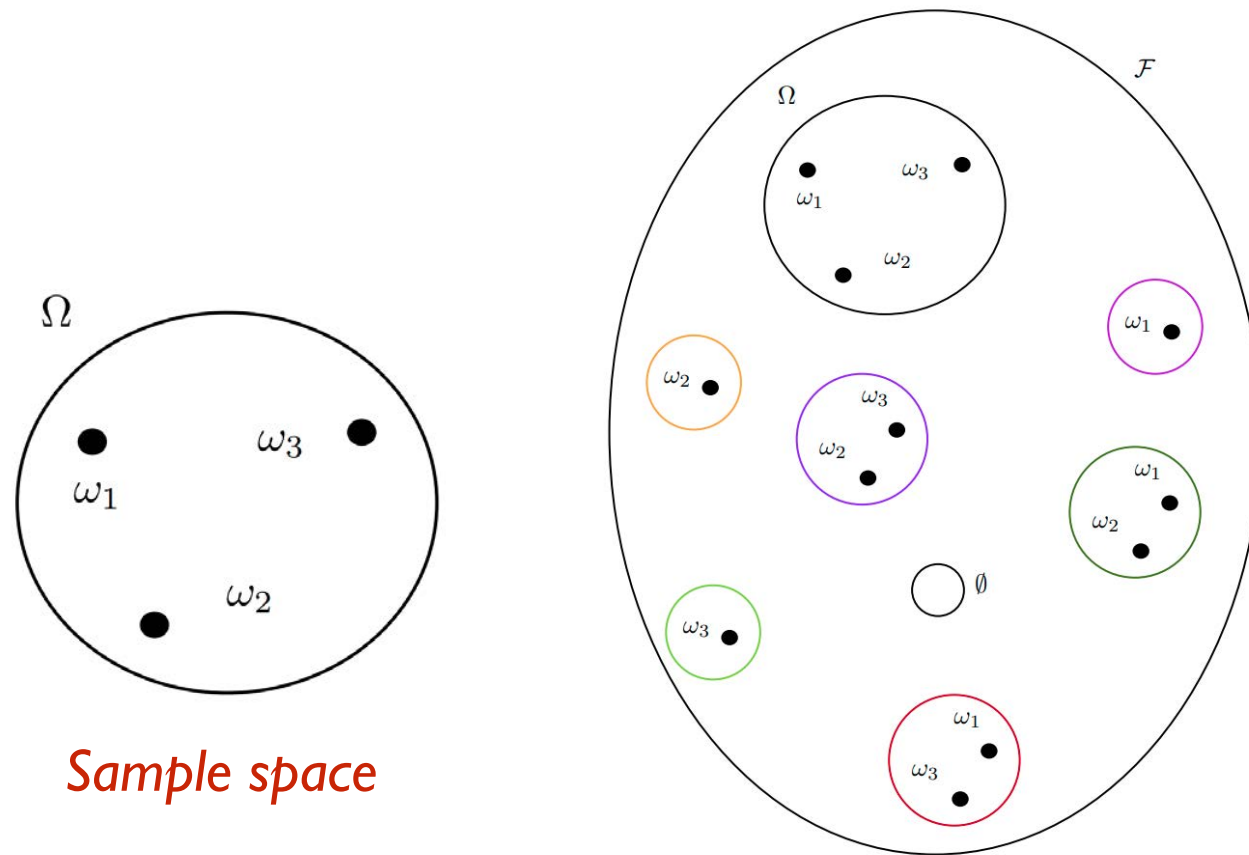


MEAM 4600: AI for Science and Engineering

*Lecture #4: Primer on Probability*

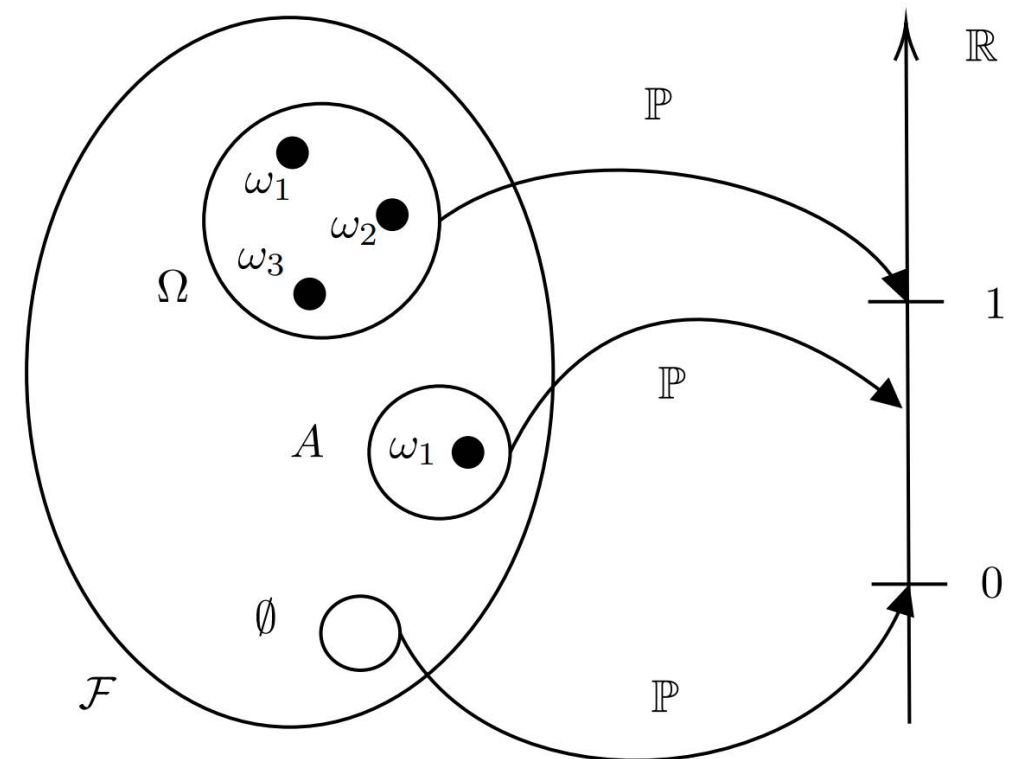


# Probability spaces & random variables



Sample space

$\sigma$ -algebra of events



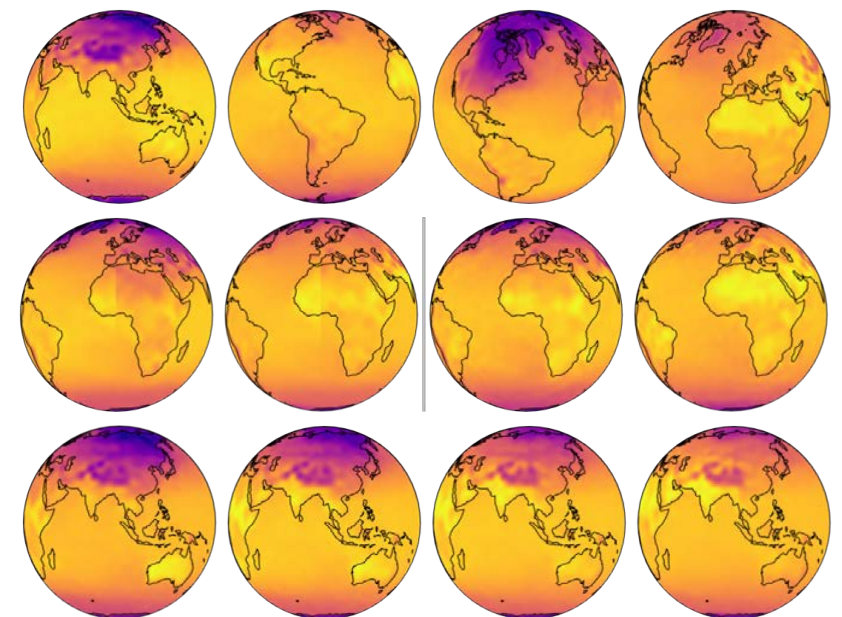
Probability measure



vectors



matrices



functions

## Recap: Continuous random variables

- A ***continuous random variable*** is one which takes an infinite number of possible values. Continuous random variables are usually measurements. Examples include height, weight, the amount of sugar in an orange, the time required to run a mile.
- A continuous random variable is not defined at specific values. Instead, it is defined over an *interval* of values, and is represented by the ***area under a curve*** (in advanced mathematics, this is known as an *integral*). The probability of observing any single value is equal to 0, since the number of values which may be assumed by the random variable is infinite.

# Probability density functions

$$\Pr[a \leq X \leq b] = \int_a^b f_X(x) dx.$$

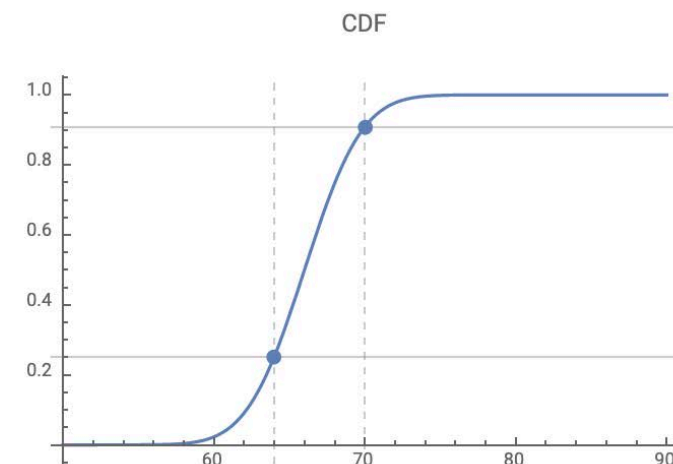
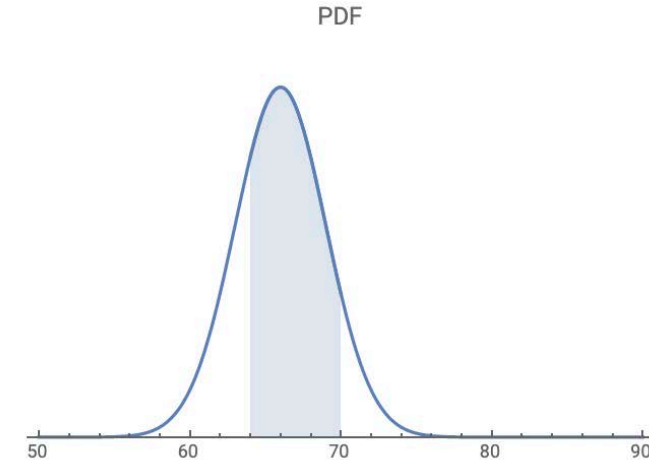
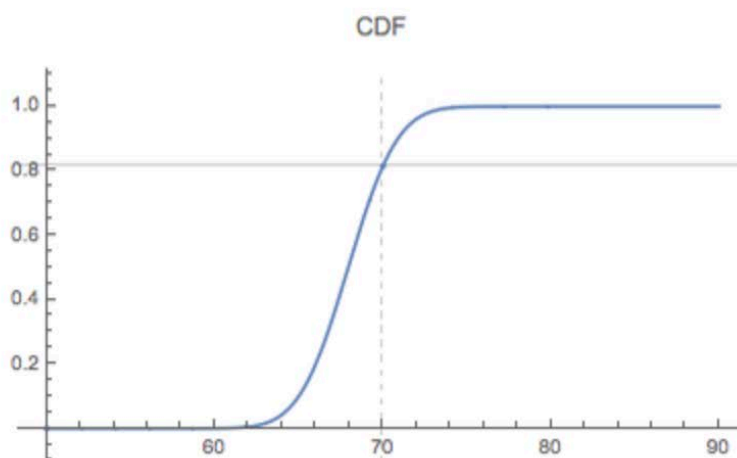
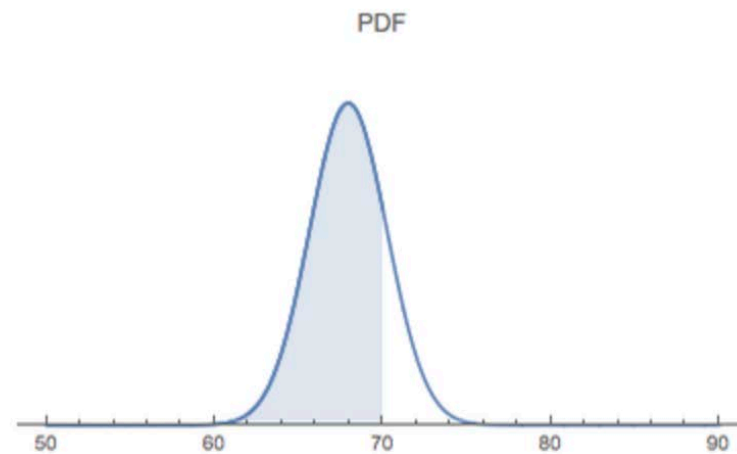
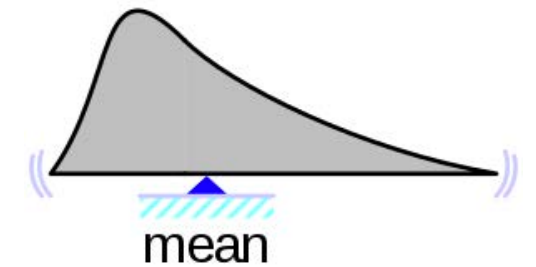
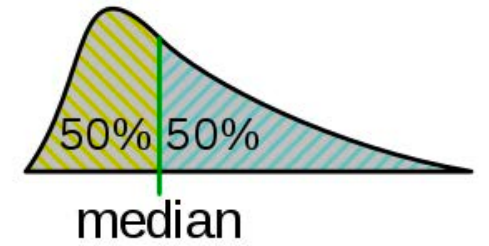
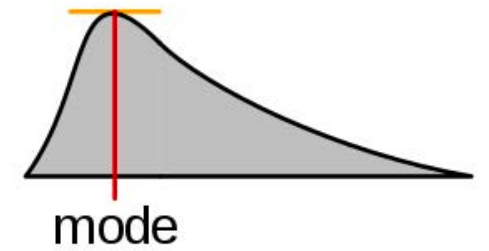
Hence, if  $F_X$  is the **cumulative distribution function** of  $X$ , then:

$$F_X(x) = \int_{-\infty}^x f_X(u) du,$$

and (if  $f_X$  is continuous at  $x$ )

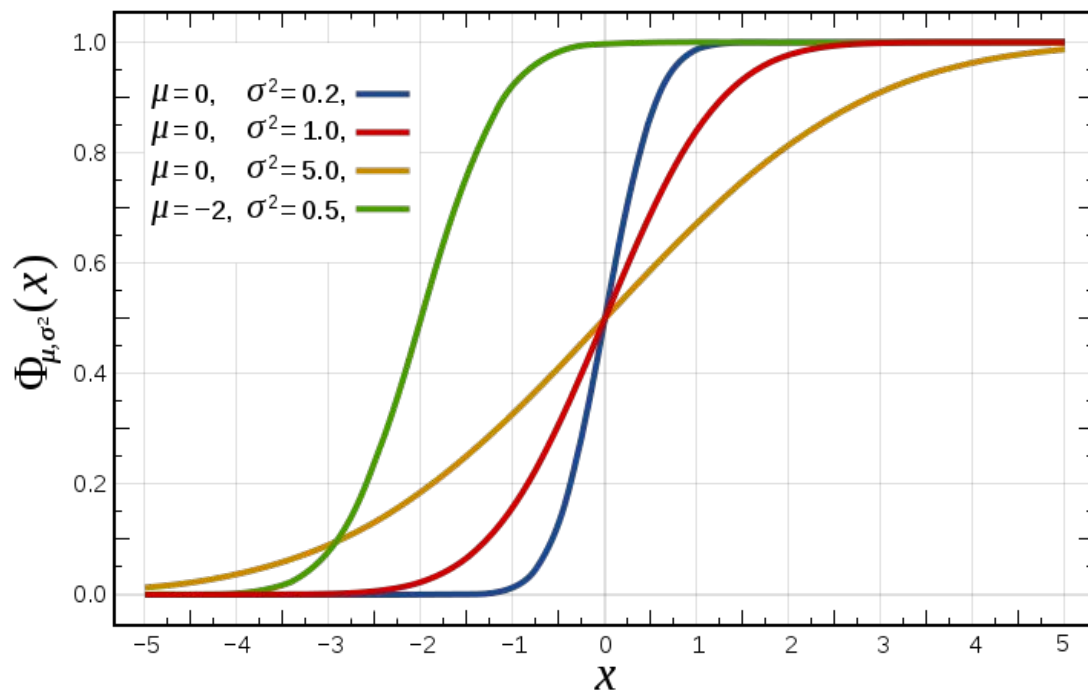
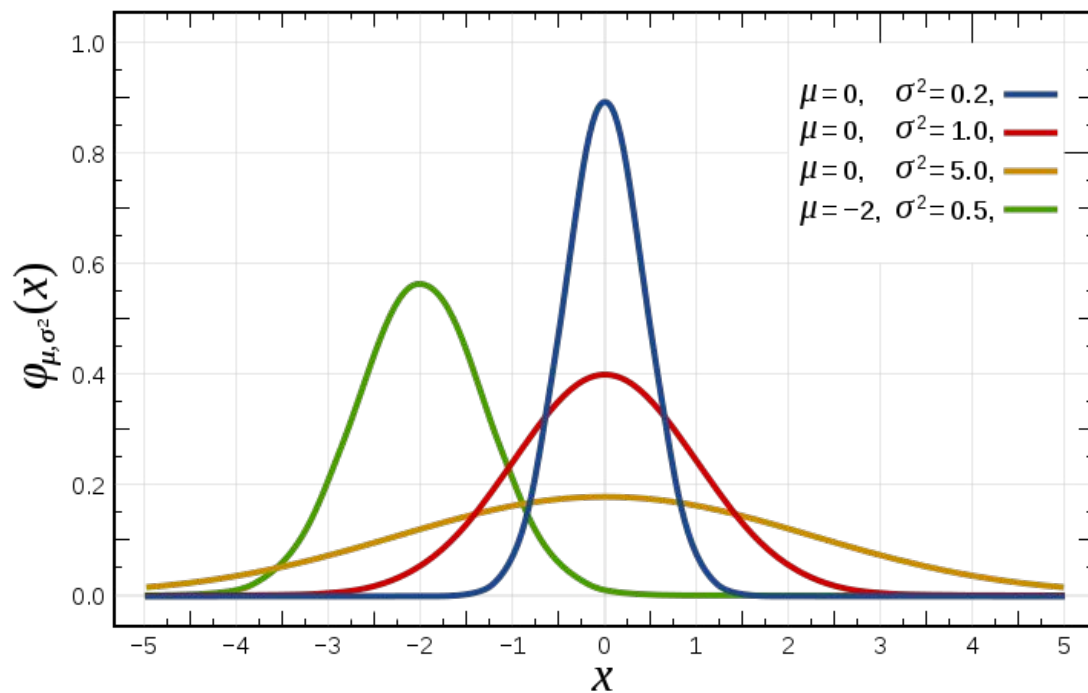
$$f_X(x) = \frac{d}{dx} F_X(x).$$

Intuitively, one can think of  $f_X(x) dx$  as being the probability of  $X$  falling within the infinitesimal **interval**  $[x, x + dx]$ .



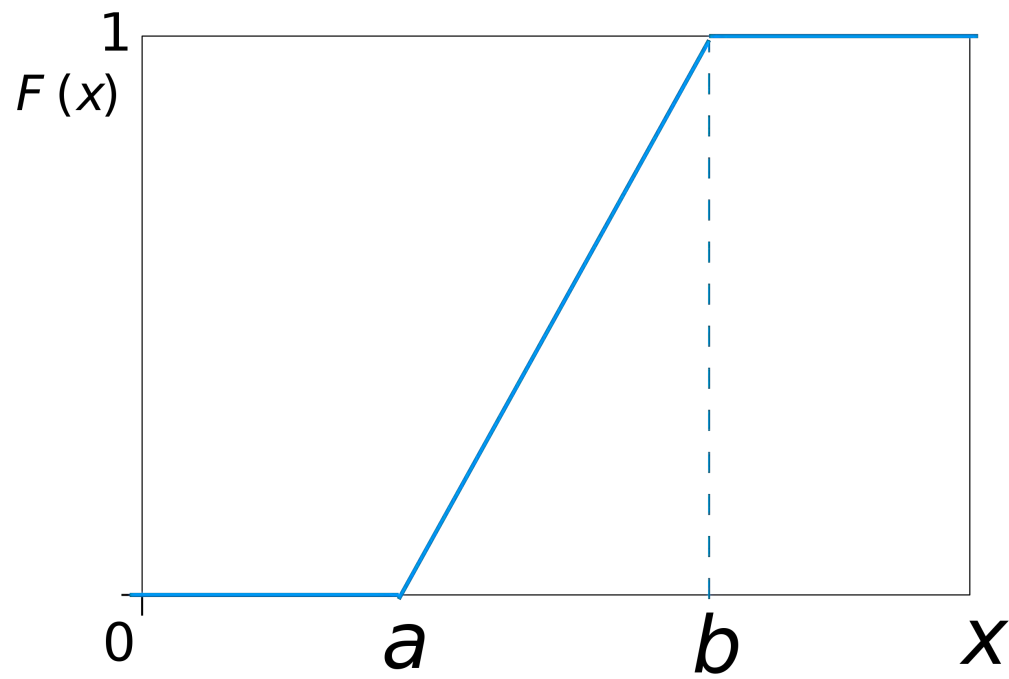
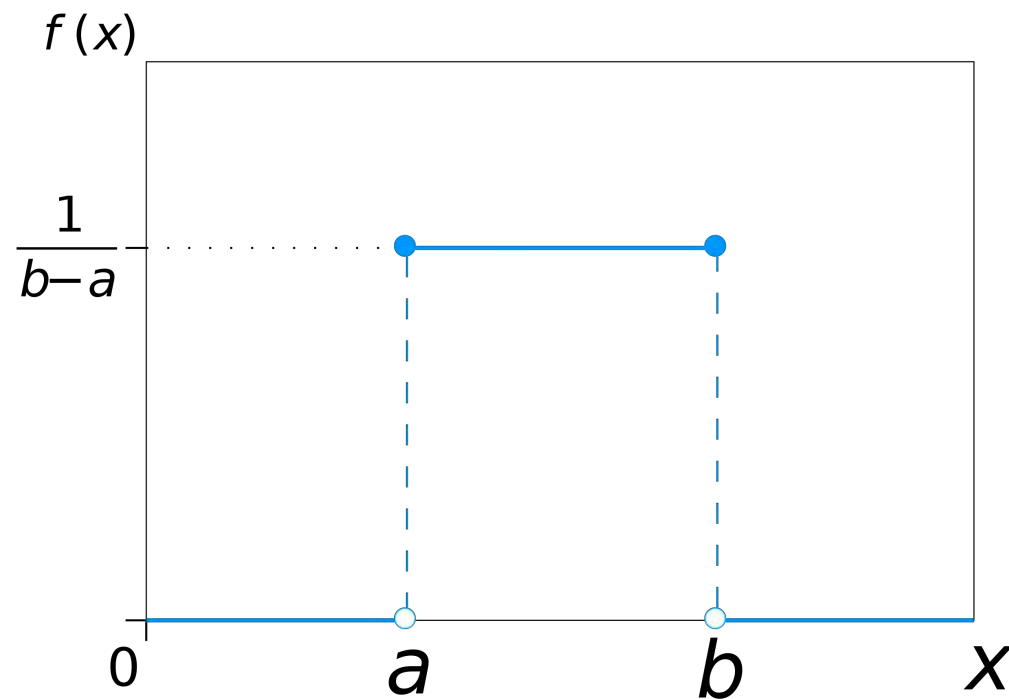


# Recap: Univariate Gaussian distribution



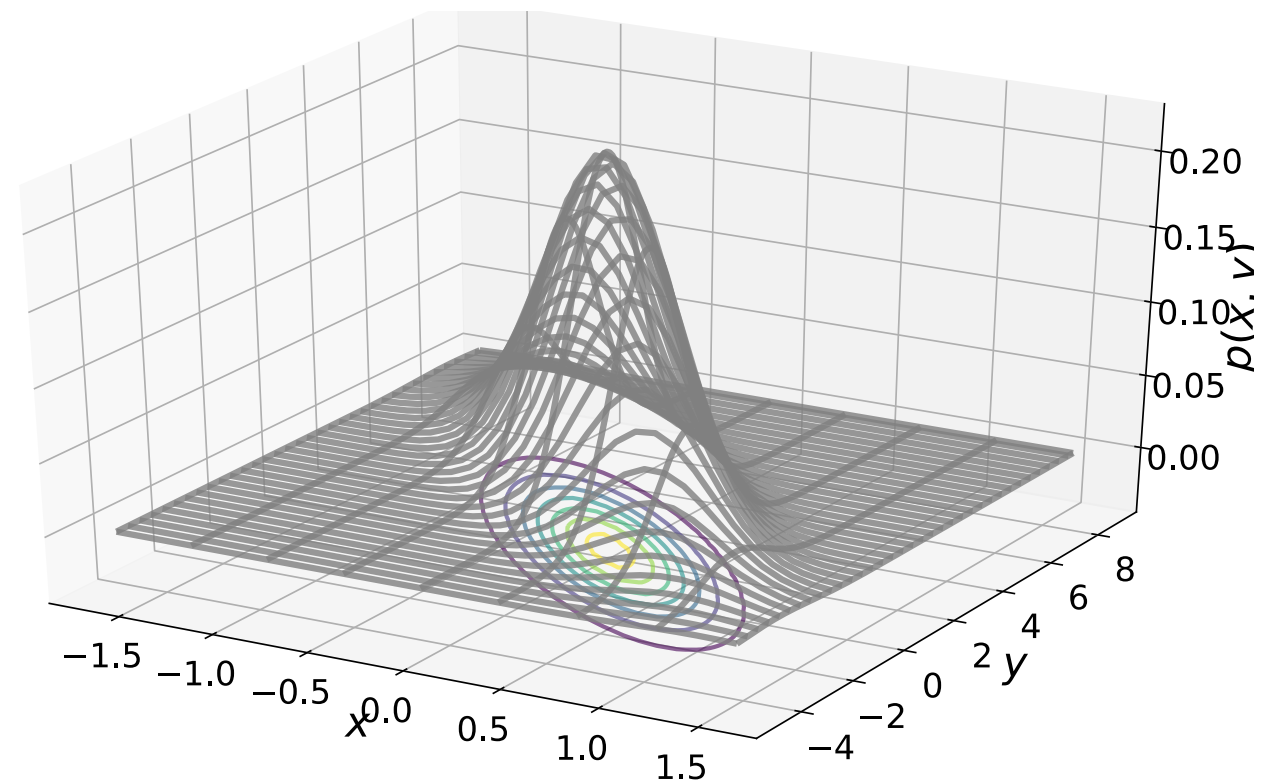
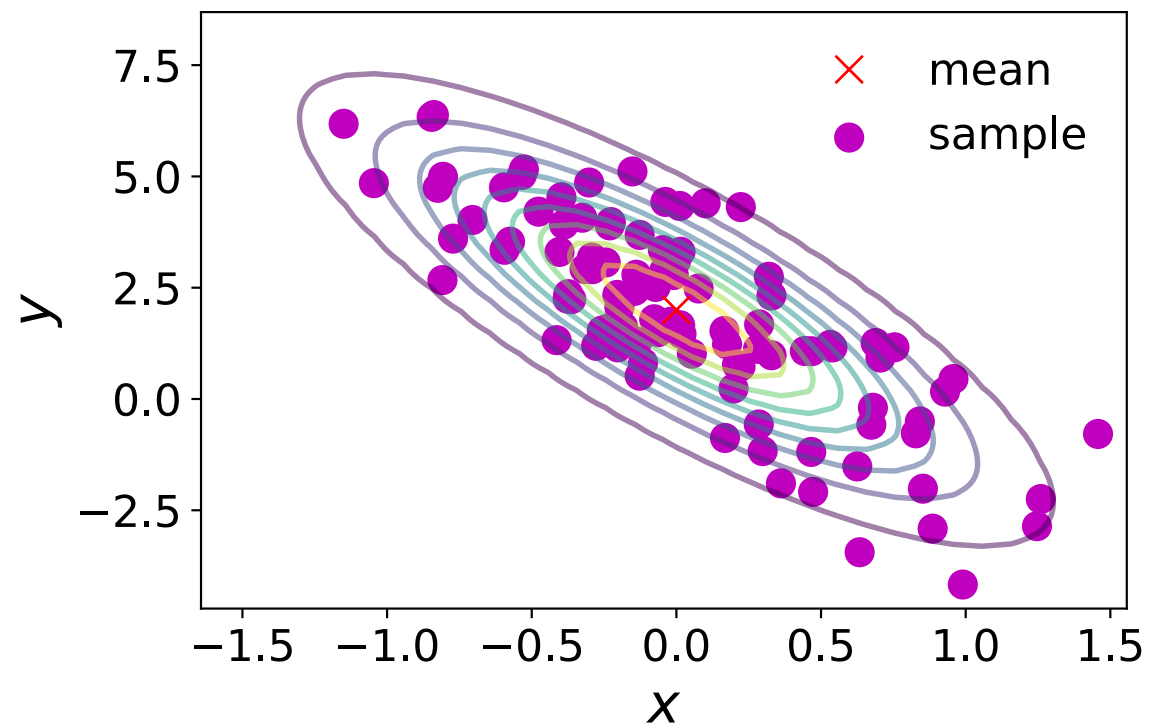
<b>Notation</b>	$\mathcal{N}(\mu, \sigma^2)$
<b>Parameters</b>	$\mu \in \mathbb{R}$ = mean (location) $\sigma^2 > 0$ = variance (squared scale)
<b>Support</b>	$x \in \mathbb{R}$
<b>PDF</b>	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
<b>CDF</b>	$\frac{1}{2} \left[ 1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right) \right]$
<b>Quantile</b>	$\mu + \sigma\sqrt{2} \operatorname{erf}^{-1}(2F - 1)$
<b>Mean</b>	$\mu$
<b>Median</b>	$\mu$
<b>Mode</b>	$\mu$
<b>Variance</b>	$\sigma^2$
<b>Skewness</b>	0
<b>Ex. kurtosis</b>	0
<b>Entropy</b>	$\frac{1}{2} \log(2\pi e \sigma^2)$
<b>MGF</b>	$\exp(\mu t + \sigma^2 t^2 / 2)$
<b>CF</b>	$\exp(i\mu t - \sigma^2 t^2 / 2)$
<b>Fisher information</b>	$\mathcal{I}(\mu, \sigma) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix} \quad \mathcal{I}(\mu, \sigma^2) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{pmatrix}$
<b>Kullback-Leibler divergence</b>	$D_{\text{KL}}(\mathcal{N}_0 \parallel \mathcal{N}_1) = \frac{1}{2} \left\{ (\sigma_0/\sigma_1)^2 + \frac{(\mu_1 - \mu_0)^2}{\sigma_1^2} - 1 + 2 \ln \frac{\sigma_1}{\sigma_0} \right\}$

# Recap: Uniform distribution



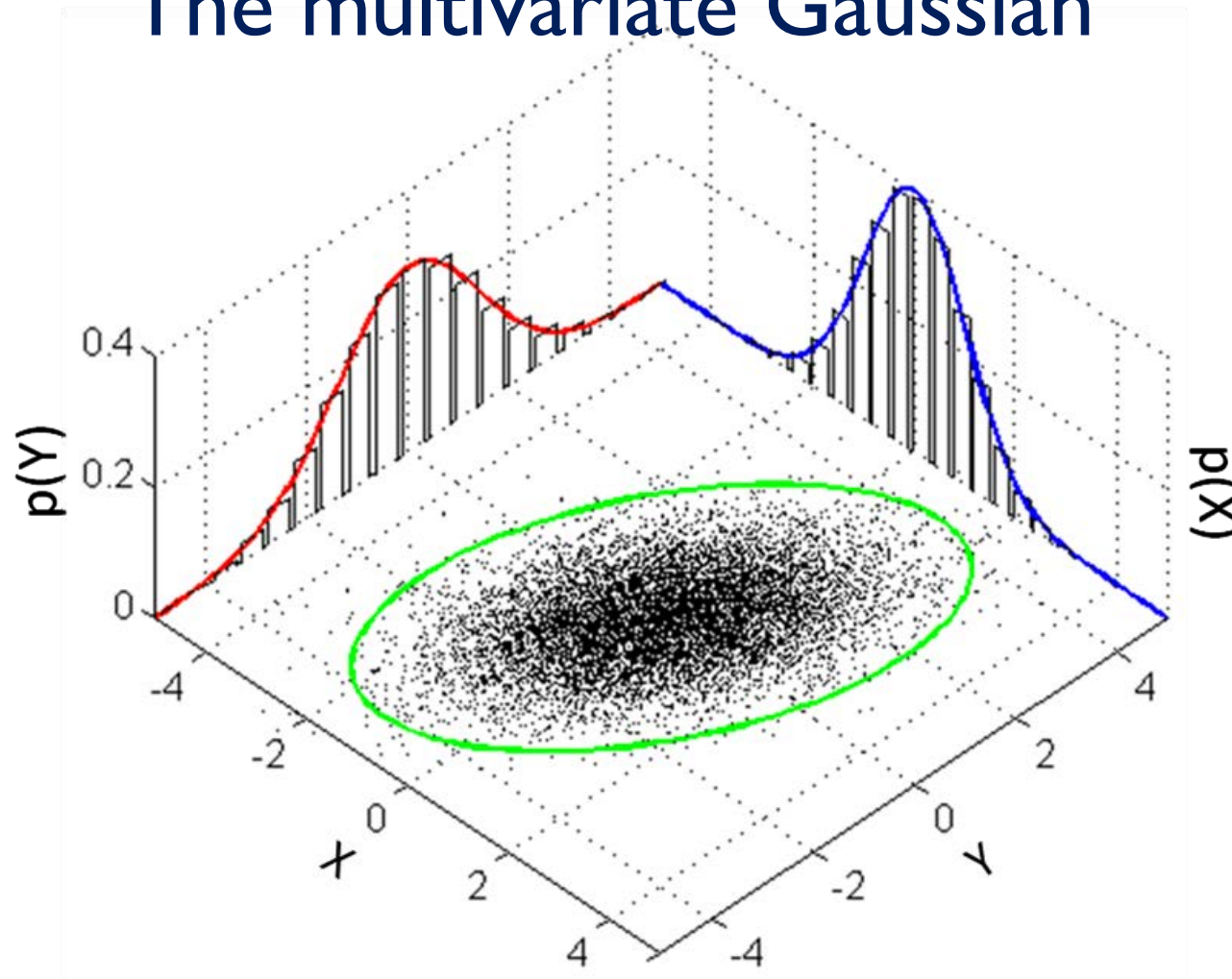
<b>Notation</b>	$\mathcal{U}(a, b)$ or $\text{unif}(a, b)$
<b>Parameters</b>	$-\infty < a < b < \infty$
<b>Support</b>	$x \in [a, b]$
<b>PDF</b>	$\begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$
<b>CDF</b>	$\begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a, b) \\ 1 & \text{for } x \geq b \end{cases}$
<b>Mean</b>	$\frac{1}{2}(a + b)$
<b>Median</b>	$\frac{1}{2}(a + b)$
<b>Mode</b>	any value in $(a, b)$
<b>Variance</b>	$\frac{1}{12}(b - a)^2$
<b>Skewness</b>	$0$
<b>Ex. kurtosis</b>	$-\frac{6}{5}$
<b>Entropy</b>	$\ln(b - a)$
<b>MGF</b>	$\begin{cases} \frac{e^{tb} - e^{ta}}{t(b-a)} & \text{for } t \neq 0 \\ 1 & \text{for } t = 0 \end{cases}$
<b>CF</b>	$\begin{cases} \frac{e^{itb} - e^{ita}}{it(b-a)} & \text{for } t \neq 0 \\ 1 & \text{for } t = 0 \end{cases}$

# The multivariate Gaussian



$$p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp \left( -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right)$$

# The multivariate Gaussian

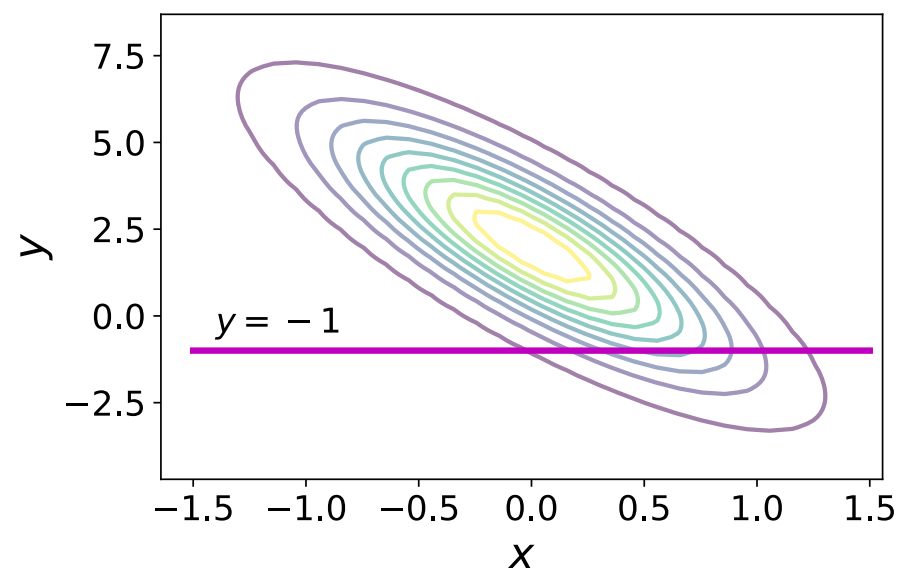


<b>Notation</b>	$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
<b>Parameters</b>	$\boldsymbol{\mu} \in \mathbf{R}^k$ — location $\boldsymbol{\Sigma} \in \mathbf{R}^{k \times k}$ — covariance (positive semi-definite matrix)
<b>Support</b>	$\mathbf{x} \in \boldsymbol{\mu} + \text{span}(\boldsymbol{\Sigma}) \subseteq \mathbf{R}^k$
<b>PDF</b>	$\det(2\pi\boldsymbol{\Sigma})^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$ , exists only when $\boldsymbol{\Sigma}$ is positive-definite
<b>Mean</b>	$\boldsymbol{\mu}$
<b>Mode</b>	$\boldsymbol{\mu}$
<b>Variance</b>	$\boldsymbol{\Sigma}$



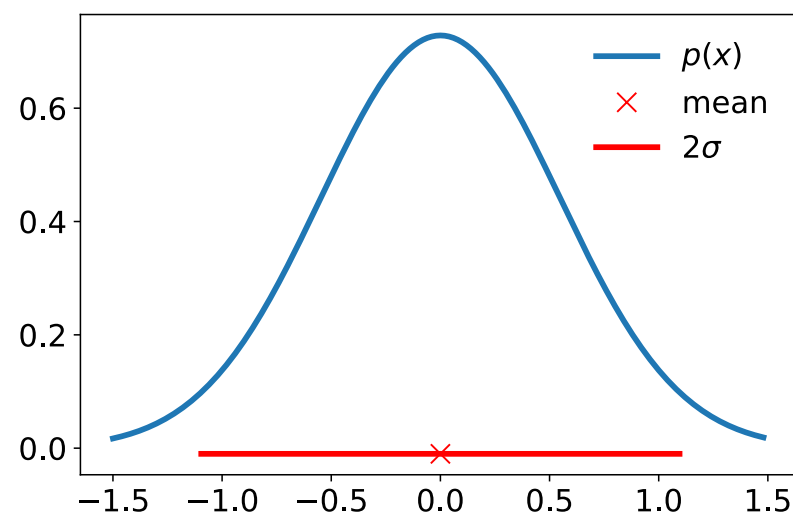
# Marginals and conditionals of a Gaussian

$$p(\mathbf{x}, \mathbf{y}) = \mathcal{N} \left( \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix} \right)$$



*Marginal distribution*

$$p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{y}) d\mathbf{y} = \mathcal{N}(\mathbf{x} \mid \mu_x, \Sigma_{xx})$$

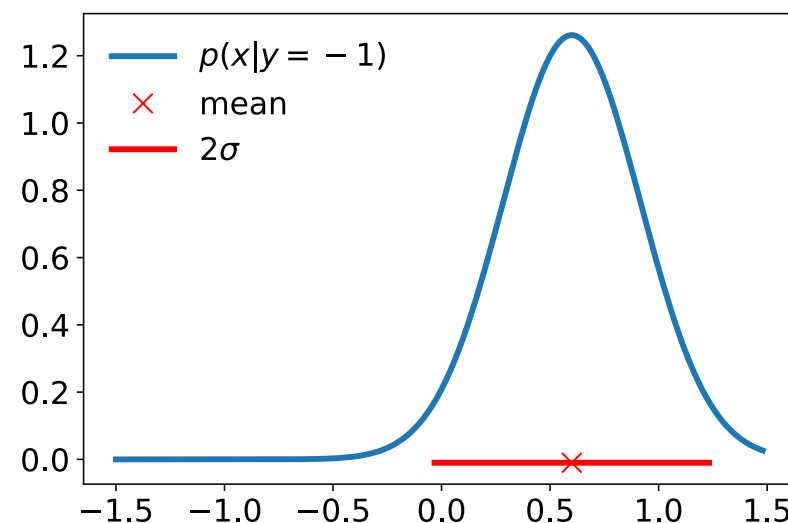


*Conditional distribution*

$$p(\mathbf{x} \mid \mathbf{y}) = \mathcal{N}(\mu_{x \mid y}, \Sigma_{x \mid y})$$

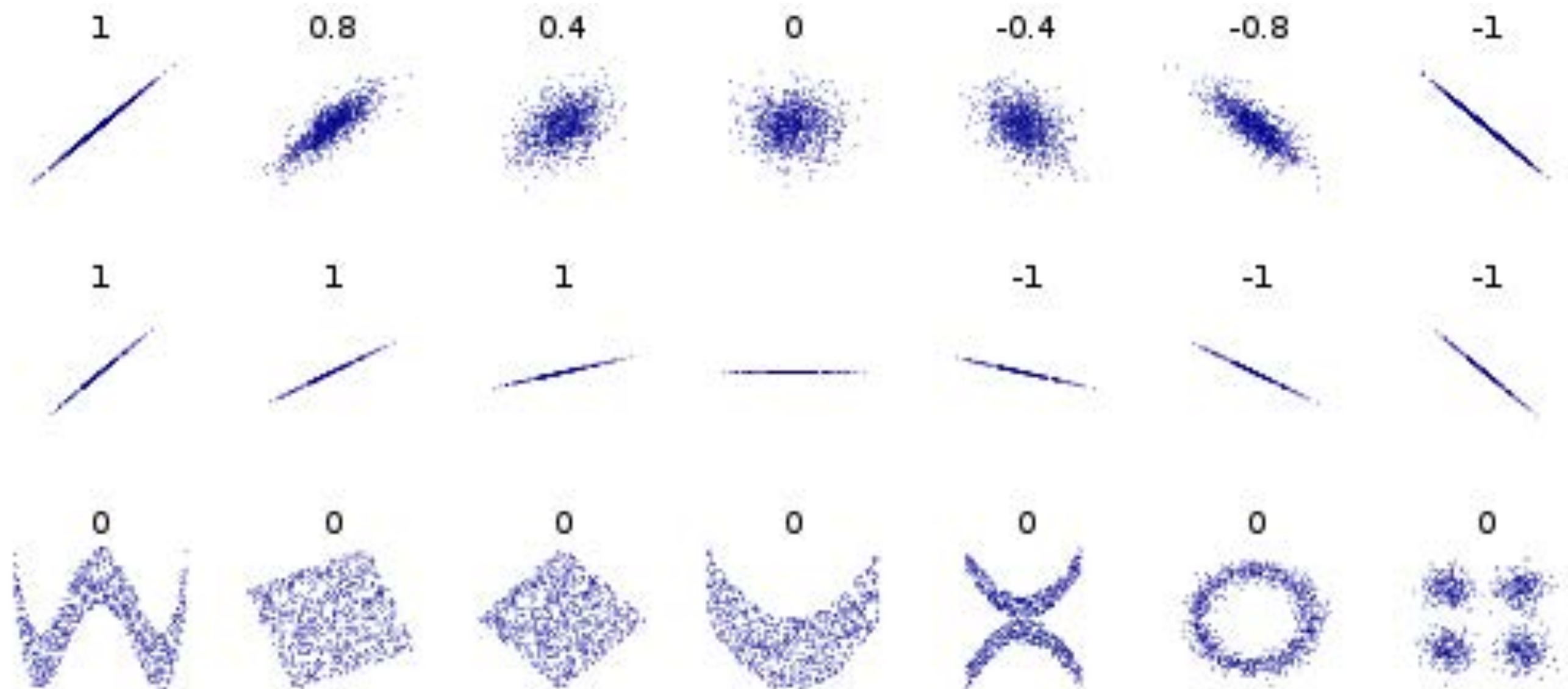
$$\mu_{x \mid y} = \mu_x + \Sigma_{xy} \Sigma_{yy}^{-1} (\mathbf{y} - \mu_y)$$

$$\Sigma_{x \mid y} = \Sigma_{xx} - \Sigma_{xy} \Sigma_{yy}^{-1} \Sigma_{yx}$$



These are unique properties that make the Gaussian distribution very simple and attractive to compute with! It is essentially our main building block for computing under uncertainty.

# Correlation and linear dependence



# Entropy and Mutual Information

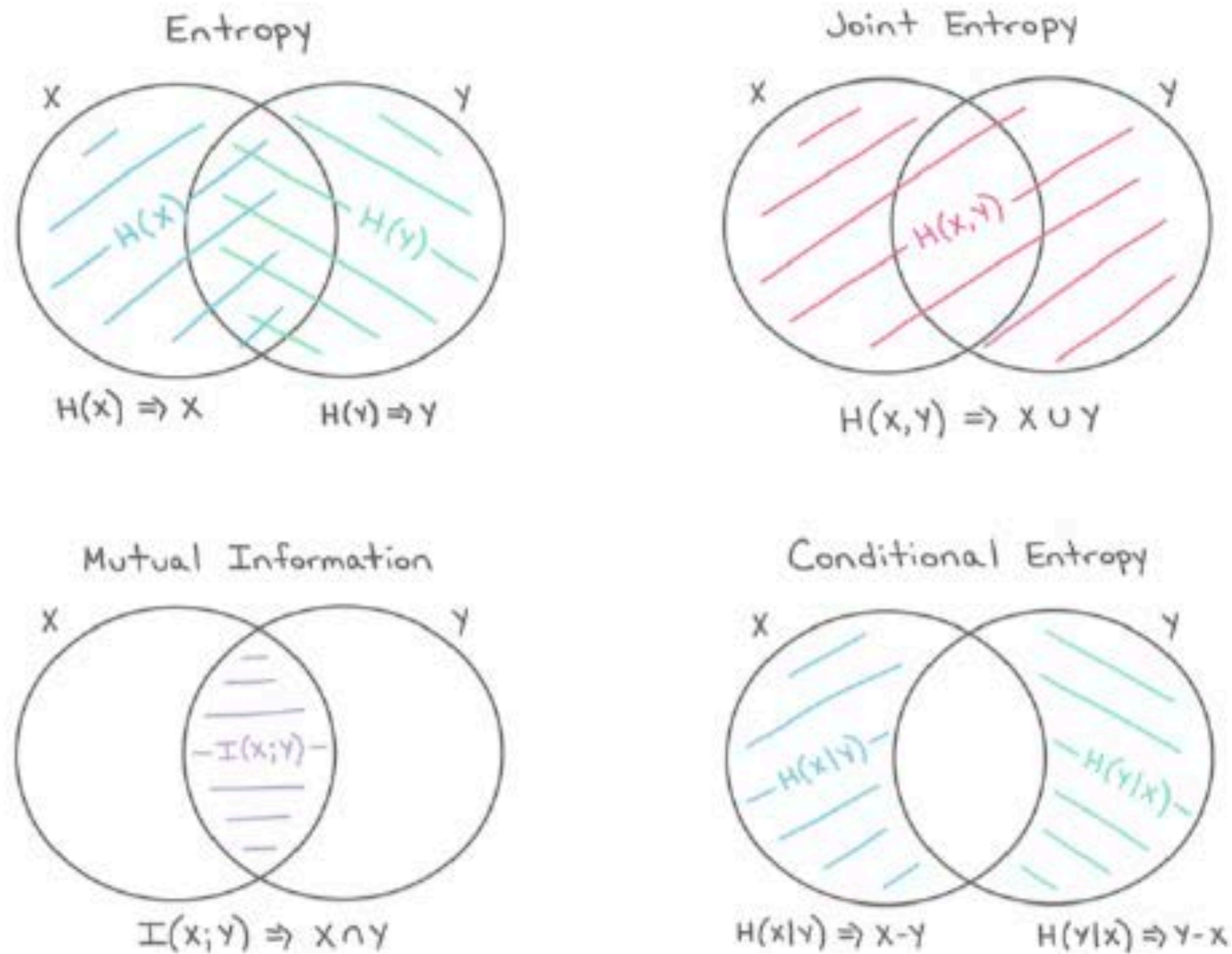


Figure 6.4: The marginal entropy, joint entropy, conditional entropy and mutual information represented as information diagrams. Used with kind permission of Katie Everett.



# Forward vs Reverse KL

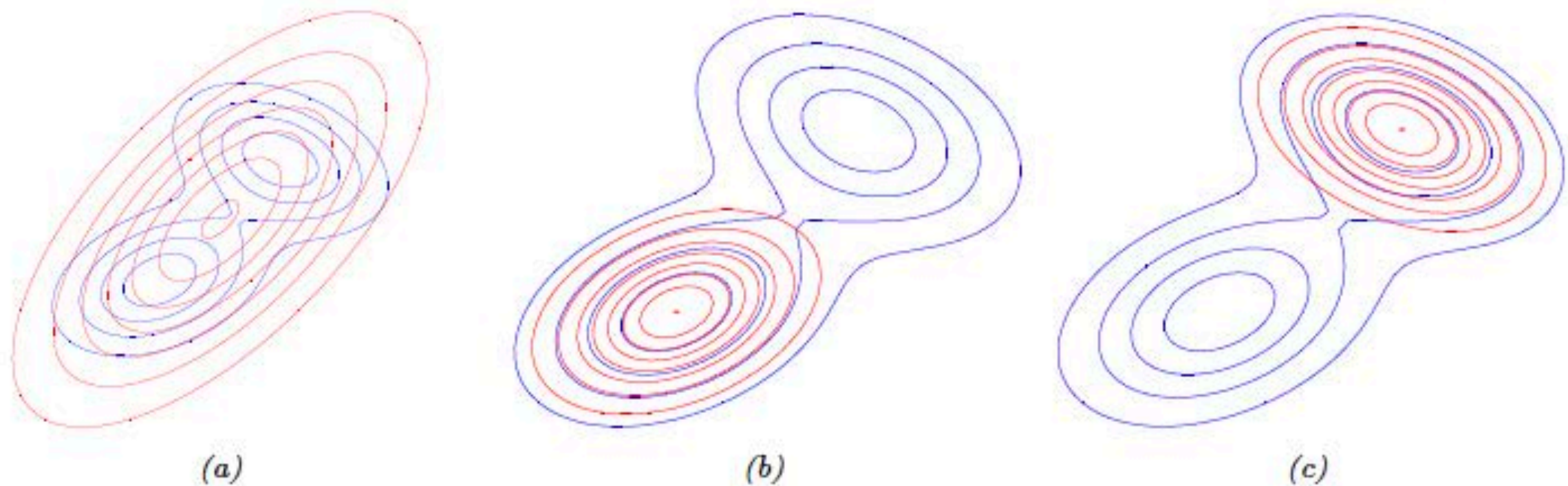
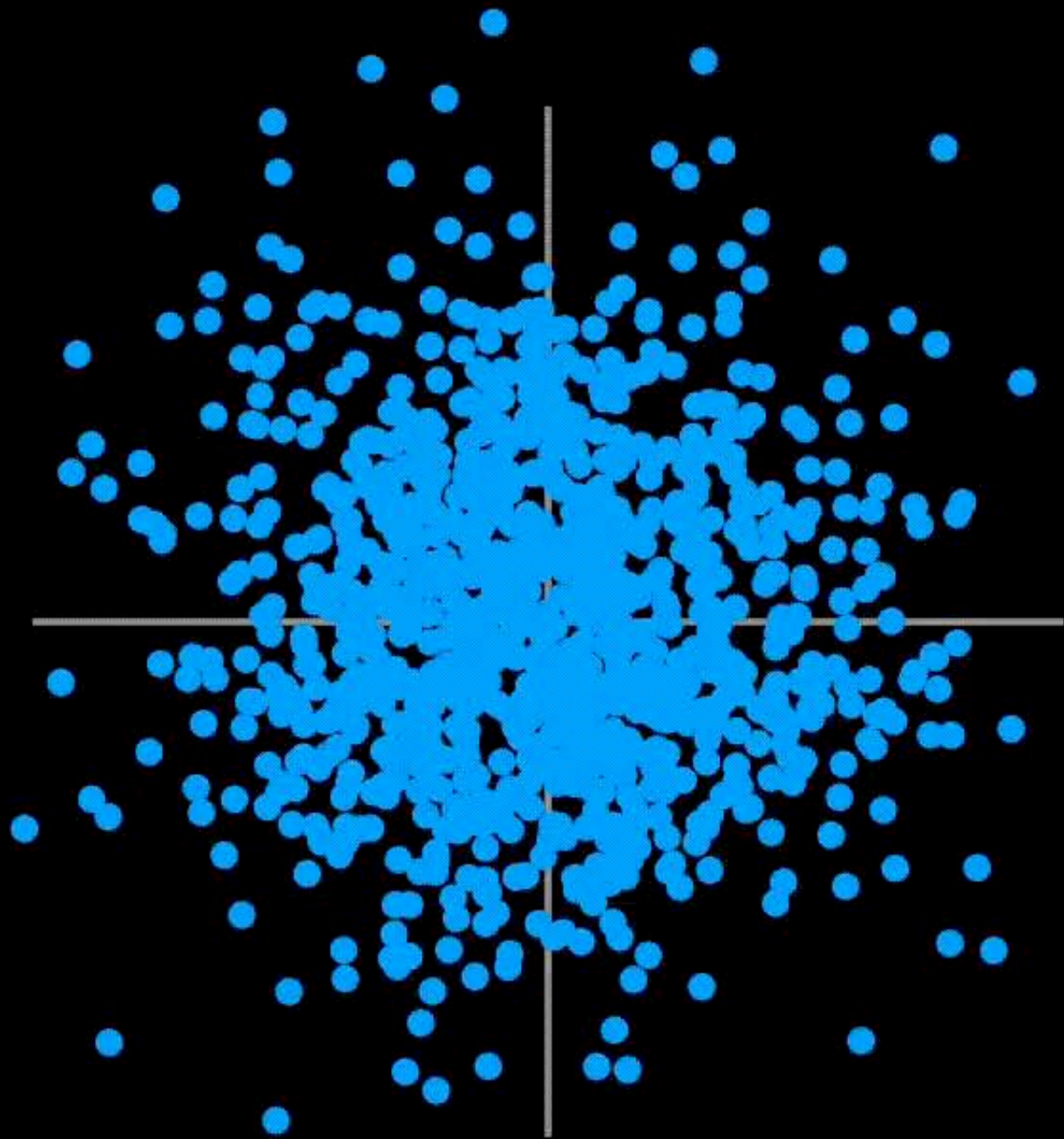


Figure 6.3: Illustrating forwards vs reverse KL on a bimodal distribution. The blue curves are the contours of the true distribution  $p$ . The red curves are the contours of the unimodal approximation  $q$ . (a) Minimizing forwards KL,  $D_{\text{KL}}(p \parallel q)$ , wrt  $q$  causes  $q$  to “cover”  $p$ . (b-c) Minimizing reverse KL,  $D_{\text{KL}}(q \parallel p)$  wrt  $q$  causes  $q$  to “lock onto” one of the two modes of  $p$ . Adapted from Figure 10.3 of [Bis06]. Generated by [KLfwdReverseMixGauss.ipynb](#).



# Covariance vs Mutual Information

$$\text{cov}(X, Y) \quad I(X; Y)$$



@ari\_seff