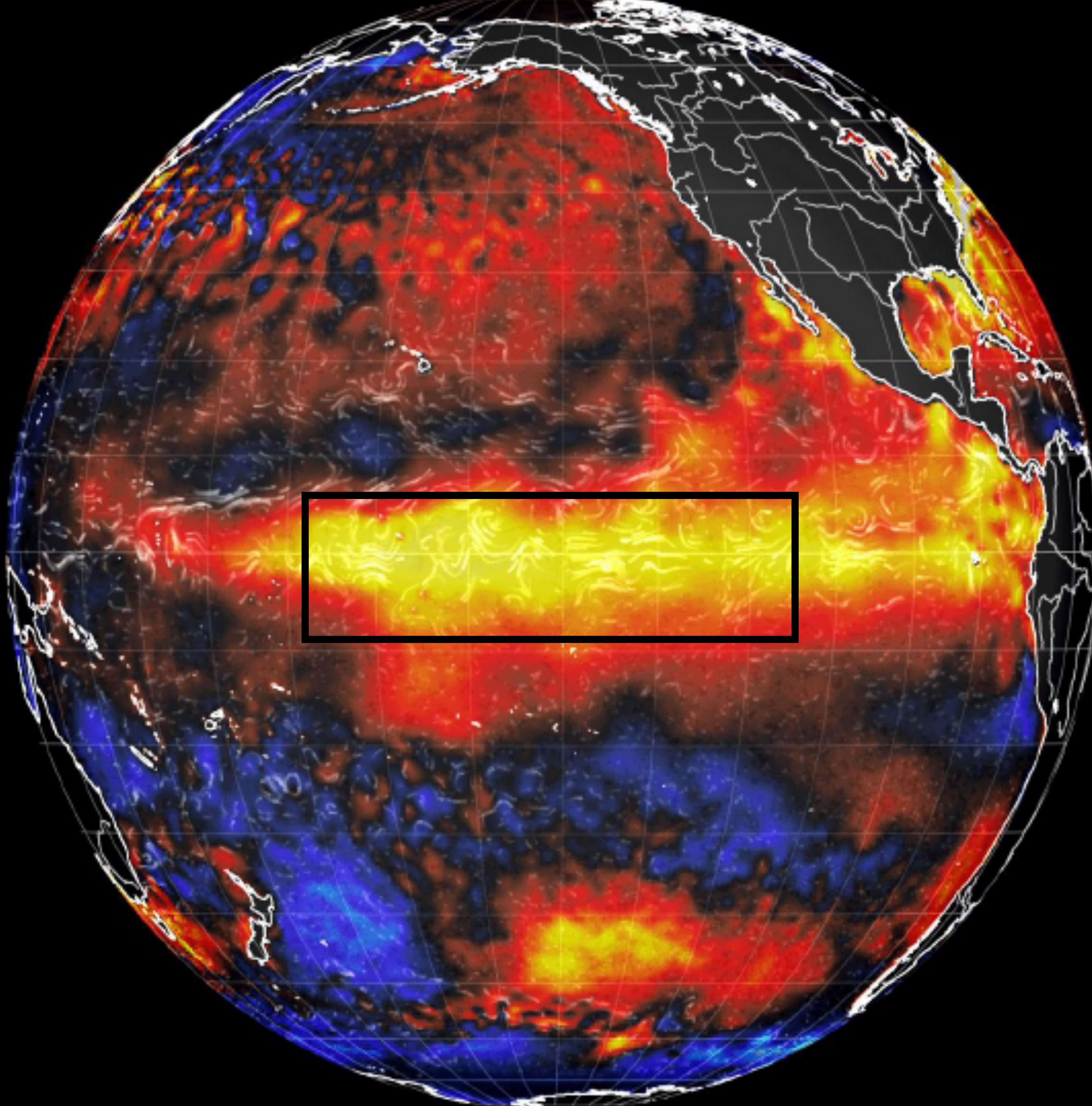
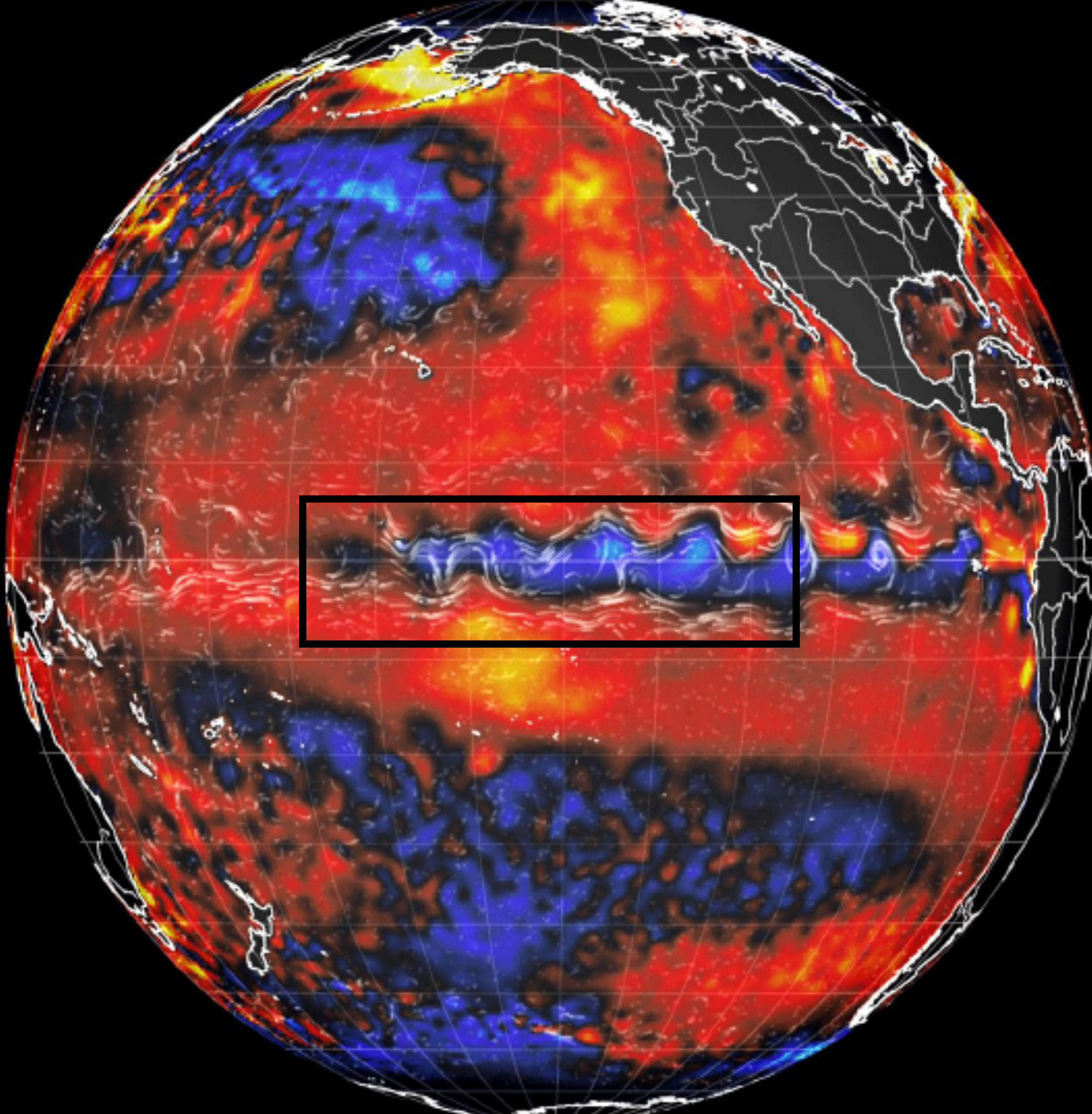
Caltech

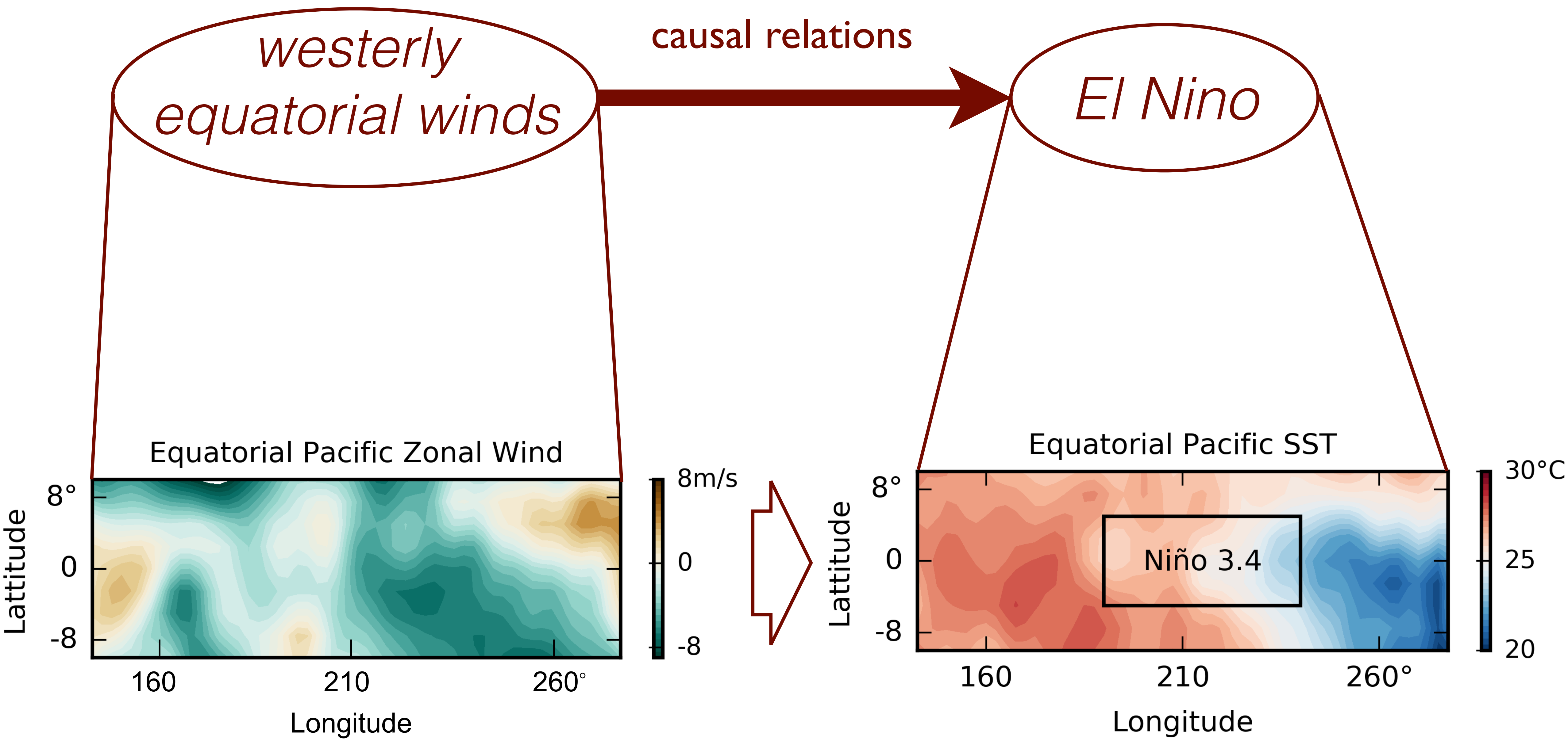
Causal Emergence:
When Distortions in the Map Obscure the Territory

Frederick Eberhardt & Lin Lin Lee





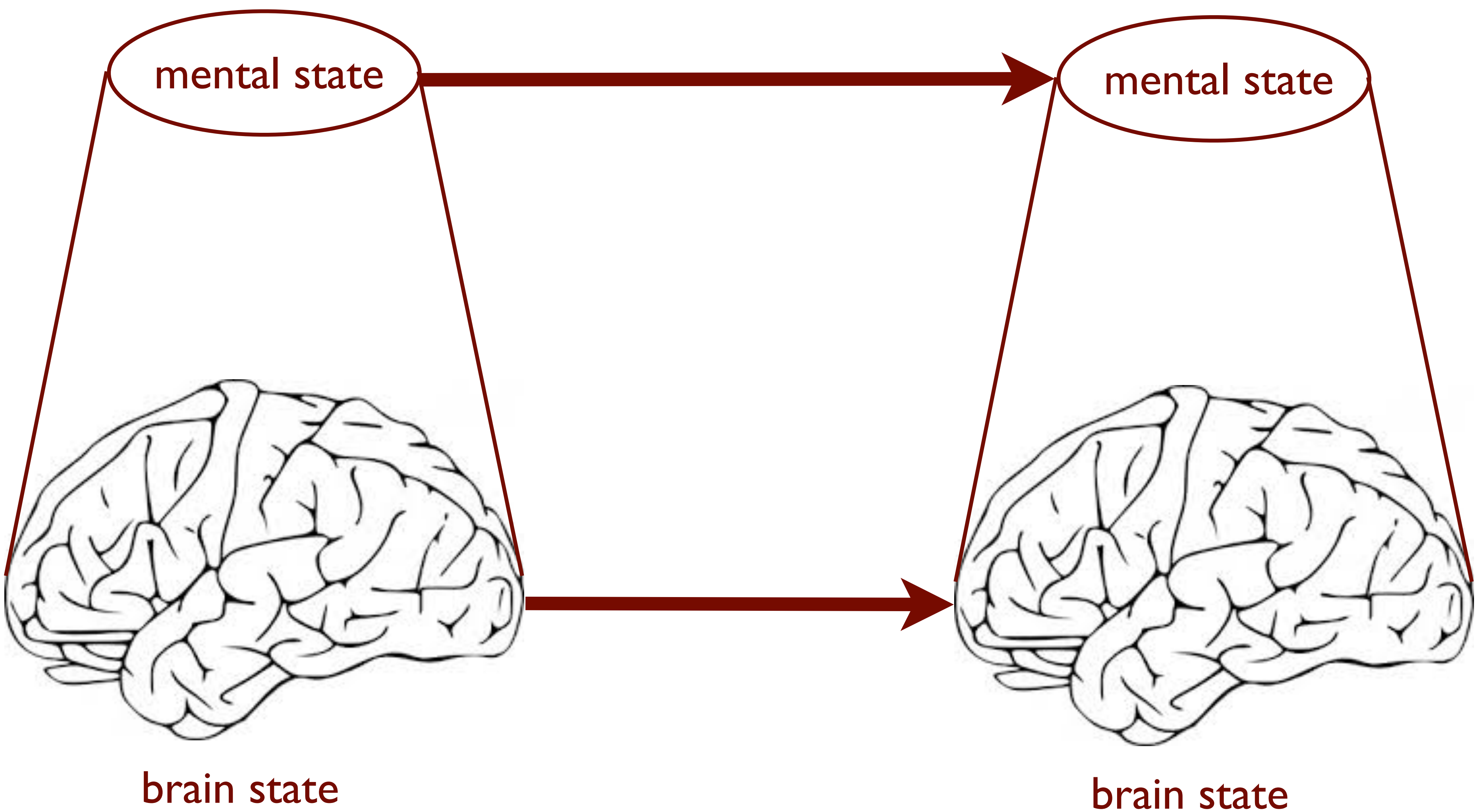
Causal Representation Learning



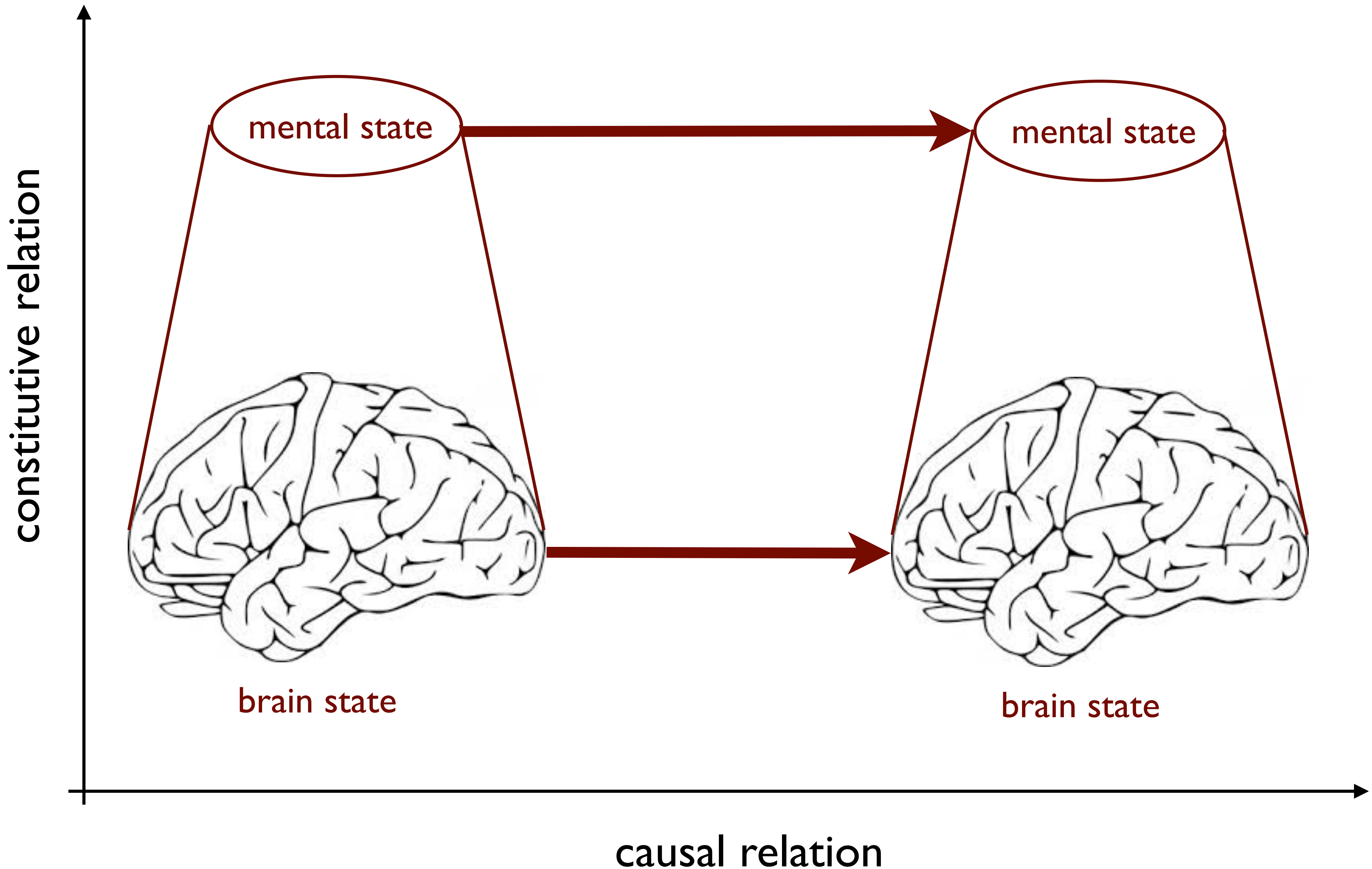
wind

sea surface temperature

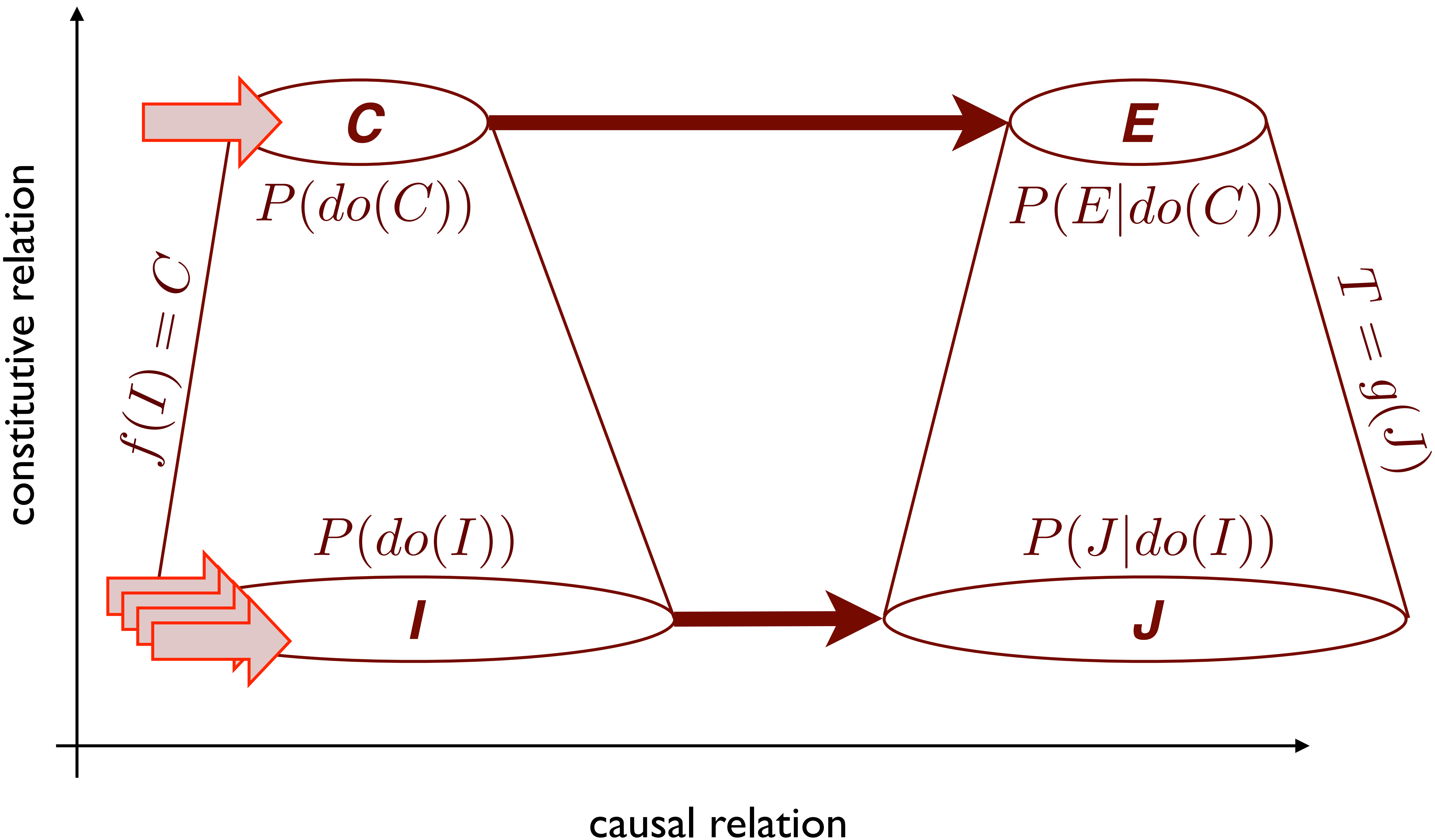
Mental Causation: Psychology vs. Neuroscience



Mental Causation: Psychology vs. Neuroscience



Micro- and Macro Causal Description

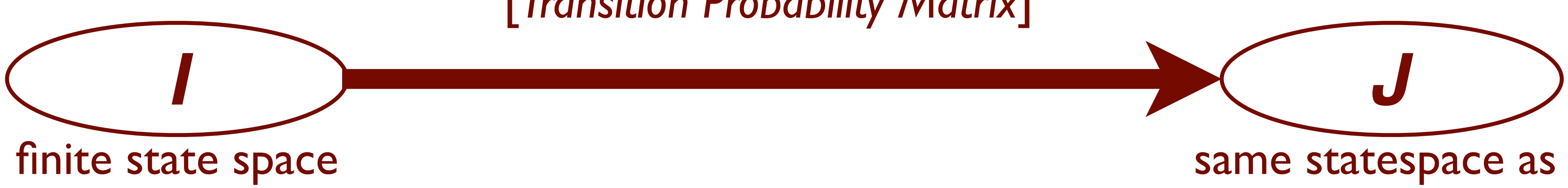


Hoel (2017): *When the Map is Better than the Territory*

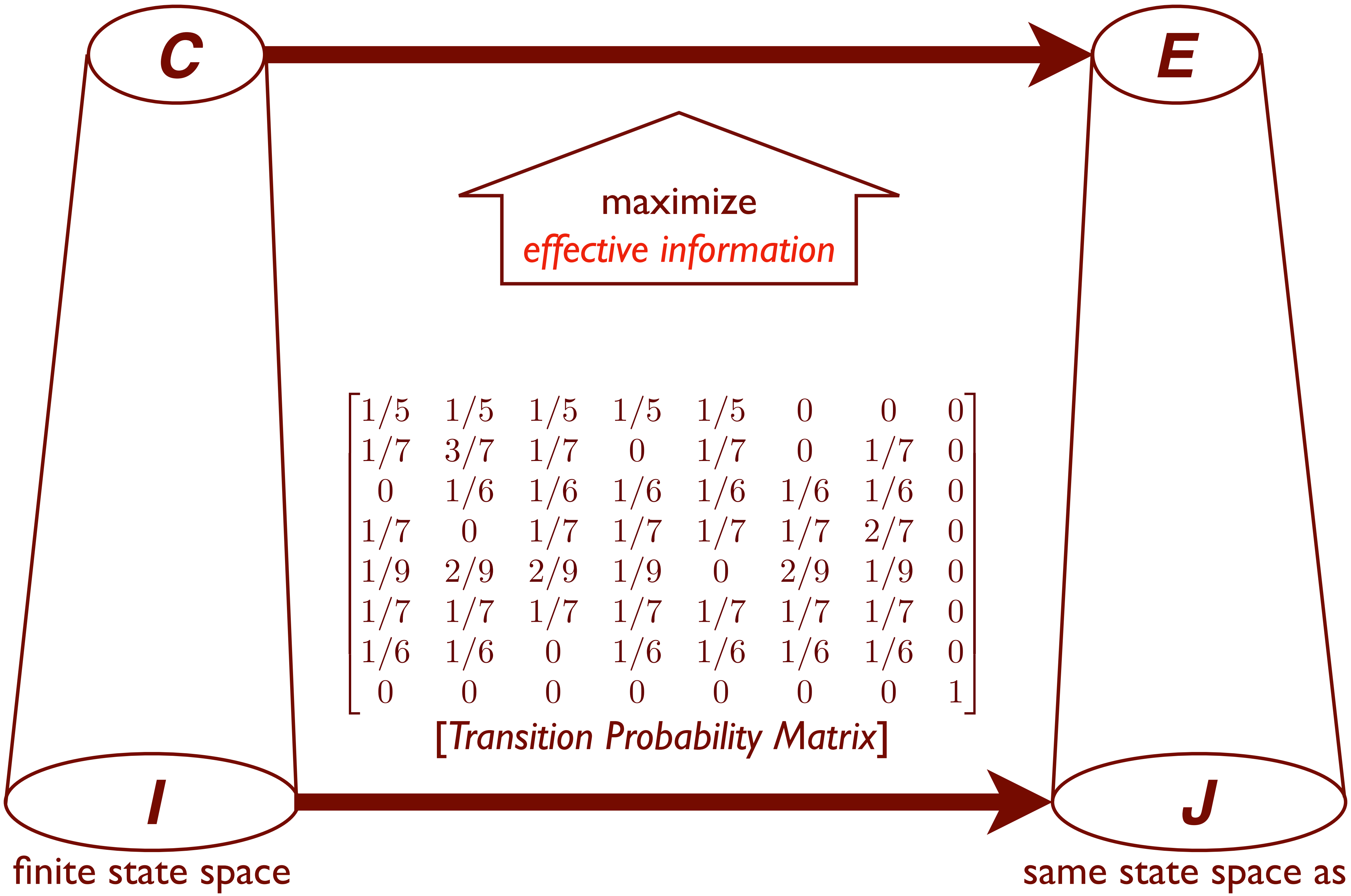
Hoel (2017): *When the Map is Better than the Territory*

$$\begin{bmatrix}
1/5 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 0 & 0 \\
1/7 & 3/7 & 1/7 & 0 & 1/7 & 0 & 1/7 & 0 \\
0 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 0 \\
1/7 & 0 & 1/7 & 1/7 & 1/7 & 1/7 & 2/7 & 0 \\
1/9 & 2/9 & 2/9 & 1/9 & 0 & 2/9 & 1/9 & 0 \\
1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\
1/6 & 1/6 & 0 & 1/6 & 1/6 & 1/6 & 1/6 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}$$

[*Transition Probability Matrix*]



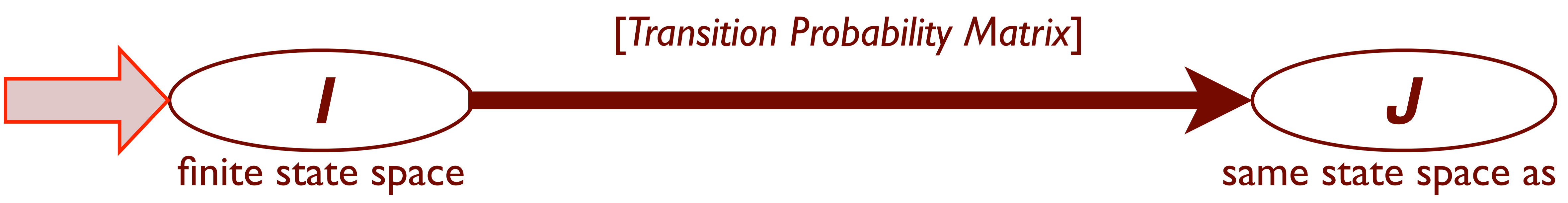
Hoel (2017): *When the Map is Better than the Territory*



$$\begin{bmatrix}
1/5 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 0 & 0 \\
1/7 & 3/7 & 1/7 & 0 & 1/7 & 0 & 1/7 & 0 \\
0 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 0 \\
1/7 & 0 & 1/7 & 1/7 & 1/7 & 1/7 & 2/7 & 0 \\
1/9 & 2/9 & 2/9 & 1/9 & 0 & 2/9 & 1/9 & 0 \\
1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\
1/6 & 1/6 & 0 & 1/6 & 1/6 & 1/6 & 1/6 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}$$

[*Transition Probability Matrix*]

Effective Information EI



Effective Information EI

intervention distribution

$$P(do(I)) = MaxEnt(I)$$

effect distribution

$$E(J) = \frac{1}{n}\sum_{I} P(J|do(I))$$

Effective Information EI

Difference between effect of
specific intervention and (maxent)
average intervention:

$$P(J|do(I=i)) \quad \text{vs.} \quad E(J)$$

intervention distribution

$$P(do(I)) = MaxEnt(I)$$

effect distribution

$$E(J) = \frac{1}{n}\sum_I P(J|do(I))$$

[*Transition Probability Matrix*]

finite state space

same state space as I

Effective Information EI

$$EI(I \to J) = \sum_I P(do(I)) D_{KL}(P(J|do(I))||E(J))$$

KL-divergence

Difference between effect of specific intervention and (maxent) average intervention:

$$P(J|do(I=i)) \quad \text{vs.} \quad E(J)$$

intervention distribution

$$P(do(I)) = MaxEnt(I)$$

effect distribution

$$E(J) = \frac{1}{n}\sum_I P(J|do(I))$$

[*Transition Probability Matrix*]

finite state space

same state space as I

*this slide has been corrected for a typo that was in the original

Effective Information EI

$$EI(I \rightarrow J) = I(I_{maxEnt}, J_E)$$

mutual information between maxEnt cause and effect

$$= \sum_I P(do(I)) \underbrace{D_{KL}(P(J|do(I))||E(J))}_{\text{KL-divergence}}$$

Difference between effect of specific intervention and (maxent) average intervention:

$$P(J|do(I=i)) \quad \text{vs.} \quad E(J)$$

intervention distribution

$$P(do(I)) = MaxEnt(I)$$

effect distribution

$$E(J) = \frac{1}{n} \sum_I P(J|do(I))$$

*this slide has been corrected for a typo that was in the original

What is great about Effective Information?

$$EI(I \to J) = I(I_{maxEnt}, J_E)$$

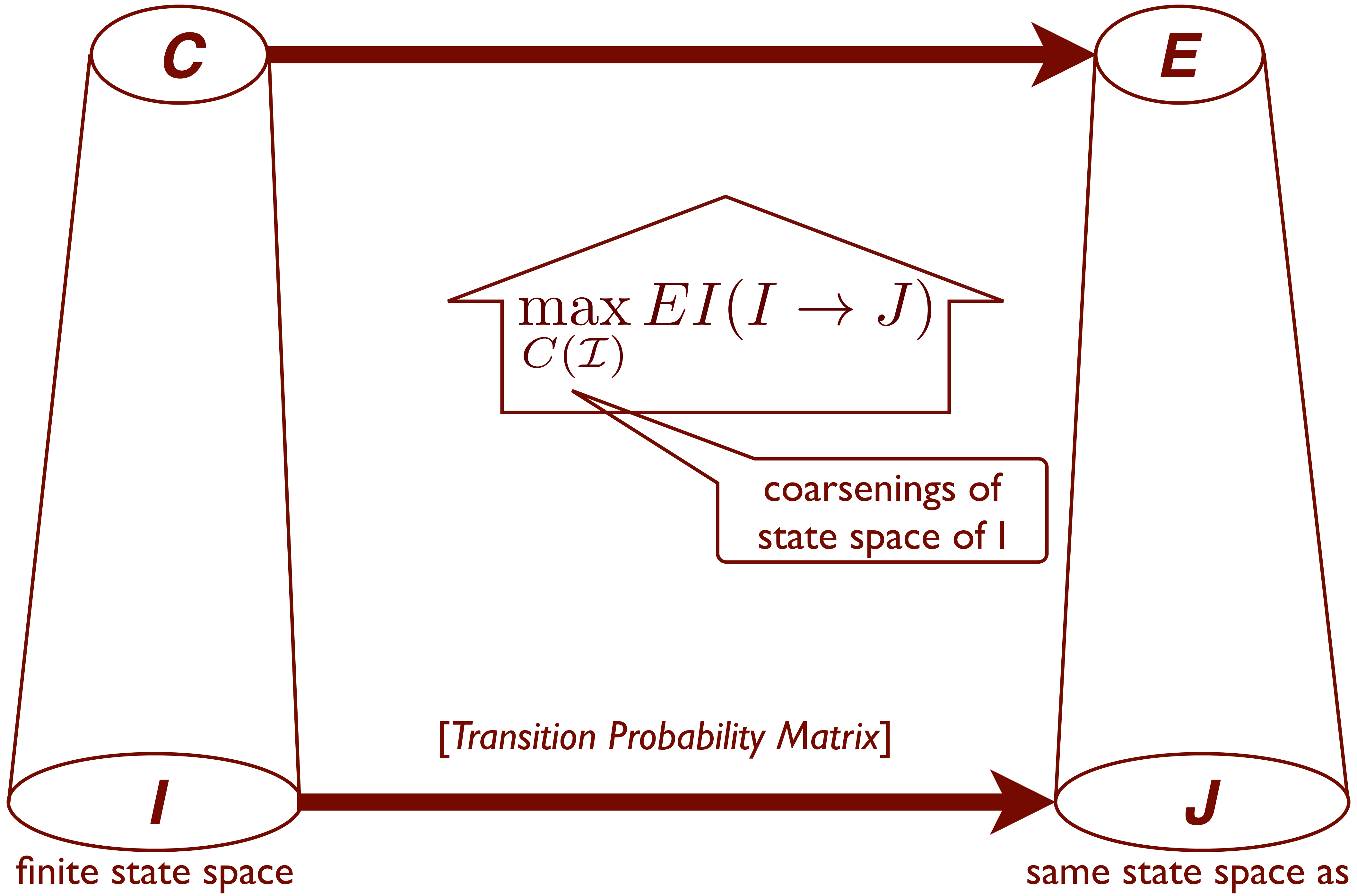
mutual information between maxEnt cause and effect

$$= \sum_I P(do(I)) D_{KL}(P(J|do(I)) || E(J))$$

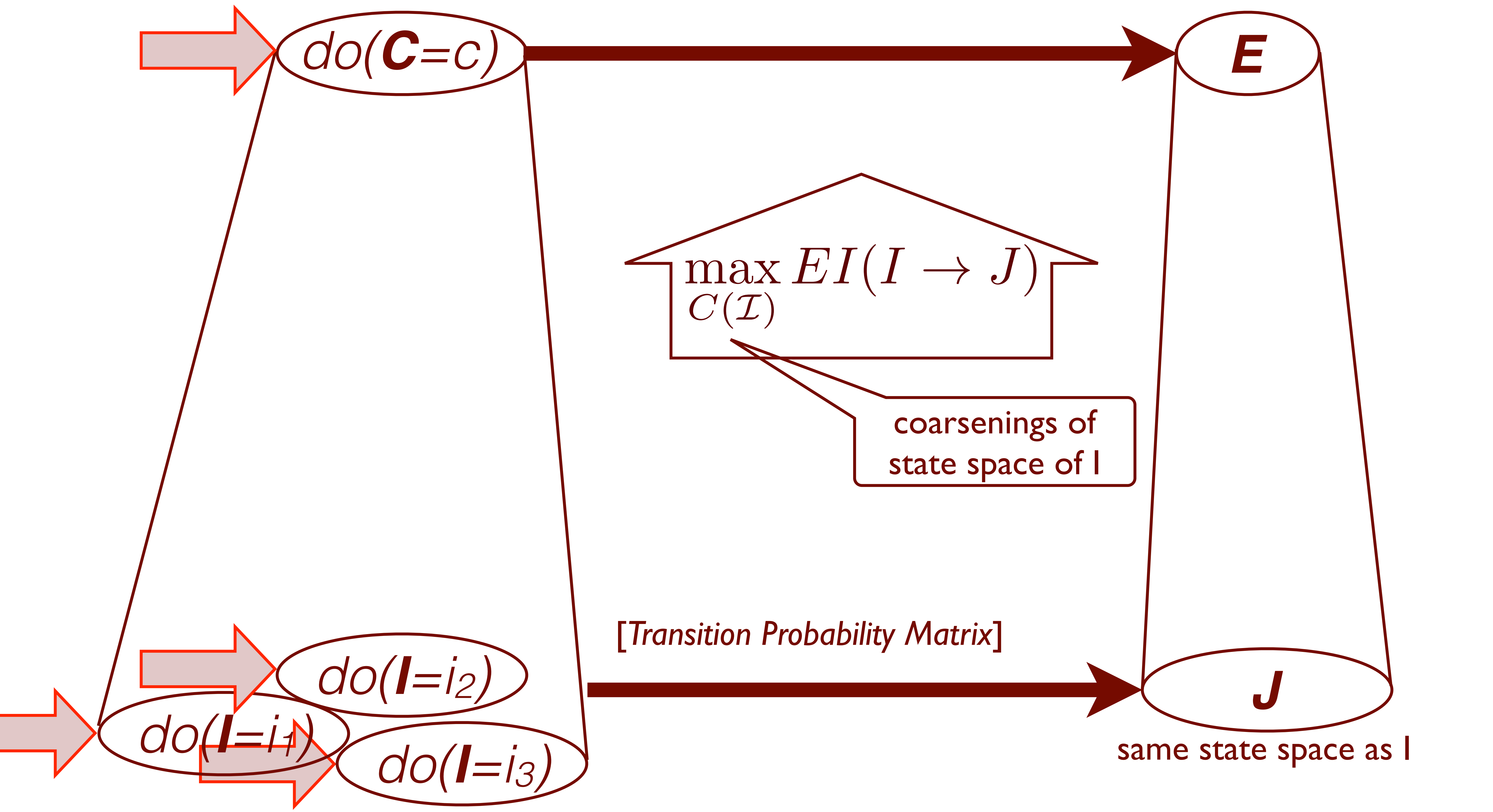
- directed information measure (defined in terms of interventions)
- connection between causality and information theory
- explores full cause space / is independent of observed P(I)
- [core feature of characterization of consciousness in Tononi's Integrated Information Theory of Consciousness]

*this slide has been corrected for a typo that was in the original

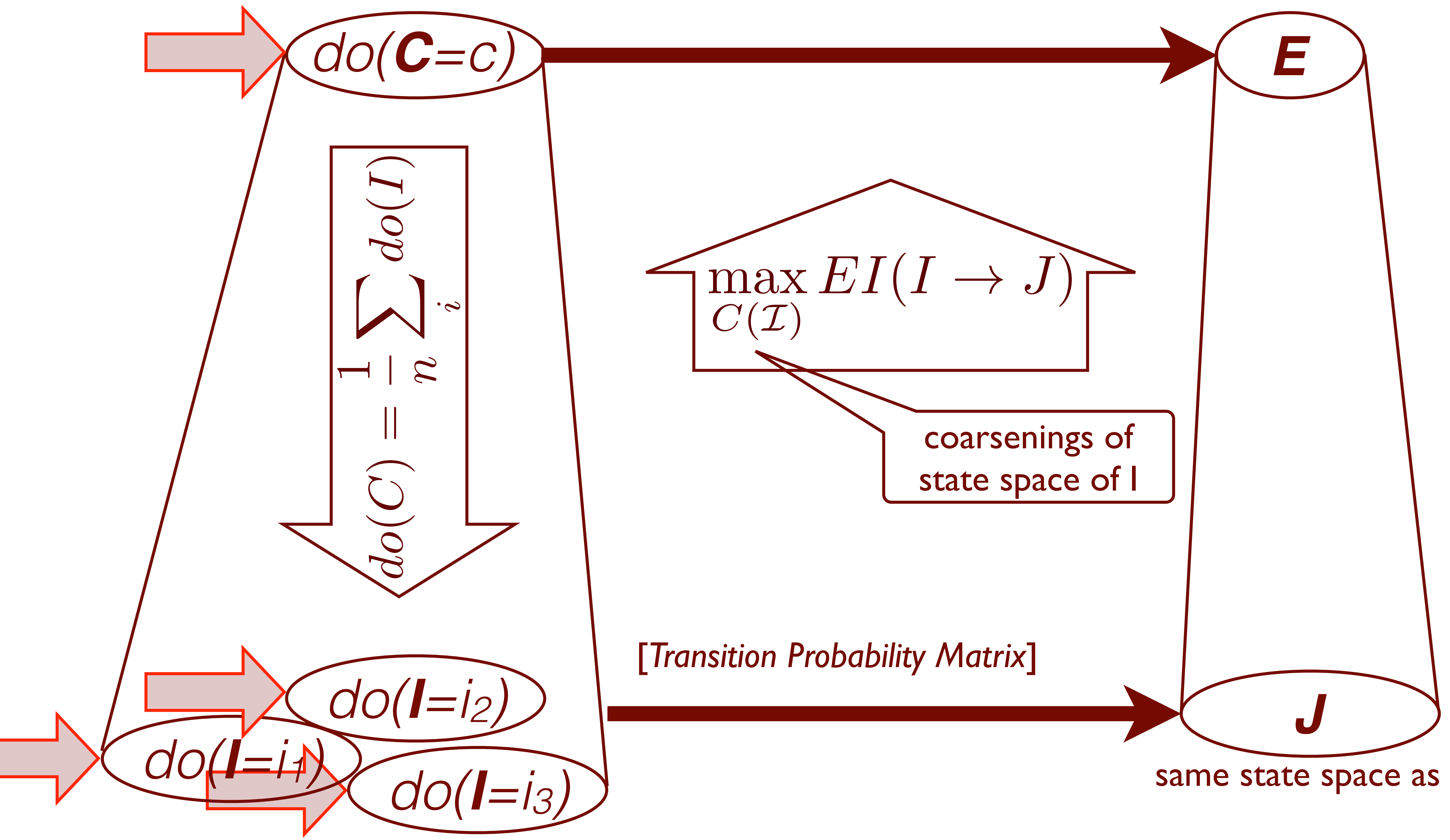
Hoel's Causal Emergence



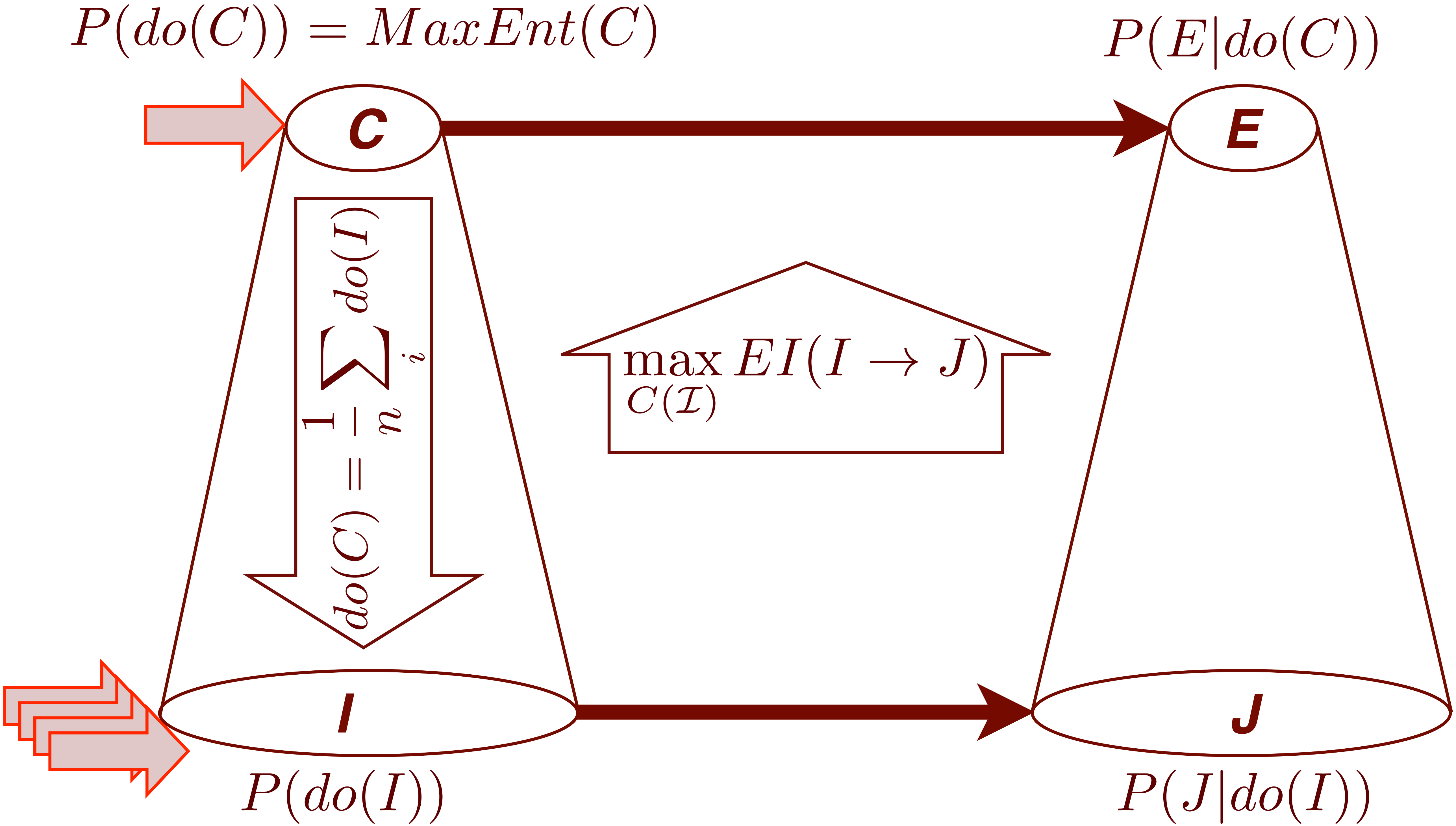
Hoel's Macro Intervention



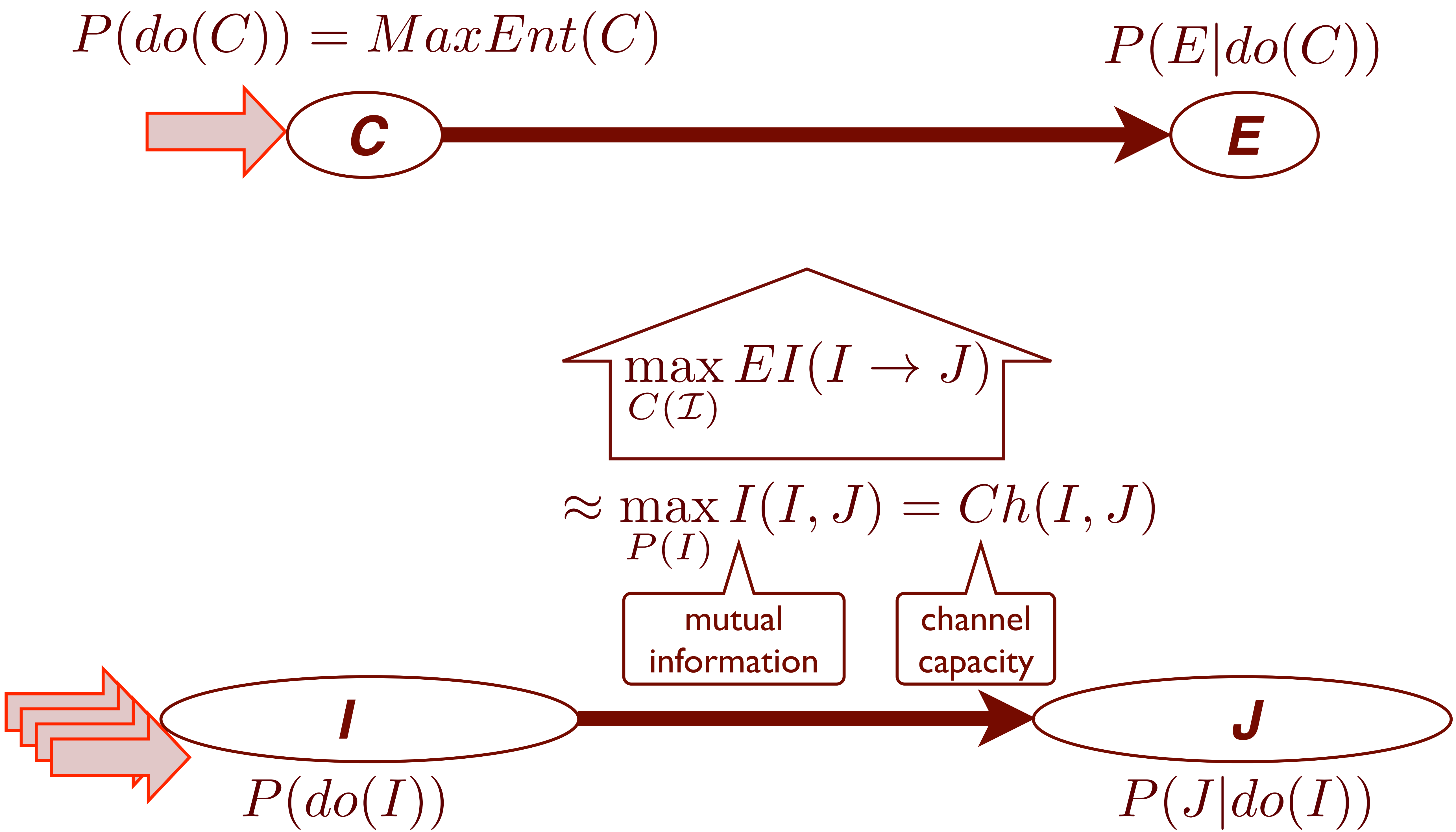
Hoel's Macro Intervention



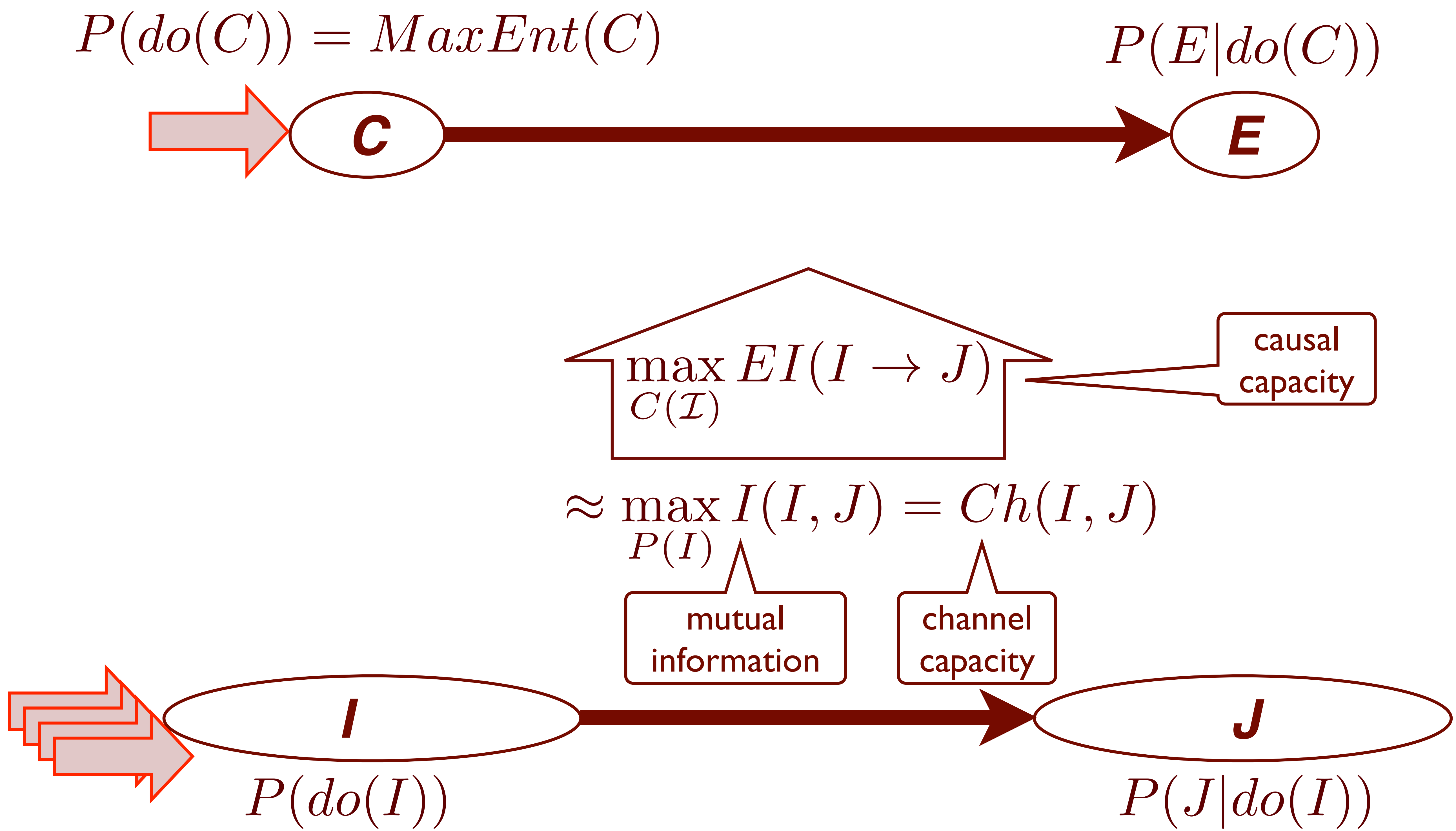
Causal Emergence in Hoel 2017



Causal Emergence in Hoel 2017



Causal Emergence in Hoel 2017



Causal Emergence in Hoel 2017

$P(do(C)) = MaxEnt(C)$

$P(E|do(C))$

BUT, the EI-maximization is subject to:

- subset of possible intervention distributions
- identical state spaces for I and J that change with the coarsening

$$\max_{C(\mathcal{I})} EI(I \to J)$$

causal capacity

$$\approx \max_{P(I)} I(I,J) = Ch(I,J)$$

mutual information

channel capacity

$P(do(I))$

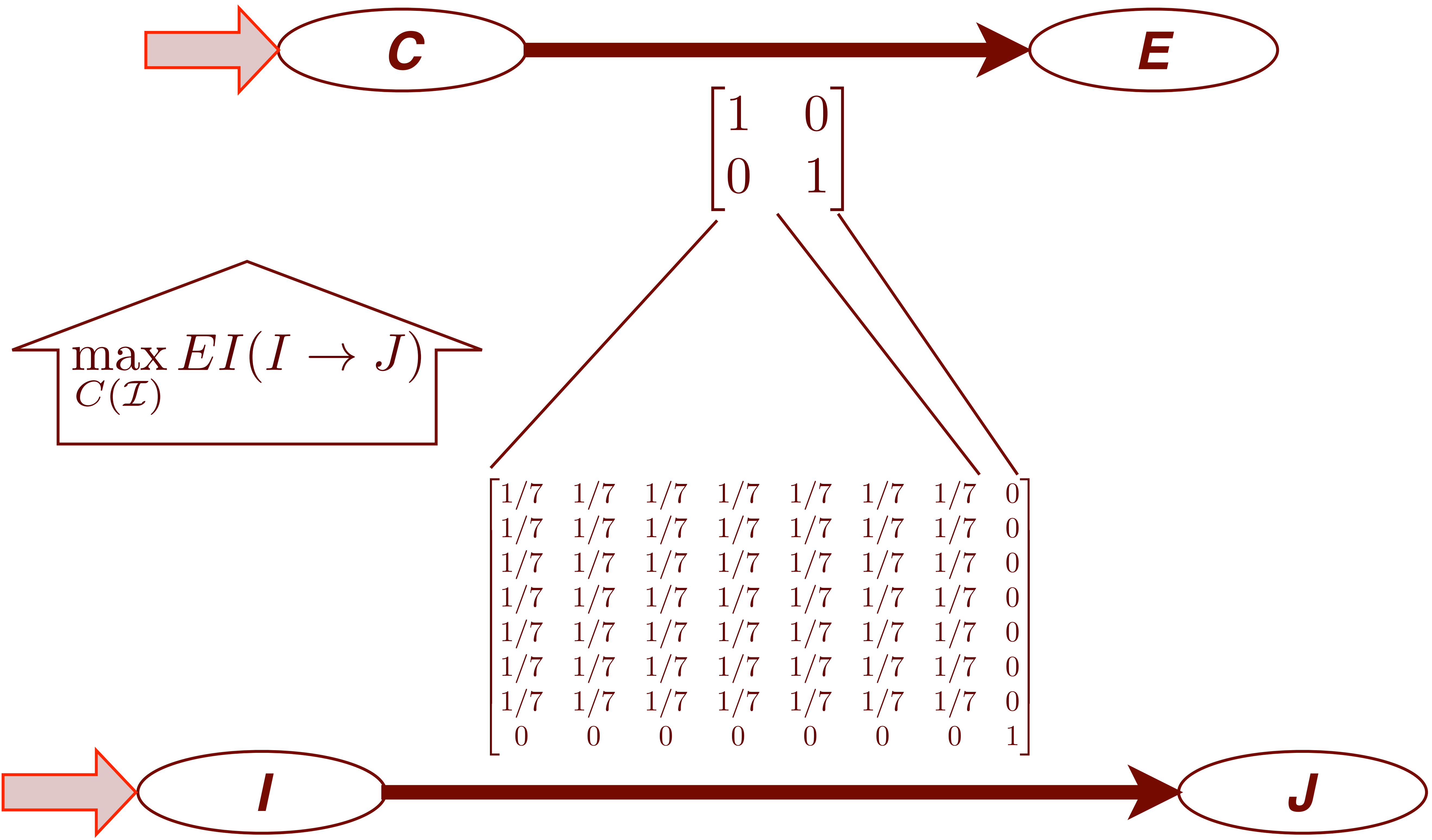
$P(J|do(I))$

Example 1

$$\max_{C(\mathcal{I})} EI(I \to J)$$

$$\begin{bmatrix} 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Example 1



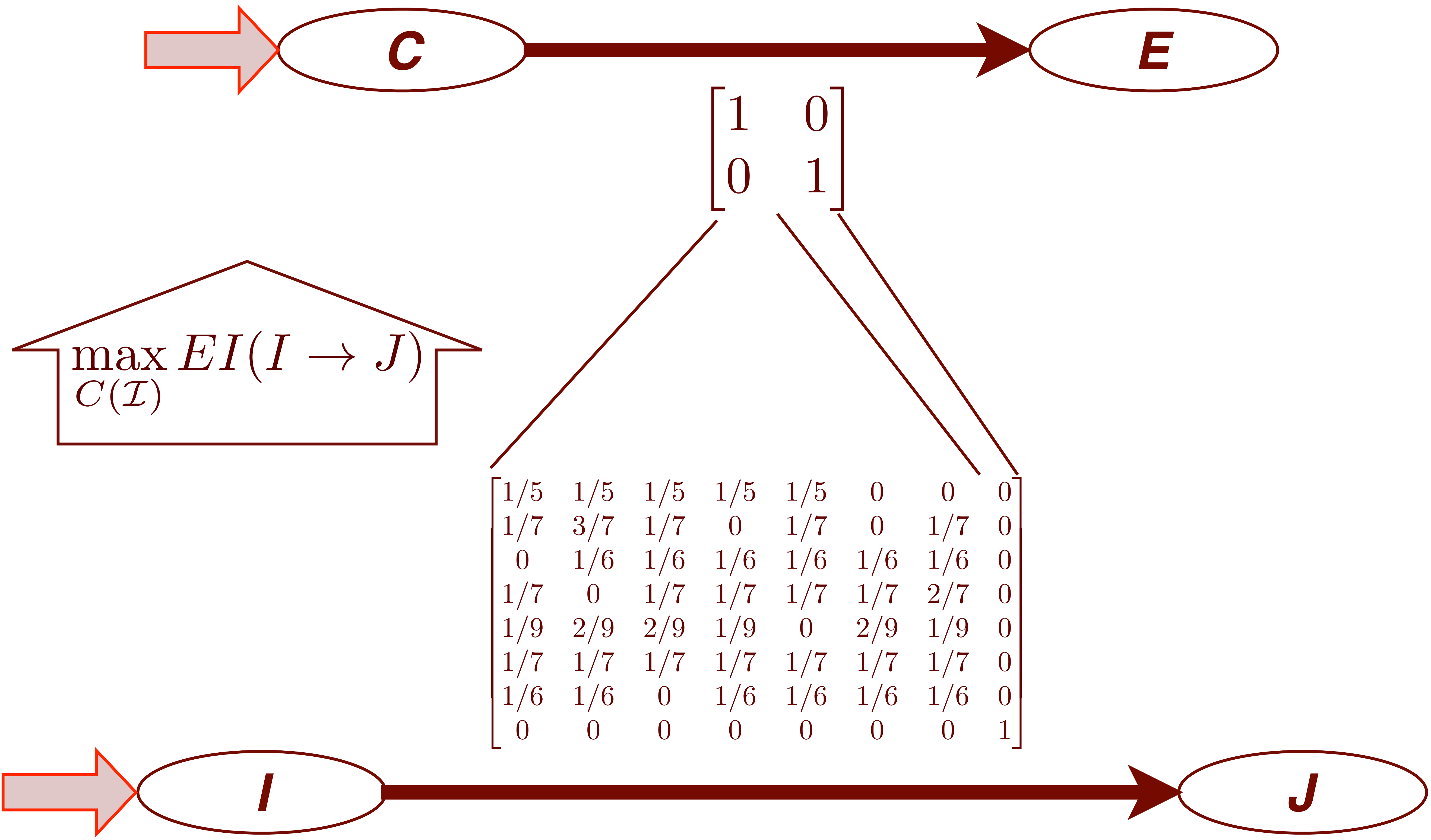
C
E
$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
$\max_{C(\mathcal{I})} EI(I \to J)$
$\begin{bmatrix} 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$
I
J

Example 2

$$\max_{C(\mathcal{I})} EI(I \to J)$$

$$\begin{bmatrix} 1/5 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 0 & 0 \\ 1/7 & 3/7 & 1/7 & 0 & 1/7 & 0 & 1/7 & 0 \\ 0 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 0 \\ 1/7 & 0 & 1/7 & 1/7 & 1/7 & 1/7 & 2/7 & 0 \\ 1/9 & 2/9 & 2/9 & 1/9 & 0 & 2/9 & 1/9 & 0 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\ 1/6 & 1/6 & 0 & 1/6 & 1/6 & 1/6 & 1/6 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Example 2

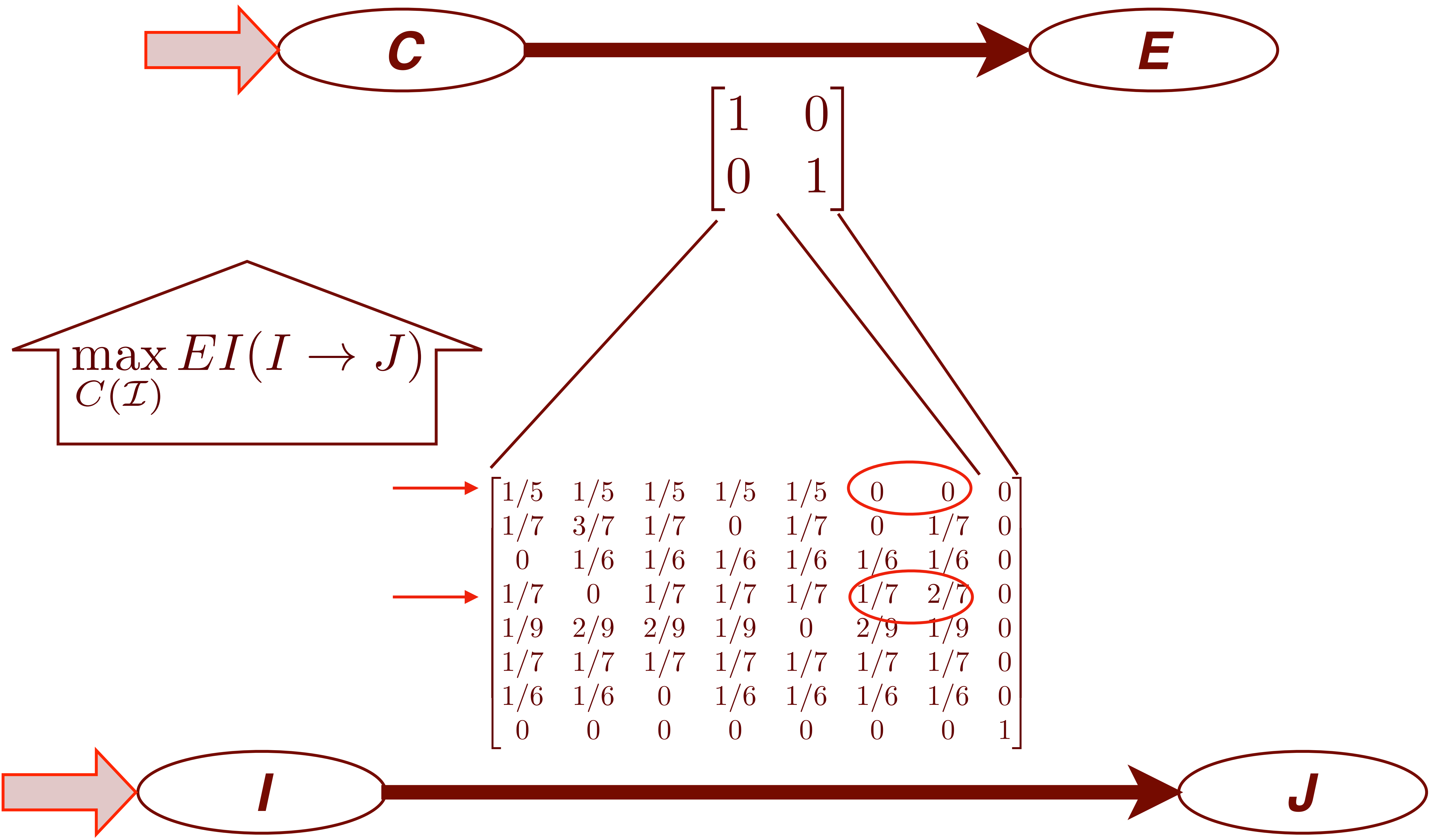


$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\max_{C(\mathcal{I})} EI(I \to J)$$

$$\begin{bmatrix}
1/5 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 0 & 0 \\
1/7 & 3/7 & 1/7 & 0 & 1/7 & 0 & 1/7 & 0 \\
0 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 0 \\
1/7 & 0 & 1/7 & 1/7 & 1/7 & 1/7 & 2/7 & 0 \\
1/9 & 2/9 & 2/9 & 1/9 & 0 & 2/9 & 1/9 & 0 \\
1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\
1/6 & 1/6 & 0 & 1/6 & 1/6 & 1/6 & 1/6 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}$$

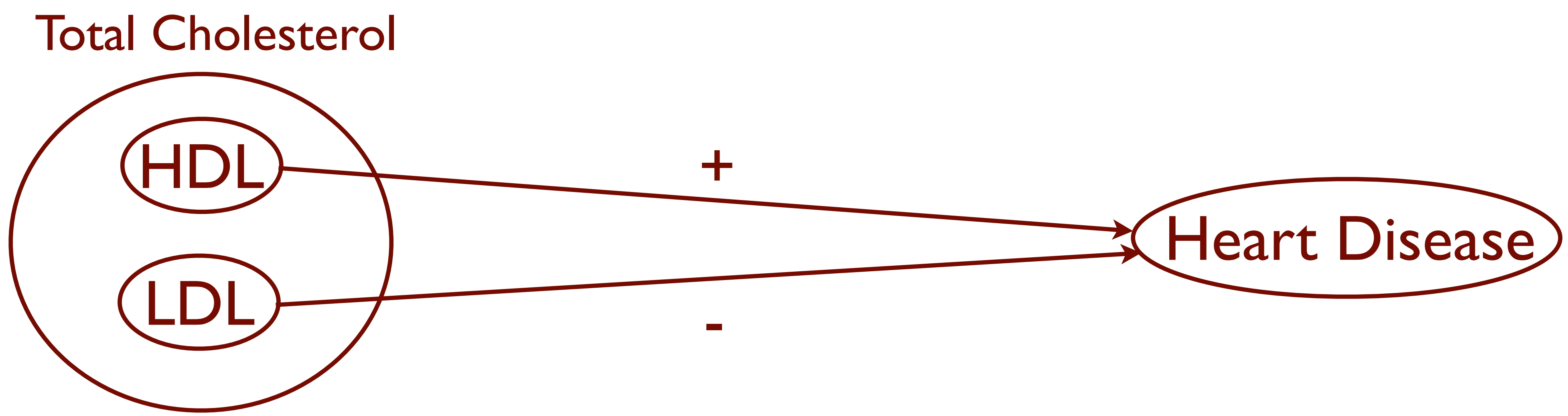
Example 2


C
E
$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$
$$\max_{C(\mathcal{I})} EI(I \to J)$$
$$\begin{bmatrix} 1/5 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 0 & 0 \\ 1/7 & 3/7 & 1/7 & 0 & 1/7 & 0 & 1/7 & 0 \\ 0 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 0 \\ 1/7 & 0 & 1/7 & 1/7 & 1/7 & 1/7 & 2/7 & 0 \\ 1/9 & 2/9 & 2/9 & 1/9 & 0 & 2/9 & 1/9 & 0 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\ 1/6 & 1/6 & 0 & 1/6 & 1/6 & 1/6 & 1/6 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$
I
J

Ambiguous Manipulation

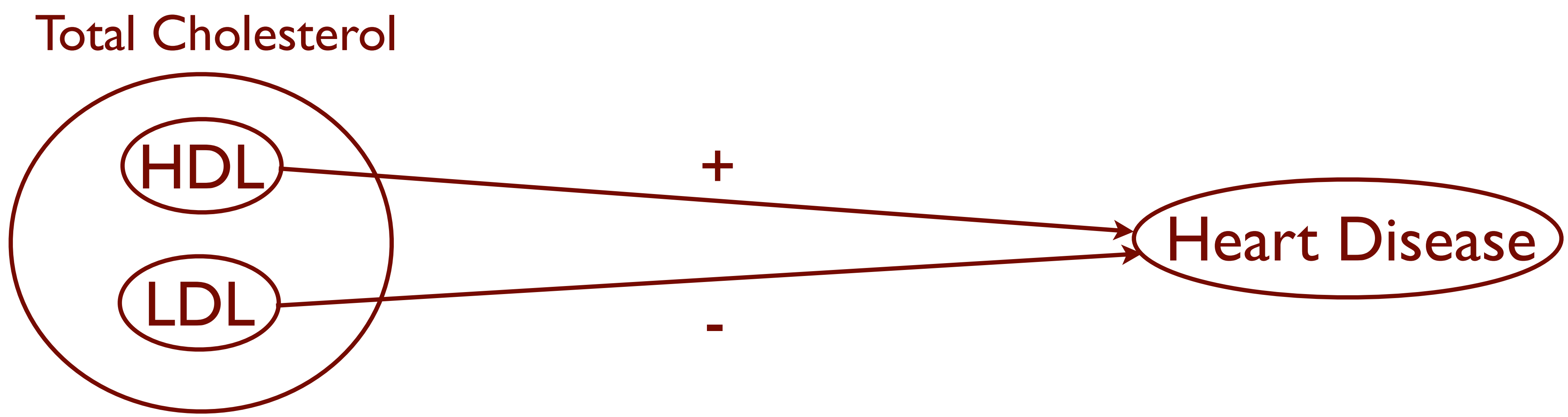
Spirtes & Scheines (2002)

Ambiguous Manipulation



Spirtes & Scheines (2002)

Ambiguous Manipulation



- the causal effect of *Total Cholesterol* on *Heart Disease* is **ambiguous**

➡ *Total Cholesterol* is over-aggregated, it cannot be described as a cause of *Heart Disease*

Spirtes & Scheines (2002)

Example 2

$C \longrightarrow E$

$$P(do(C)) \quad \begin{array}{c} 1/2 \\ 1/2 \end{array} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

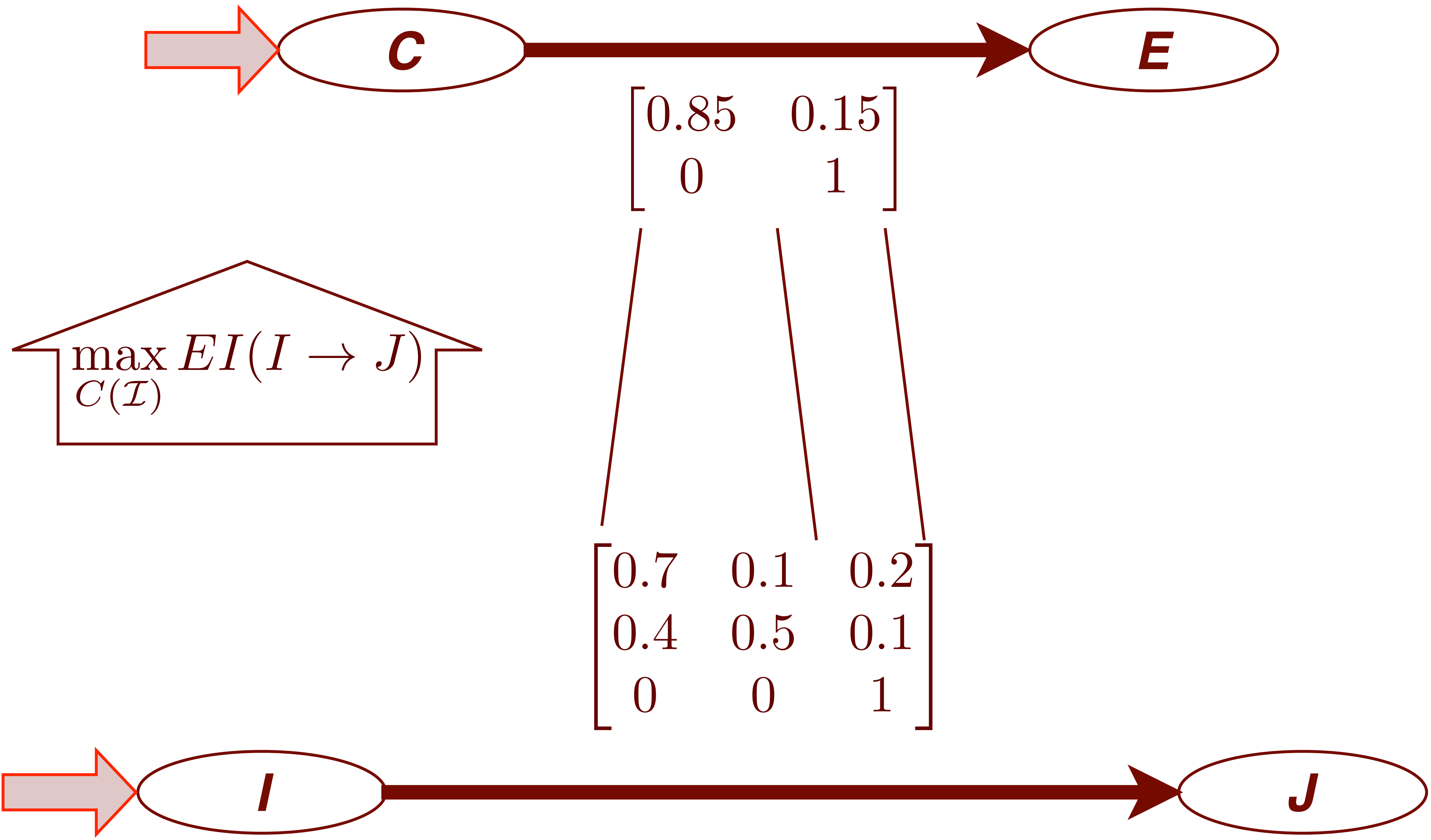
How different can the causal effects of two micro states be such that they still get mapped to the same macro state?

$$\max_{C(\mathcal{I})} EI(I \to J)$$

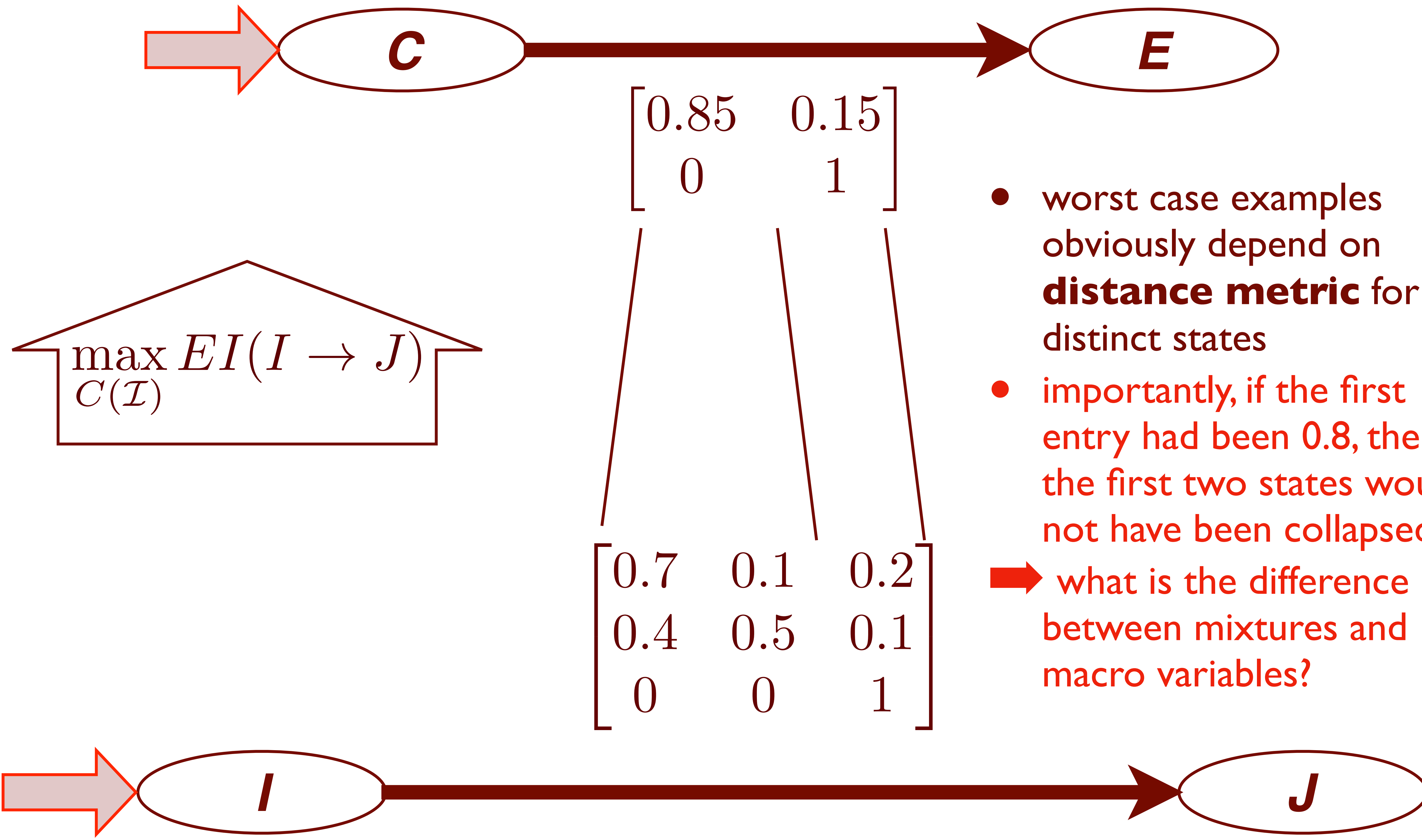
$$P(do(I)) \quad 1/2 \left[\begin{array}{c} 1/14 \\ 1/14 \\ 1/14 \\ 1/14 \\ 1/14 \\ 1/14 \\ 1/14 \end{array}\right. \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ 1/2 \end{array} \begin{bmatrix} 1/5 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 0 & 0 \\ 1/7 & 3/7 & 1/7 & 0 & 1/7 & 0 & 1/7 & 0 \\ 0 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 0 \\ 1/7 & 0 & 1/7 & 1/7 & 1/7 & 1/7 & 2/7 & 0 \\ 1/9 & 2/9 & 2/9 & 1/9 & 0 & 2/9 & 1/9 & 0 \\ 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 1/7 & 0 \\ 1/6 & 1/6 & 0 & 1/6 & 1/6 & 1/6 & 1/6 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$I \longrightarrow J$

Example 3: collapsing micro states with different causal effects



Example 3: collapsing micro states with different causal effects



$$\max_{C(\mathcal{I})} EI(I \to J)$$

$$\begin{bmatrix} 0.85 & 0.15 \\ 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 0.7 & 0.1 & 0.2 \\ 0.4 & 0.5 & 0.1 \\ 0 & 0 & 1 \end{bmatrix}$$

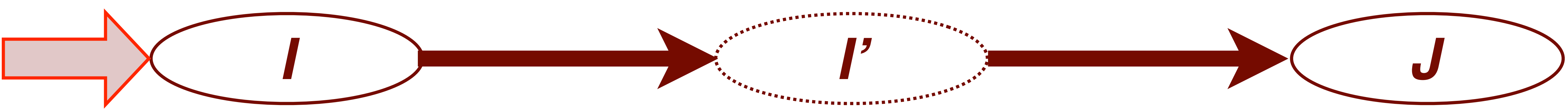
- worst case examples obviously depend on **distance metric** for distinct states
- importantly, if the first entry had been 0.8, then the first two states would not have been collapsed

➡ what is the difference between mixtures and macro variables?

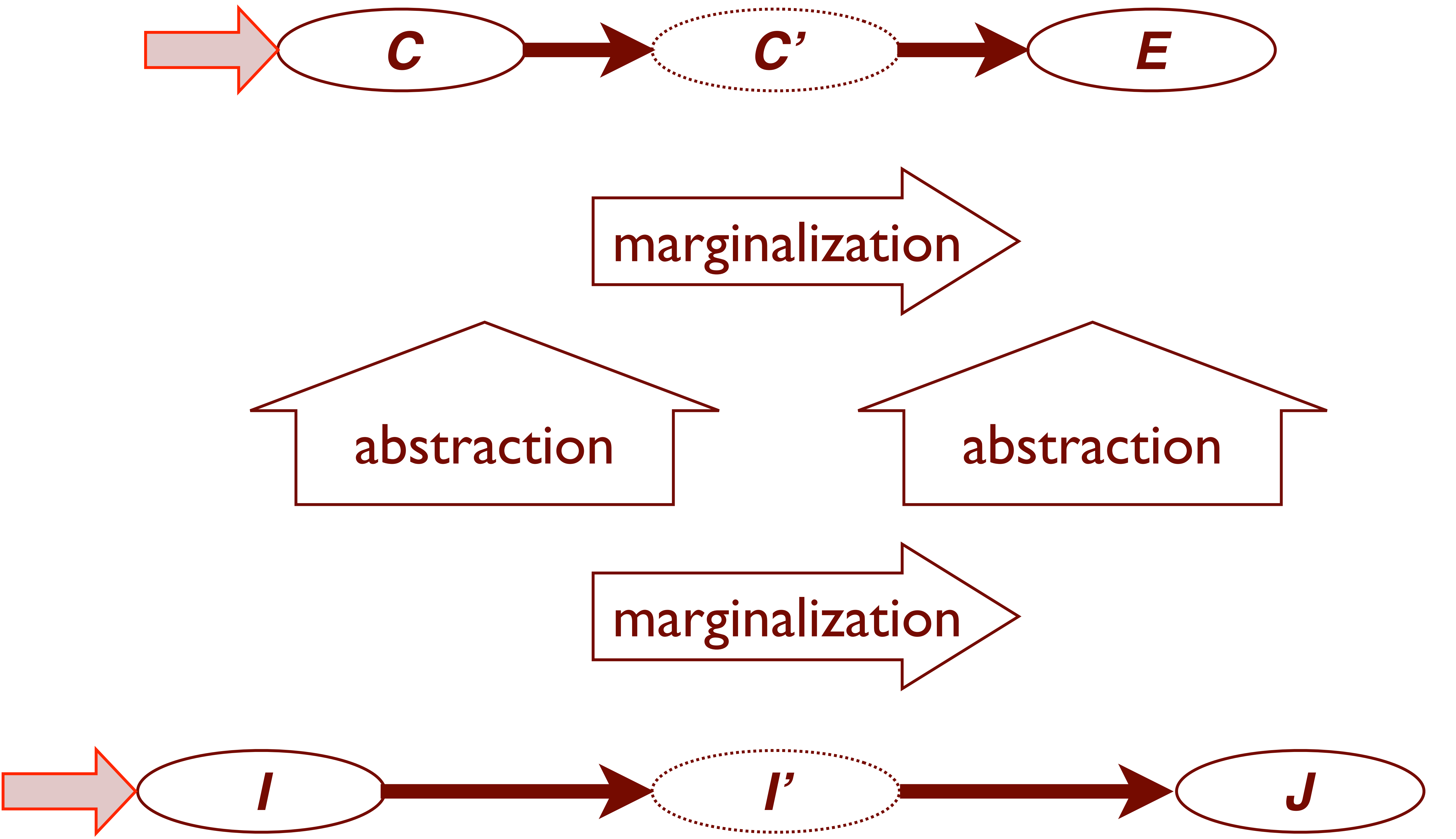
Marginalization

Marginalization

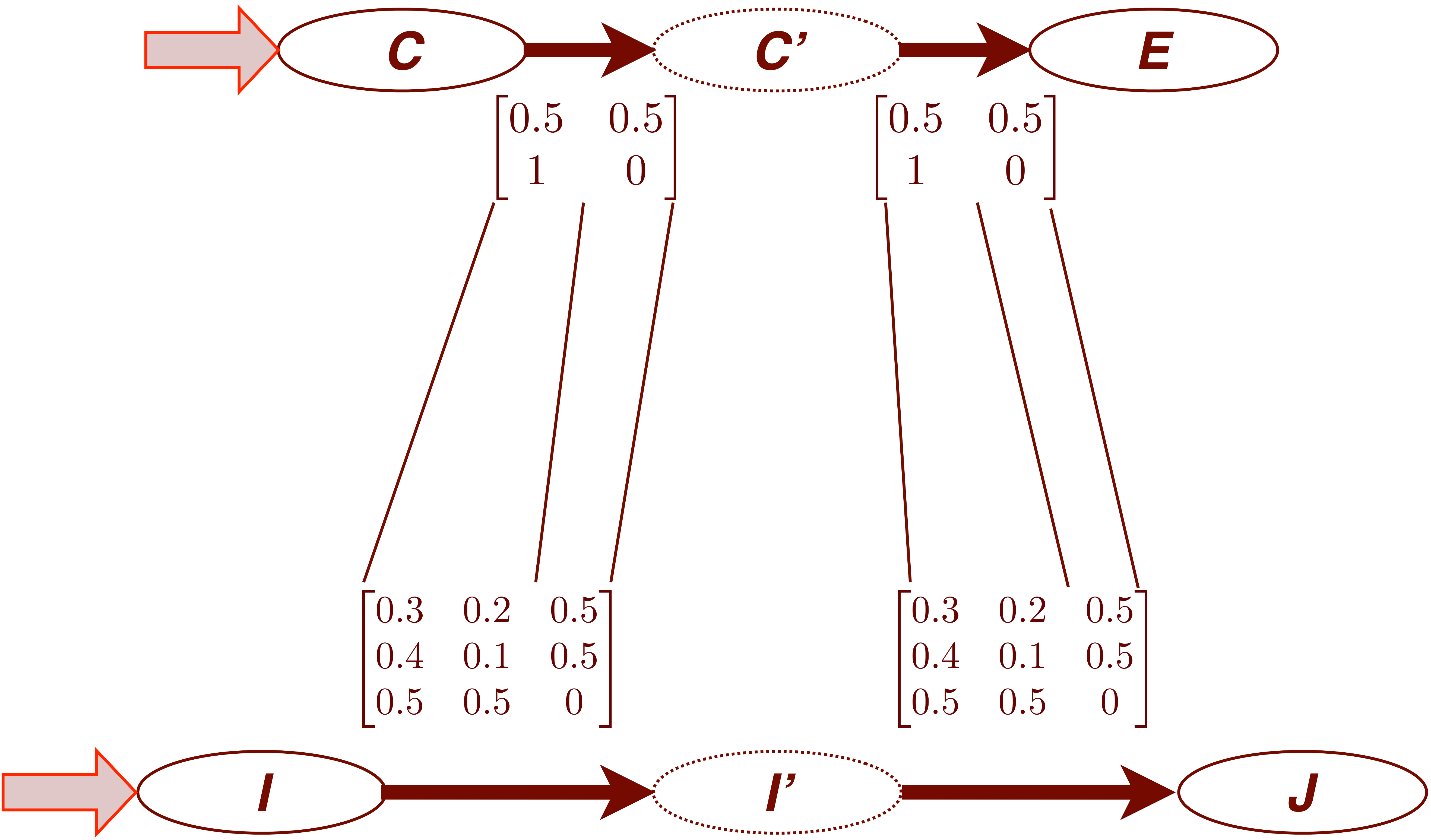
Marginalization



Abstraction and Marginalization should commute



Abstraction and Marginalization DO NOT commute in Hoel 2017



Abstraction and Marginalization DO NOT commute in Hoel 2017

$$\begin{bmatrix} 0.75 & 0.25 \\ 0.5 & 0.5 \end{bmatrix}$$

$$\begin{bmatrix} 0.3 & 0.2 & 0.5 \\ 0.4 & 0.1 & 0.5 \\ 0.5 & 0.5 & 0 \end{bmatrix} \qquad \begin{bmatrix} 0.3 & 0.2 & 0.5 \\ 0.4 & 0.1 & 0.5 \\ 0.5 & 0.5 & 0 \end{bmatrix}$$

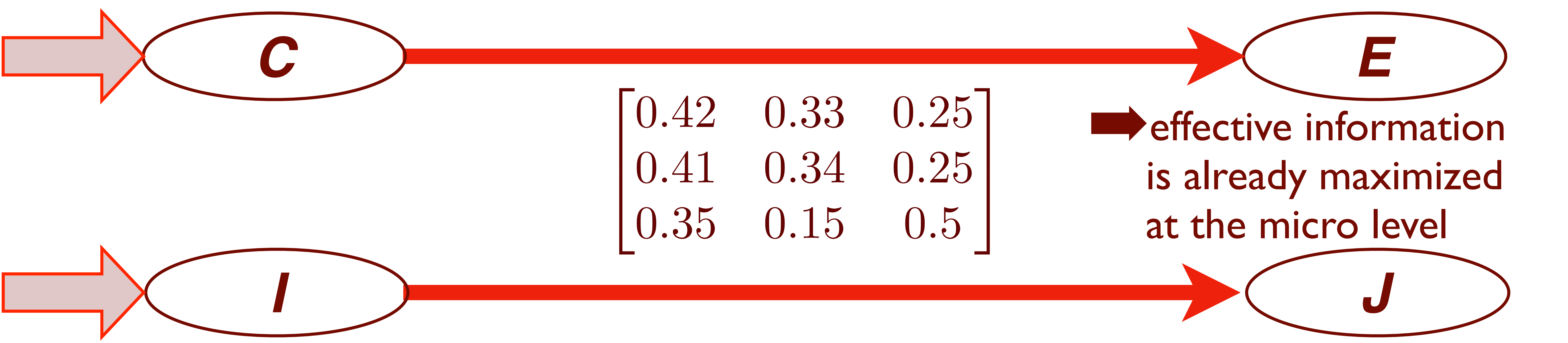
Abstraction and Marginalization DO NOT commute in Hoel 2017

$$I \xrightarrow{\begin{bmatrix} 0.3 & 0.2 & 0.5 \\ 0.4 & 0.1 & 0.5 \\ 0.5 & 0.5 & 0 \end{bmatrix}} I' \xrightarrow{\begin{bmatrix} 0.3 & 0.2 & 0.5 \\ 0.4 & 0.1 & 0.5 \\ 0.5 & 0.5 & 0 \end{bmatrix}} J$$

Abstraction and Marginalization DO NOT commute in Hoel 2017

$$\Rightarrow \; I \xrightarrow{\begin{bmatrix} 0.42 & 0.33 & 0.25 \\ 0.41 & 0.34 & 0.25 \\ 0.35 & 0.15 & 0.5 \end{bmatrix}} J$$

Abstraction and Marginalization DO NOT commute in Hoel 2017



$$\begin{bmatrix} 0.42 & 0.33 & 0.25 \\ 0.41 & 0.34 & 0.25 \\ 0.35 & 0.15 & 0.5 \end{bmatrix}$$

➡ effective information is already maximized at the micro level

Abstraction and Marginalization DO NOT commute in Hoel 2017

$$C \longrightarrow E$$

$$\begin{bmatrix} 0.75 & 0.25 \\ 0.5 & 0.5 \end{bmatrix}$$

$$\neq$$

$$C \longrightarrow E$$

$$\begin{bmatrix} 0.42 & 0.33 & 0.25 \\ 0.41 & 0.34 & 0.25 \\ 0.35 & 0.15 & 0.5 \end{bmatrix}$$

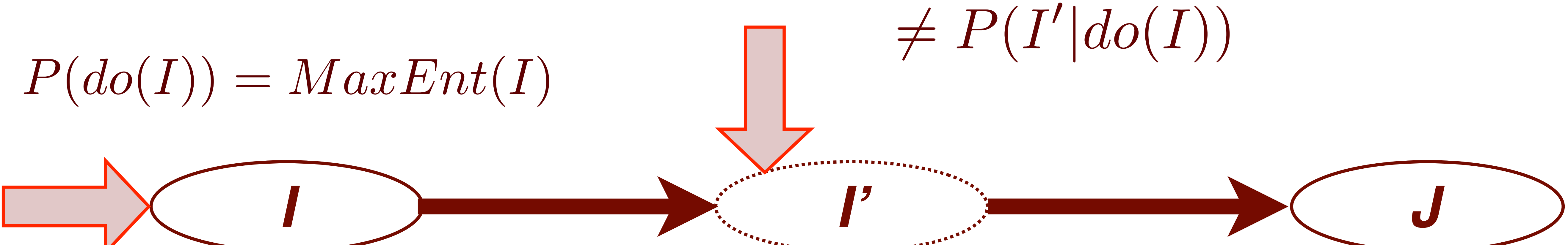
$$I \longrightarrow J$$

➡ effective information is already maximized at the micro level

The Problem: Introducing MaxEnt distributions

$$P(do(I')) = MaxEnt(I')$$
$$\neq P(I'|do(I))$$

$$P(do(I)) = MaxEnt(I)$$



Upshots

- it is worth distinguishing between macro level causes (or causal representations) and mixtures of causal effects
- whether or not there are macro-level causal descriptions is an empirical question determined by $P(E \mid do(C))$, independent of $P(do(C))$

Upshots

- it is worth distinguishing between macro level causes (or causal representations) and mixtures of causal effects
- whether or not there are macro-level causal descriptions is an empirical question determined by $P(E \mid do(C))$, independent of $P(do(C))$

➡ this also ensures that abstraction and marginalization commute

Upshots

- it is worth distinguishing between macro level causes (or causal representations) and mixtures of causal effects
- whether or not there are macro-level causal descriptions is an empirical question determined by P(E | do(C)), independent of P(do(C))

  ➡ this also ensures that abstraction and marginalization commute
- (although I have not discussed this in detail here) there is a distinction between how one determines the macro cause and how one determines the macro effect, though of course they are related

Specifically for Hoel's account

- the suggested relation between information theory and causality via effective information is tenuous and suggestive at best

Specifically for Hoel's account

- the suggested relation between information theory and causality via effective information is tenuous and suggestive at best
- channel capacity is a normative concept; whether or not it is exhausted is an empirical question; so the described causal emergence here is a *possible* emergence that may never be exhibited by the system in question
- effective information is uniquely maximized, but it is not clear that the implied partition of the state space is unique; this cuts both ways: either one wants uniqueness, or one wants non-uniqueness but not in the way implied by this theory: one wants many very different levels of aggregation

References

- Erik P Hoel. When the map is better than the territory. Entropy, 19(5):188, 2017.

Other useful references

- Joe Dewhurst. Causal emergence from effective information: Neither causal nor emergent? Thought: A Journal of Philosophy, 2021.
- Paul Rubenstein, Sebastian Weichwald et al. Causal consistency of structural equation models. UAI 2017
- Cosma Shalizi & Cristopher Moore. What is a macrostate? Subjective observations and objective dynamics. arXiv preprint cond-mat/0303625, 2003.
- Peter Spirtes & Richard Scheines. Causal inference of ambiguous manipulations. Philosophy of Science, 71(5):833–845, 2004.
- Sander Beckers & Joseph Halpern. Abstracting causal models. AAAI, 2019.
- Scott Aaronson, Higher-level causation exists (but I wish it didn't). https://www.scottaaronson.com/blog/?p=3294 (and reply by Hoel)
- F.E. Rosas, et al. Reconciling emergences: An information-theoretic approach to identify causal emergence in multivariate data. PLOS Computational Biology, 2020.

Attemps at an alternative account (i.e. shameless self-promotion)

- Krzysztof Chalupka, Pietro Perona, & Frederick Eberhardt. Visual causal feature learning. UAI, 2015.
- Krzysztof Chalupka, Frederick Eberhardt, and Pietro Perona. Multi-level cause-effect systems. AISTATS, 2016.

Thank you!