When do neurons* represent `True`?

Joey Velez-Ginorio[1],

in conversation with Nada Amin[3], Steve Zdancewic[1], Konrad Körding[1,2]

[1]Department of Computer and Information Science; University of Pennsylvania
[2]Department of Neuroscience; University of Pennsylvania
[3]Harvard John A. Paulson School Of Engineering And Applied Sciences; Harvard University

When do ~~neurons~~ neurons* represent ~~True~~ X?

Joey Velez-Ginorio[1],

in conversation with Nada Amin[3], Steve Zdancewic[1], Konrad Körding[1,2]

[1]Department of Computer and Information Science; University of Pennsylvania
[2]Department of Neuroscience; University of Pennsylvania
[3]Harvard John A. Paulson School Of Engineering And Applied Sciences; Harvard University

Penn University of Pennsylvania

Harvard University

When do neurons* represent `True`?

Joey Velez-Ginorio[1],

in conversation with Nada Amin[3], Steve Zdancewic[1], Konrad Körding[1,2]

[1]Department of Computer and Information Science; University of Pennsylvania
[2]Department of Neuroscience; University of Pennsylvania
[3]Harvard John A. Paulson School Of Engineering And Applied Sciences; Harvard University

When do neurons* represent `True`?

Response 1: "I often say neural *activity* represents some behavior or aspect of behavior if there's a strong enough correlation e.g. neural activity represents monkey hopping if it can be linearly decoded. So I guess the logic would be a neuron represents X when its firing pattern strongly relates to X."

When do neurons* represent `True`?

Response 1: “I often say neural *activity* represents some behavior or aspect of behavior if there’s a strong enough correlation e.g. neural activity represents monkey hopping if it can be linearly decoded. So I guess the logic would be a neuron represents X when its firing pattern strongly relates to X.”

Response 2: “Hm. Maybe something like – you get a consistent membrane potential across the neuron when X is in the representation. Invariant to any other features of the representation.”

When do neurons* represent `True`?

Response 1: "I often say neural *activity* represents some behavior or aspect of behavior if there's a strong enough correlation e.g. neural activity represents monkey hopping if it can be linearly decoded. So I guess the logic would be a neuron represents X when its firing pattern strongly relates to X."

Response 2: "Hm. Maybe something like – you get a consistent membrane potential across the neuron when X is in the representation. Invariant to any other features of the representation."

Response 3: "...The classic neuro/ml approach tends to focus on there being some decodable correlate of X, e.g. if X is linearly decodable from neural activity. I would add that *how a putative representation is used to drive some behavior* is also an important aspect."

When do neurons* represent `True`?

When do neurons* represent **`True`**?

Step 1: Describe **`True`** as a vector.

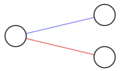
$$\texttt{True} \mapsto \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}$$

When do neurons* represent **True**?

Step 1: Describe **True** as a vector.

$$\texttt{True} \mapsto \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}$$

Step 2: Describe neurons as a vector.



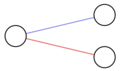
$$\mapsto \begin{bmatrix} 3.98 \\ 0.3 \end{bmatrix}$$

When do neurons* represent **`True`**?

Step 1: Describe **`True`** as a vector.

$$\texttt{True} \mapsto \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}$$

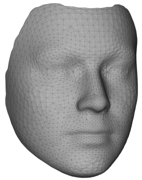
Step 2: Describe neurons as a vector.

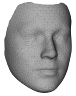


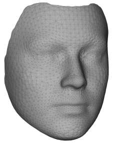
$$\mapsto \begin{bmatrix} 3.98 \\ 0.3 \end{bmatrix}$$

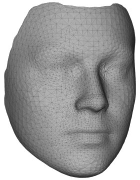
Step 3: Relate the descriptions.

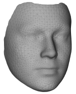
$$Corr^*\left(\begin{bmatrix} 3.98 \\ 0.3 \end{bmatrix}, \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}\right) \approx \underline{0.99}$$

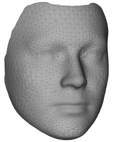
When do neurons* represent  ?

Step 1: Describe  as a vector.

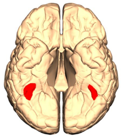
 $\mapsto \begin{bmatrix} 1.3 & \dots & 8.23 \\ 4.2 & \dots & 0.53 \end{bmatrix}$

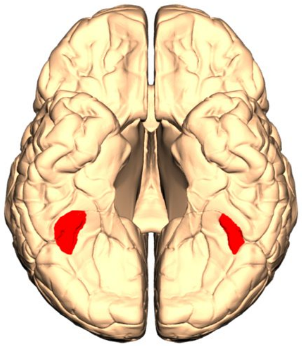
When do neurons* represent  ?

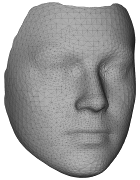
Step 1: Describe  as a vector.

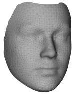
 $\mapsto \begin{bmatrix} 1.3 & \dots & 8.23 \\ 4.2 & \dots & 0.53 \end{bmatrix}$

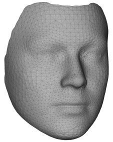
Step 2: Describe neurons as a vector.

 $\mapsto \begin{bmatrix} 1.3 & \dots & 8.23 \\ 4.2 & \dots & 0.53 \end{bmatrix}$

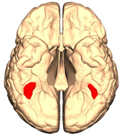


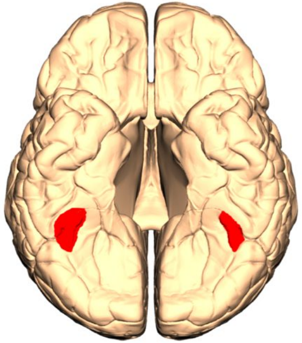
When do neurons* represent  ?

Step 1: Describe  as a vector.

 $\mapsto \begin{bmatrix} 1.3 & \dots & 8.23 \\ 4.2 & \dots & 0.53 \end{bmatrix}$

Step 2: Describe neurons as a vector.

 $\mapsto \begin{bmatrix} 1.3 & \dots & 8.23 \\ 4.2 & \dots & 0.53 \end{bmatrix}$



Step 3: Relate the descriptions.

$$Corr^*\left(\begin{bmatrix} 0.49 & \dots & 4.5 \\ 3.7 & \dots & 0.04 \end{bmatrix}, \begin{bmatrix} 1.3 & \dots & 8.23 \\ 4.2 & \dots & 0.53 \end{bmatrix}\right) \approx \underline{0.99}$$

But...

When do neurons* represent **`True`**?

Step 1: Describe **`True`** as a vector.

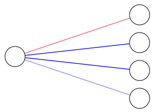
$$\texttt{True} \mapsto \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}$$

When do neurons* represent **`True`**?

Step 1: Describe **`True`** as a vector.

$$\texttt{True} \mapsto \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}$$

Step 2: Describe neurons as a vector.



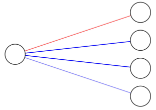
$$\mapsto \begin{bmatrix} 0.3 \\ 13.1 \\ 10.3 \\ 3.4 \end{bmatrix}$$

When do neurons* represent **`True`**?

Step 1: Describe **`True`** as a vector.

$$\texttt{True} \mapsto \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}$$

Step 2: Describe neurons as a vector.



$$\mapsto \begin{bmatrix} 0.3 \\ 13.1 \\ 10.3 \\ 3.4 \end{bmatrix}$$

Step 3: Relate the descriptions.

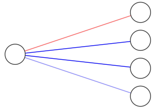
$$Corr^*\left(\begin{bmatrix} 0.3 \\ 13.1 \\ 10.3 \\ 3.4 \end{bmatrix}, \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix} \right) \approx \underline{0.12}$$

When do neurons* represent **True**?

Step 1: Describe **`True`** as a vector.

$$\texttt{True} \mapsto \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}$$

Step 2: Describe neurons as a vector.



$$\mapsto \begin{bmatrix} 0.3 \\ 13.1 \\ 10.3 \\ 3.4 \end{bmatrix}$$

Step 3: Relate the descriptions.

$$Corr^*\left(\begin{bmatrix} 0.3 \\ 13.1 \\ 10.3 \\ 3.4 \end{bmatrix}, \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix} \right) \approx \underline{0.12}$$

When do neurons* represent `True`?

Step 1: Describe `True` as a vector.

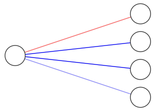
$$\texttt{True} \mapsto \begin{bmatrix} 0.3 \\ 2.1 \\ 2.8 \\ 0.8 \end{bmatrix}$$

When do neurons* represent **`True`**?

Step 1: Describe **`True`** as a vector.

$$\texttt{True} \mapsto \begin{bmatrix} 0.3 \\ 2.1 \\ 2.8 \\ 0.8 \end{bmatrix}$$

Step 2: Describe neurons as a vector.



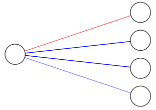
$$\mapsto \begin{bmatrix} 0.3 \\ 13.1 \\ 10.3 \\ 3.4 \end{bmatrix}$$

When do neurons* represent **`True`**?

Step 1: Describe **`True`** as a vector.

$$\texttt{True} \mapsto \begin{bmatrix} 0.3 \\ 2.1 \\ 2.8 \\ 0.8 \end{bmatrix}$$

Step 2: Describe neurons as a vector.



$$\mapsto \begin{bmatrix} 0.3 \\ 13.1 \\ 10.3 \\ 3.4 \end{bmatrix}$$

Step 3: Relate the descriptions.

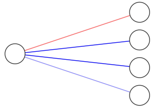
$$Corr^*\left(\begin{bmatrix} 0.3 \\ 13.1 \\ 10.3 \\ 3.4 \end{bmatrix}, \begin{bmatrix} 0.3 \\ 2.1 \\ 2.8 \\ 0.8 \end{bmatrix} \right) \approx \underline{0.99}$$

When do neurons* represent **`True`**?

Step 1: Describe **`True`** as a vector.

$$\texttt{True} \mapsto \begin{bmatrix} 0.3 \\ 2.1 \\ 2.8 \\ 0.8 \end{bmatrix}$$

Step 2: Describe neurons as a vector.



$$\mapsto \begin{bmatrix} 0.3 \\ 13.1 \\ 10.3 \\ 3.4 \end{bmatrix}$$

Step 3: Relate the descriptions.

$$Corr^*\left(\begin{bmatrix} 0.3 \\ 13.1 \\ 10.3 \\ 3.4 \end{bmatrix}, \begin{bmatrix} 0.3 \\ 2.1 \\ 2.8 \\ 0.8 \end{bmatrix} \right) \approx \underline{0.99}$$

The maps problem and the mapping problem: Two challenges for a cognitive neuroscience of speech and language

David Poeppel

Department of Psychology, New York University, New York, NY, USA

The maps problem and the $\boxed{\text{mapping problem:}}$ Two challenges for a cognitive neuroscience of speech and language

David Poeppel

Department of Psychology, New York University, New York, NY, USA

Step 1: Describe `True` as a vector.

$$\texttt{True} \mapsto \begin{bmatrix} 0.3 \\ 2.1 \\ 2.8 \\ 0.8 \end{bmatrix}$$

The maps problem and the $\boxed{\text{mapping problem:}}$ Two challenges for a cognitive neuroscience of speech and language

David Poeppel

Department of Psychology, New York University, New York, NY, USA

"...there is absolutely no mapping to date that we understand in even the most vague sense."

The maps problem and the $\boxed{\text{mapping problem:}}$ Two challenges for a cognitive neuroscience of speech and language

David Poeppel

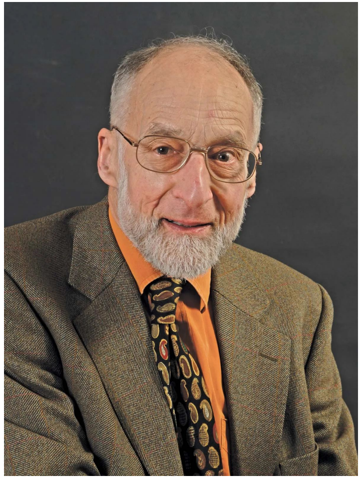
Department of Psychology, New York University, New York, NY, USA

1. What properties should this mapping have?
2. What's the domain of this map?

"Look at the unknown! And try to think of a familiar problem having the same or a similar unknown."

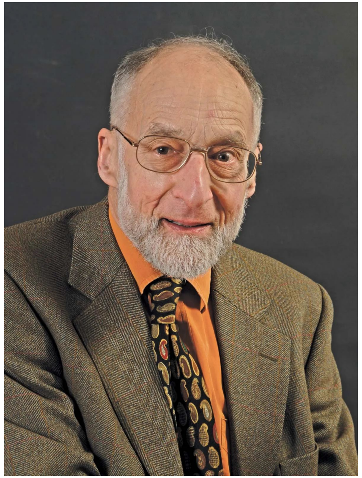
– George Polya, *How to solve it*

Robin Milner, 1991 Turing Award



"…he formulated and strongly advanced full abstraction, the study of the relationship between operational and denotational semantics."

Robin Milner, 1991 Turing Award



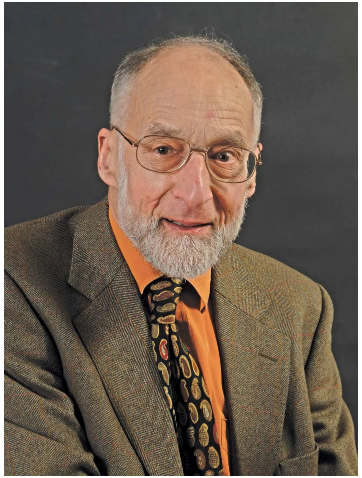
"…he formulated and strongly advanced full abstraction, ~~the study of the relationship between operational and denotational semantics."~~

a solution to the mapping problem.

Step 1: Describe **`True`** as a vector.

$$\texttt{True} \mapsto \begin{bmatrix} 0.3 \\ 2.1 \\ 2.8 \\ 0.8 \end{bmatrix}$$

Robin Milner, 1991 Turing Award



"…he formulated and strongly advanced full abstraction, ~~the study of the relationship between operational and denotational semantics."~~
a solution to the mapping problem.

1. What properties should this mapping have?
   Answer: It should be fully abstract

2. What's the domain of this map?
   Answer: A programming language

1. What properties should this mapping have?

Step 1: Describe **`True`** as a vector using a fully abstract map.

$$\texttt{True} \mapsto \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}$$

1. What properties should this mapping have?

Step 1: Describe `True` as a vector using a fully abstract map.

$$\texttt{True} \mapsto \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}$$

Definition: A fully abstract map $[\![-]\!]$ preserves equivalences and distinctions i.e.

$$f \simeq_s g \implies [\![f]\!] \simeq_t [\![g]\!]$$

$$f \not\simeq_s g \implies [\![f]\!] \not\simeq_t [\![g]\!]$$

1. What properties should this mapping have?

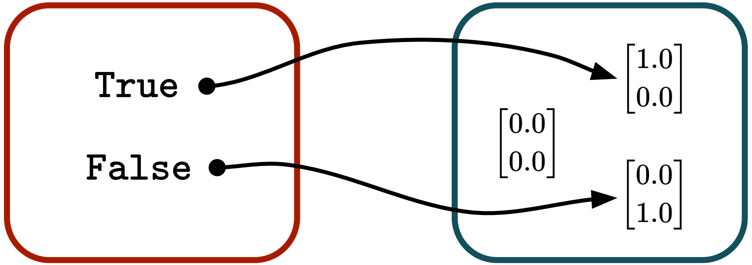
Step 1: Describe **`True`** as a vector using a fully abstract map.

$$\texttt{True} \mapsto \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}$$

Definition: A fully abstract map $[\![-]\!]$ preserves equivalences and distinctions i.e.

$$f \simeq_s g \implies [\![f]\!] \simeq_t [\![g]\!]$$

$$f \not\simeq_s g \implies [\![f]\!] \not\simeq_t [\![g]\!]$$



1. What properties should this mapping have?

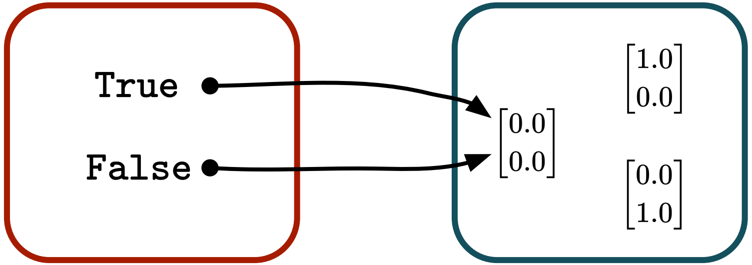
Step 1: Describe **`True`** as a vector using a fully abstract map.

$$\texttt{True} \mapsto \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}$$

Definition: A fully abstract map $[\![-]\!]$ preserves equivalences and distinctions i.e.

$$f \simeq_s g \implies [\![f]\!] \simeq_t [\![g]\!]$$

$$f \not\simeq_s g \implies [\![f]\!] \not\simeq_t [\![g]\!]$$



2. What’s the domain of this map?

2. What's the domain of this map?

Bootstrapping in a language of thought: A formal model of numerical concept learning

Steven T. Piantadosi [a,*], Joshua B. Tenenbaum [b], Noah D. Goodman [c]

[a] *Department of Brain and Cognitive Sciences, University of Rochester, United States*
[b] *Department of Brain and Cognitive Sciences, MIT, United States*
[c] *Department of Psychology, Stanford University, United States*

One-knower

```
λ S . (if (singleton? S)
          "one"
          undef)
```

Two-knower

```
λ S . (if (singleton? S)
          "one"
          (if (doubleton? S)
              "two"
              undef))
```

Three-knower

```
λ S . (if (singleton? S)
          "one"
          (if (doubleton? S)
              "two"
              (if (tripleton? S)
                  "three"
                  undef))
```

CP-knower

```
λ S . (if (singleton? S)
          "one"
          (next (L (set-difference S
                                   (select S)))))
```

2. What's the domain of this map?

Bootstrapping in a language of thought: A formal model of numerical concept learning

Steven T. Piantadosi [a,*], Joshua B. Tenenbaum [b], Noah D. Goodman [c]

[a] *Department of Brain and Cognitive Sciences, University of Rochester, United States*
[b] *Department of Brain and Cognitive Sciences, MIT, United States*
[c] *Department of Psychology, Stanford University, United States*

One-knower

```
λ S . (if (singleton? S)
          "one"
          undef)
```

Two-knower

```
λ S . (if (singleton? S)
          "one"
          (if (doubleton? S)
              "two"
              undef))
```

Three-knower

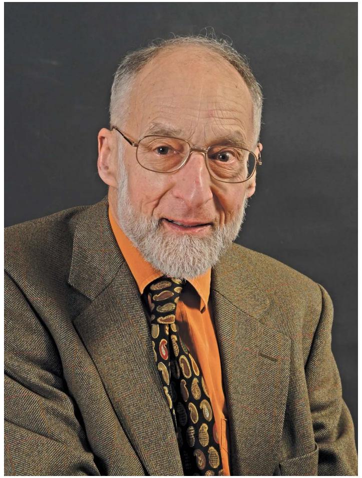
```
λ S . (if (singleton? S)
          "one"
          (if (doubleton? S)
              "two"
              (if (tripleton? S)
                  "three"
                  undef))
```

CP-knower

```
λ S . (if (singleton? S)
          "one"
          (next (L (set-difference S
                                   (select S)))))
```

Programming languages are sufficiently expressive to describe the abstract relations that cognitive scientists are interested in!

Robin Milner, 1991 Turing Award



"...he formulated and strongly advanced full abstraction, ~~the study of the relationship between operational and denotational semantics."~~
a solution to the mapping problem.

1. What properties should this mapping have?
   Answer: It should be fully abstract

2. What's the domain of this map?
   Answer: A programming language

But...

```
Modelling terms
---------------

                                                 (Bool-I1)
[[Γ ⊢ true : Bool]] ≔  tru : [[Γ]]→ [[Bool]]

                                                 (Bool-I2)
[[Γ ⊢ false : Bool]] ≔  fls   : [[Γ]]→ [[Bool]]

[[Γ₁ ⊢ b  : Bool]] ≔  f : [[Γ₁]] → [[Bool]]
[[1,Γ₂ ⊢ s  : A]] ≔  g : [[1,Γ₂]] → [[A]]
[[1,Γ₂ ⊢ t  : A]] ≔  h : [[1,Γ₂]] → [[A]]
                                                 (Bool-E)
[[Γ₁,Γ₂ ⊢ if b then s else t ]] ≔ (π₁ + π₂) ∘ (g ⊕ h) ∘ δₗ ∘ (f ⊗ id)  : [[Γ₁,Γ₂]]→[[A]]

                                                 (Unit-I)
[[Γ ⊢ unit : ()]] ≔  ⟨⟩ : [[Γ]]→ [[()]]

                                                 (Var-I)
[[x:A ⊢ x : A]] ≔  id : [[A]]→ [[A]]

[[Γ₁ ⊢ s : A]] ≔  f : [[Γ₁]] → [[A]]
[[Γ₂ ⊢ t : B]] ≔  g : [[Γ₂]] → [[B]]
                                                 (MProd-I)
[[Γ₁,Γ₂ ⊢ ⟨s ⊗ t⟩ : A ⊗ B]] ≔  f ⊗ g :[[Γ₁,Γ₂]] → [[A ⊗ B]]

[[Γ,x:A ⊢ t : B]] ≔  f :[[Γ,x:A]]→[[B]]
                                                 (LFun-I)
[[Γ ⊢ λx:A.t : A ⊸ B]] ≔  cur f : [[Γ]]→ [[A ⊸ B]]

[[Γ₁ ⊢ s : 1]] ≔  f :[[Γ₁]]→[[1]]
[[Γ₂ ⊢ t : A]] ≔  g :[[Γ₂]]→[[A]]
                                                 (Unit-E)
[[Γ₁,Γ₂ ⊢ let () = s in t : A ]] ≔  l ∘ (f      [[A]]

[[Γ₁          ⊢ s : A ⊗ B]] ≔  f :[[
[[Γ₂,x:A,y:B ⊢ t : C]] ≔  g :[[
                                                 (MProd-E)
[[Γ₁,Γ₂ ⊢ let (x ⊗ y)      ≔  g ∘ a ∘ s ∘ (f ⊗ id) :[[Γ₁,Γ₂]]→[[C]]

[[Γ₁ ⊢ s : A          →[[A ⊸ B]]
[[Γ₂ ⊢ t           ]]→[[A]]
                                                 (LFun-E)
            B]] ≔ ev ∘ (f ⊗ g) :[[Γ₁,Γ₂]]→[[B]]
```

1 Syntax

Defining syntax of expressions, values, types, and contexts.

1.1 Expressions

$e ::= ()$		*unit*	(…ION FORMS)
$\mid$	x	*variables*	
$\mid$	$\lambda x{:}A.e$	*function*	
$\mid$	$\langle e_1 \otimes e_2 \rangle$	*mu…*	
$\mid$	$\mathtt{inj}_1\ e$		
$\mid$	$\mathtt{inj}_2\ e$	*…njection*	
$\mid$	$e_1 e_2$	*…pplication*	(ELIMINATION FORMS)
$\mid$	$\mathtt{case}\ e_1\ \mathtt{of}$ …	*case-unit*	
$\mid$	$\mathtt{case}$ … $\Rightarrow e_2$	*case-prod*	
	… $\Rightarrow e_1; y \Rightarrow e_2\}$	*case-sum*	

…lues

$v ::= ()$		*unit*
$\mid$	$\lambda x{:}A.e$	*function*
$\mid$	$\langle e_1 \otimes e_2 \rangle$	*multiplicative product*
$\mid$	$\mathtt{inj}_1\ e$	*first injection*
$\mid$	$\mathtt{inj}_2\ e$	*second injection*

1.3 Types

$A, B ::= 1$		*unit*
$\mid$	$A \multimap B$	*linear function*
$\mid$	$A \otimes B$	*multiplicative product*
$\mid$	$A \oplus B$	*additive sum*

1.4 Contexts

$\Gamma ::= \cdot$		*empty*
$\mid$	$\Gamma, x{:}A$	*binding*

UNDER CONSTRUCTION UNDER CONSTRUCTION UNDER CONSTRUCTION UNDER CONSTRUCTION UNDER CON

When do neurons* represent `True`?

When do neurons* represent **`True`**?

Step 1: Describe **`True`** as a vector using a fully abstract map.

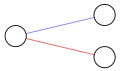
$$\texttt{True} \;\mapsto\; \lambda x.\, \lambda y.\, x \;\mapsto\; \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}$$

When do neurons* represent `True`?

Step 1: Describe `True` as a vector using a fully abstract map.

$$\texttt{True} \mapsto \lambda x.\, \lambda y.\, x \mapsto \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}$$

Step 2: Describe neurons as a vector.



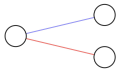
$$\mapsto \begin{bmatrix} 3.98 \\ 0.3 \end{bmatrix}$$

When do neurons* represent **`True`**?

Step 1: Describe **`True`** as a vector using a fully abstract map.

$$\texttt{True} \mapsto \lambda x.\, \lambda y.\, x \mapsto \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}$$

Step 2: Describe neurons as a vector.



$$\mapsto \begin{bmatrix} 3.98 \\ 0.3 \end{bmatrix}$$

Step 3: Relate the descriptions.

$$Corr^{*}\left(\begin{bmatrix} 3.98 \\ 0.3 \end{bmatrix}, \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}\right) \approx \underline{0.99}$$