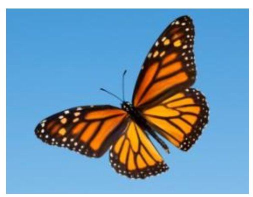
# Mutual Information

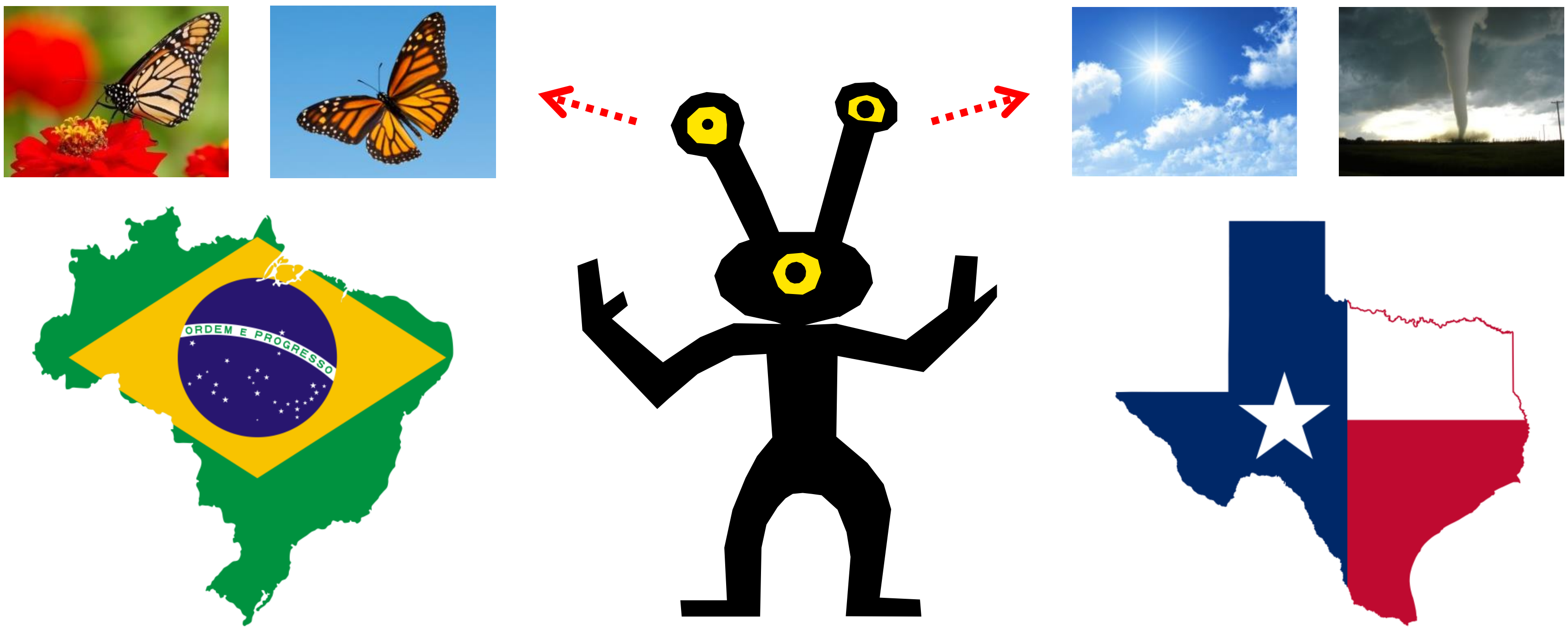



**Hiroki Sayama**
sayama@binghamton.edu

# Relationship between multiple probabilistic systems

- Self-information was defined for an individual event
- Information entropy was defined for a single probabilistic system
- How can we capture the relationship between multiple probabilistic systems using information measurements?

# Multiple variables




- X: A butterfly in Brazil flaps its wings, or not
- Y: A tornado appears in Texas, or not

# Considering multiple variables simultaneously

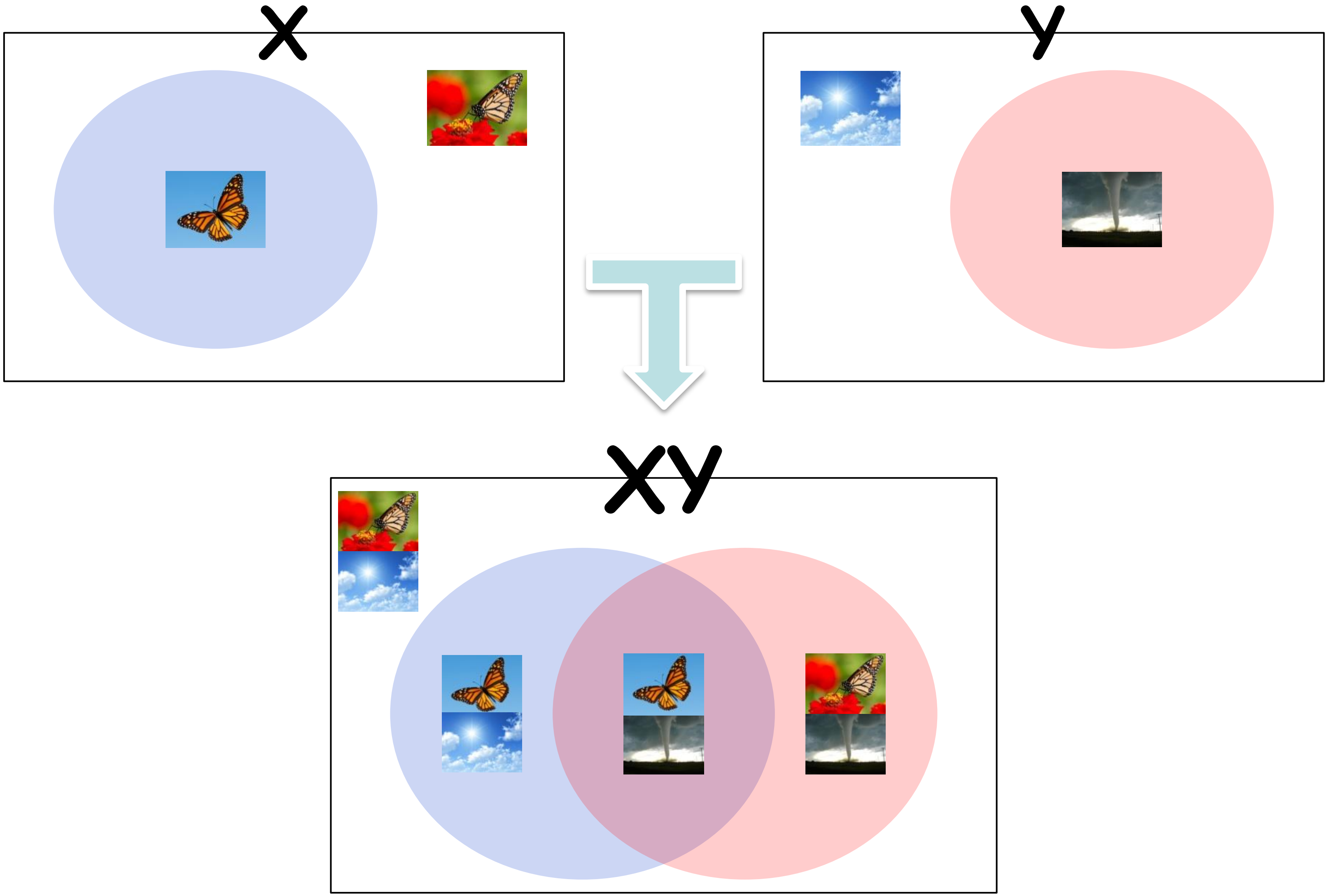
- X: A butterfly in Brazil flaps its wings, or not
- Y: A tornado appears in Texas, or not

## Considering their relationship means considering their composites

XY:

A butterfly in Brazil flaps its wings *and* a tornado appears in Texas, or
A butterfly in Brazil flaps its wings *and* a tornado does <u>not</u> appear in Texas, or
A butterfly in Brazil does <u>not</u> flap its wings *and* a tornado appears in Texas, or
A butterfly in Brazil does <u>not</u> flap its wings *and* a tornado does <u>not</u> appear in Texas

# Considering multiple variables simultaneously



# Composite events

- $X : \{ x_1, x_2 \}$
- $Y : \{ y_1, y_2 \}$

- $XY : \{ (x_1, y_1), (x_1, y_2), (x_2, y_1), (x_2, y_2) \}$

$$XY : \{ (x_1, y_1),\ (x_1, y_2),\ (x_2, y_1),\ (x_2, y_2) \}$$

(This works even if the numbers of events are not the same between X and Y)

# Product probability space

- Prob. space X: $\{x_1, x_2\}$, $\{p(x_1), p(x_2)\}$
- Prob. space Y: $\{y_1, y_2\}$, $\{p(y_1), p(y_2)\}$

- Product probability space XY:

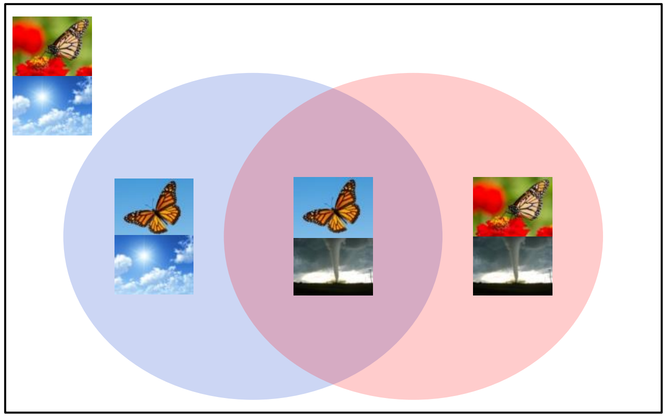
$\{(x_1, y_1), (x_1, y_2), (x_2, y_1), (x_2, y_2)\}$,
$\{p(x_1, y_1), p(x_1, y_2), p(x_2, y_1), p(x_2, y_2)\}$

Composite events

# Probability of composite events

- Probability of composite event $(x, y)$:

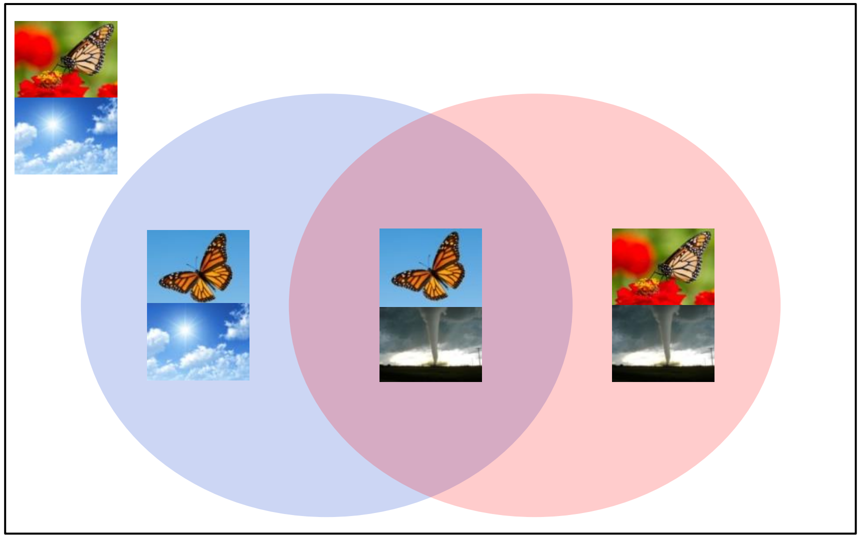
$$
\begin{aligned}
p(x, y) &= p(y, x) \\
&= p(x \mid y)\, p(y) \\
&= p(y \mid x)\, p(x)
\end{aligned}
$$



$p(x \mid y)$: Conditional probability for x to occur when y already occurred

# Some important properties

- $p(x) = \Sigma_y \, p(x, y)$
- $p(y) = \Sigma_x \, p(x, y)$



- $p(x \mid y) = p(x)$
- $P(y \mid x) = p(y)$

  if X and Y are independent from each other

# Exercise: Bayes' theorem

- Define $p(x \mid y)$ using $p(y \mid x)$ and $p(x)$
  - Use the following formula as needed

$$p(x) = \Sigma_y \, p(x, y)$$

$$p(y) = \Sigma_x \, p(x, y)$$

$$\begin{aligned} p(x, y) &= p(x)\, p(y \mid x) \\ &= p(y)\, p(x \mid y) \end{aligned}$$

# Exercise

- Using the data given on the right, calculate:
  - p(Rays)
  - p(Rays, Dodgers)
  - p(Rays | Dodgers)
  - p(Dodgers | Rays)

| Year | X (American) | Y (National) |
|---|---|---|
| 2001 | Yankees | Diamondbacks |
| 2002 | Angels | Giants |
| 2003 | Yankees | Marlins |
| 2004 | Red Sox | Cardinals |
| 2005 | White Sox | Astros |
| 2006 | Tigers | Cardinals |
| 2007 | Red Sox | Rockies |
| 2008 | Rays | Phillies |
| 2009 | Yankees | Phillies |
| 2010 | Rangers | Giants |
| 2011 | Rangers | Cardinals |
| 2012 | Tigers | Giants |
| 2013 | Red Sox | Cardinals |
| 2014 | Royals | Giants |
| 2015 | Royals | Mets |
| 2016 | Indians | Cubs |
| 2017 | Astros | Dodgers |
| 2018 | Red Sox | Dodgers |
| 2019 | Astros | Nationals |
| 2020 | Rays | Dodgers |

# Information Entropy and Multiple Probability Spaces

# Joint entropy

- Entropy of product probability space XY:

$$H(XY) = -\sum_x \sum_y p(x, y) \log p(x, y)$$

  - $H(XY) = H(YX)$
  - If X and Y are independent:

    $H(XY) = H(X) + H(Y)$
  - If Y completely depends on X:

    $H(XY) = H(X) \quad ( >= H(Y) )$

# Exercise

- Using the data given on the right, calculate the **joint entropy H(XY)**
  - Data is available on myCourses in csv format

| Year | X (American) | Y (National) |
|---|---|---|
| 2001 | Yankees | Diamondbacks |
| 2002 | Angels | Giants |
| 2003 | Yankees | Marlins |
| 2004 | Red Sox | Cardinals |
| 2005 | White Sox | Astros |
| 2006 | Tigers | Cardinals |
| 2007 | Red Sox | Rockies |
| 2008 | Rays | Phillies |
| 2009 | Yankees | Phillies |
| 2010 | Rangers | Giants |
| 2011 | Rangers | Cardinals |
| 2012 | Tigers | Giants |
| 2013 | Red Sox | Cardinals |
| 2014 | Royals | Giants |
| 2015 | Royals | Mets |
| 2016 | Indians | Cubs |
| 2017 | Astros | Dodgers |
| 2018 | Red Sox | Dodgers |
| 2019 | Astros | Nationals |
| 2020 | Rays | Dodgers |

# Conditional entropy

- Expected entropy of Y when a specific event occurred in X:

$$H(Y \mid X) = \Sigma_x \, p(x) \, H(Y \mid X{=}x)$$
$$= - \Sigma_x \, p(x) \, \Sigma_y \, p(y \mid x) \log p(y \mid x)$$
$$= - \Sigma_x \Sigma_y \, p(y, x) \log p(y \mid x)$$

  - If X and Y are independent:

    $H(Y \mid X) = H(Y)$

  - If Y completely depends on X:

    $H(Y \mid X) = 0$

# Exercise

- Using the data given on the right, calculate the **conditional entropy $H(Y \mid X)$**

| Year | X (American) | Y (National) |
|---|---|---|
| 2001 | Yankees | Diamondbacks |
| 2002 | Angels | Giants |
| 2003 | Yankees | Marlins |
| 2004 | Red Sox | Cardinals |
| 2005 | White Sox | Astros |
| 2006 | Tigers | Cardinals |
| 2007 | Red Sox | Rockies |
| 2008 | Rays | Phillies |
| 2009 | Yankees | Phillies |
| 2010 | Rangers | Giants |
| 2011 | Rangers | Cardinals |
| 2012 | Tigers | Giants |
| 2013 | Red Sox | Cardinals |
| 2014 | Royals | Giants |
| 2015 | Royals | Mets |
| 2016 | Indians | Cubs |
| 2017 | Astros | Dodgers |
| 2018 | Red Sox | Dodgers |
| 2019 | Astros | Nationals |
| 2020 | Rays | Dodgers |

# Exercise

- Prove the following:

$$H(Y \mid X) = H(YX) - H(X)$$

# Mutual Information

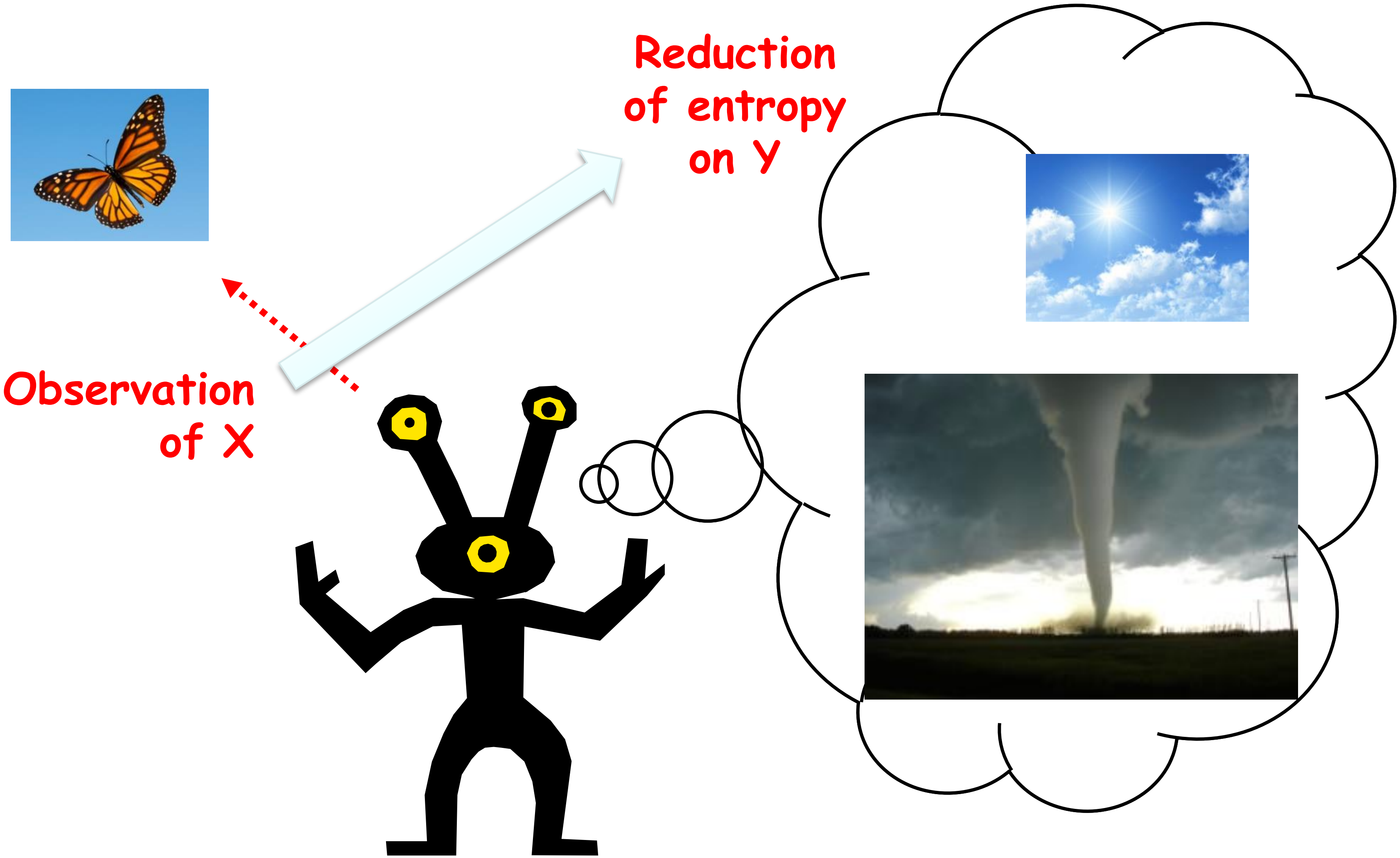
# Mutual information

- **Conditional entropy measures** how much ambiguity still remains on Y after observing an event on X
- Average reduction of ambiguity on Y by one observation on X is written as:

$$I(Y; X) = H(Y) - H(Y \mid X)$$

Mutual information

# Intuitive meaning of I(Y; X)

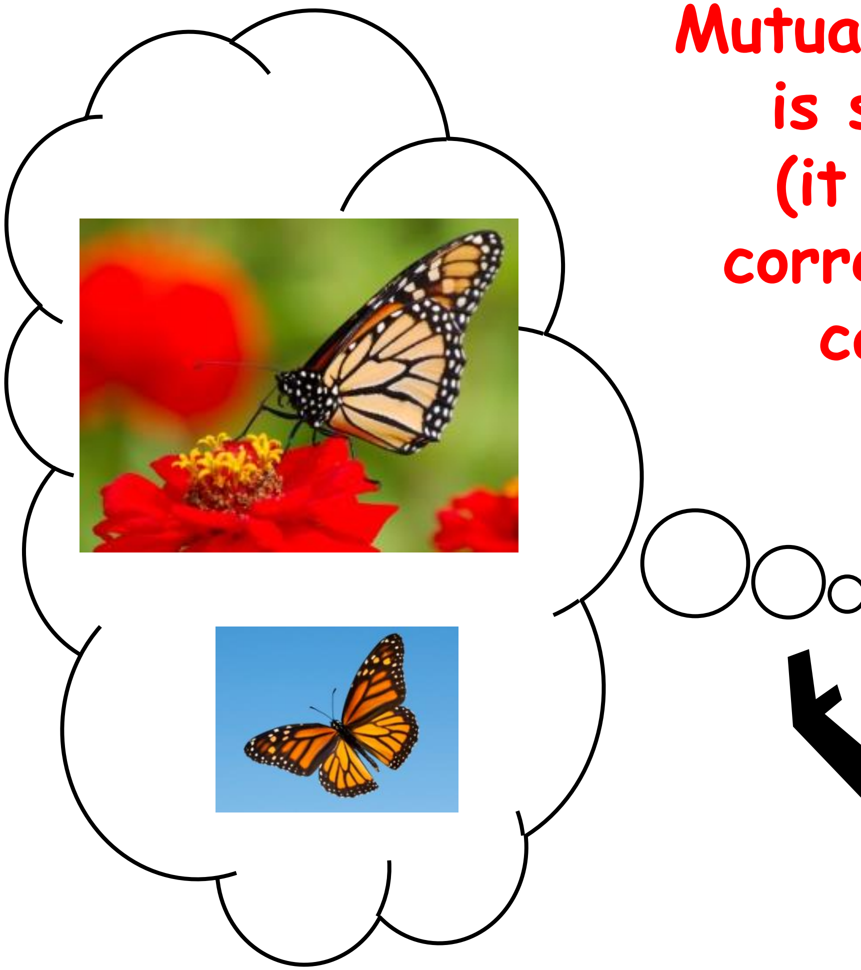


# Symmetry of mutual information

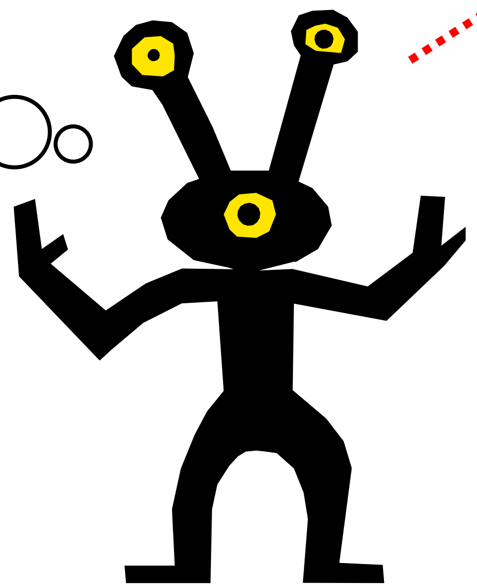
$$
\begin{aligned}
I(Y;X) &= H(Y) - H(Y \mid X) \\
&= H(Y) + H(X) - H(YX) \\
&= H(X) + H(Y) - H(XY) \\
&= I(X;Y)
\end{aligned}
$$

**Mutual information is symmetric in terms of X and Y**

# Symmetry of mutual information



Mutual information
is symmetric
(it measures
correlation, not
causality)



# Exercise

- Using the data given on the right, calculate the **mutual information** $I(X; Y)$

| Year | X (American) | Y (National) |
|---|---|---|
| 2001 | Yankees | Diamondbacks |
| 2002 | Angels | Giants |
| 2003 | Yankees | Marlins |
| 2004 | Red Sox | Cardinals |
| 2005 | White Sox | Astros |
| 2006 | Tigers | Cardinals |
| 2007 | Red Sox | Rockies |
| 2008 | Rays | Phillies |
| 2009 | Yankees | Phillies |
| 2010 | Rangers | Giants |
| 2011 | Rangers | Cardinals |
| 2012 | Tigers | Giants |
| 2013 | Red Sox | Cardinals |
| 2014 | Royals | Giants |
| 2015 | Royals | Mets |
| 2016 | Indians | Cubs |
| 2017 | Astros | Dodgers |
| 2018 | Red Sox | Dodgers |
| 2019 | Astros | Nationals |
| 2020 | Rays | Dodgers |

# Exercise

- Prove the following:
  - If X and Y are independent:
    $I(X; Y) = 0$
  - If Y completely depends on X:
    $I(X; Y) = H(Y)$

# Use of mutual information

- Mutual information can be used to measure how much correlation exists between two subsystems in a complex system
  - Traditional statistical correlation only works for quantitative measures and detects only linear relationships
  - Mutual information works for qualitative measures (discrete, categorical) and nonlinear relationships as well

# Exercise

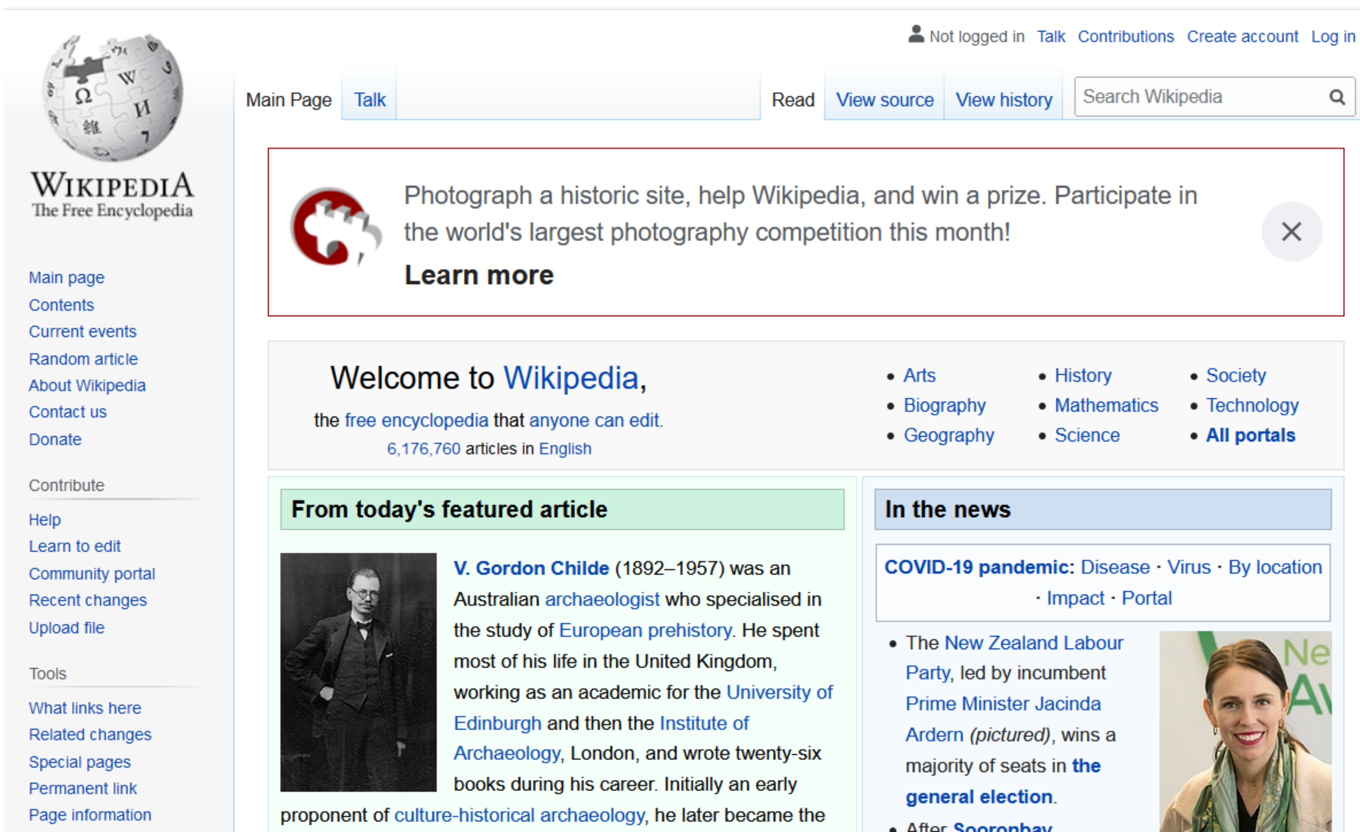
- Choose two discrete variables that may be influencing each other
  - E.g., people's first name initials vs. last name initials

- Obtain data about their values

- Calculate mutual information between them

# Exercise

- Calculate the mutual information between the first letter of a word (X) and its case (Y) for all the words on the top page of English Wikipedia



# FYI: Pointwise mutual information

$$
\begin{aligned}
\mathrm{pmi}(x; y) \\
&= -\log p(x) - \log p(y) + \log p(x,y) \\
&= \log \frac{p(x,y)}{p(x)\, p(y)}
\end{aligned}
$$

- PMI measures the association between two single events (it can be either positive or negative)

# Mutual Information for Continuous Variables

# Definition of mutual information

$$I(Y;X) = H(Y) - H(Y \mid X)$$
$$= H(Y) + H(X) - H(YX)$$
$$= H(X) + H(Y) - H(XY)$$
$$= I(X;Y)$$

# … holds for continuous variables

$$
\begin{aligned}
I(Y;\,X) &= H_{dif}(Y) - H_{dif}(Y \mid X) \\
&= H_{dif}(Y) + H_{dif}(X) - H_{dif}(YX) \\
&= H_{dif}(X) + H_{dif}(Y) - H_{dif}(XY) \\
&= I(X;\,Y)
\end{aligned}
$$

$$H_{dif}(Y \mid X) = - \int_x \int_y \mathrm{pdf}(y,\, x) \log \mathrm{pdf}(y \mid x)\, dx\, dy$$

$$\mathrm{pdf}(y \mid x) = \mathrm{pdf}(y,\, x) \,/\, p(x)$$

$$p(x) = \int_{y'} \mathrm{pdf}(y',\, x)\, dy'$$

$$H_{dif}(XY) = - \int_x \int_y \mathrm{pdf}(x,\, y) \log \mathrm{pdf}(x,\, y)\, dx\, dy$$

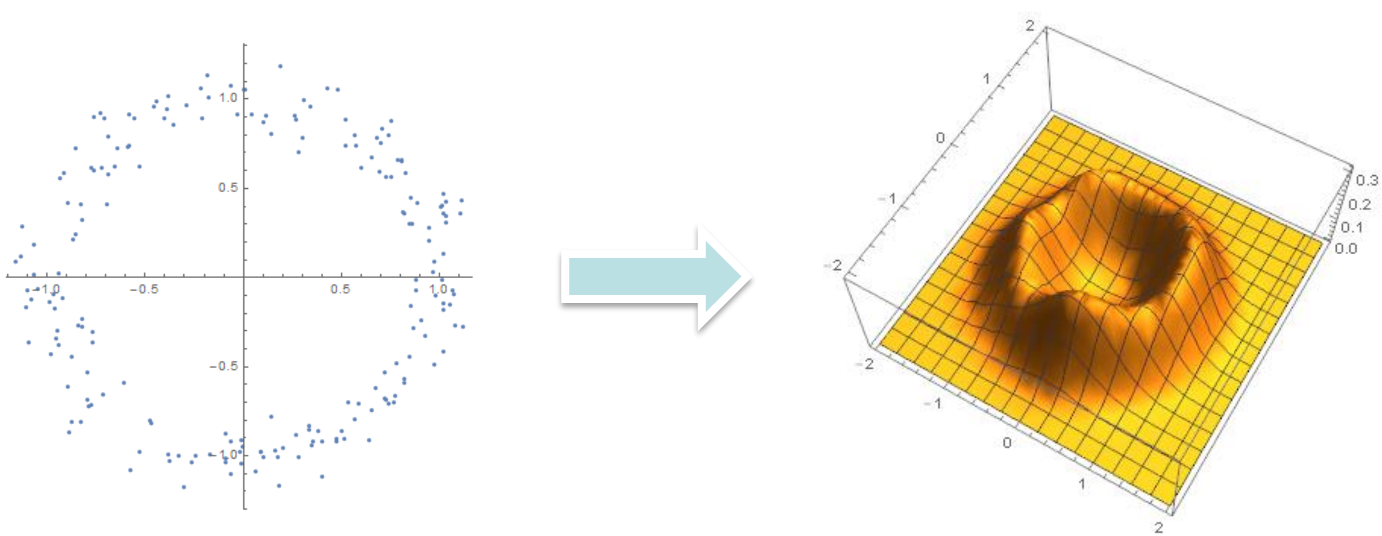
# Note on mutual information for continuous variables

$$I(Y; X) = H_{dif}(Y) - H_{dif}(Y \mid X)$$

- Both $H_{dif}$ originally contained the same "infinity" term, which cancel out
  - → No infinity is ignored in the definition of $I(Y; X)$
  - → The value of $I(Y; X)$ has actual meaning; the amount of information shared between X and Y
  - → $I(Y; X)$ is always non-negative

# Calculating mutual information from data points of continuous values

- Create a smooth PDF using, e.g., Gaussian kernel method
- Calculate
  $I(Y;X) = H_{dif}(Y) + H_{dif}(X) - H_{dif}(YX)$

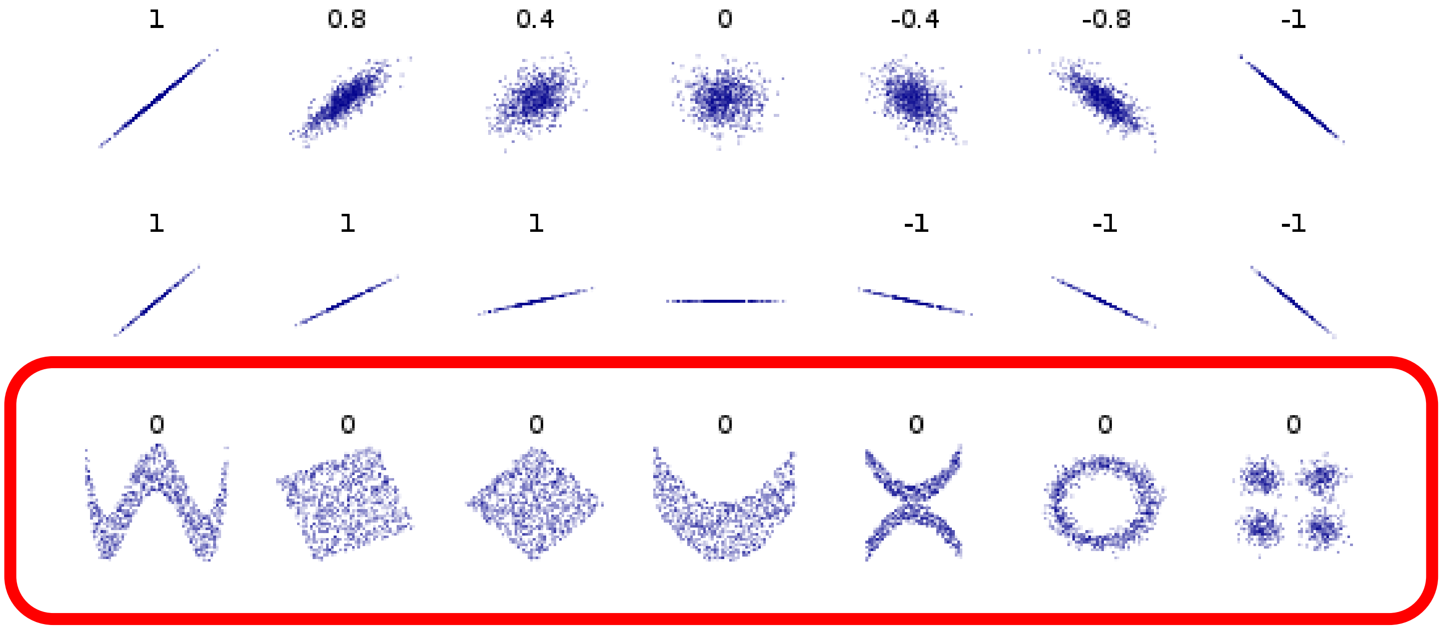


# Exercise

- Choose two continuous variables that may be influencing each other
  - E.g., people's height vs. latitude of their addresses
- Obtain data about their values
- Calculate mutual information between them

# Mutual information as a tool to detect nonlinear correlation

- **Mutual information can detect nonlinear correlations that simple correlation metrics cannot**



# FYI: Kullback-Leibler divergence

$D_{KL}(p(x) \,||\, q(x))$
$= (- \int_x p(x) \log q(x)\, dx) - H_{dif}(p(x))$

- This measures how distribution p(x) is different from q(x)

- It is known that:

$$I(X; Y) = D_{KL}(p(x, y) \,||\, p(x)p(y))$$

  - Mutual Information tells you how p(x, y) is different from a hypothetical PDF with independent X and Y

# Exercise

- Prove this:

$$I(X;\ Y) = D_{KL}(p(x,\ y)\ ||\ p(x)p(y)\ )$$