Lecture 4
Principles of Information Theory and Thermodynamics

- Statistics
- Information
- Entropy
- Free Energy
- Statistical Ensembles
- Boltzmann Distribution
- Target Functions
- Optimisation

Models and Reality

Essentially, all models are wrong, but some are useful. (George E. P. Box)

Observation / Question — world

↓

Model ⇔ Algorithm

↕

Data ⇔ Representation — formalisation, computer

↓

Description / Solution — principle, knowledge

Model and Algorithm

Example Alignment Model:
Sequence evolution by substitution and indel events.

```
GA-GTGA
GAGGCGA
```

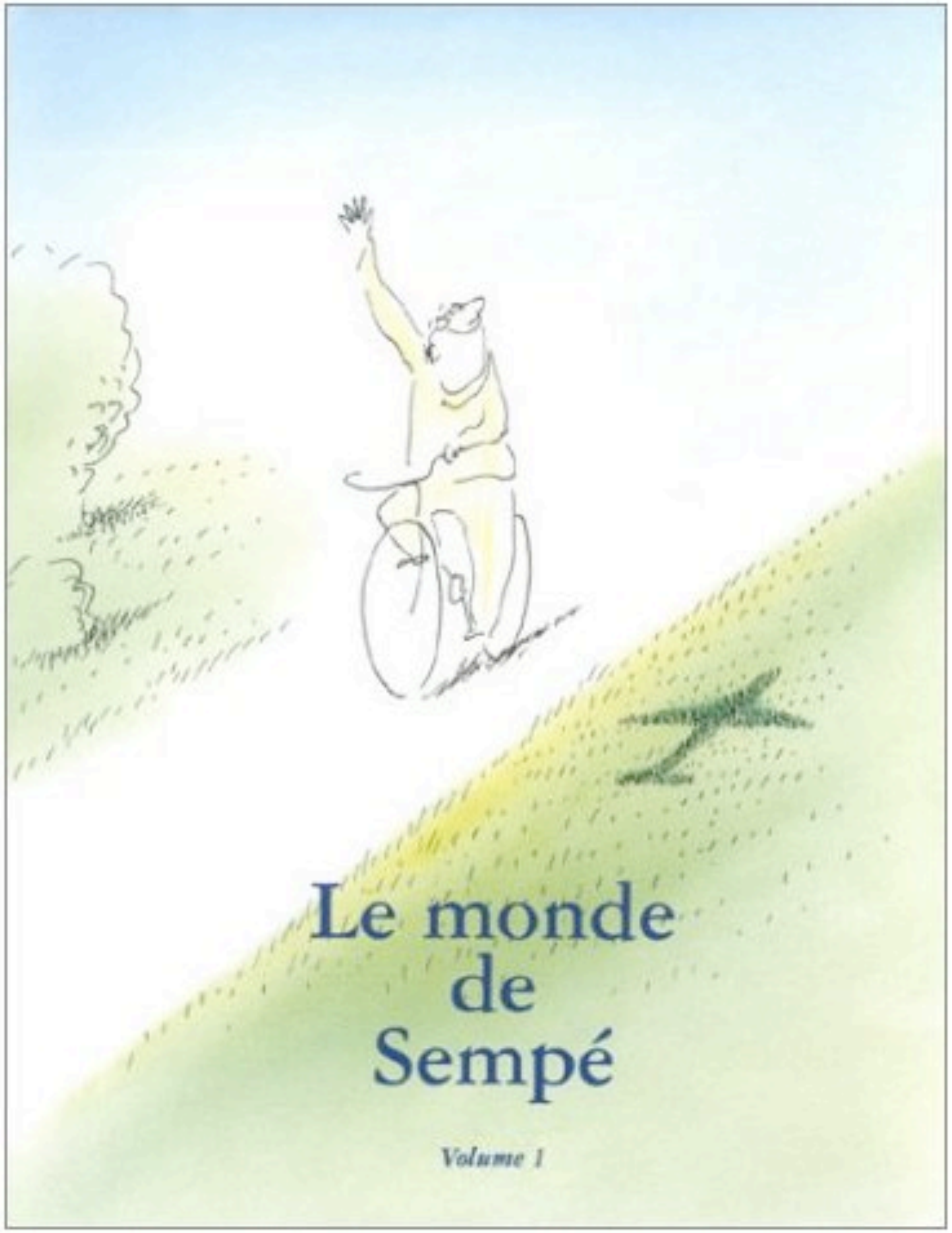
Algorithm:
Dynamic Programming

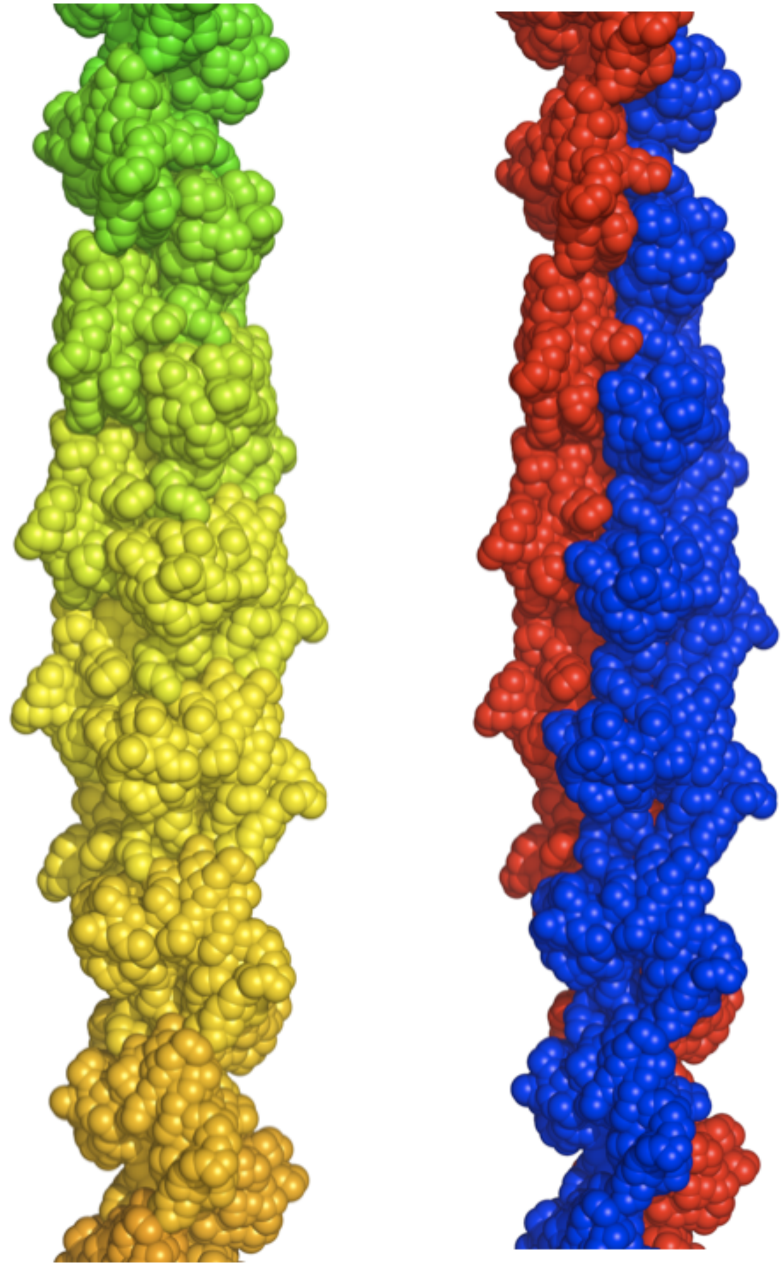
	-	G	A	G	T	G	A
-	0	-2	-4	-6	-8	-10	-12
G	-2	1	-1	-3	-5	-7	-9
A	-4	-1	2	0	-2	-4	-6
G	-6	-3	0	3	1	-1	-3
G	-8	-5	-2	1	2	2	0
C	-10	-7	-4	-1	0	1	1
G	-12	-9	-6	-3	-2	1	0
A	-14	-11	-8	-5	-4	-1	2

Not included in Model:
Inversions

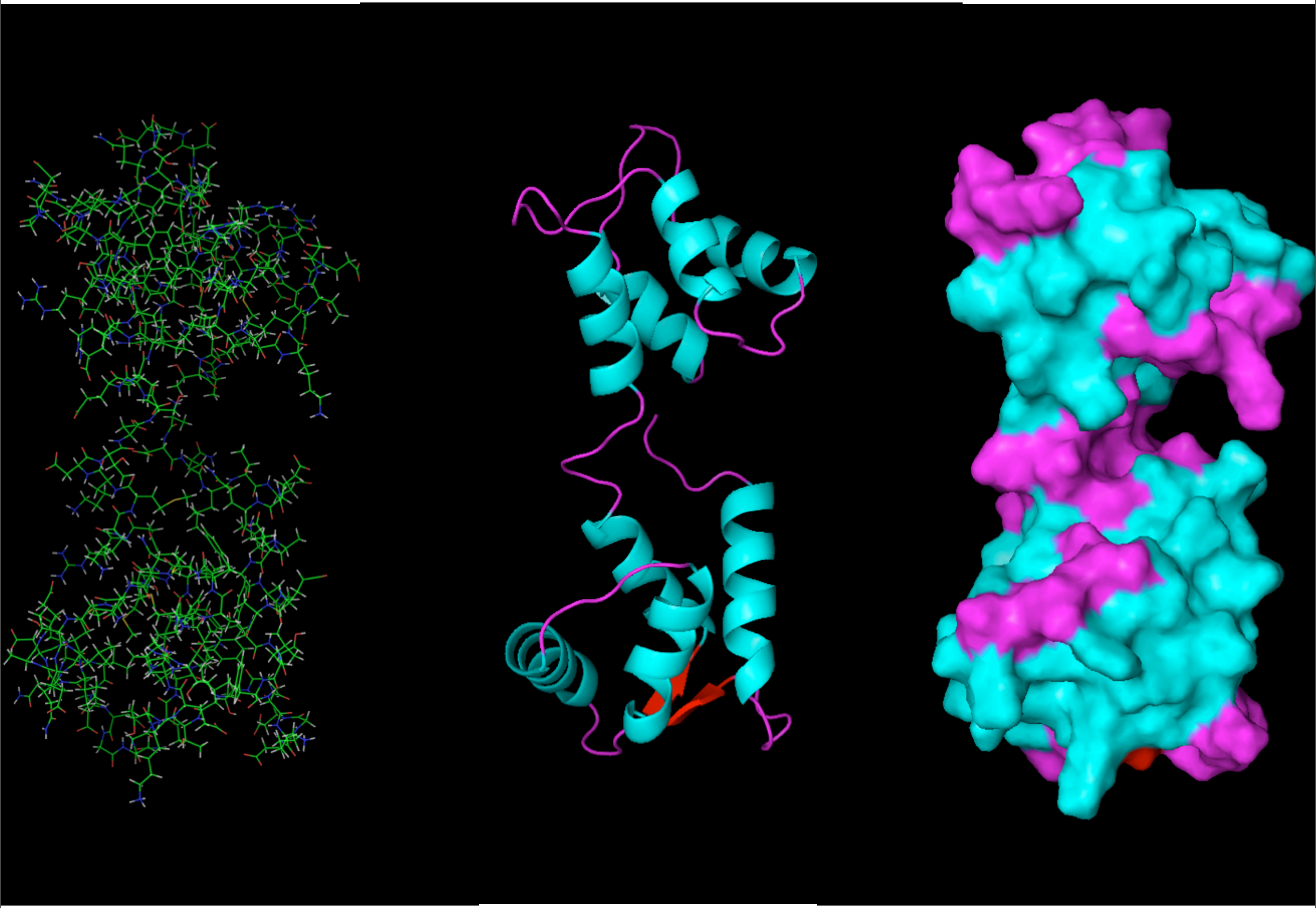
Dynamic Programming fails on Inversions.

Data and Representation





4



Probability

A probability p is a quantitative estimate for the occurrence of an uncertain event E.

Frequency probability

p(E) = Number of times an event occurred / Total number of opportunities for the event to occur

Probability range

The probability of an event ranges from non-occurring to certain:

$$0 <= p <= 1 \qquad \sum_i p_i = 1$$

Joint probability

Two independent events E1 and E2 occur together with probability:

p(E1, E2) = p(E1) * p(E2)

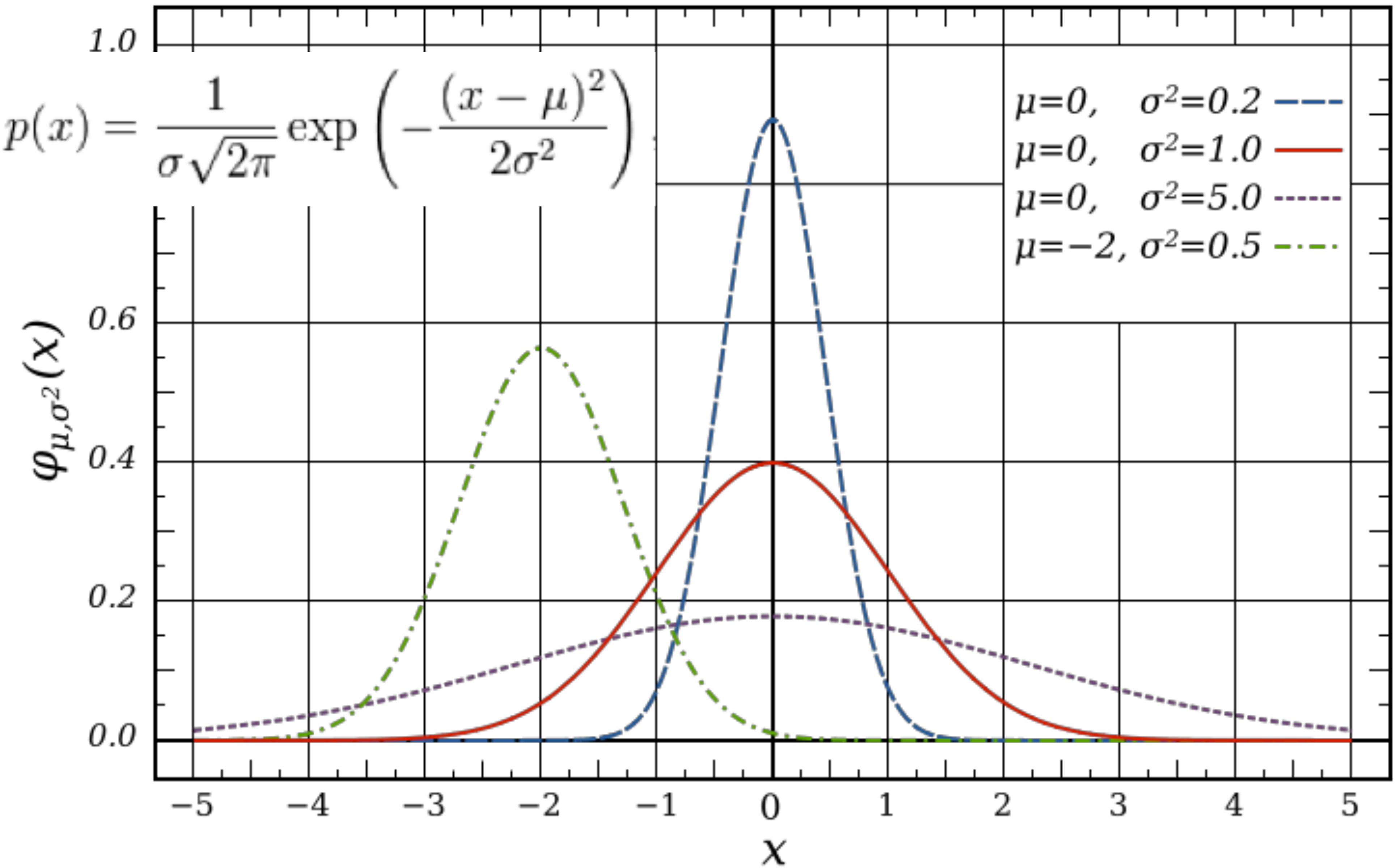
Conditional probability

Two dependent events E1 and E2 occur together with probability:

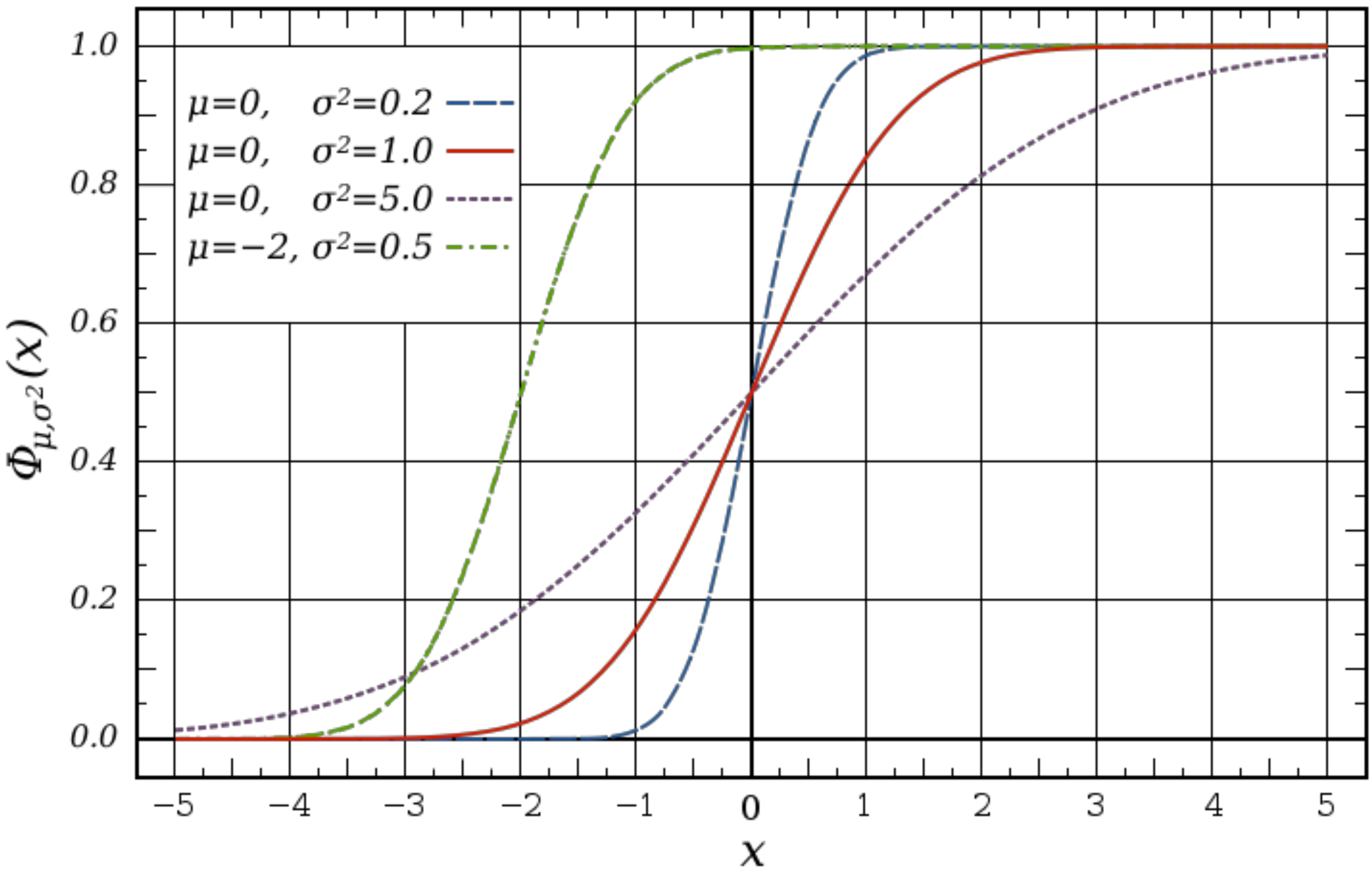
P(E1 and E2) = P(E1) * P(E2 | E1)

p(E2 | E1) [Read: Conditional probability of E2 given E1.]

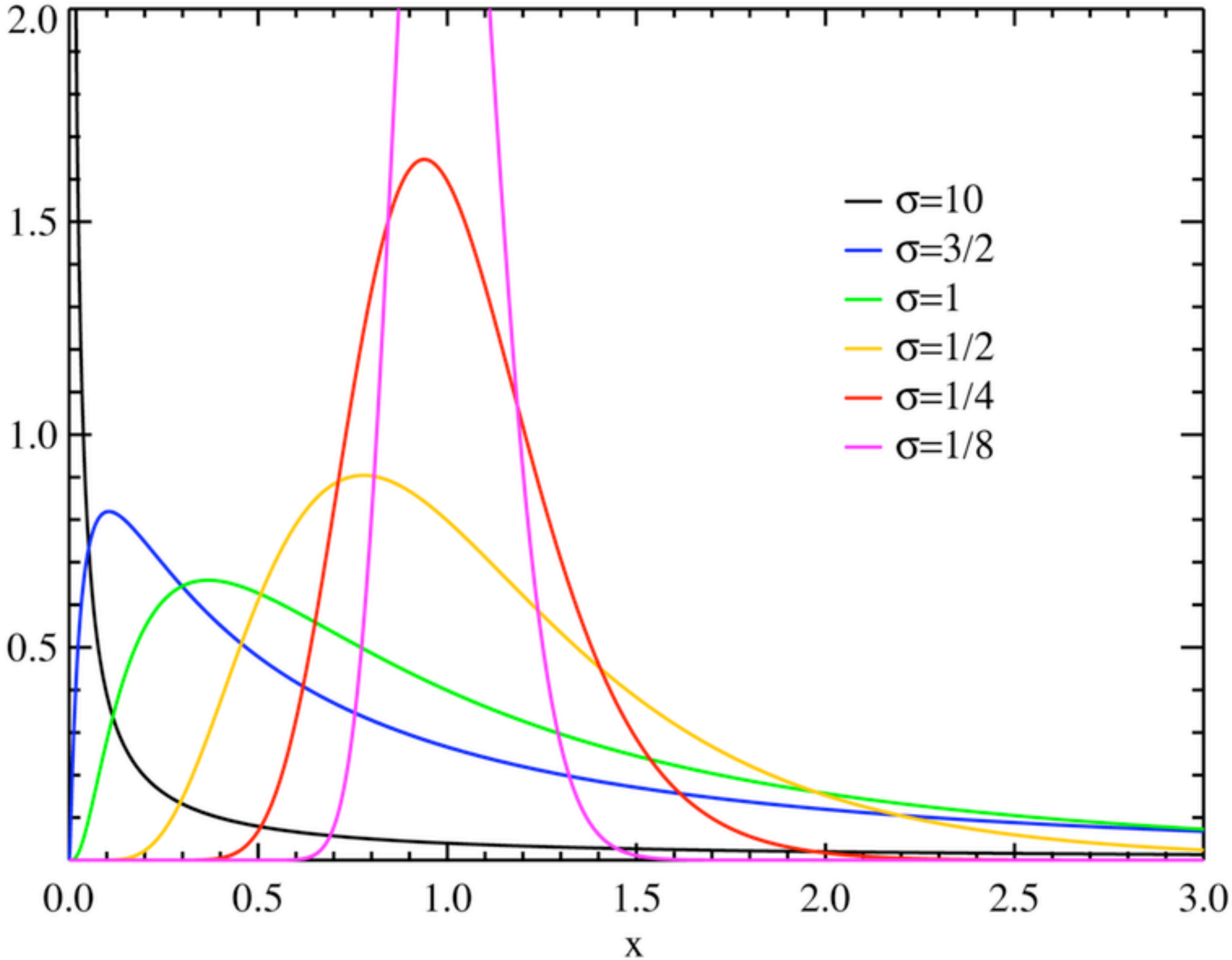
Normal Distribution: Probability Density Function



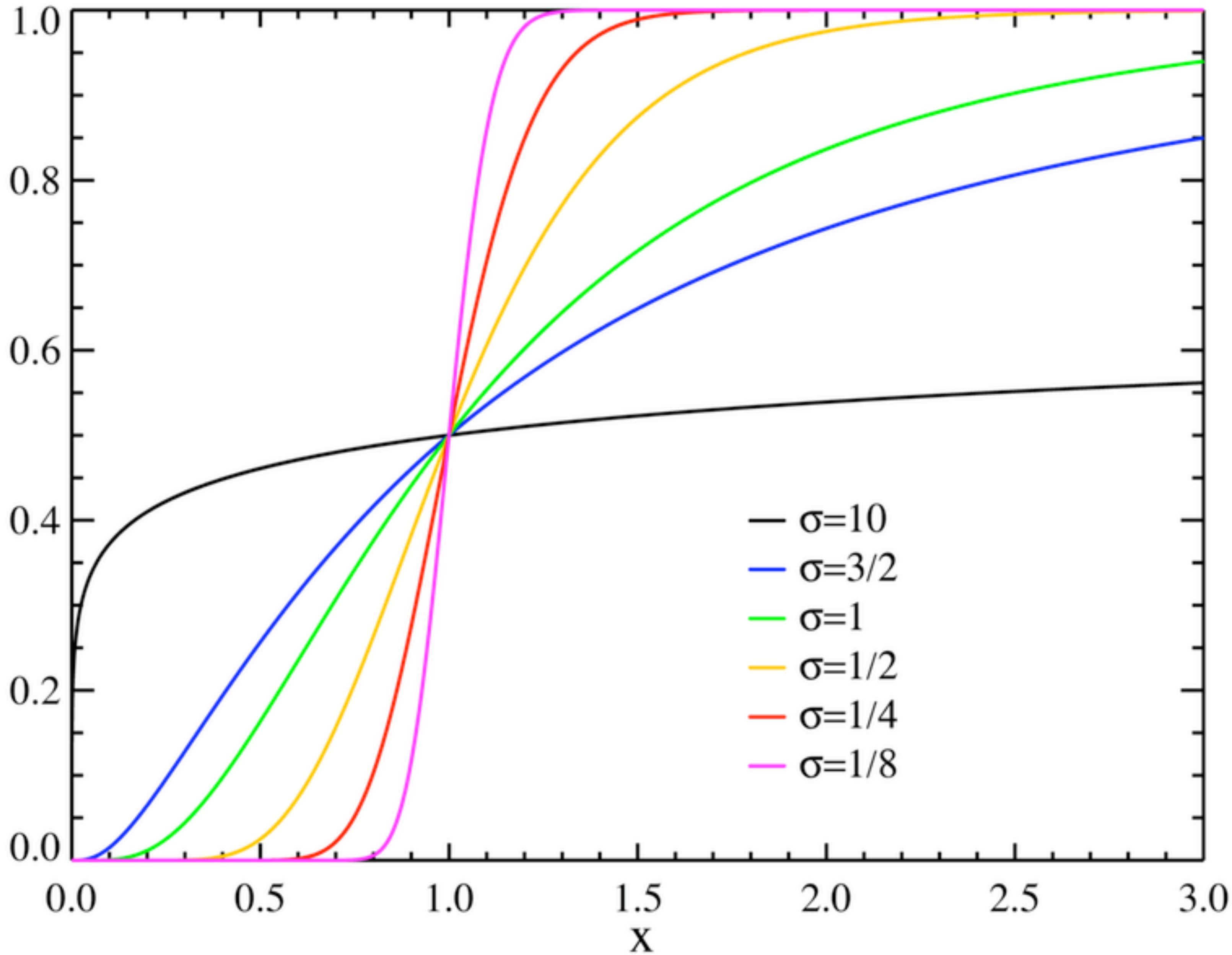
Normal Distribution: Cumulative Distribution Function



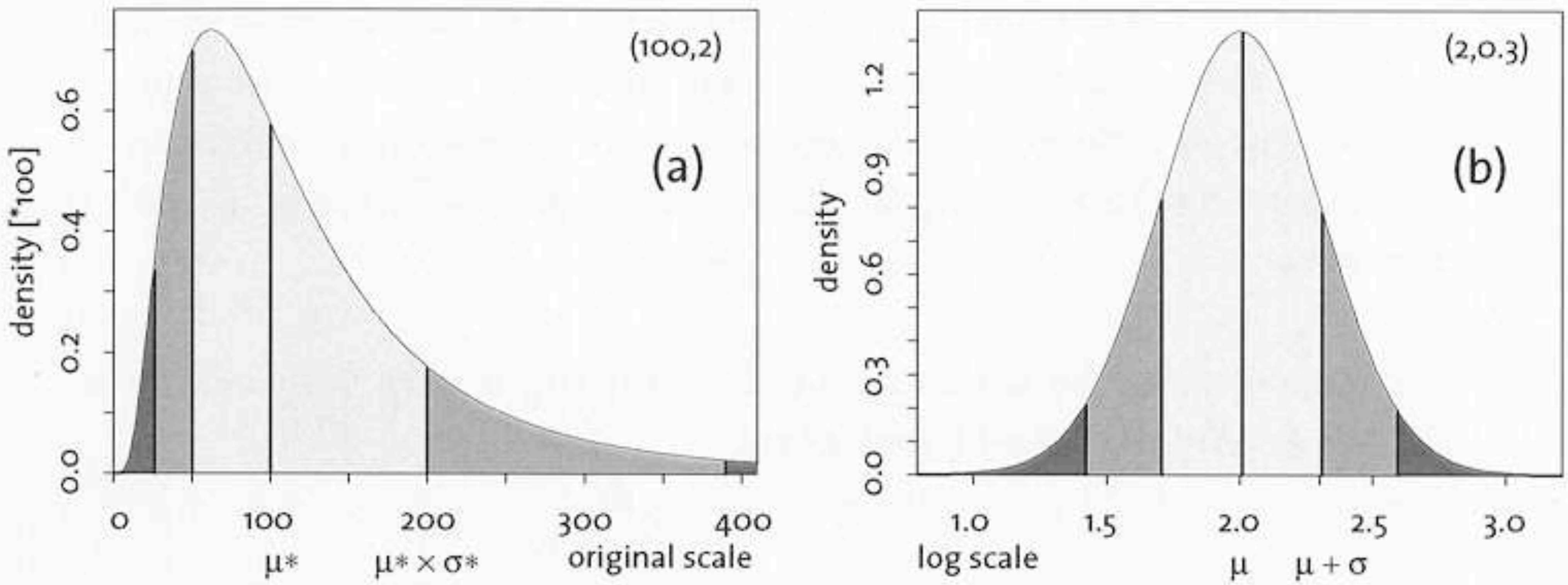
Log-normal Distribution: Probability Density Function



Log-normal Distribution: Cumulative Distribution Function



Life is log-normal !



Disciplines		mu*	sigma*
Medicine	Latent periods of infectious diseases	Months to years	3
Environment	Rainfall	80-200 m3 (x10^3)	4-5
Linguistics	Lengths of spoken words	3-5 letters	1.5

Central Measures of Probability Distributions

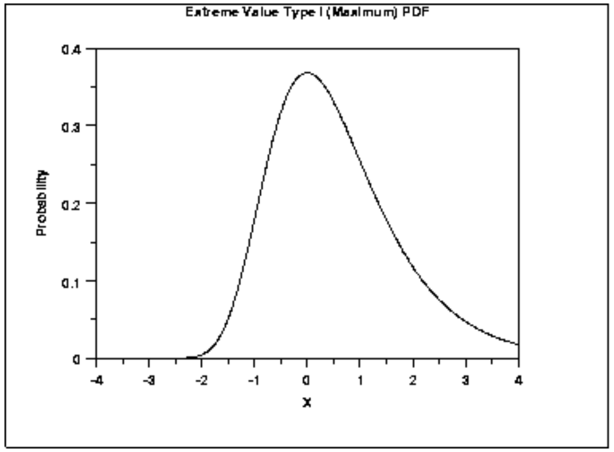
mean η (average) : the centre of the distribution

standard deviation σ : the spread of the distribution

variance σ^2 : the squared deviation from the mean, also called ‘error square’ or ‘fluctuation’

median : the most frequent value, equal to mean in symmetric distributions

Extreme Value Distribution



Distribution peaks at

$$U = \ln(Kmn) / \lambda$$

Cumulative Distribution Function

$$P(S < x) = \exp(-e^{-\lambda(x-U)})$$

Probability of $S \geq x$

$$P(S \geq x) = 1 - \exp(-e^{-\lambda(x-U)})$$

Substitute U

$$P(S \geq x) = 1 - \exp(-Kmne^{-\lambda x})$$

Probabilistic Modelling

Modelling of biological systems usually requires to make assumptions about certain functions and their parameters.

Taking the simple equation

$$y = m * x + n$$

we would denote x,y as 'data', and m,n as 'parameters'.
The equation is a mathematical formulation of a model.

Data are values from measurements or calculations, while 'parameters' describe the relation between data (or model features). Parameters can be determined for known sets of data pairs x,y.

Parameters values for m,n that fit best the data x,y are the most probable ones. However, an uncertainty remains and the values are strictly probability distributions around these values.

Modelling and Inference

The general situation in biology is that we have a property 'y' that is depending on many parameters, or the parameter space 'S'.

$$p(y) = f(y \mid S)$$

This equation presents two distinguishable aspects:

Modelling

The choice of functional form for 'f', which should appropriately describe the dependency of 'y' from 'S' (for example a force field function).
This requires creative and inductive thought.

Parametrisation (reference data) / Inference (test data)

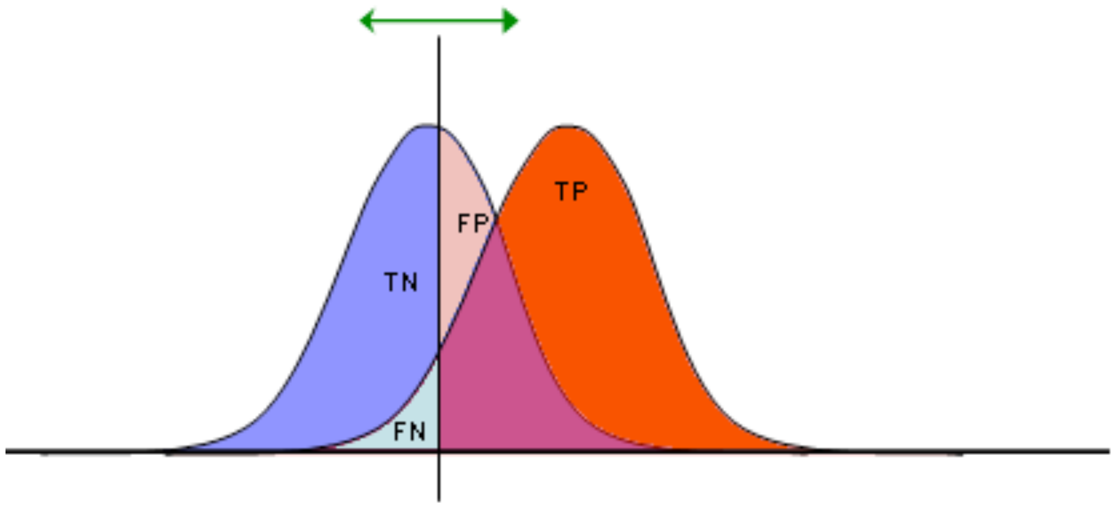
Given the functional form of 'f' is assumed true, model parameters can be specified by mathematical conclusions. Once parametrised on reference data, the model can be applied for inference using test data.

Variable Correlation

Type of Statistics	Type of response variable	Type of correlation analysis
Regression (normal theory)	continuous response variable	linear regression, non-linear regression
Classification (binomial theory)	discrete/categorical response variable	logistic regression, Bayesian classifier, machine learning

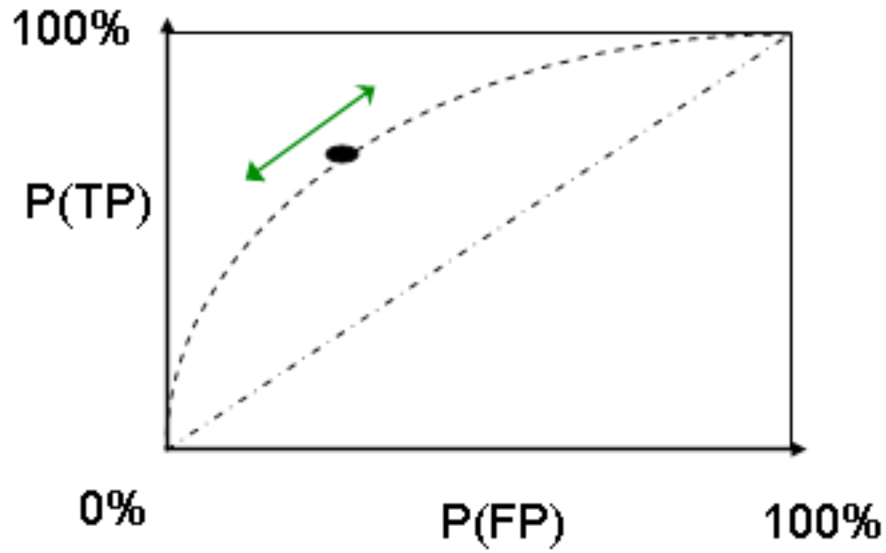
Inference / Prediction Statistics

Random and Target Distributions



Contingency Table

TP	FP
FN	TN
1	1



ROC or benchmark curve

Information Theory

Information theory is a very elegant, well founded and rigorous framework to quantify information and information processing. It is remarkable that this framework was essentially the solo project of one single creative person, Claude Shannon (who launched information theory with a publication in 1948, building on work he did during the war on radar communication), and at the time came almost entirely out of the blue, much like Einstein's general relativity theory.

Measuring information at first sight appears to be a somewhat vague and strange concept. Although especially those of us who are reasonably familiar with using computers will be used to thinking about information content (in terms of bits), we will find out that measuring information is not simply counting the number of bits in a data file, but involves somewhat surprisingly probability concepts. (Text: ACC Coolen)

Claude Shannon

Probabilities, Entropy and Information

Information is not a physical quantity, therefore we can define it in an arbitrary way. The Shannon definition is very similar to the thermodynamic entropy, which is a logarithmic measure of the number of possible states.

Shannon equation of uncertainty (information entropy)

$$H = -\sum p_i \, log_2 \, p_i$$

(unit: bits per symbol)

Information is the decrease of uncertainty (entropy).

$$I = -H$$

Grouping symbols (characters) into groups can change the probability distribution.

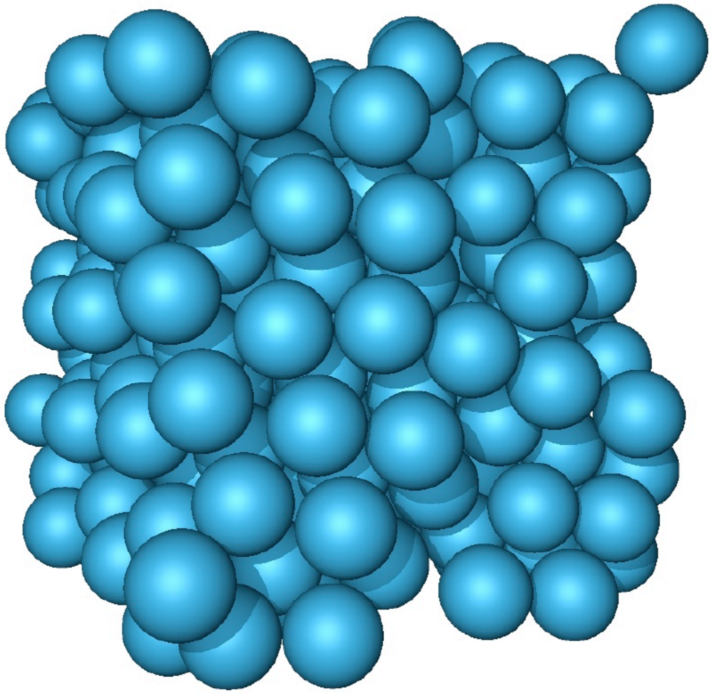
Example: Amino Acid Frequencies

Frequencies deviate from expected p = 1/20 = 0.05 .

We can assign an information content to each character:

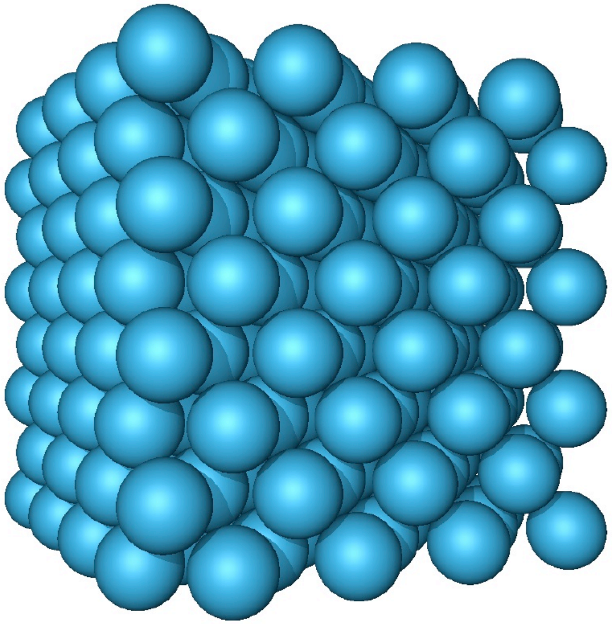
$$I = -ln_2\, p_i$$

	p	-ln2		p	-ln2
A	0.087	3.5	M	0.015	6.1
C	0.033	4.9	N	0.040	4.6
D	0.047	4.4	P	0.051	4.3
E	0.050	4.3	Q	0.038	4.7
F	0.040	4.6	R	0.041	4.6
G	0.089	3.5	S	0.070	3.8
H	0.034	4.9	T	0.058	4.1
I	0.037	4.8	V	0.065	3.9
K	0.080	3.6	W	0.010	6.6
L	0.085	3.6	Y	0.030	5.1



Hard-sphere liquid

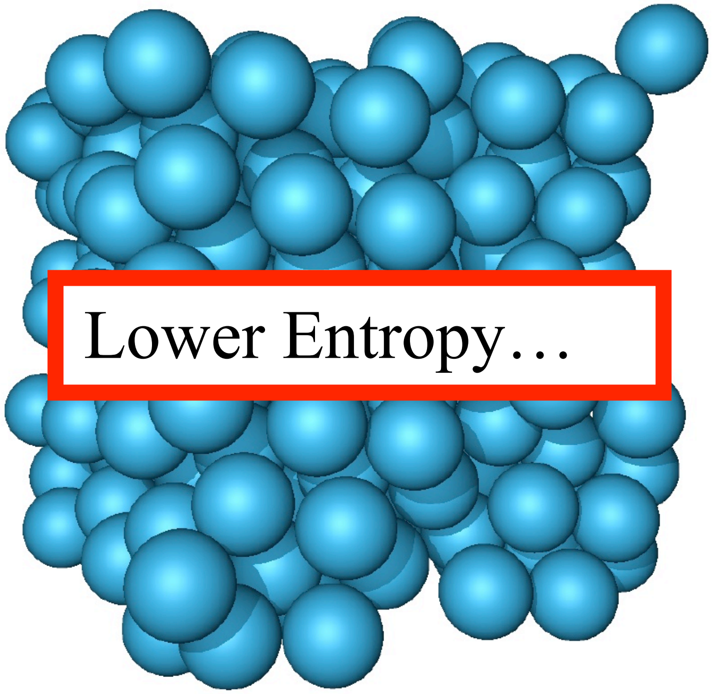
Hard-sphere freezing is driven by entropy !



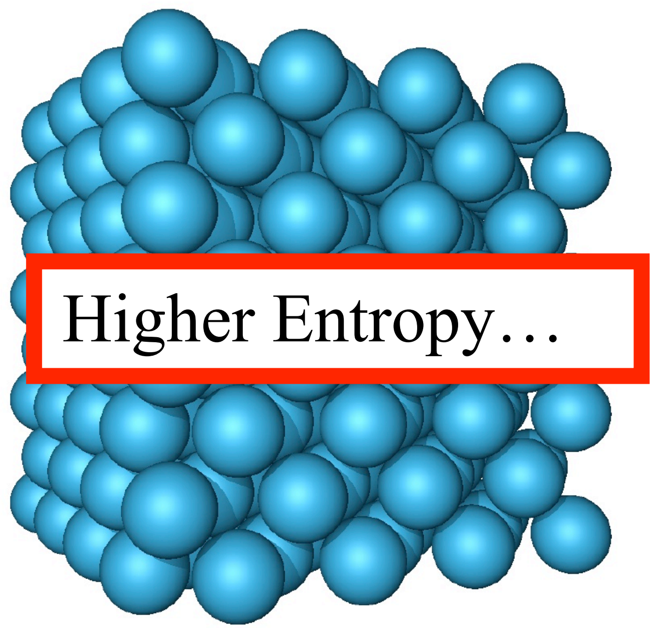
Hard-sphere crystal

Slide by Daan Frenkel and Dima Lukatsky

Hard-sphere liquid



Hard-sphere freezing is driven by entropy !



Hard-sphere crystal

Slide by Daan Frenkel and Dima Lukatsky

Information Theoretic Properties

Mutual Information

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log\left(\frac{p(x,y)}{p_1(x)\,p_2(y)}\right)$$

How much information contains X about Y?

Kullback-Leibler Distance (Relative Entropy)

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

How similar are the probability distributions P and Q?

Akaike Information Criterion

$$AIC = 2k - 2\ln(L)$$

How good is the model with likelihood L and k parameters?

Comparison Information Theory and Thermodynamics

$\Delta E = RT \ln K$

	information-theoretic	thermo-dynamic
Probability	$p_i = n_i / N$	$p_i = e^{-\Delta Ei/kT}$
Entropy	$H = p_i \log p_i$	$S = k \log \Omega$


S = k. log W
LVDWIG
BOLTZMANN
1844 – 1906
Dr PHIL. PAULA
BOLTZMANN
GEB. CHIARI
ARTHUR
BOLTZMANN
LVDWIG
BOLTZMANN
HENRIETTE
BOLTZMANN

Learning Outcomes

- Probability distributions
- Probabilistic models
- Definition of Information
- Information Theoretic Properties
- Shannon Entropy and Thermodynamic Entropy
- Connection between Energy and Probability
- Boltzmann equation