

Mutual Information

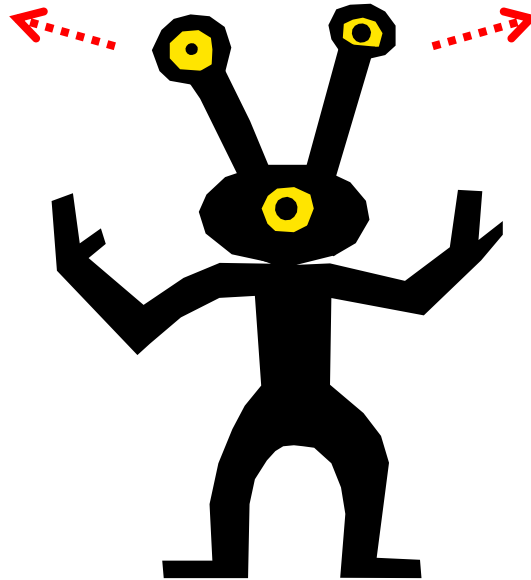


Hiroki Sayama
sayama@binghamton.edu

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

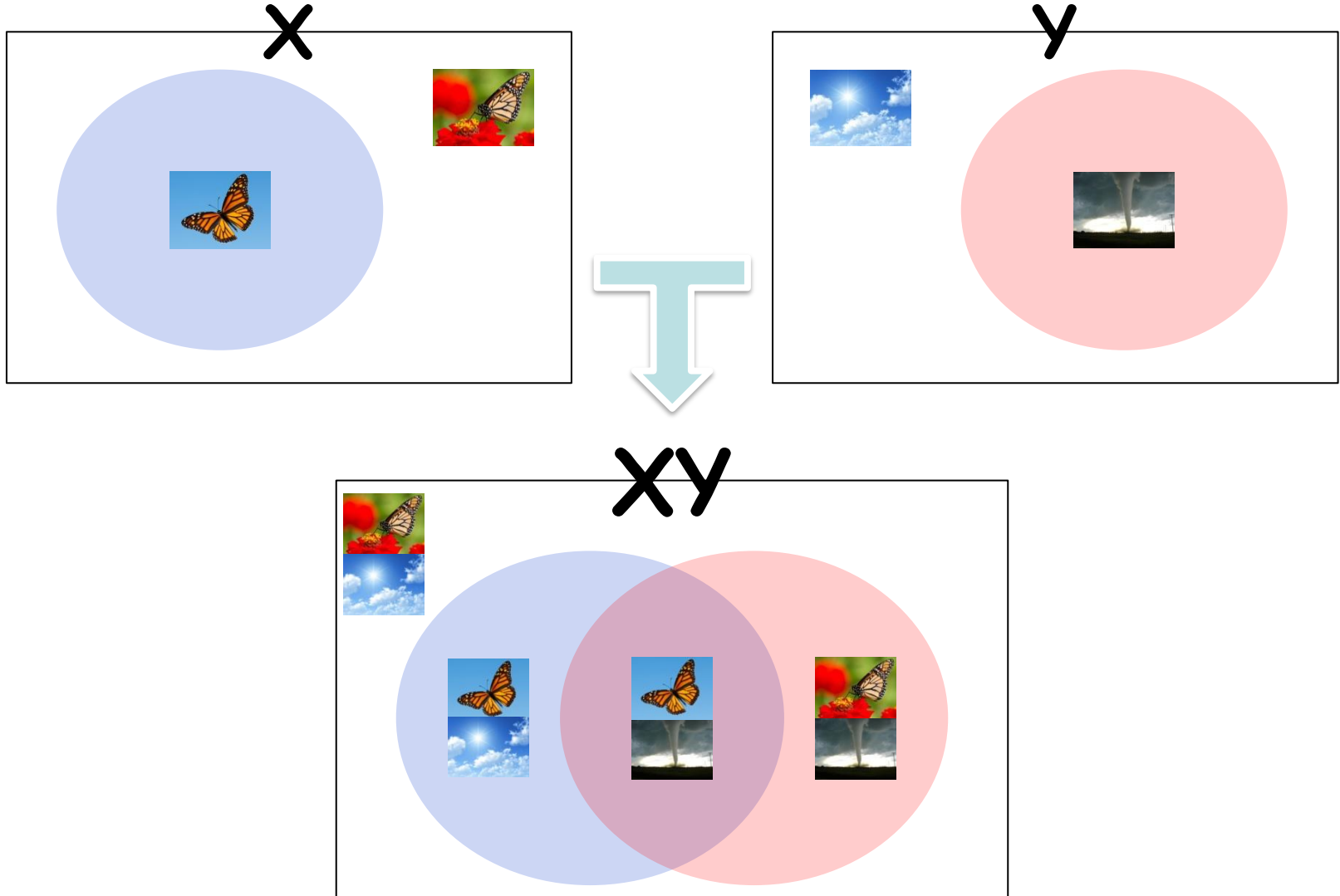
- X: A butterfly in Brazil flaps its wings, or not
- Y: A tornado appears in Texas, or not

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 not appear in Texas, or
A butterfly in Brazil does not flap its wings *and* a tornado appears in Texas, or
A butterfly in Brazil does not flap its wings *and* a tornado does not appear in Texas

Considering multiple variables simultaneously

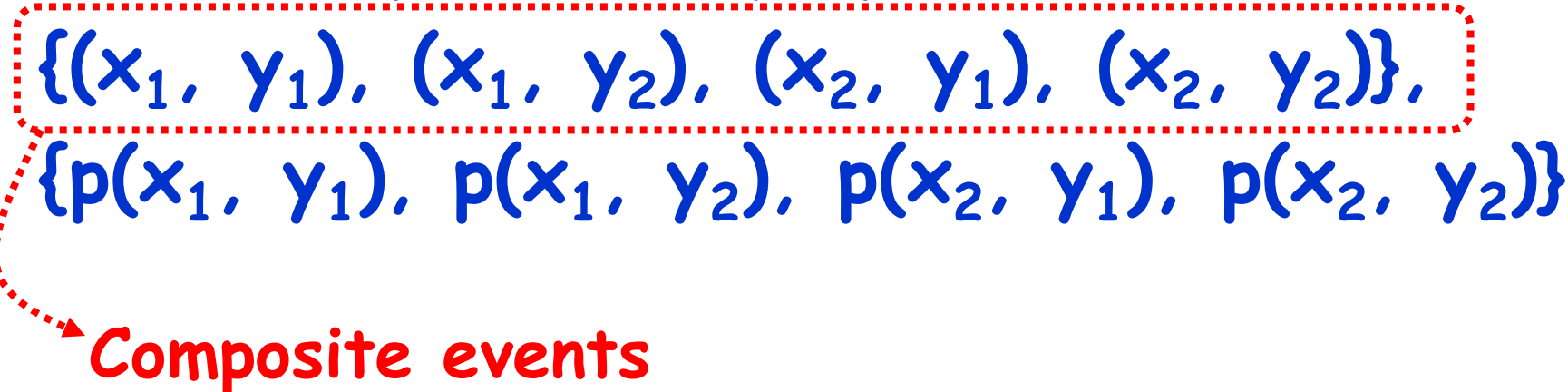


Composite events

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

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

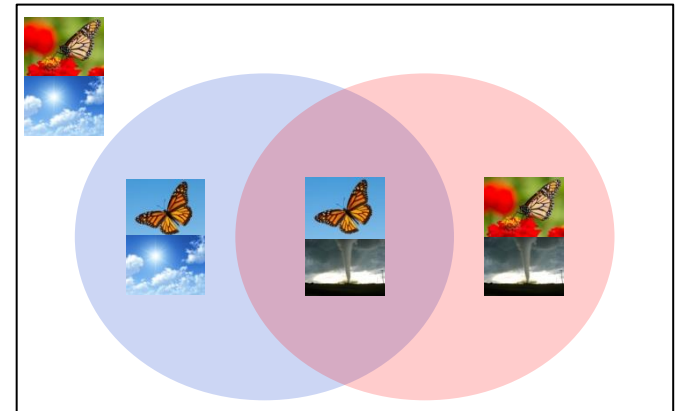
Product probability space

- 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) :

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

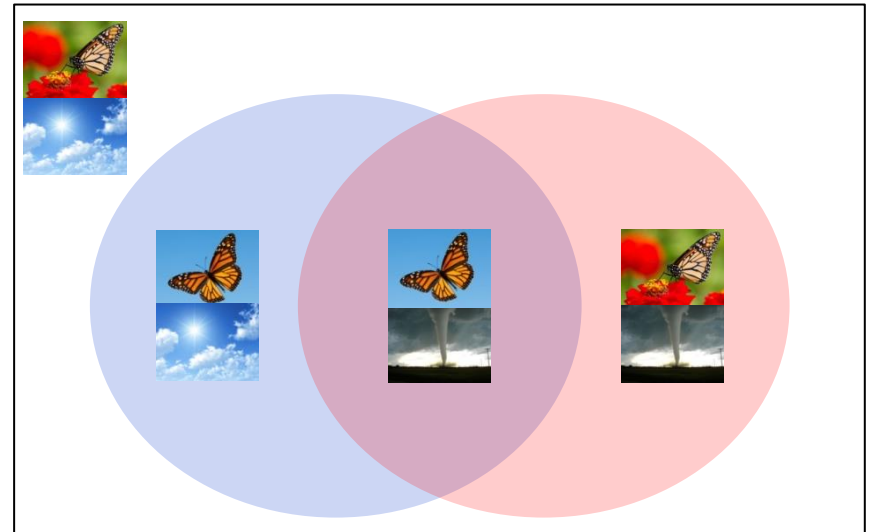


$p(x \mid y)$: Conditional probability for x to occur when y already occurred

Some important properties

- $p(x) = \sum_y p(x, y)$
- $p(y) = \sum_x p(x, y)$
- $p(x \mid y) = p(x)$
- $P(y \mid x) = p(y)$

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

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

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

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

Exercise

- Using the data given on the right, calculate:
 - $p(\text{Rays})$
 - $p(\text{Rays}, \text{Dodgers})$
 - $p(\text{Rays} \mid \text{Dodgers})$
 - $p(\text{Dodgers} \mid \text{Rays})$

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 :
$$H(XY) = H(X) \quad (\geq H(Y))$$

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 :

$$\begin{aligned} H(Y \mid X) &= \sum_x p(x) H(Y \mid X=x) \\ &= - \sum_x p(x) \sum_y p(y \mid x) \log p(y \mid x) \\ &= - \sum_x \sum_y p(y, x) \log p(y \mid x) \end{aligned}$$

- 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 $H(Y | X)$

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

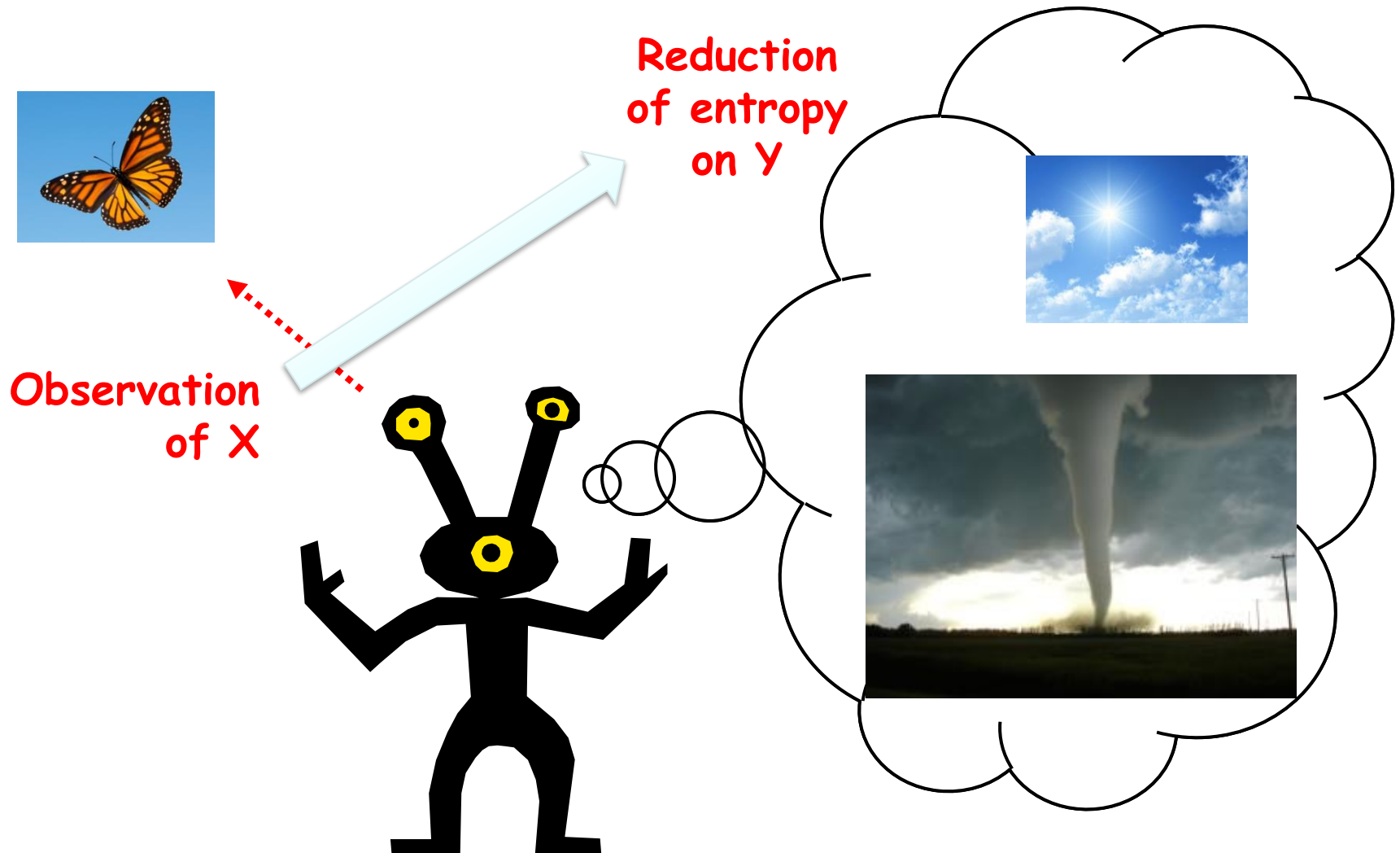
Mutual information

- 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:

$$I(Y; X) = H(Y) - H(Y | X)$$

Mutual information

Intuitive meaning of $I(Y; X)$



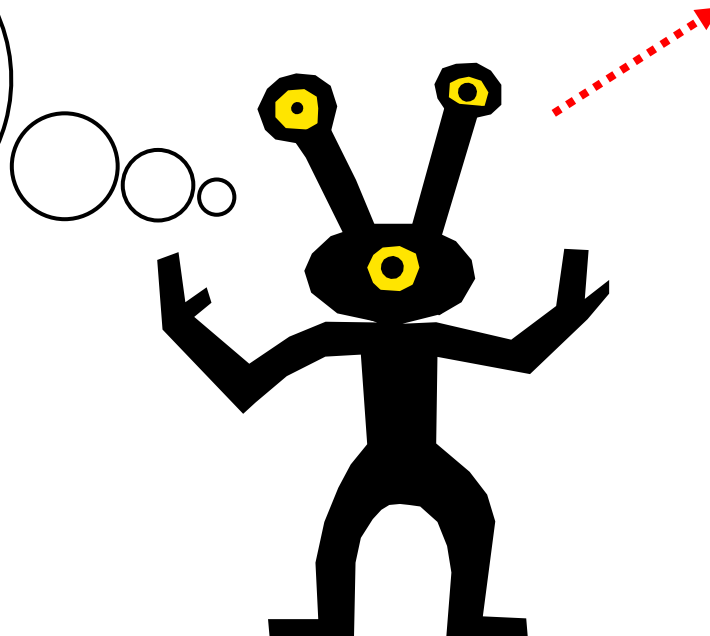
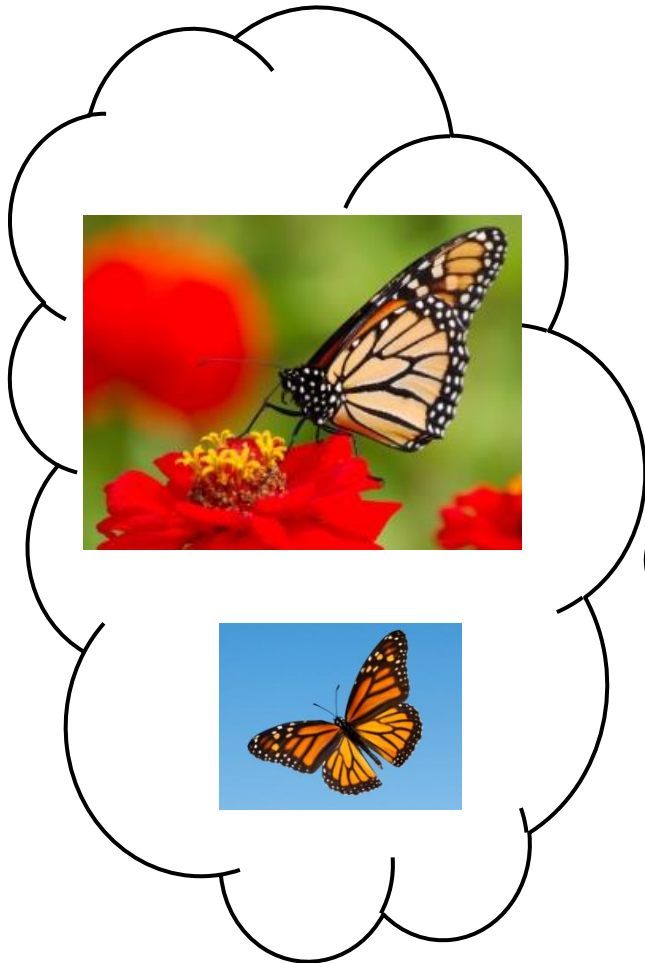
Symmetry of mutual information

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

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



The screenshot shows the English Wikipedia homepage. At the top left is the Wikipedia logo, a globe with various characters, and the text "WIKIPEDIA The Free Encyclopedia". Below this is a sidebar with links: Main page, Contents, Current events, Random article, About Wikipedia, Contact us, Donate, Contribute, Help, Learn to edit, Community portal, Recent changes, Upload file, Tools, What links here, Related changes, Special pages, Permanent link, and Page information. At the top right, there is a user status bar showing "Not logged in" and links for Talk, Contributions, Create account, and Log in. Below this is a navigation bar with "Main Page" and "Talk" tabs, and a search bar labeled "Search Wikipedia". A large red-bordered box in the center contains a puzzle piece icon and text: "Photograph a historic site, help Wikipedia, and win a prize. Participate in the world's largest photography competition this month! Learn more". Below this is a "Welcome to Wikipedia" section with the text "the free encyclopedia that anyone can edit. 6,176,760 articles in English". To the right of this is a list of portals: Arts, Biography, Geography, History, Mathematics, Science, Society, Technology, and All portals. Below the welcome section is a "From today's featured article" box featuring a portrait of V. Gordon Childe and text about his life and work. To the right of this is an "In the news" box with a headline about the COVID-19 pandemic and a list of news items, including one about the New Zealand Labour Party and its leader Jacinda Ardern, who is pictured in a small image.

WIKIPEDIA
The Free Encyclopedia


Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute
Help
Learn to edit
Community portal
Recent changes
Upload file

Tools
What links here
Related changes
Special pages
Permanent link
Page information

Not logged in [Talk](#) [Contributions](#) [Create account](#) [Log in](#)


Main Page [Talk](#) [Read](#) [View source](#) [View history](#)

 Photograph a historic site, help Wikipedia, and win a prize. Participate in the world's largest photography competition this month! [Learn more](#)

Welcome to **Wikipedia**,
the free encyclopedia that anyone can edit.
6,176,760 articles in English

- Arts
- Biography
- Geography
- History
- Mathematics
- Science
- Society
- Technology
- **All portals**


From today's featured article

 **V. Gordon Childe** (1892–1957) was an Australian archaeologist who specialised in the study of [European prehistory](#). He spent most of his life in the United Kingdom, working as an academic for the [University of Edinburgh](#) and then the [Institute of Archaeology](#), London, and wrote twenty-six books during his career. Initially an early proponent of [culture-historical archaeology](#), he later became the

In the news

COVID-19 pandemic: [Disease](#) · [Virus](#) · [By location](#) · [Impact](#) · [Portal](#)

- The [New Zealand Labour Party](#), led by incumbent Prime Minister [Jacinda Ardern](#) (*pictured*), wins a majority of seats in **the general election**.
- After [Sooronbay](#)



FYI: Pointwise mutual information

$$\text{pmi}(x; y)$$

$$= -\log p(x) - \log p(y) + \log p(x, y)$$

$$= \log \frac{p(x, y)}{p(x) p(y)}$$

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

Mutual Information for Continuous Variables

Definition of mutual information

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

... holds for continuous variables

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

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

$$\text{pdf}(y | x) = \text{pdf}(y, x) / p(x)$$

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

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

Note on mutual information for continuous variables

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

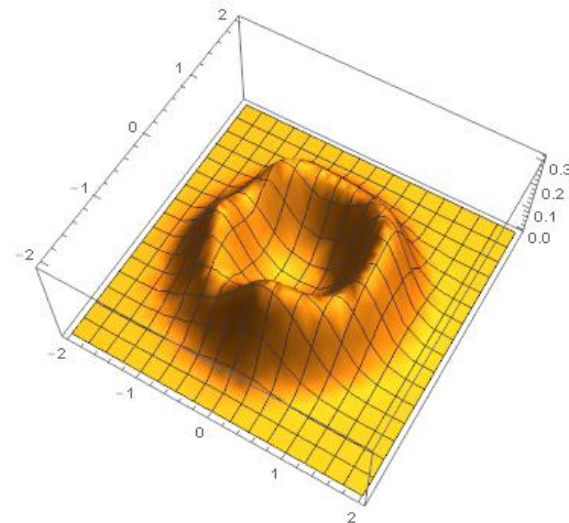
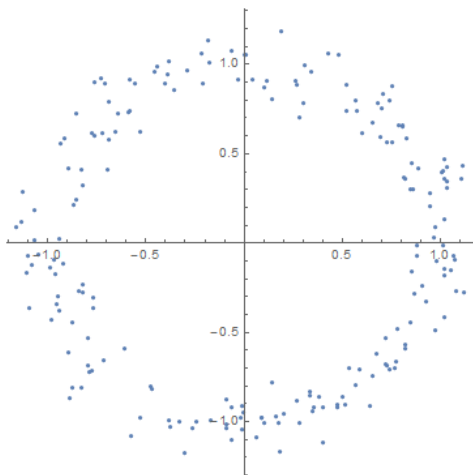
- Both H_{dif} originally contained the same “infinity” term, which cancel out
 - No infinity is ignored in the definition of $I(Y; X)$
 - The value of $I(Y; X)$ has actual meaning; the amount of information shared between X and Y
 - $I(Y; X)$ is always non-negative

Calculating mutual information from data points of continuous values

- Create a smooth PDF using, e.g., Gaussian kernel method

- Calculate

$$I(Y; X) = H_{\text{dif}}(Y) + H_{\text{dif}}(X) - H_{\text{dif}}(YX)$$

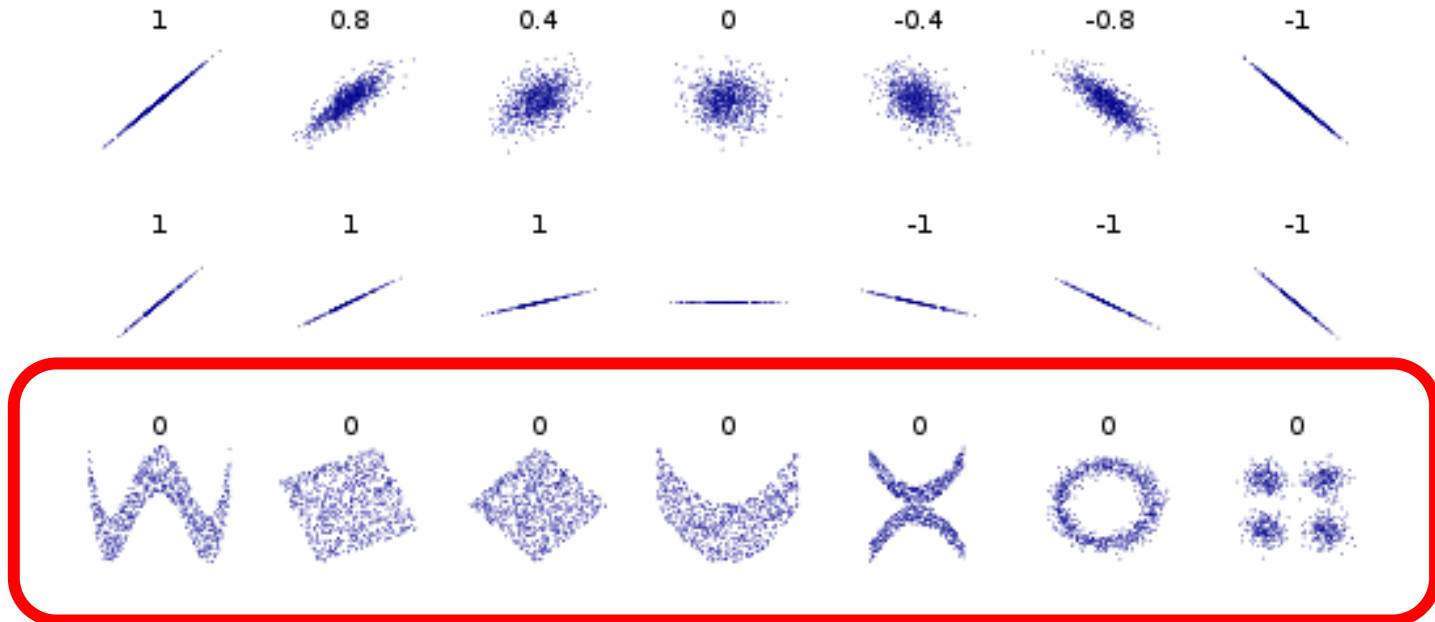


Exercise

- 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

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

- This measures how distribution $p(x)$ is different from $q(x)$
- It is known that:

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

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

Exercise

- Prove this:

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