ABNMS 2013

Synergy and Redundancy

Erik Nyberg

erik.nyberg@monash.edu

Kevin Korb

kevin.korb@monash.edu

Recent controversy

Recently, there has been controversy over how to measure synergistic and redundant information.

- Williams & Beer (2010)
- Griffith & Koch (2012)
- Harder, Salge & Polani (2013)

But today, I'm not going to explain other people's proposals, because they're all wrong! I'll just explain our approach.

- Nyberg & Korb (manuscript)

Non-information-theoretic concepts

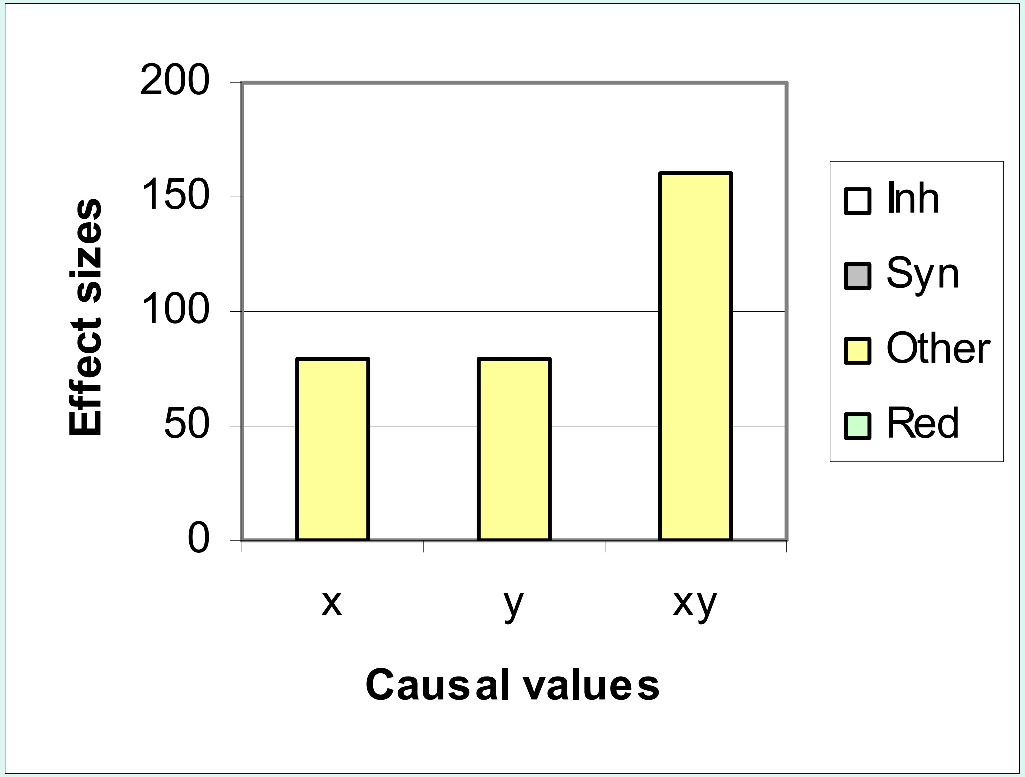
Two jobs example

Suppose that Joe, who currently earns $0 per day, must decide whether to accept or reject two different part-time jobs. If he sells seashells by the seashore, then he will earn $80 per day. If he sells e-books via eBay, then he will also earn $80 per day. The point of interest is how much he will earn if he accepts *both* jobs.

Non-information-theoretic concepts

Independence

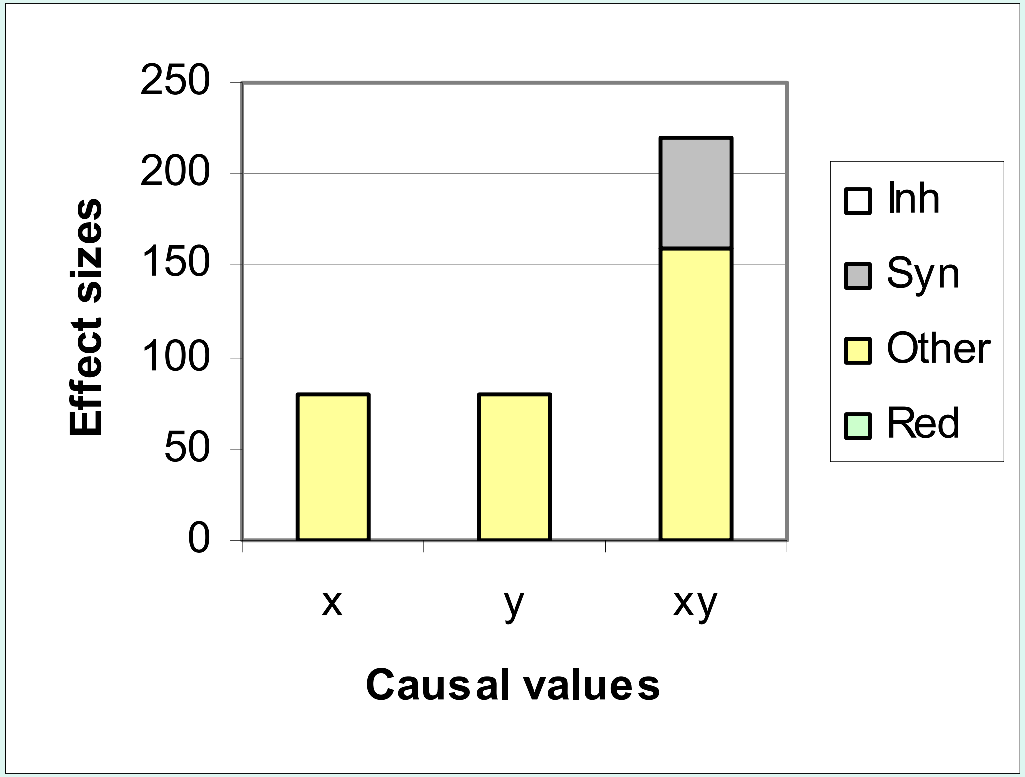
If the two effects are ‘independent’, then the joint effect will be to earn $160 per day.



Non-information-theoretic concepts

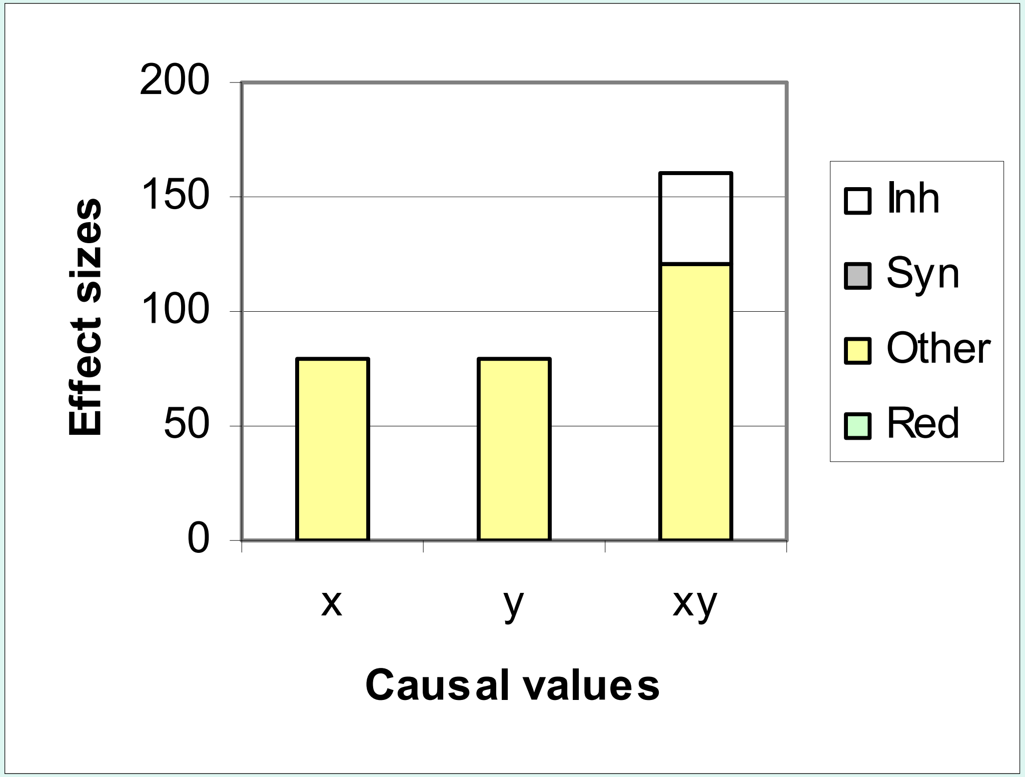
Synergy

If he earns $220 per day, then the joint effect involves 'synergy' for $60 more per day.



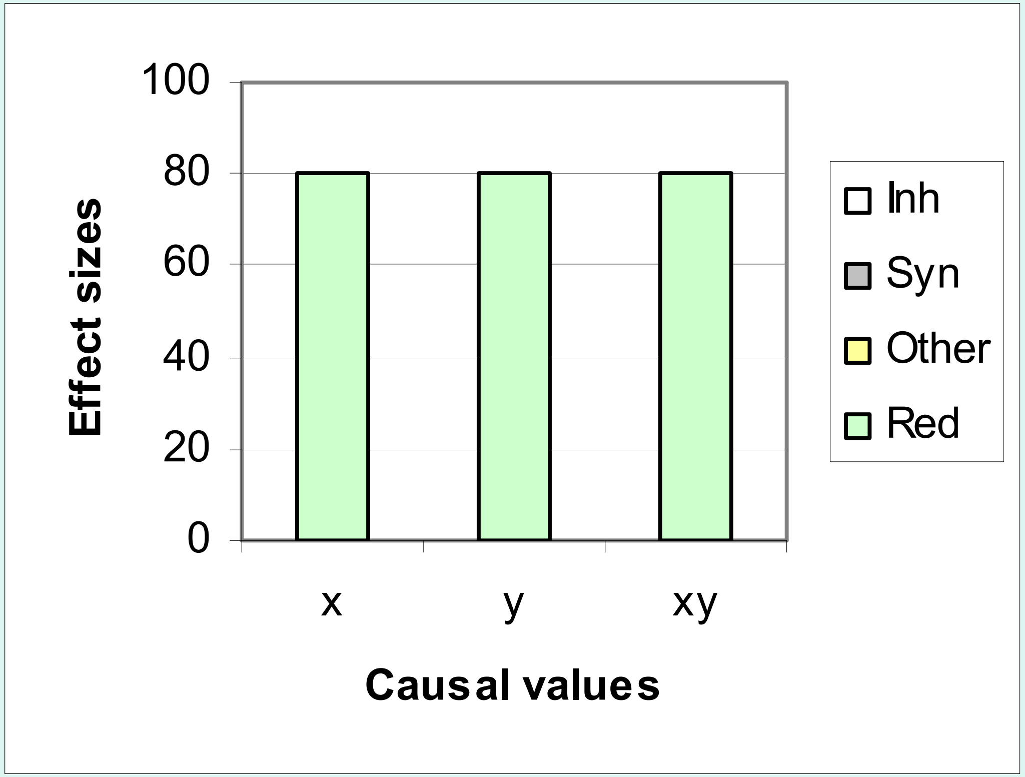
Inhibition

If he only earns $120 per day, then the joint effect involves 'inhibition' for $40 less per day.



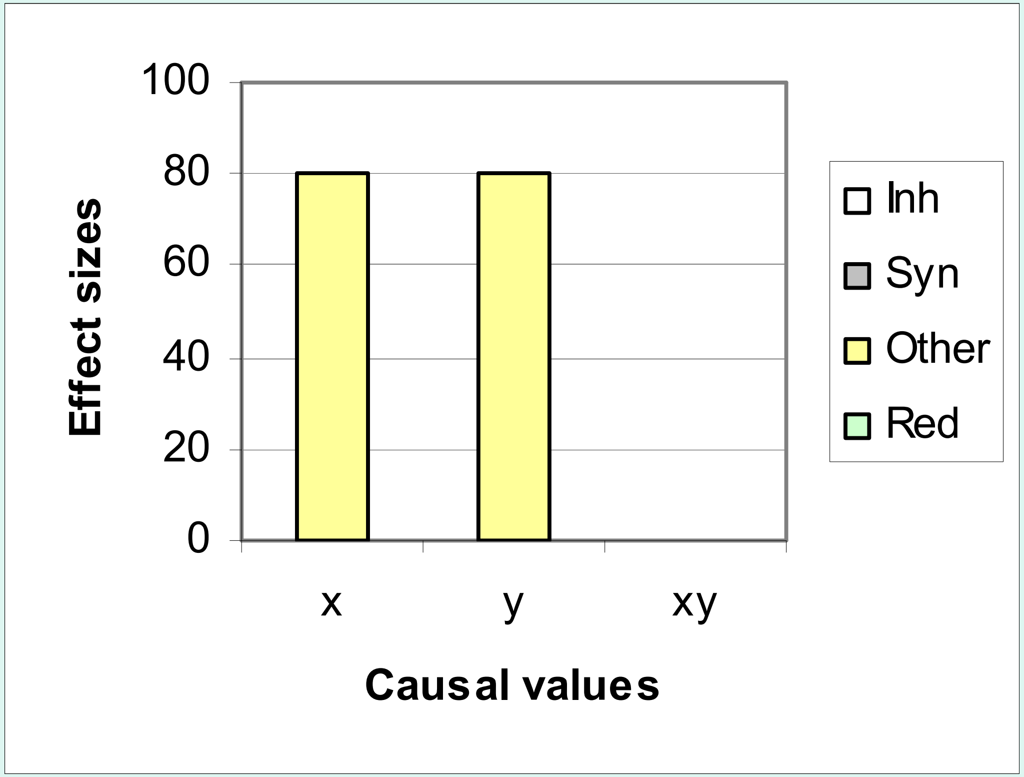
Redundancy

And if Joe earns only $80 per day, then the joint effect involves ‘redundancy’: he could quit either job and continue to earn the same amount.



Similarity

Since he earns $80 with either job alone, the two individual effects are ‘similar’. Redundancy also entails a *similar joint* effect. And redundancy entails the same *joint inhibition*.



Our measurement approach

Find largest triple similarity with equal inhibition

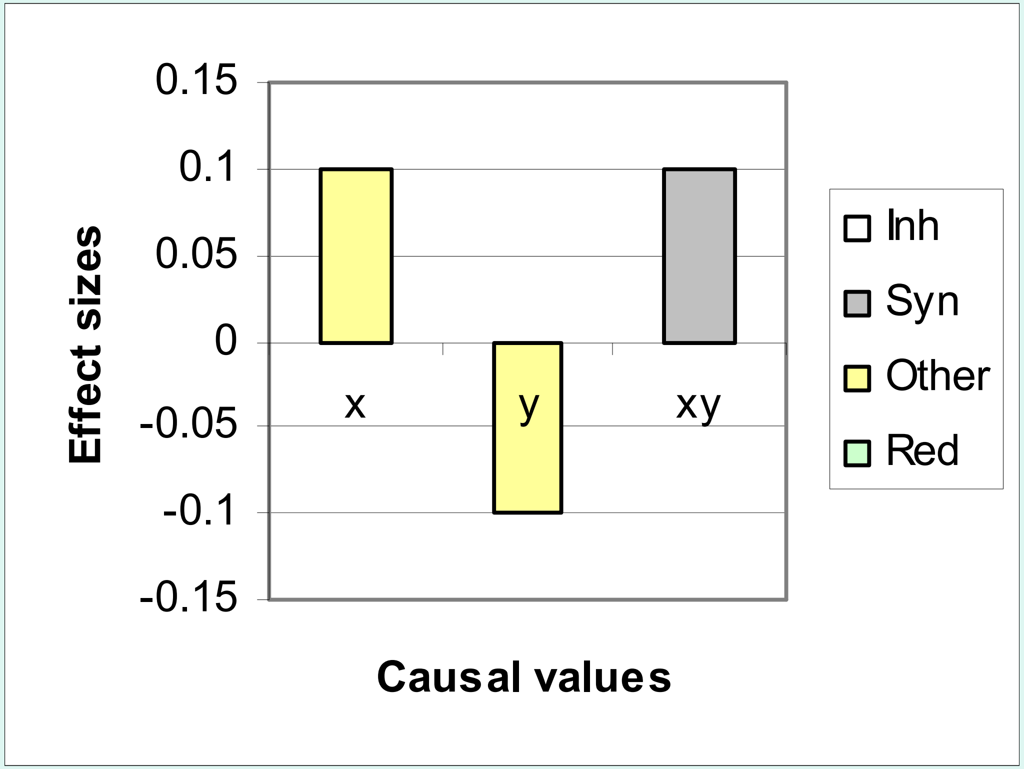
The previous examples are neat: the joint effect can be interpreted as all independence, or all synergy, or all inhibition, or all redundancy.

But we want to measure the redundant component in messy examples: the joint effect must be interpreted as a mixture. We look for the largest possible component that has the characteristic properties of redundancy: a triple similarity between the individual and joint effects, and an equal amount of inhibition.

Our measurement approach

Same signs

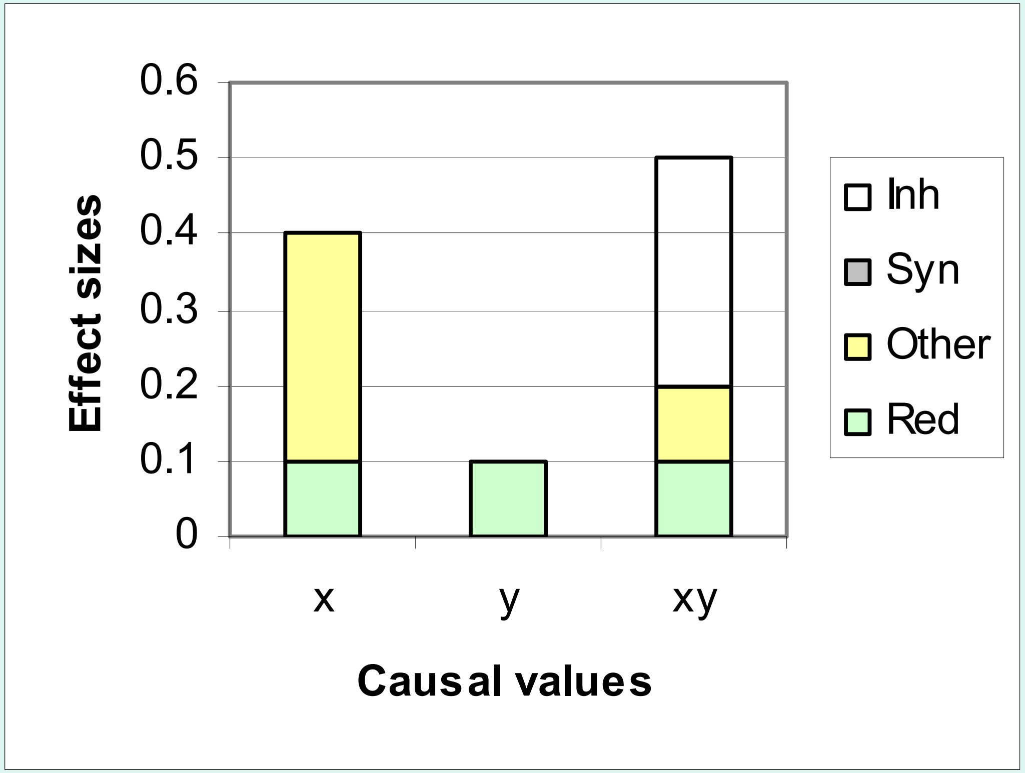
There is no redundant effect here. *All the effects must have the same sign*.



Our measurement approach

No larger than smallest effect

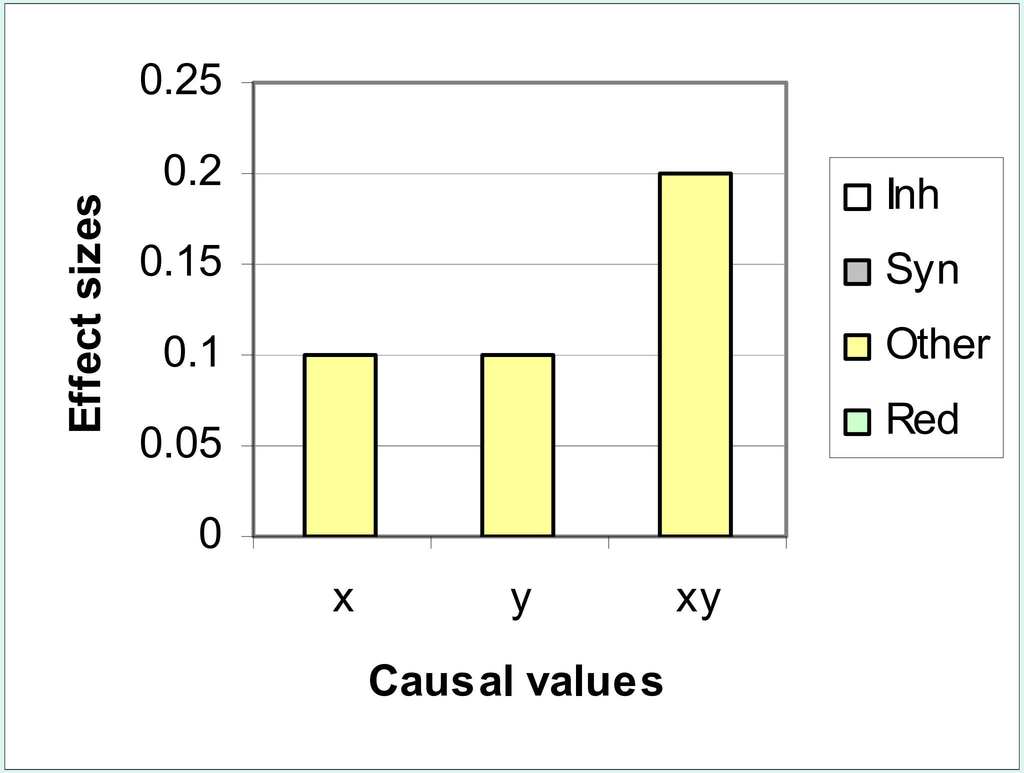
The redundant effect here must be at most 0.1. *The redundant effect is no larger than the smallest effect*.



Our measurement approach

No larger than inhibition

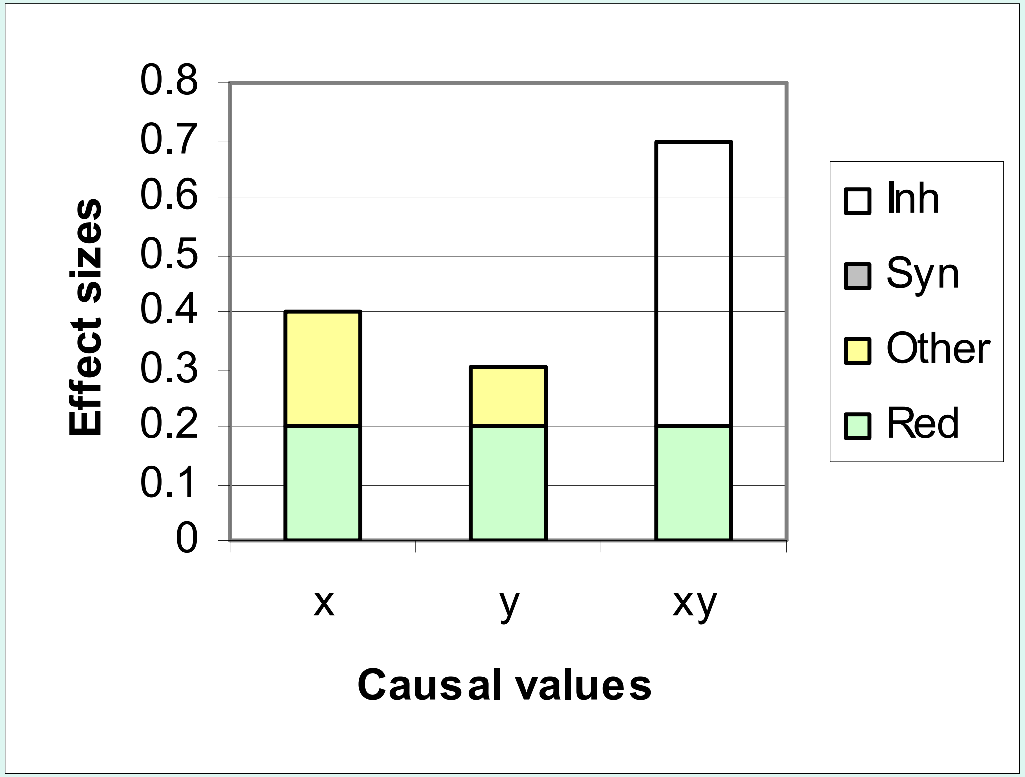
There is no redundant effect here. *The redundant effect is no larger than the inhibition*.



Our measurement approach

Residuals are unique info and inhibition

The maximum redundant effect we can attribute here is 0.2. *Residual individual effects are unique, and residual joint effects are (additional) inhibition*.



Pointwise mutual information

$$\text{pmi}(x; z) = \log_2 \frac{\Pr(z \mid x)}{\Pr(z)}$$

$$\text{pmi}(x, y; z) = \log_2 \frac{\Pr(z \mid x, y)}{\Pr(z)}$$

Mutual information

$$\text{MI}(X; Z) = \mathrm{E}_{X,Z}[\text{pmi}(x; z)]$$

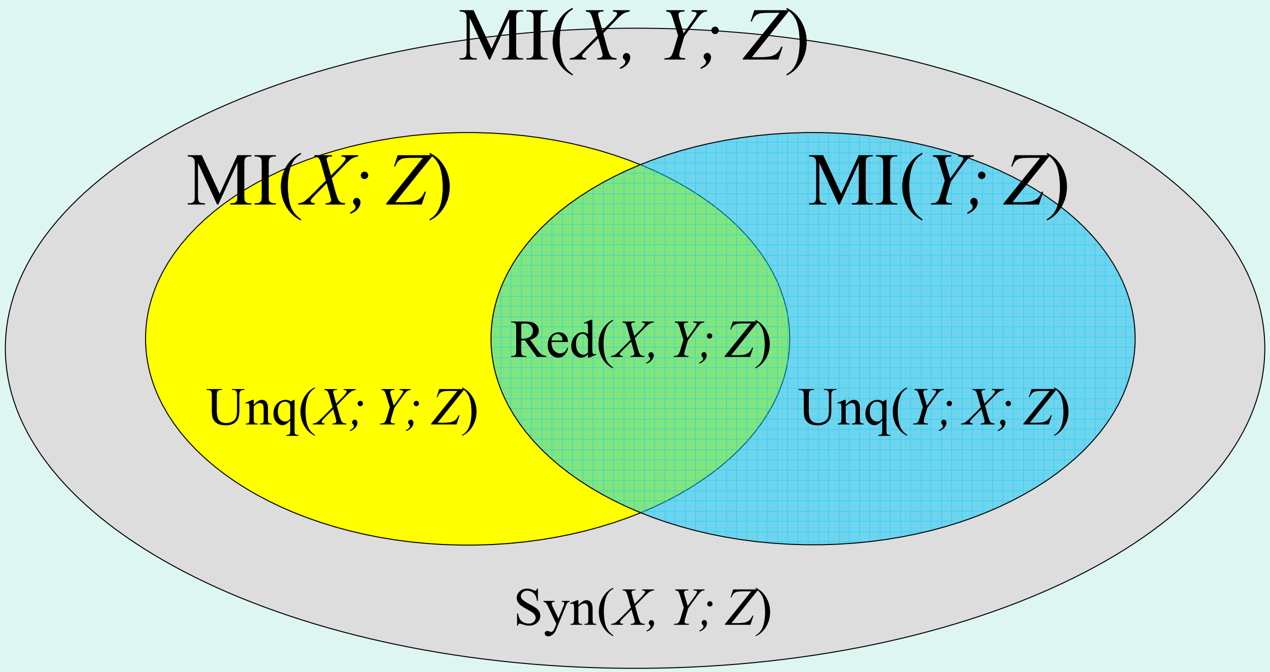
$$= \sum_{x,z} \Pr(x) \Pr(z \mid x) \log_2 \frac{\Pr(z \mid x)}{\Pr(z)}$$

$$\text{MI}(X, Y; Z) = \mathrm{E}_{X,Y,Z}[\text{pmi}(x, y; z)]$$

$$= \sum_{x,y,z} \Pr(x, y) \Pr(z \mid x, y) \log_2 \frac{\Pr(z \mid x, y)}{\Pr(z)}$$

Proposed partition of MI

Venn diagram of Williams & Beer (2010)



Problem of measuring redundancy

Eq.1 $\mathrm{MI}(X,\ Y;\ Z) = \mathrm{Unq}(X;\ Z) + \mathrm{Unq}(Y;\ Z) + \mathrm{Red}(X,\ Y;\ Z) + \mathrm{Syn}(X,\ Y;\ Z)$

Eq.2 $\mathrm{MI}(X;\ Z) = \mathrm{Unq}(X;\ Z) + \mathrm{Red}(X,\ Y;\ Z)$

Eq.3 $\mathrm{MI}(Y;\ Z) = \mathrm{Unq}(Y;\ Z) + \mathrm{Red}(X,\ Y;\ Z)$

For variables, some redundancy *and* some synergy may be present. But with four unknowns and only three linear equations, this leaves one degree of freedom.

Redundant pointwise mutual information

```
If        $0 < \mathrm{pmi}(x; z), \mathrm{pmi}(y; z), \mathrm{pmi}(x, y; z), [\mathrm{pmi}(x; z) + \mathrm{pmi}(y; z) - \mathrm{pmi}(x, y; z)]$
then      $\mathrm{red}(x, y; z) = \min\{\mathrm{pmi}(x; z), \mathrm{pmi}(y; z), \mathrm{pmi}(x, y; z), [\mathrm{pmi}(x; z) + \mathrm{pmi}(x; z) - \mathrm{pmi}(x, y; z)]\}$
else if   $0 > \mathrm{pmi}(x; z), \mathrm{pmi}(y; z), \mathrm{pmi}(x, y; z), [\mathrm{pmi}(x; z) + \mathrm{pmi}(y; z) - \mathrm{pmi}(x, y; z)]$
then      $\mathrm{red}(x, y; z) = \max\{\mathrm{pmi}(x; z), \mathrm{pmi}(y; z), \mathrm{pmi}(x, y; z), [\mathrm{pmi}(x; z) + \mathrm{pmi}(x; z) - \mathrm{pmi}(x, y; z)]\}$
else      $\mathrm{red}(x, y; z) = 0$
```

Our approach to MI

Redundant mutual information

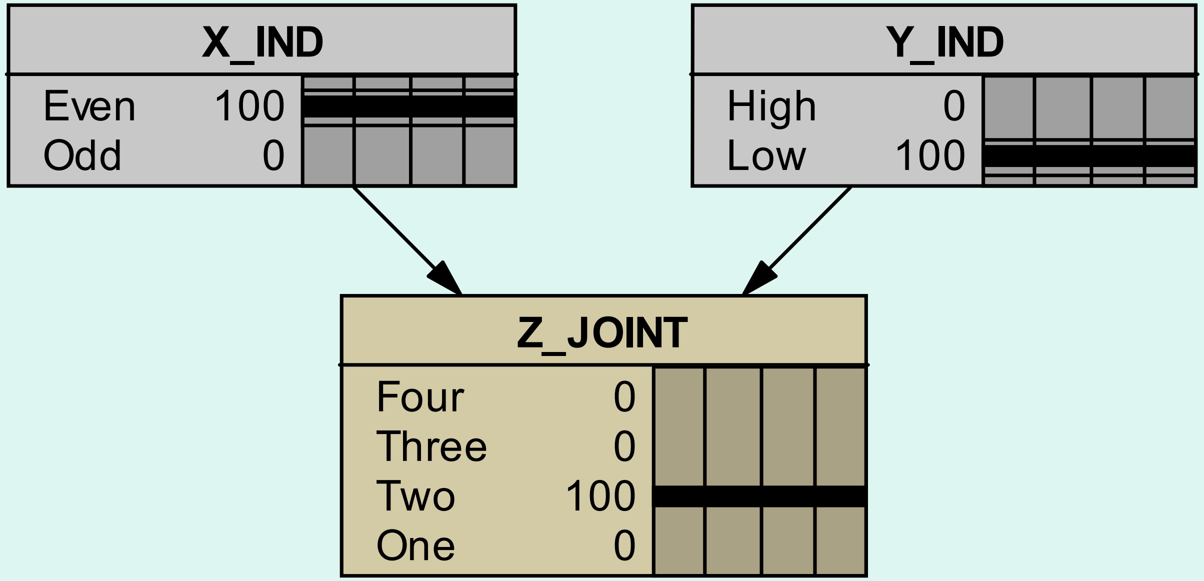
$$\mathrm{Red}(X,\ Y;\ Z) = \mathrm{E}_{X,Y,Z}\,[\mathrm{red}(x,\ y;\ z)]$$

$$= \sum_{x,y,z} \Pr(x, y) \Pr(z \mid x, y)\,\mathrm{red}(x,\ y;\ z)$$

Our results with test cases

UNQ example = all unique information

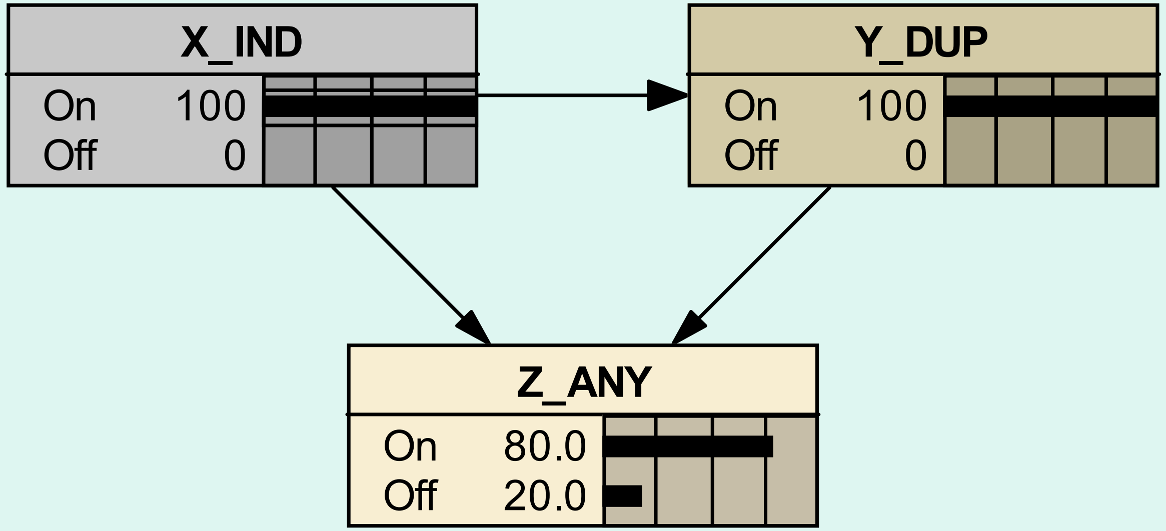
X and *Y* are independent, and there is an isomorphism between the values of *Z* and the pairs of values (*x*, *y*) such that each *z* has the same probability as the corresponding (*x*, *y*).



Our results with test cases

RDN example = all redundant information

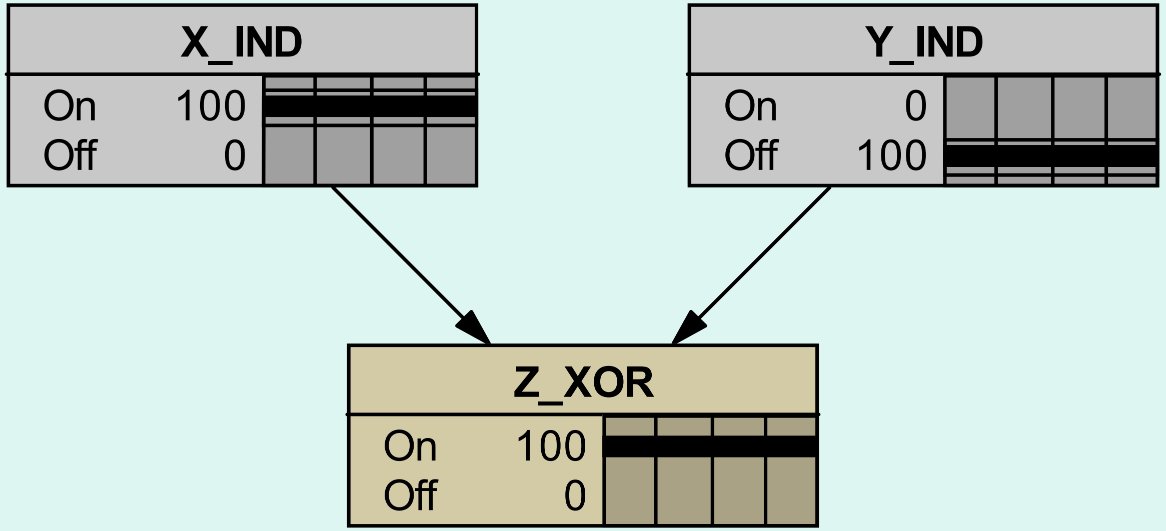
X and *Y* are dependent, and there is an isomorphism between the values of *X* and the values of *Y* such that each *x* has the same probability as the corresponding *y*.



Our results with test cases

XOR example = all synergistic information

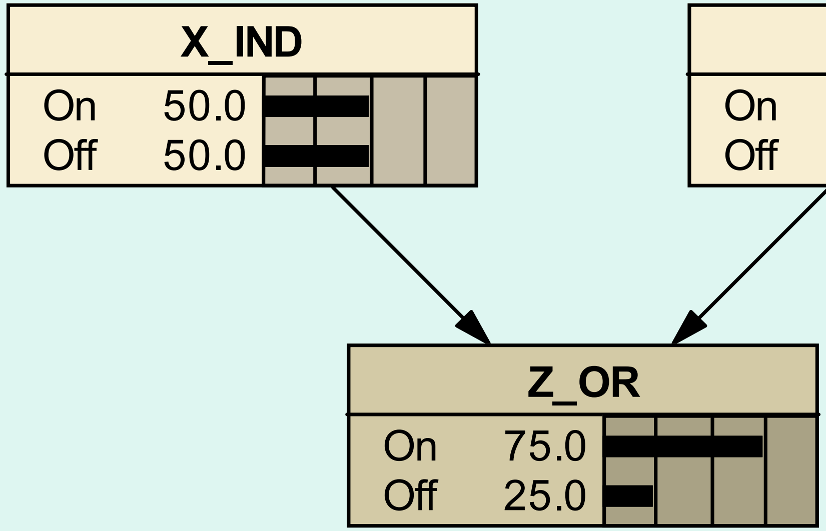
X and *Y* are independent with uniform distributions. If (x_1,y_0) or (x_0,y_1), then z_1, else z_0.



Our results with test cases

OR example = a mixture of all three

X and Y are independent. If x_1 or y_1, then z_1, else z_0.



Our results with test cases

Comparative table

		Test Cases						
		UNQ	*RDN*	*XOR*	*OR*	*AVG*	*DIE*	*GHO*
	Nett	0	Red	Syn	Syn	Syn	Syn	0
Redundancy Measures	*WB10*	∀ Red = Syn	∀ Red 0 Syn	0 Red ∀ Syn	∀ Red > Syn	∀ Red > Syn	∀ Red > Syn	∀ Red = Syn
	GK12	0 Red 0 Syn	∀ Red 0 Syn	0 Red ∀ Syn	0 Red ∃ Syn	0 Red 0 Syn	0 Red ?	0 Red 0 Syn
	HSP13	0 Red 0 Syn	∀ Red 0 Syn	0 Red ∀ Syn	∀ Red > Syn	∀ Red > Syn	∀ Red > Syn	∀ Red = Syn
	NK13	0 Red 0 Syn	∀ Red 0 Syn	0 Red ∀ Syn	∃ Red > Syn	∃ Red ∃ Syn	0 Red ∃ Syn	0 Red 0 Syn
	NK13$_{KL}$	0 Red 0 Syn	∀ Red 0 Syn	0 Red ∀ Syn	∃ Red > Syn			0 Red 0 Syn

Conclusion

Could redundancy or synergy be useful to you?

Our approach is the best available way of measuring redundancy and synergy. It has the best theoretical rationale and the best results in the standard test cases. In further work, we'd like to apply this measure to some real cases where it's useful.

Williams & Beer (2010)

WillBeer10 measure of redundancy

– Williams & Beer (2010) don't simply say that *X* and *Y* are redundant if they yield the same quantity of information about *Z*. This would obviously make UNQ a counterexample, because in that case *X* and *Y* affect *Z* to the same extent but in completely different ways, so it's qualitatively different information.

– They do say that for each value of *Z*, the minimum quantity of information this yields about either *X* or *Y* is redundant. The assumption is that the (reverse) information encoded in *X* and *Y* says the same thing about each value of *Z* (so it's both quantitatively and qualitatively identical).

Williams & Beer (2010)

Counterexample to WillBeer10

- UNQ (e.g. *Quarters*): Unfortunately, this is still a counterexample, because each value of *Z* has two independent aspects, one of which is encoded in *X* and the other in *Y*. For example, 3 is both even and greater than 1. *X* and *Y* affect *Z* to the same extent but in completely different ways, so it's qualitatively different information.
- WillBeer10 = similar quantities of information

Griffith & Koch (2012)

GrifKoch12 measure of redundancy

- GrifKoch12 had the neat idea of building a different Bayesian network, based on the old network and satisfying some new constraints, which would show one of the quantities we are interested in.
- Specifically, to each original network they add a child $Z \rightarrow Z^*$, and perform a search for a conditional probability function Pr($Z^*|Z$) that filters out all and only the synergistic information. Thus, X and Y should give the same amount of independent information about Z^* as they do about Z, but the joint amount of information that X and Y together give about Z^* should be minimised. The loss of joint information is taken to be the synergy, which allows them to also calculate a corresponding redundancy.

Griffith & Koch (2012)

Counterexample to GrifKoch12

- Unfortunately, sometimes the desired filter doesn't exist. That is, there is no conditional probability function Pr(*Z**|*Z*) that filters out all and only the synergistic information. I can prove this using the AVG counterexample. Here, I can show that it isn't possible to reduce the joint information at all, and yet in the joint information there is nett synergy.
- AVG (e.g. *Supervisors*): *X* and *Y* are independent with uniform distributions, and the probability that *Z* is *On* is the mean of the probability that *X* is *On* and the probability that *Y* is *On* (1 if both are *On*, 0.5 if only one is *On*, and 0 if neither is *On*).

Griffith & Koch (2012)

Counterexample to GrifKoch12 (cont.)

- AVG (e.g. *Supervisors*): Incorrectly reports no synergy. But the nett synergy is ?? bits.
- Therefore, the measure defined by GrifKoch12 can be an underestimate of the amount of synergy, contrary to their assertion. At best, they have provided a way to calculate a lower limit to the amount of synergy in a relationship.
- GrifKoch12 = possibly filtering synergistic information

Harder, Salge & Polani (2013)

HardSalgPola13 measure of redundancy

- HardSalgPola13 take an interesting pseudo-geometrical approach to this problem, although the information geometry plays a quite superficial role. They measure pseudo-distances between probability distributions using KL divergence, which distinguishes appropriately between raising and lowering the probability of each value of *Z*.

Harder, Salge & Polani (2013)

HardSalgPola13 measure of redundancy (cont.)

1. D_{KL1} = For each value *x*, they measure the amount of information this gives about *Z*, which is captured by the KL divergence between the two distributions Pr(*Z*) and Pr(*Z*|*x*). Roughly, this is how far the *x*-distribution is from the original distribution.
2. D_{KL2} = Then they search for the distribution Pr*(*Y*) that would have the most similar effect on *Z*, i.e. yield the least KL divergence between the distributions Pr(*Z*|*x*) and Pr(*Z*|Pr*(*Y*)). Roughly, this is how far the Pr*(*Y*)-distribution is from the *x*-distribution.
3. I(*X*;*Z*) in I(*Y*;*Z*) = The amount of information from *x* that can be expressed by *Y* is taken to be the difference between these KL divergences. Roughly, subtract one distance from the other to get the shared distance from the original distribution.

Harder, Salge & Polani (2013)

HardSalgPola13 measure of redundancy (cont.)

$$\begin{tikzcd}
\mathrm{Pr}(Z|x) \arrow[rr, "D_{\mathrm{KL1}}"] \arrow[dr, "D_{\mathrm{KL2}}"'] & & \mathrm{Pr}(Z) \\
& \mathrm{Pr}(Z|\mathrm{Pr}^*(Y)) &
\end{tikzcd}$$

Harder, Salge & Polani (2013)

Counterexample to HardSalgPola13

- It's not clear that this pseudo-geometric calculation will always give us the best measure of the common effect of *x* and Pr*(*Y*).
- But I'm not going to explore this here, because there are bigger problems with this proposal:

1. Pr*(*Y*) is merely a *possible* distribution, there is no requirement that it actually occurs. So, *Y* need not ever actually have a similar effect to *X*.
2. There is no discussion of whether *x* and Pr*(*Y*) ever occur at the same time, and what the resulting distribution is over *Z*.

Harder, Salge & Polani (2013)

Counterexample to HardSalgPola13 (cont.)

- We think redundancy would be where *x* and Pr*(*Y*) occur at the same time, and have the same common effect (but no more) that either of them would provide alone. So, HardSalgPola13 fail to measure what they said they wanted to measure. They are measuring something like 'possible similar information' instead. We can vividly illustrate the difference with our GHO example.
- GHO: As for UNQ, except that *X* has two additional values with zero probability that could produce the same effect on *Z* as the two values of *Y*, and similarly *X* has two additional values with zero probability that could produce the same effect on *Z* as the two values of *Y*.

Harder, Salge & Polani (2013)

Counterexample to HardSalgPola13 (cont.)

- GHO: Incorrectly reports complete redundancy. These "ghost" values entail that *X* and *Y* could possibly have the same effects as each other, and this has a huge impact on their measure. Yet these "ghost" values never actually occur, so they have no impact in calculating the individual and joint information. The presence and effects of "ghost" values can be altered arbitrarily to affect their measure, without ever affecting the individual and joint information. This is implausible.
- HardSalgPola13 = possible similar information