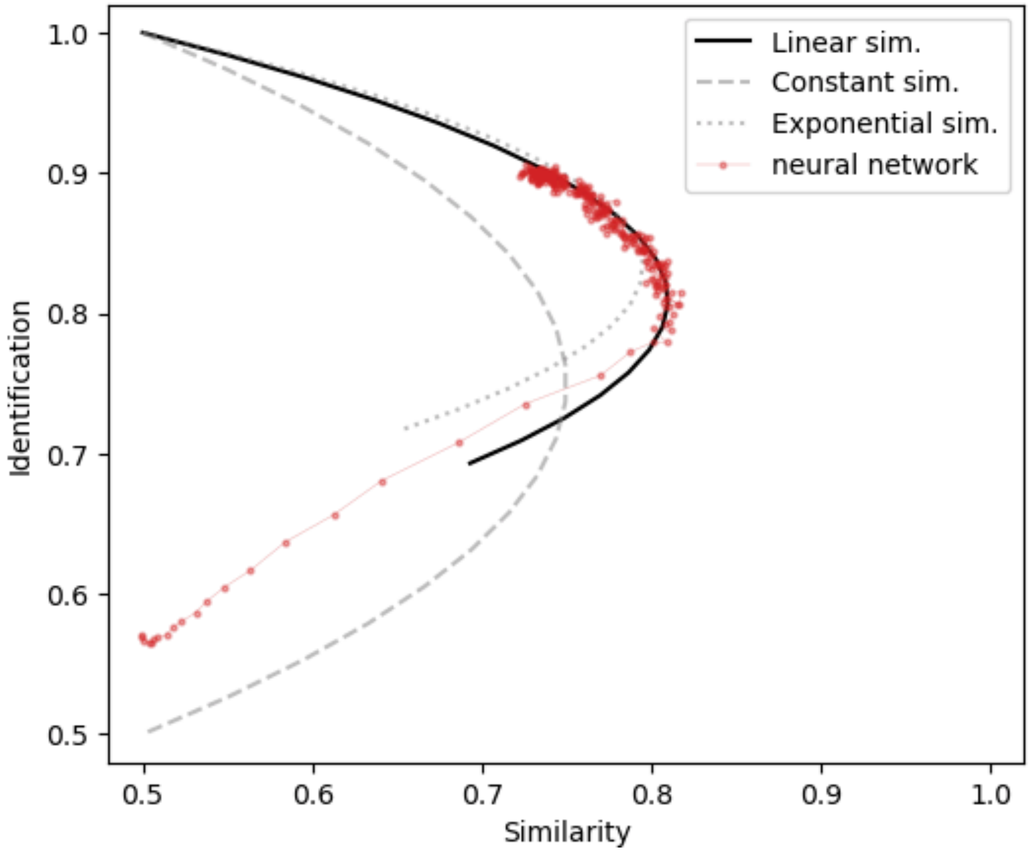
If we assume linear "hat" similarities  we can still find the formulae, at least for the circle

$$G(\varepsilon) = \frac{1}{2} + 2\varepsilon - 2(3 - 2\log 2)\varepsilon^2$$

$$I(\varepsilon) = 1 - 2(1 - \log 2)\varepsilon$$

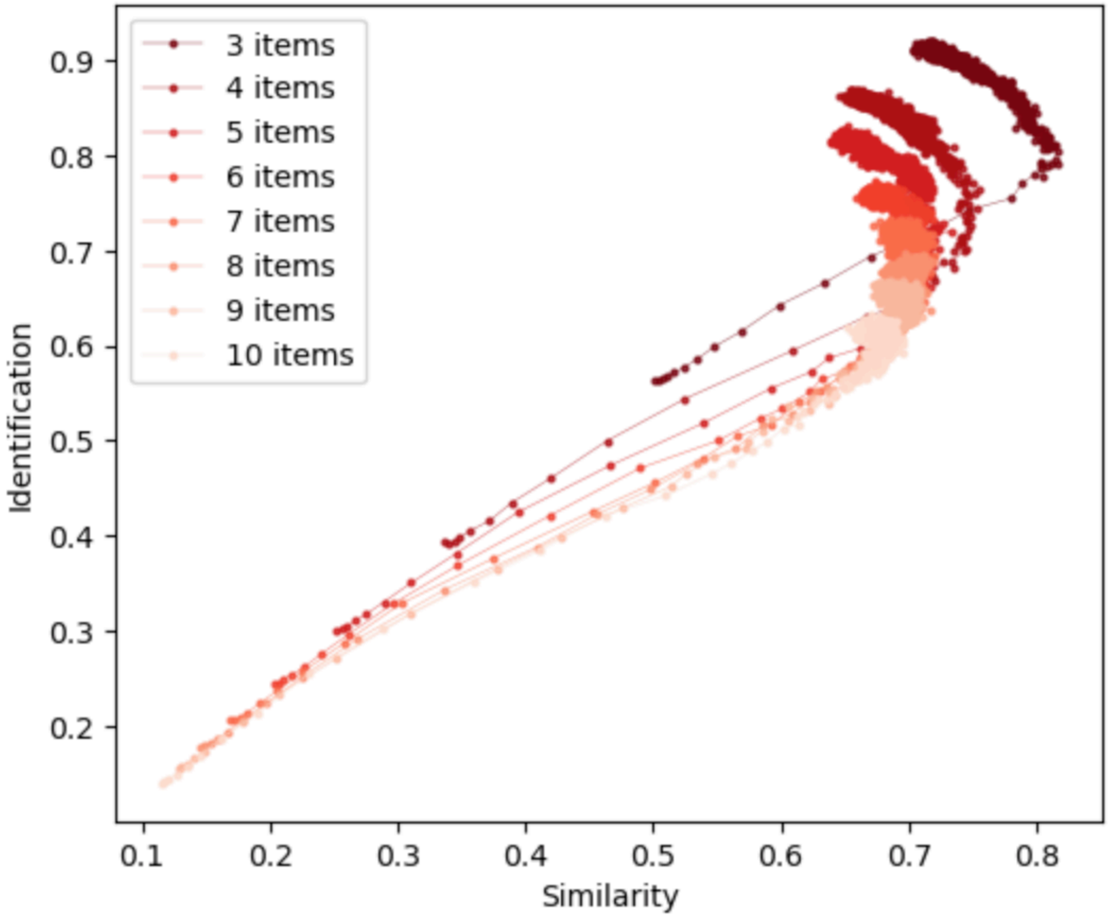
Linear decay gives us a great fit!

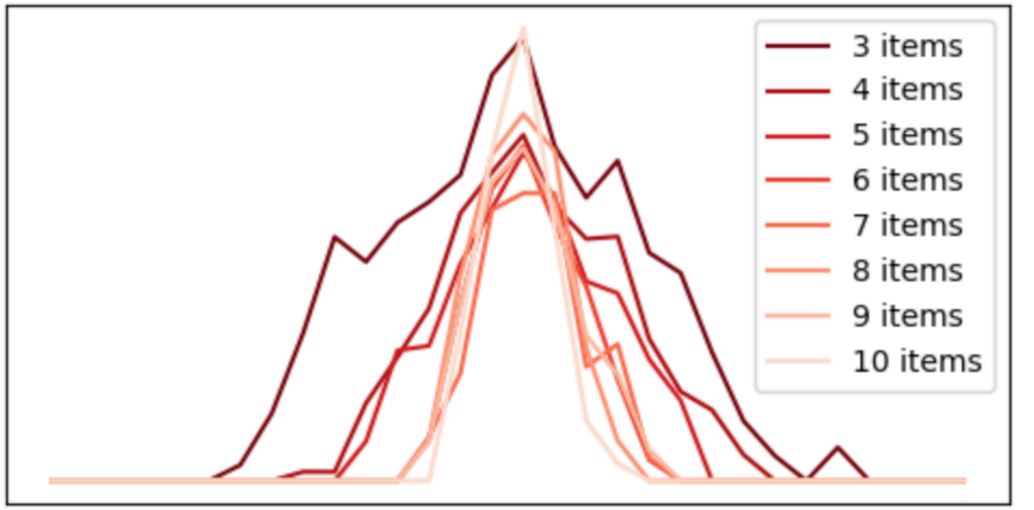
Notice that $\varepsilon_* = \dfrac{1}{2(3 - 2\log 2)} \approx 0.31$ is the radius maximizing G



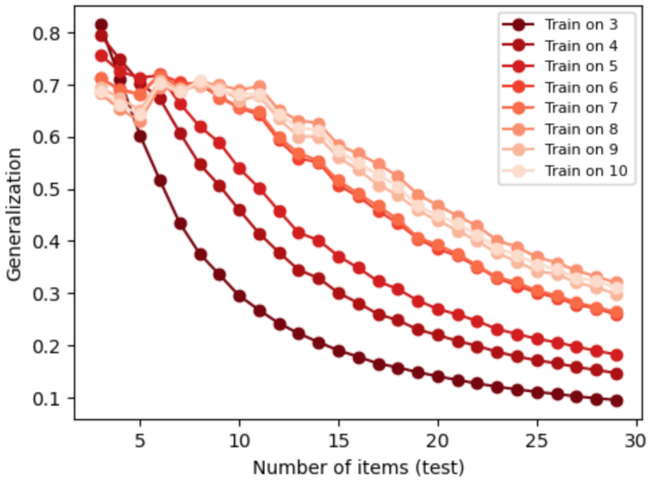
Multi-item analysis

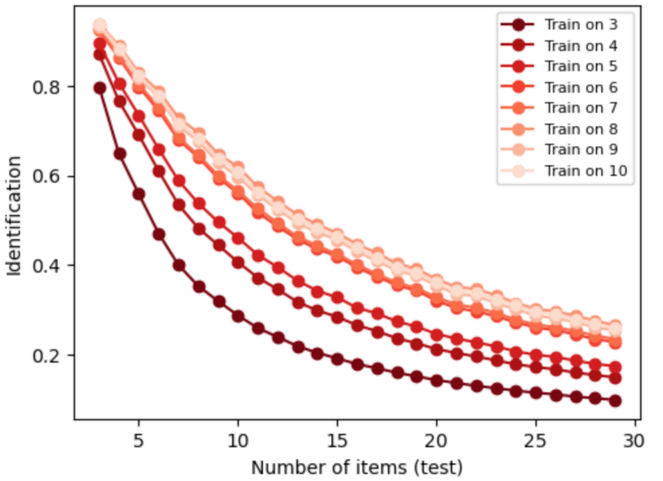
30 inputs, circle structure, 10 latents, 2000 samples, 1000 epochs, 128 batch size, init scale [0,2] uniform, 0.007 lr, 0 wd, $\lambda_{sim} = 0.1$, $\lambda_{rec} = 0$
Note: higher number of items means that the model sees proportionally more data in the training process. *Probably not the best idea*





Smaller sim. functions as the number of items gets larger

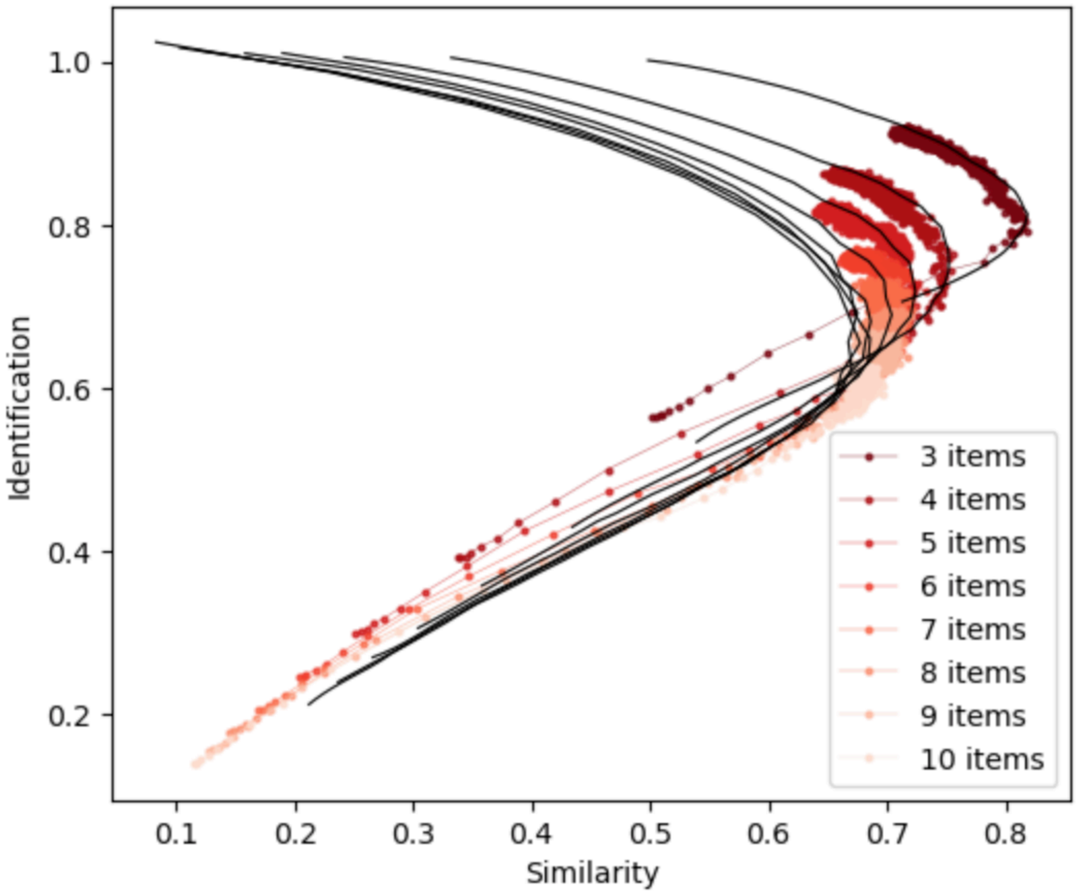




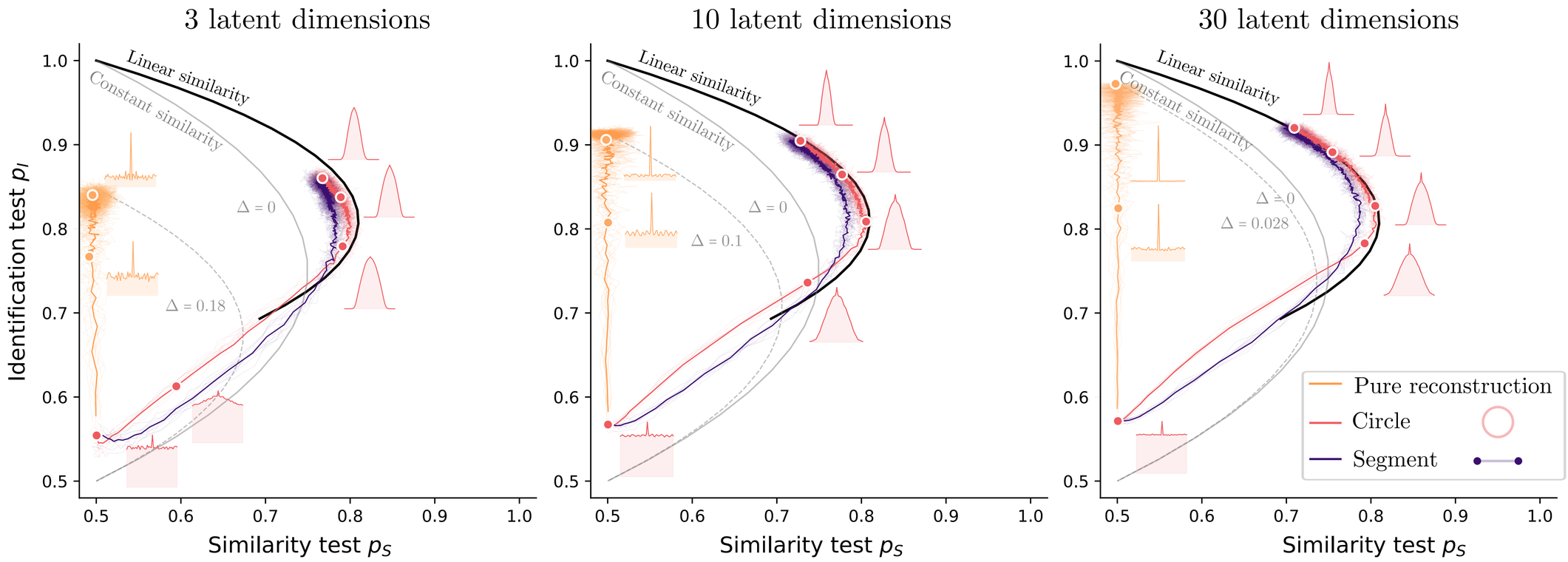
Multi-item analysis

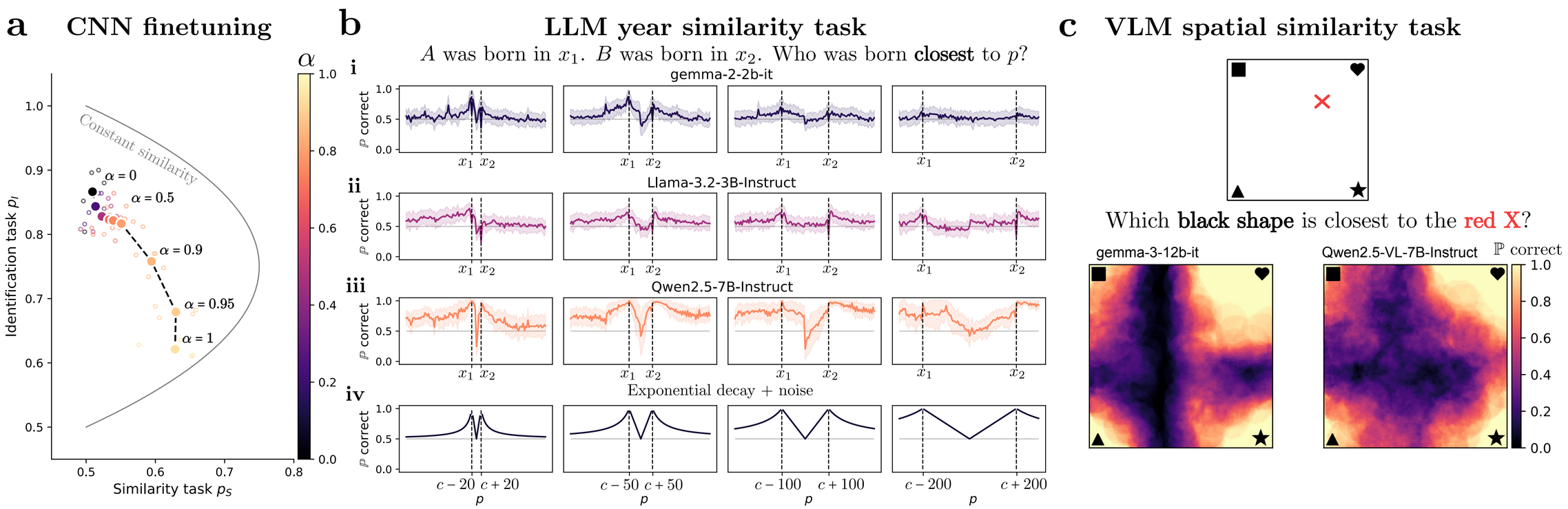
If I plot the multi-item curves for the linear decaying similarity function (by simulation + smoothing), I get the following

Not perfect but still quite good



This means that the similarity function learned can be approximated by linear for small n but changes when we increase.


3 latent dimensions
10 latent dimensions
30 latent dimensions
Linear similarity
Constant similarity
Δ = 0
Δ = 0.18
Δ = 0.1
Δ = 0.028
Identification test p_I
Similarity test p_S
1.0
0.9
0.8
0.7
0.6
0.5
0.5
0.6
0.7
0.8
0.9
1.0
Pure reconstruction
Circle
Segment


a
CNN finetuning
Constant similarity
α = 0
α = 0.5
α = 0.9
α = 0.95
α = 1
Identification task p_I
Similarity task p_S
α
b
LLM year similarity task
A was born in x_1. B was born in x_2. Who was born closest to p?
i
gemma-2-2b-it
ii
Llama-3.2-3B-Instruct
iii
Qwen2.5-7B-Instruct
iv
Exponential decay + noise
ℙ correct
x_1 x_2
$c-20$ $c+20$
$c-50$ $c+50$
$c-100$ $c+100$
$c-200$ $c+200$
p
c
VLM spatial similarity task
Which black shape is closest to the red X?
gemma-3-12b-it
Qwen2.5-VL-7B-Instruct
ℙ correct