CS 466
Introduction to Bioinformatics

Instructor: Jian Peng

Probability and Statistics

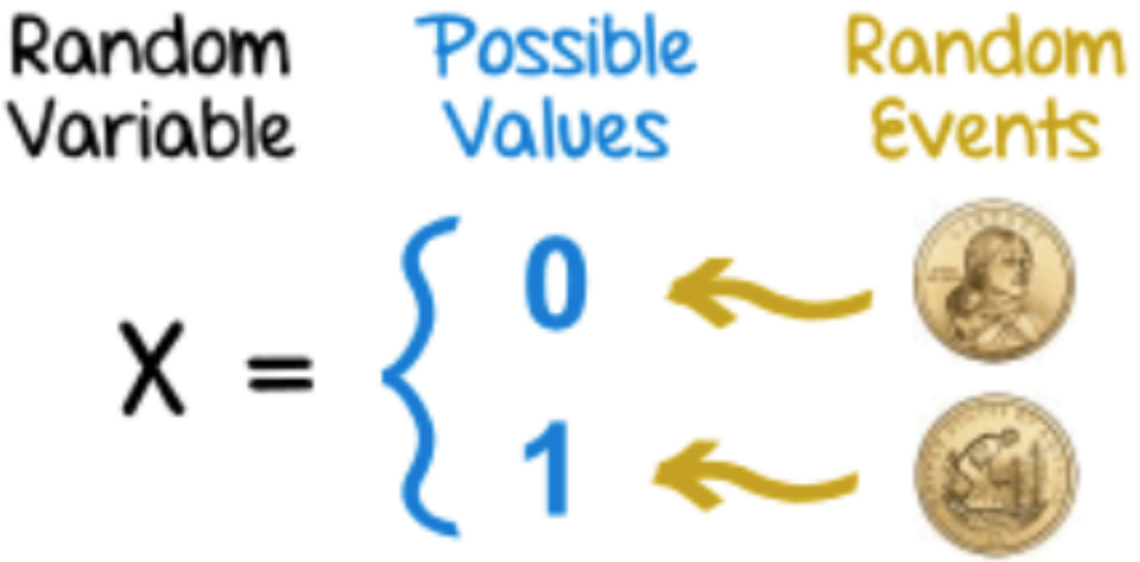
Random Variables and Expectations

Random Variable

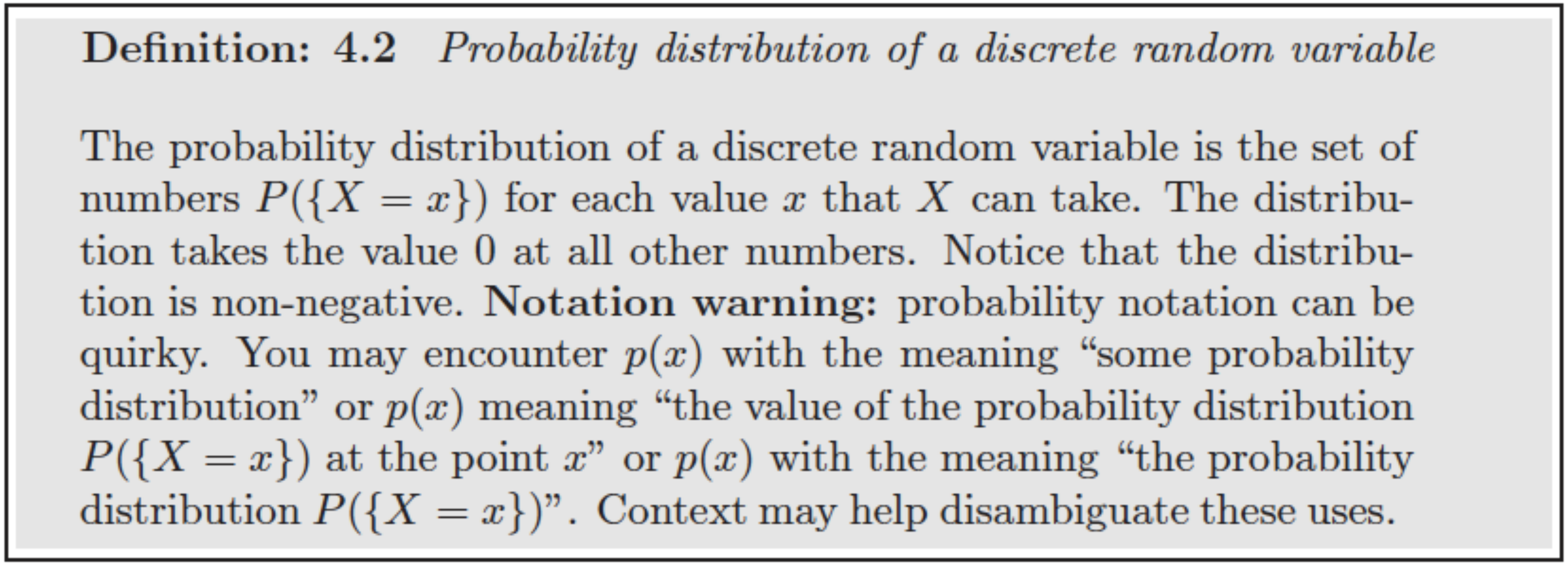
Quite commonly, we would like to deal with numbers that are random. We can do so by linking numbers to the outcome of an experiment. We define a **random variable**:

Definition: 4.1 *Discrete random variable*

Given a sample space Ω, a set of events $\mathcal{F}$, a probability function P, and a countable set of of real numbers D, a discrete random variable is a function with domain Ω and range D.



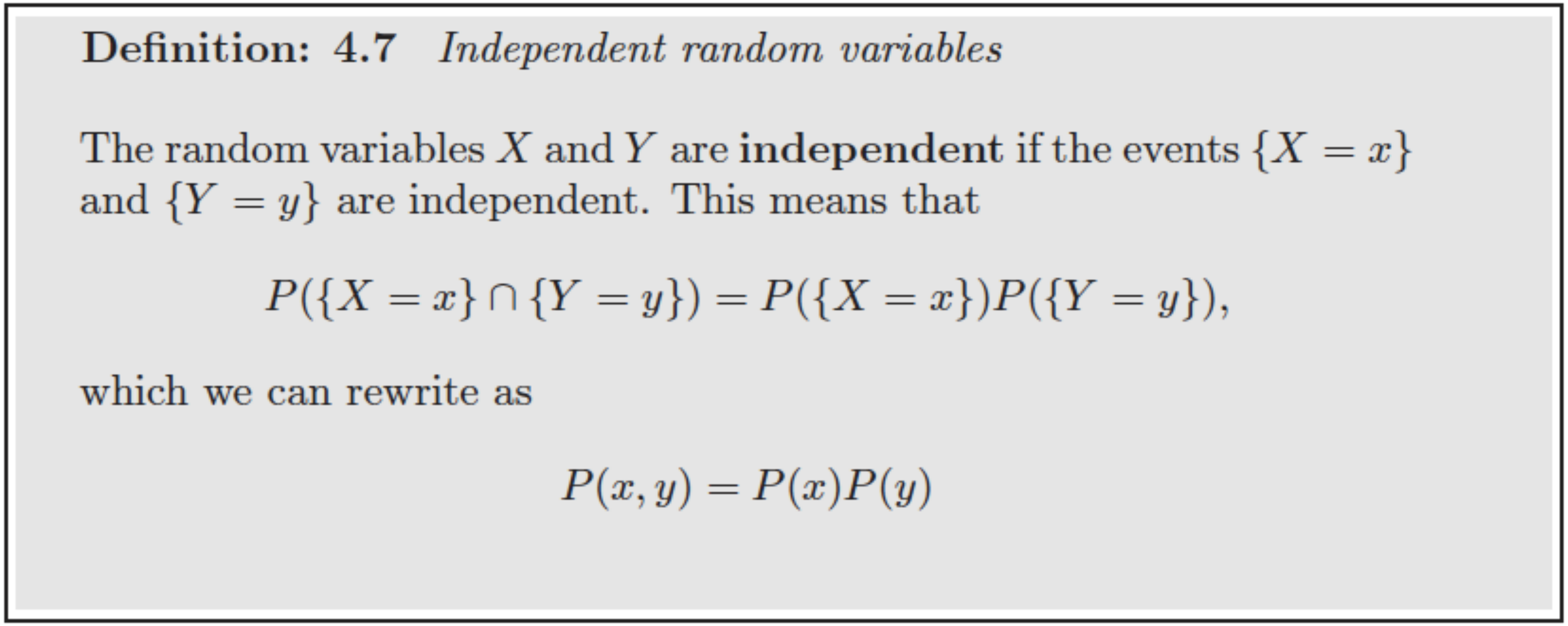
Probability distribution



Definition: 4.2 *Probability distribution of a discrete random variable*

The probability distribution of a discrete random variable is the set of numbers $P(\{X = x\})$ for each value x that X can take. The distribution takes the value 0 at all other numbers. Notice that the distribution is non-negative. **Notation warning:** probability notation can be quirky. You may encounter $p(x)$ with the meaning "some probability distribution" or $p(x)$ meaning "the value of the probability distribution $P(\{X = x\})$ at the point x" or $p(x)$ with the meaning "the probability distribution $P(\{X = x\})$". Context may help disambiguate these uses.

Independent variables



Definition: 4.7 *Independent random variables*

The random variables X and Y are **independent** if the events $\{X = x\}$ and $\{Y = y\}$ are independent. This means that

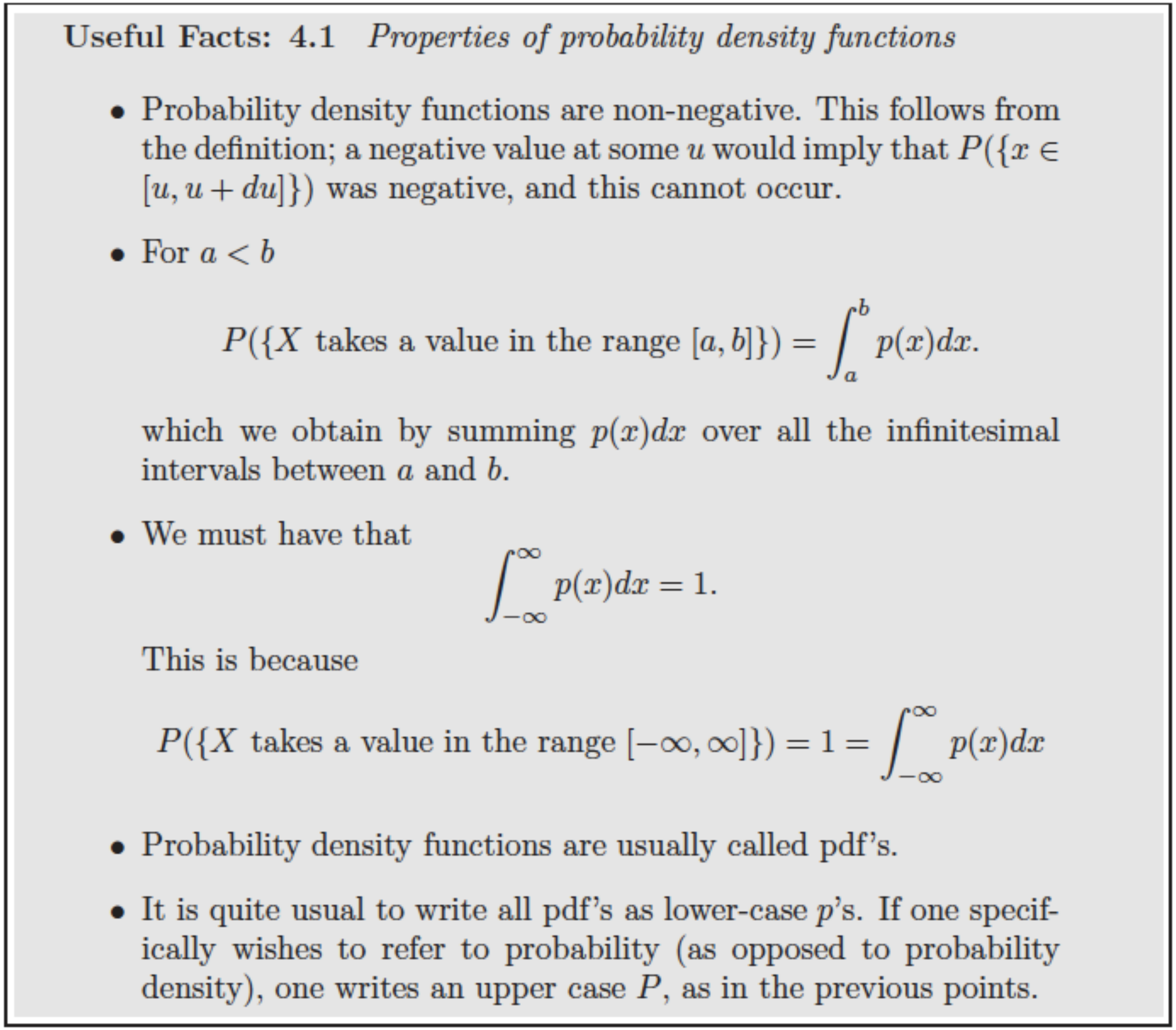
$$P(\{X = x\} \cap \{Y = y\}) = P(\{X = x\})P(\{Y = y\}),$$

which we can rewrite as

$$P(x, y) = P(x)P(y)$$

Continuous distribution: density function

$$p(x)dx = P(\{\text{event that } X \text{ takes a value in the range } [x, x + dx]\}).$$



Useful Facts: 4.1 *Properties of probability density functions*

- Probability density functions are non-negative. This follows from the definition; a negative value at some u would imply that $P(\{x \in [u, u + du]\})$ was negative, and this cannot occur.
- For $a < b$

  $$P(\{X \text{ takes a value in the range } [a, b]\}) = \int_a^b p(x)dx.$$

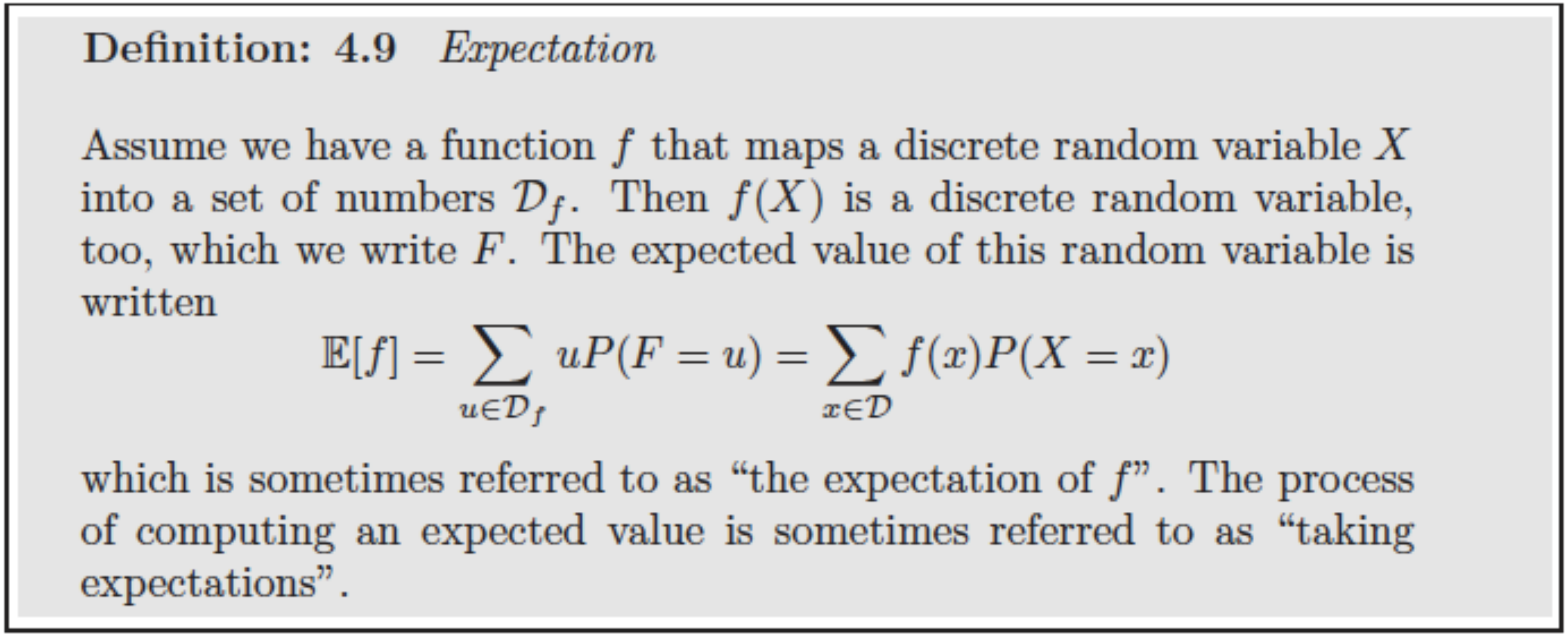
  which we obtain by summing $p(x)dx$ over all the infinitesimal intervals between a and b.
- We must have that

  $$\int_{-\infty}^{\infty} p(x)dx = 1.$$

  This is because

  $$P(\{X \text{ takes a value in the range } [-\infty, \infty]\}) = 1 = \int_{-\infty}^{\infty} p(x)dx$$
- Probability density functions are usually called pdf's.
- It is quite usual to write all pdf's as lower-case p's. If one specifically wishes to refer to probability (as opposed to probability density), one writes an upper case P, as in the previous points.

Expectation

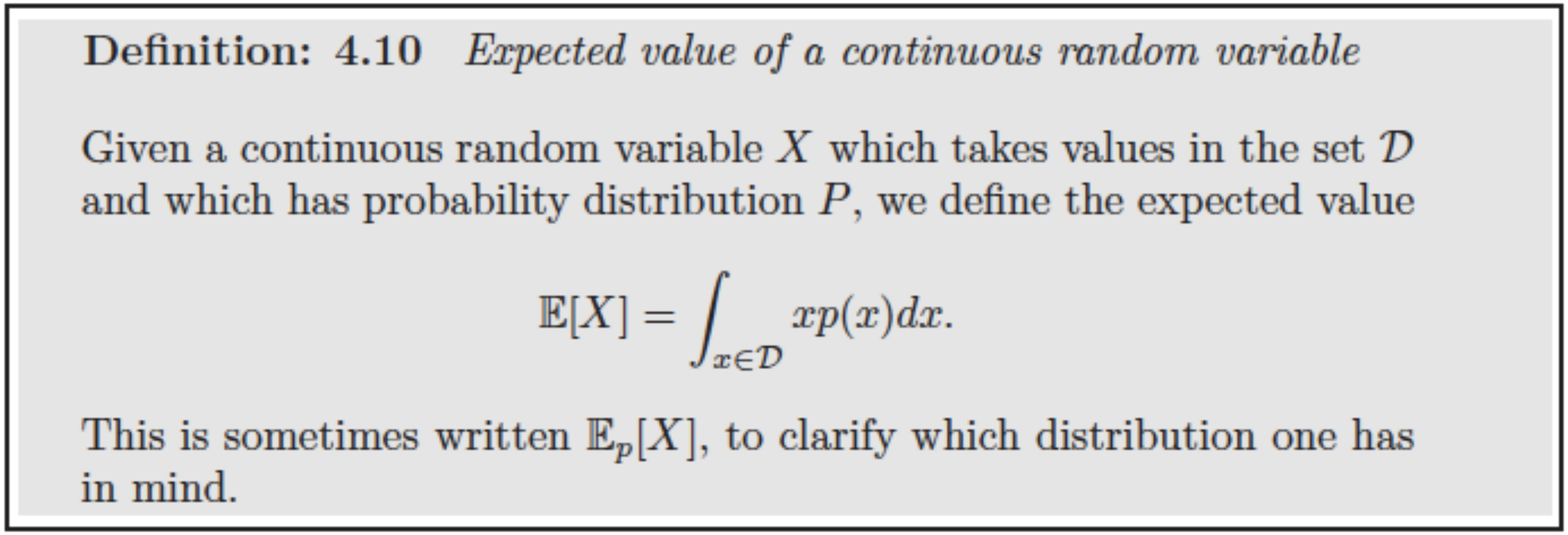


Definition: 4.9 *Expectation*

Assume we have a function f that maps a discrete random variable X into a set of numbers $\mathcal{D}_f$. Then $f(X)$ is a discrete random variable, too, which we write F. The expected value of this random variable is written

$$\mathbb{E}[f] = \sum_{u \in \mathcal{D}_f} uP(F = u) = \sum_{x \in \mathcal{D}} f(x)P(X = x)$$

which is sometimes referred to as "the expectation of f". The process of computing an expected value is sometimes referred to as "taking expectations".



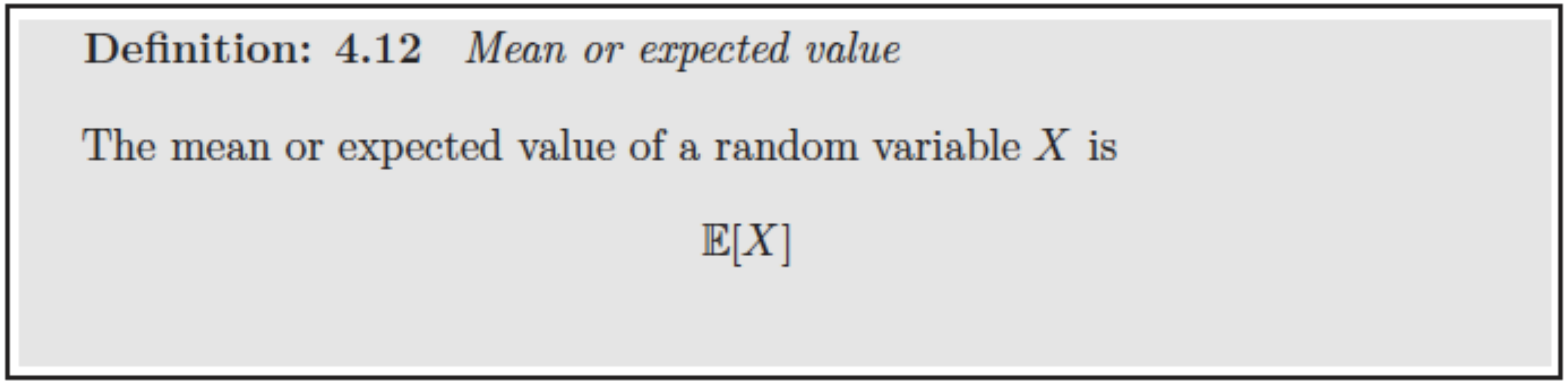
Definition: 4.10 *Expected value of a continuous random variable*

Given a continuous random variable X which takes values in the set $\mathcal{D}$ and which has probability distribution P, we define the expected value

$$\mathbb{E}[X] = \int_{x \in \mathcal{D}} xp(x)dx.$$

This is sometimes written $\mathbb{E}_p[X]$, to clarify which distribution one has in mind.

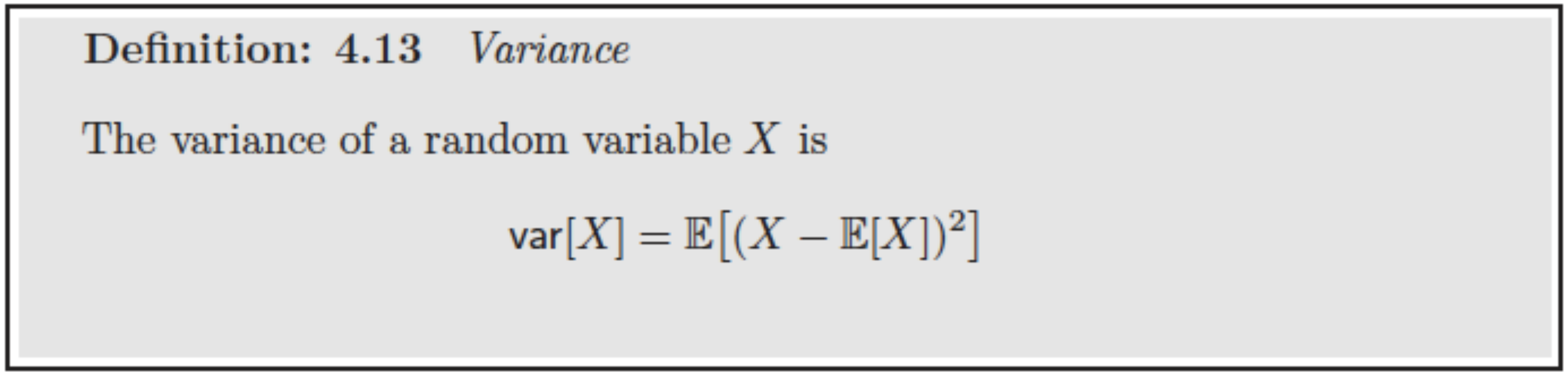
Mean, Variance and Covariance



Definition: 4.12 *Mean or expected value*

The mean or expected value of a random variable X is

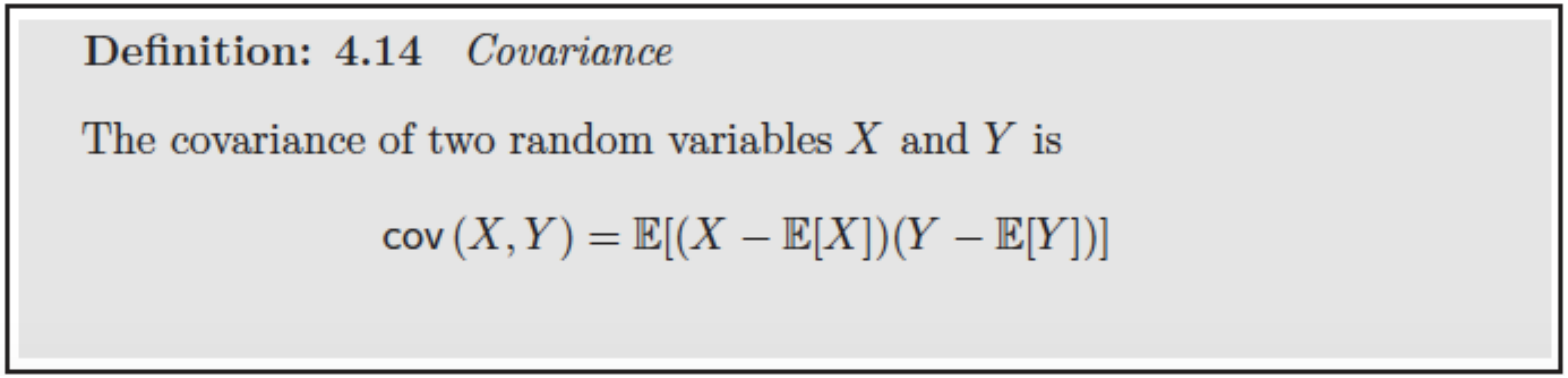
$$\mathbb{E}[X]$$



Definition: 4.13 *Variance*

The variance of a random variable X is

$$\mathsf{var}[X] = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big]$$



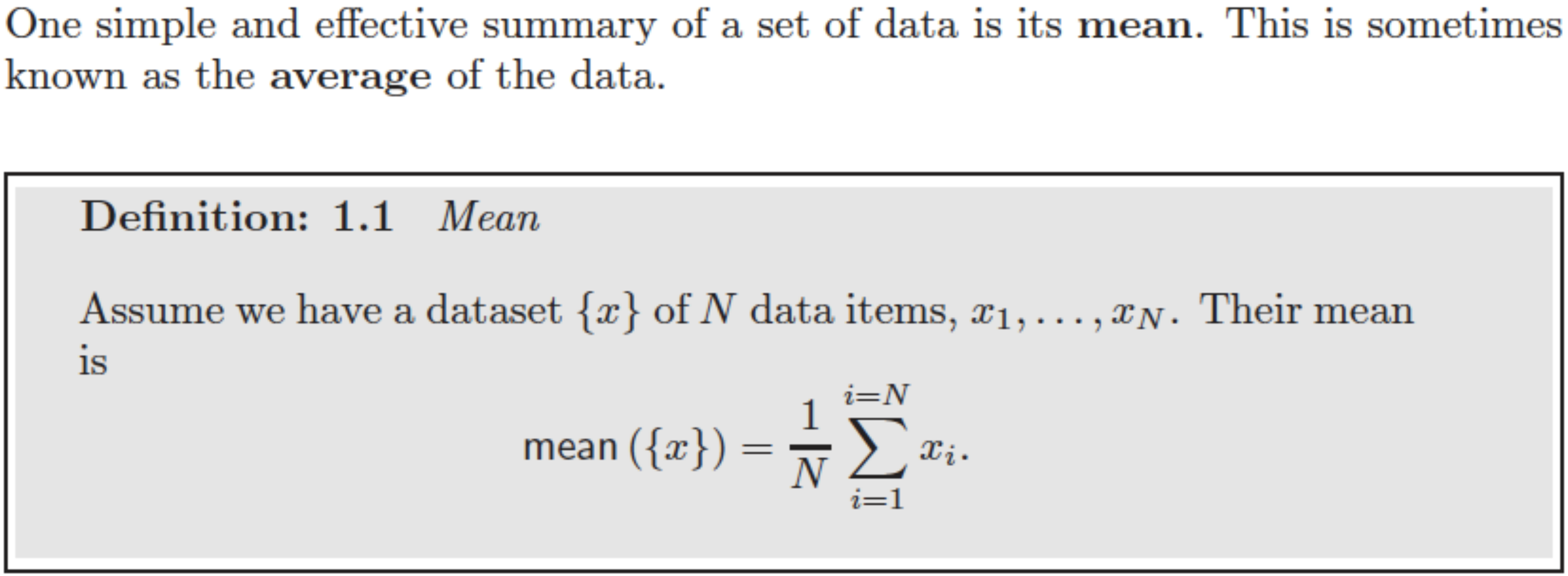
Definition: 4.14 *Covariance*

The covariance of two random variables X and Y is

$$\mathsf{cov}\,(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

Statistics

Mean



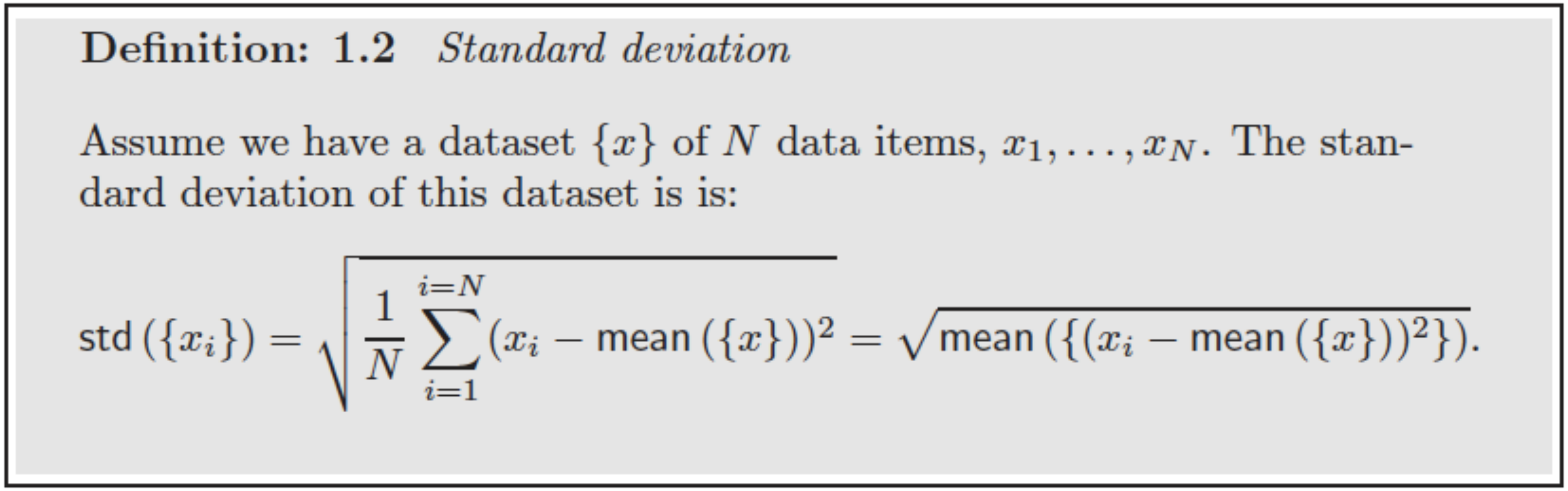
One simple and effective summary of a set of data is its **mean**. This is sometimes known as the **average** of the data.

Definition: 1.1 *Mean*

Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. Their mean is

$$\mathsf{mean}\left(\{x\}\right) = \frac{1}{N}\sum_{i=1}^{i=N} x_i.$$

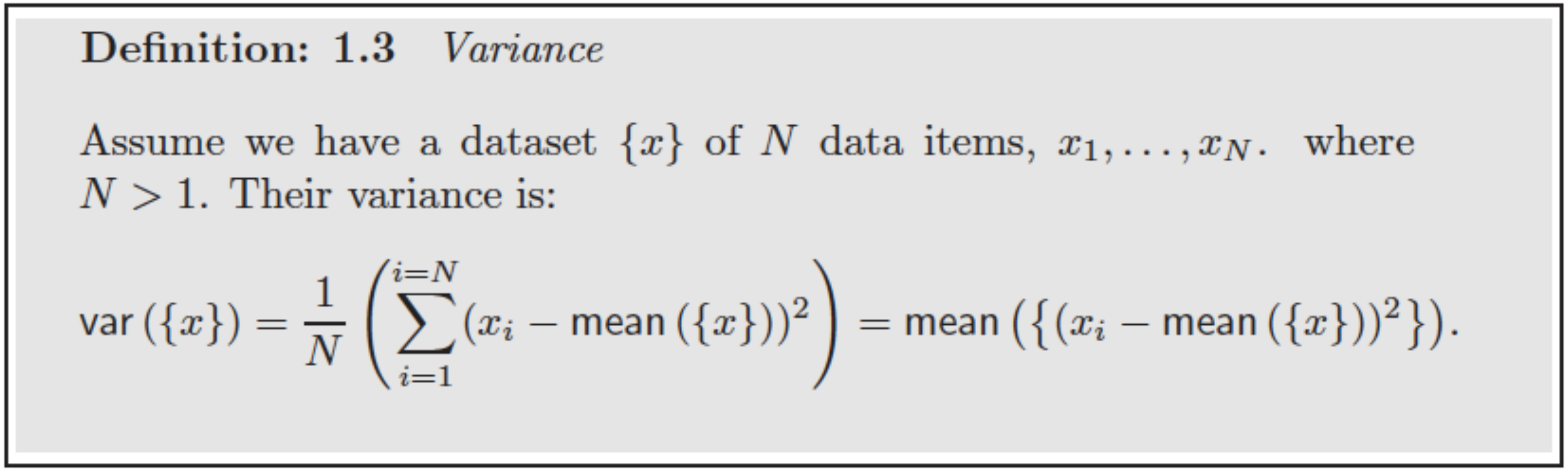
Standard deviation and Variance



Definition: 1.2 *Standard deviation*

Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. The standard deviation of this dataset is is:

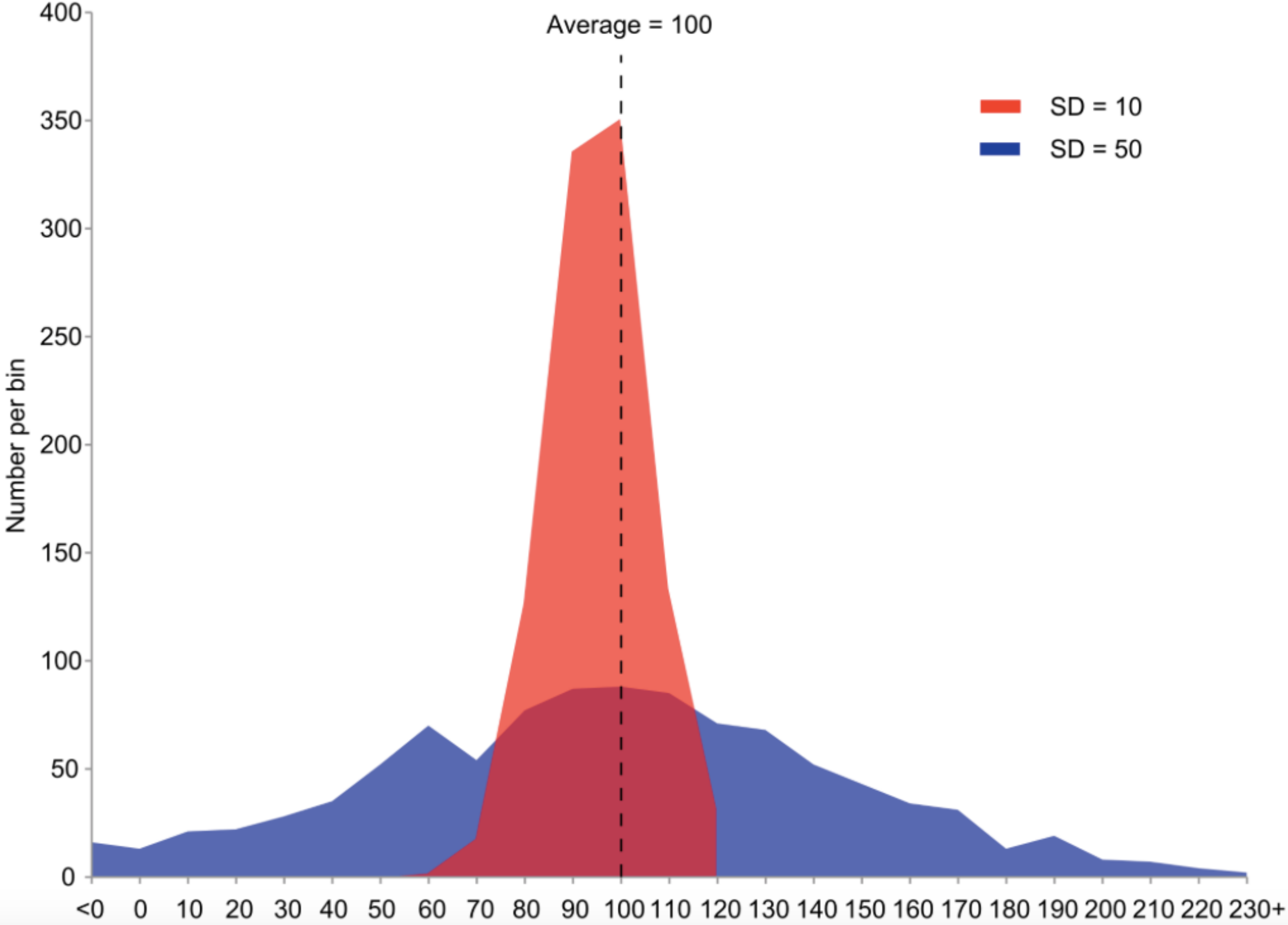
$$\mathsf{std}\left(\{x_i\}\right) = \sqrt{\frac{1}{N}\sum_{i=1}^{i=N}(x_i - \mathsf{mean}\left(\{x\}\right))^2} = \sqrt{\mathsf{mean}\left(\{(x_i - \mathsf{mean}\left(\{x\}\right))^2\}\right)}.$$



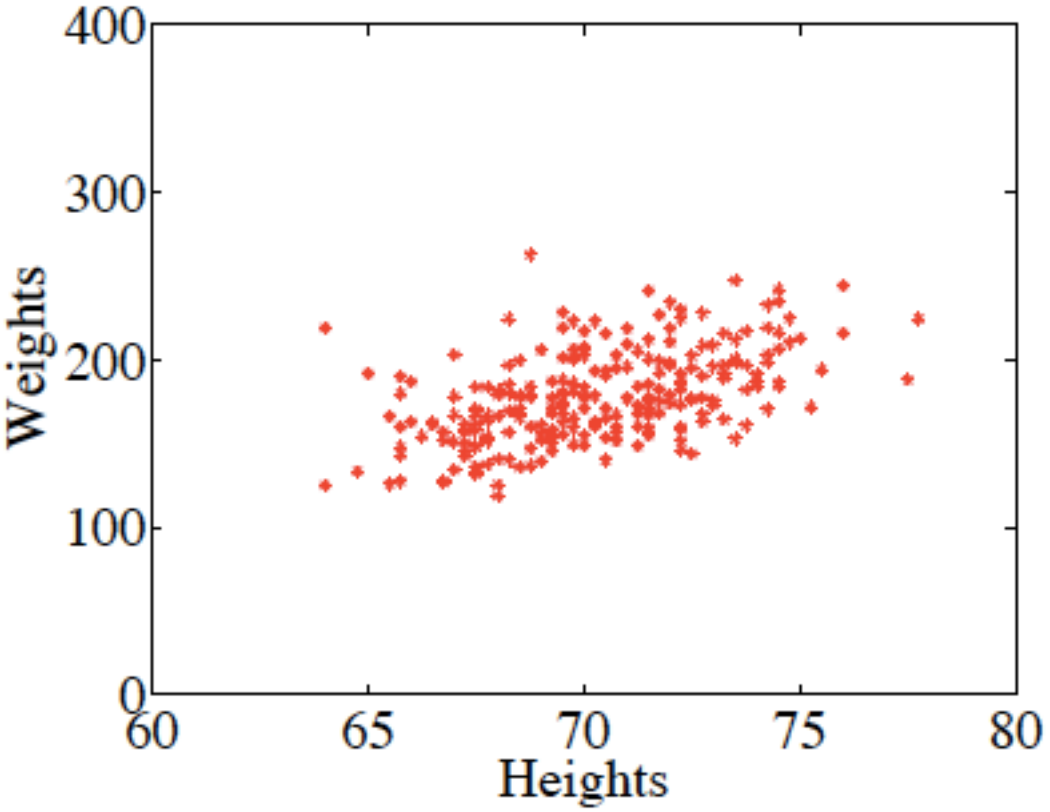
Definition: 1.3 *Variance*

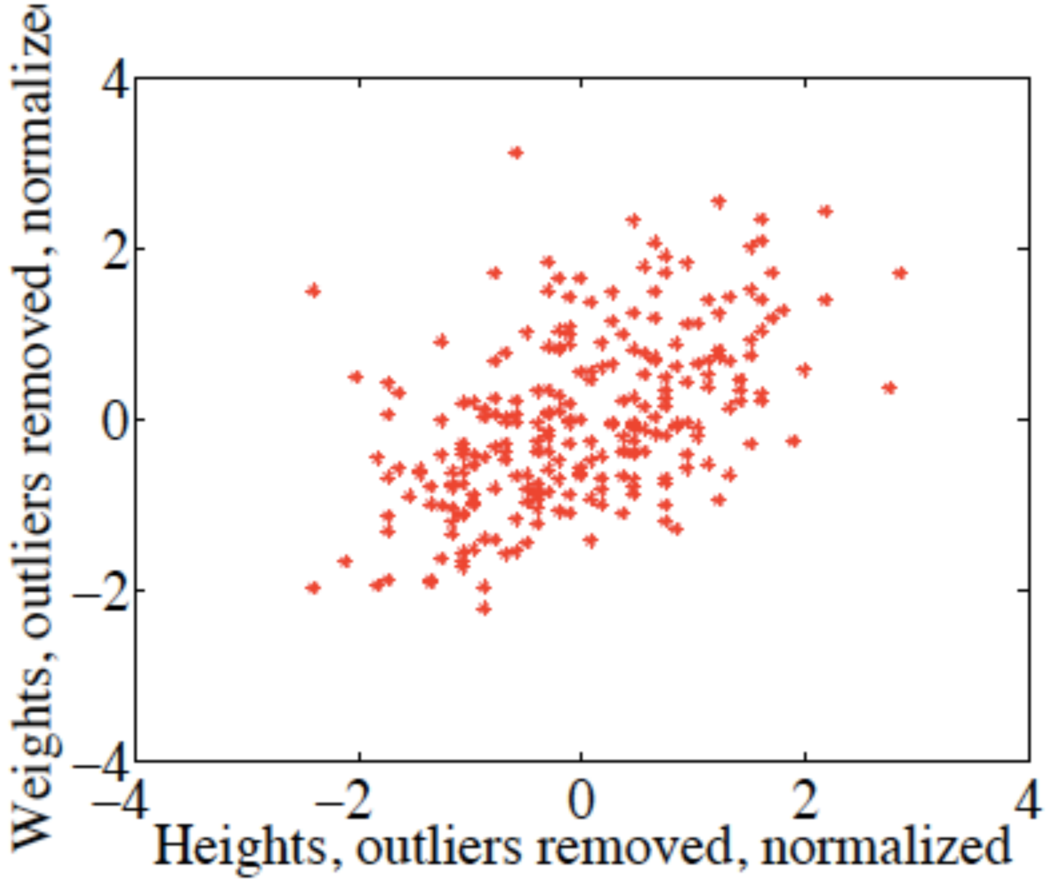
Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. where $N > 1$. Their variance is:

$$\mathsf{var}\left(\{x\}\right) = \frac{1}{N}\left(\sum_{i=1}^{i=N}(x_i - \mathsf{mean}\left(\{x\}\right))^2\right) = \mathsf{mean}\left(\left\{(x_i - \mathsf{mean}\left(\{x\}\right))^2\right\}\right).$$


Average = 100
SD = 10
SD = 50
Number per bin
400
350
300
250
200
150
100
50
0
<0
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220
230+

Normalization





$$\hat{x}_i = \frac{(x_i - \text{mean}(\{x\}))}{\text{std}(\{x\})}.$$

Correlation

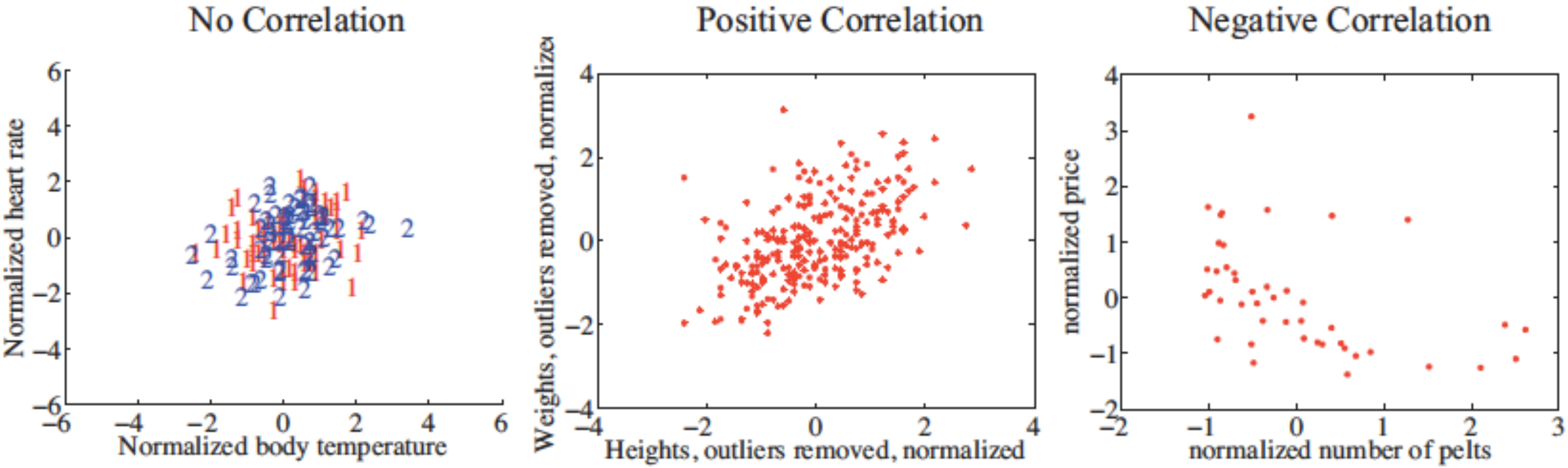
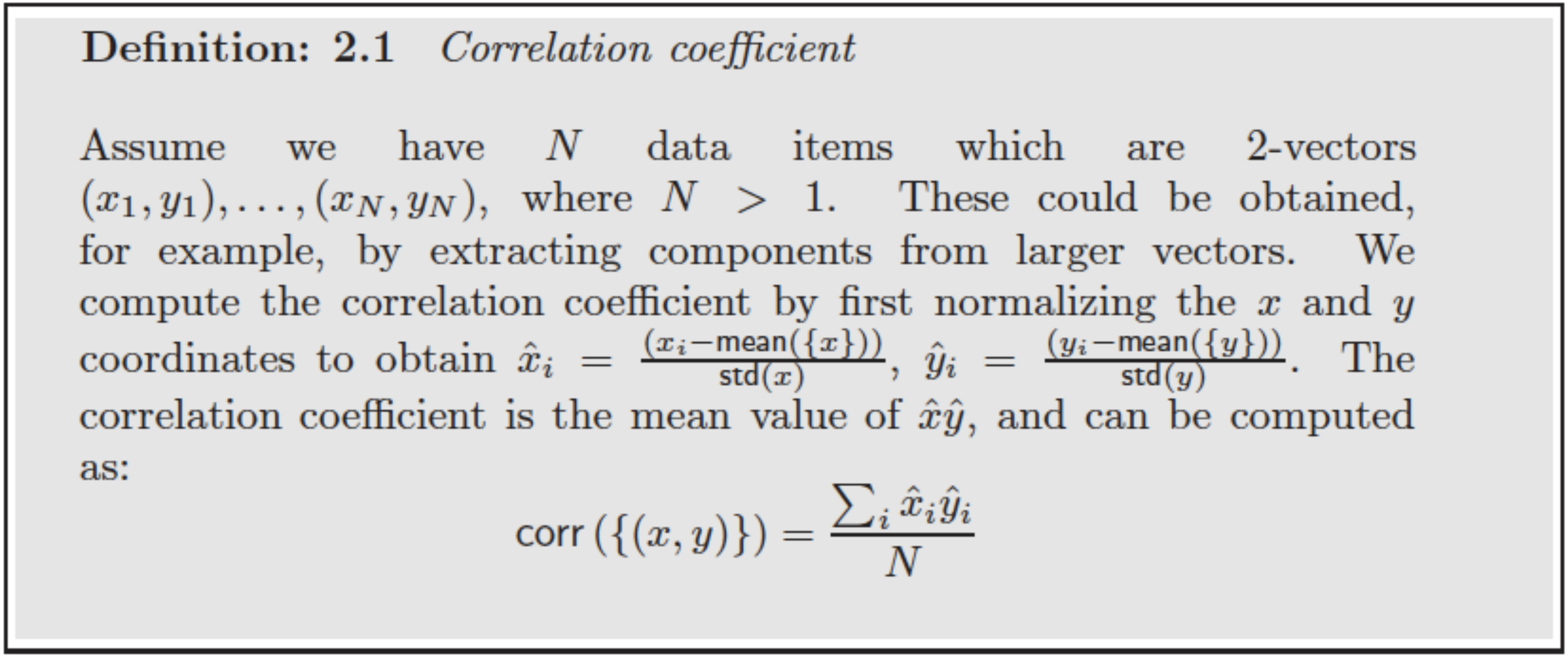


FIGURE 2.16: *The three kinds of scatter plot are less clean for real data than for our idealized examples. Here I used the body temperature vs heart rate data for the zero correlation; the height-weight data for positive correlation; and the lynx data for negative correlation. The pictures aren't idealized — real data tends to be messy — but you can still see the basic structures.*

Correlation coefficient

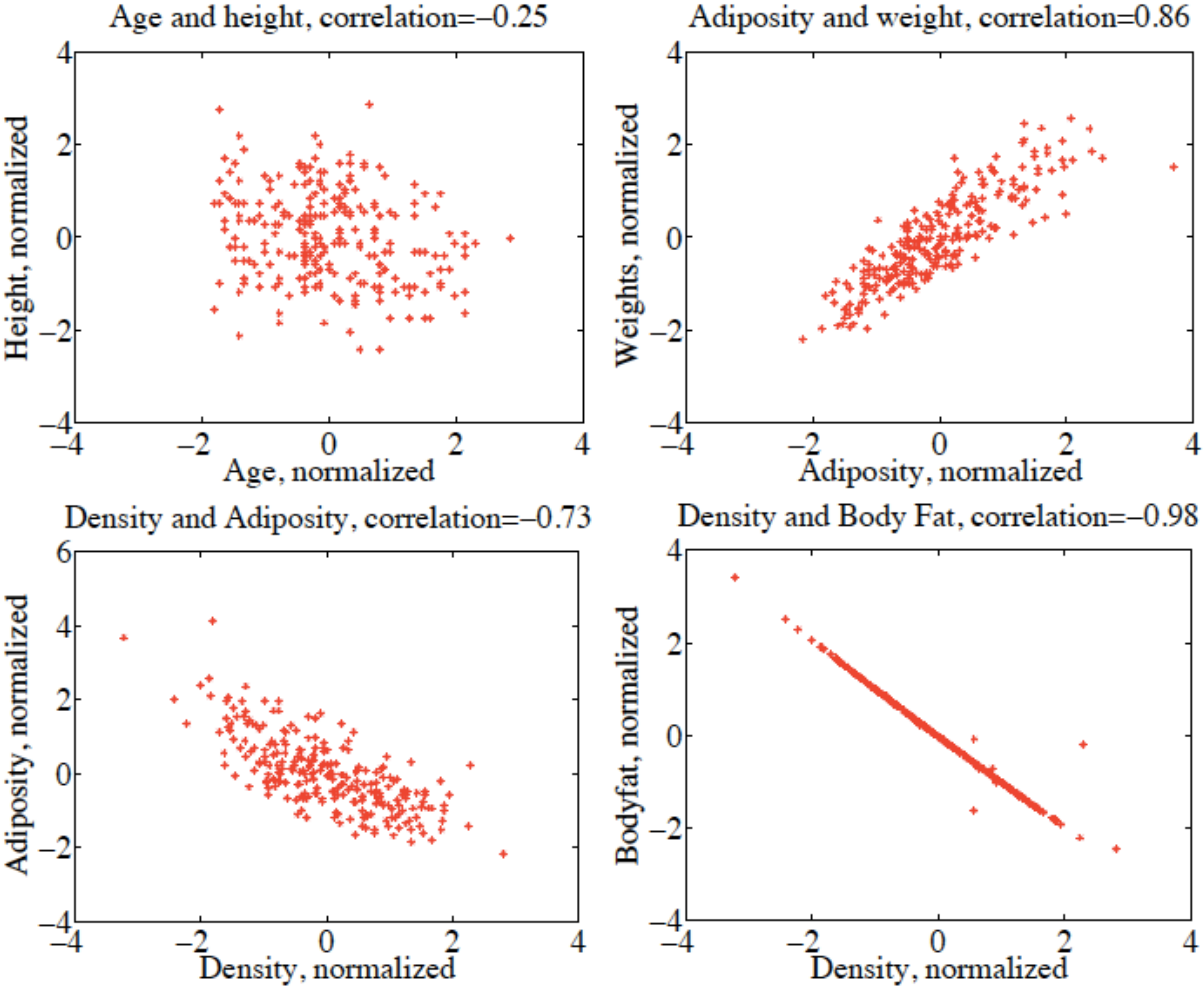


Definition: 2.1 *Correlation coefficient*

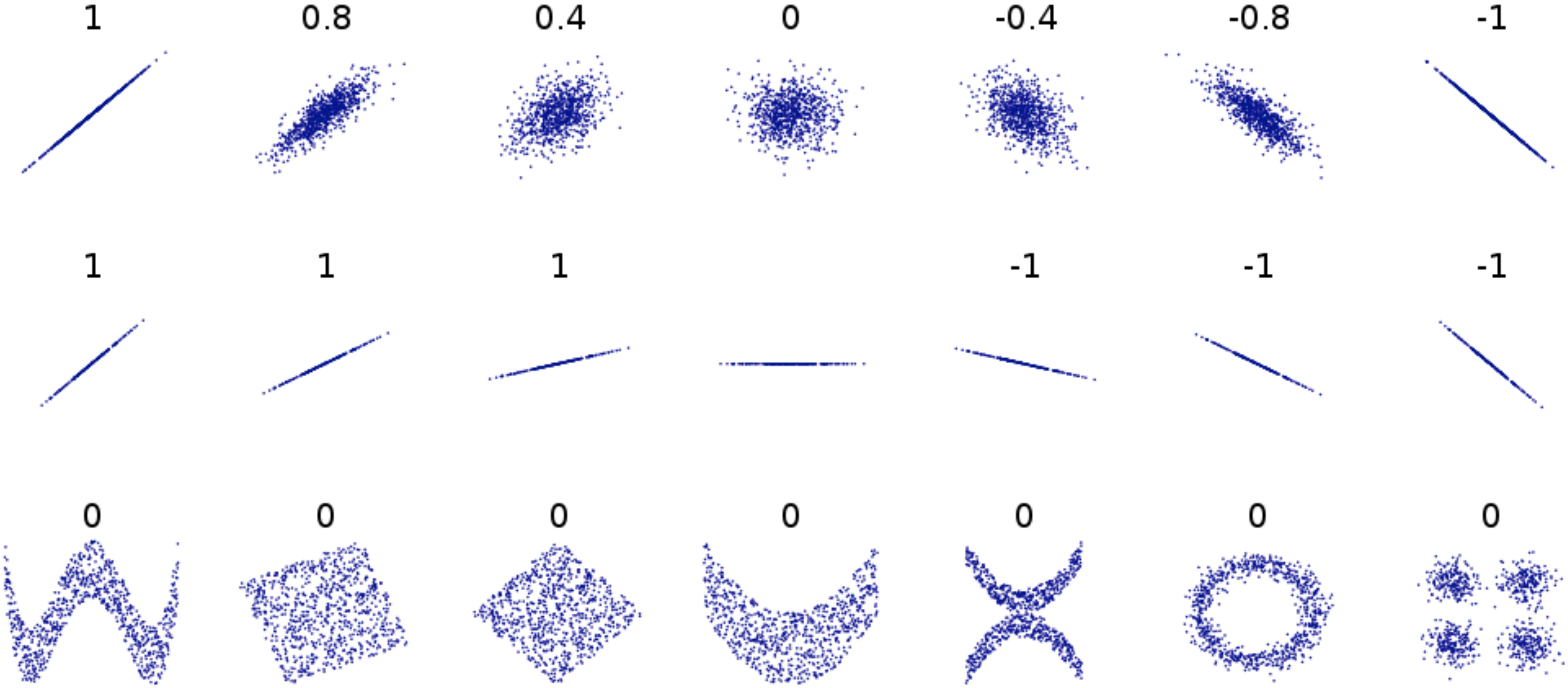
Assume we have N data items which are 2-vectors $(x_1, y_1), \ldots, (x_N, y_N)$, where $N > 1$. These could be obtained, for example, by extracting components from larger vectors. We compute the correlation coefficient by first normalizing the x and y coordinates to obtain $\hat{x}_i = \frac{(x_i - \mathsf{mean}(\{x\}))}{\mathsf{std}(x)}$, $\hat{y}_i = \frac{(y_i - \mathsf{mean}(\{y\}))}{\mathsf{std}(y)}$. The correlation coefficient is the mean value of $\hat{x}\hat{y}$, and can be computed as:

$$\mathsf{corr}\left(\{(x, y)\}\right) = \frac{\sum_i \hat{x}_i \hat{y}_i}{N}$$

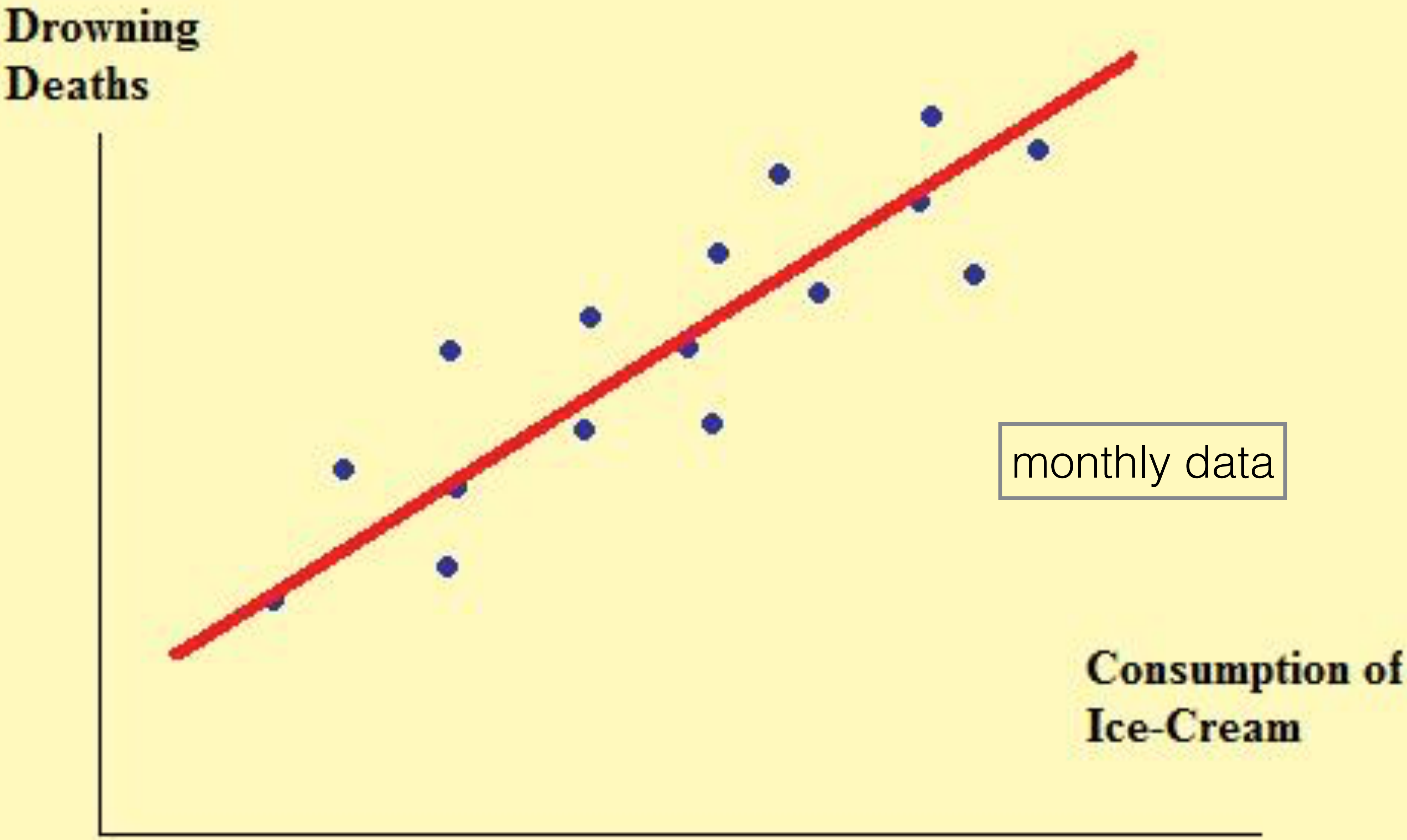
Also called **Pearson Correlation Coefficient**


Age and height, correlation=–0.25
Height, normalized
Age, normalized
Adiposity and weight, correlation=0.86
Weights, normalized
Adiposity, normalized
Density and Adiposity, correlation=–0.73
Adiposity, normalized
Density, normalized
Density and Body Fat, correlation=–0.98
Bodyfat, normalized
Density, normalized
–4
–2
0
2
4
6

Correlation coefficient vs Relationship



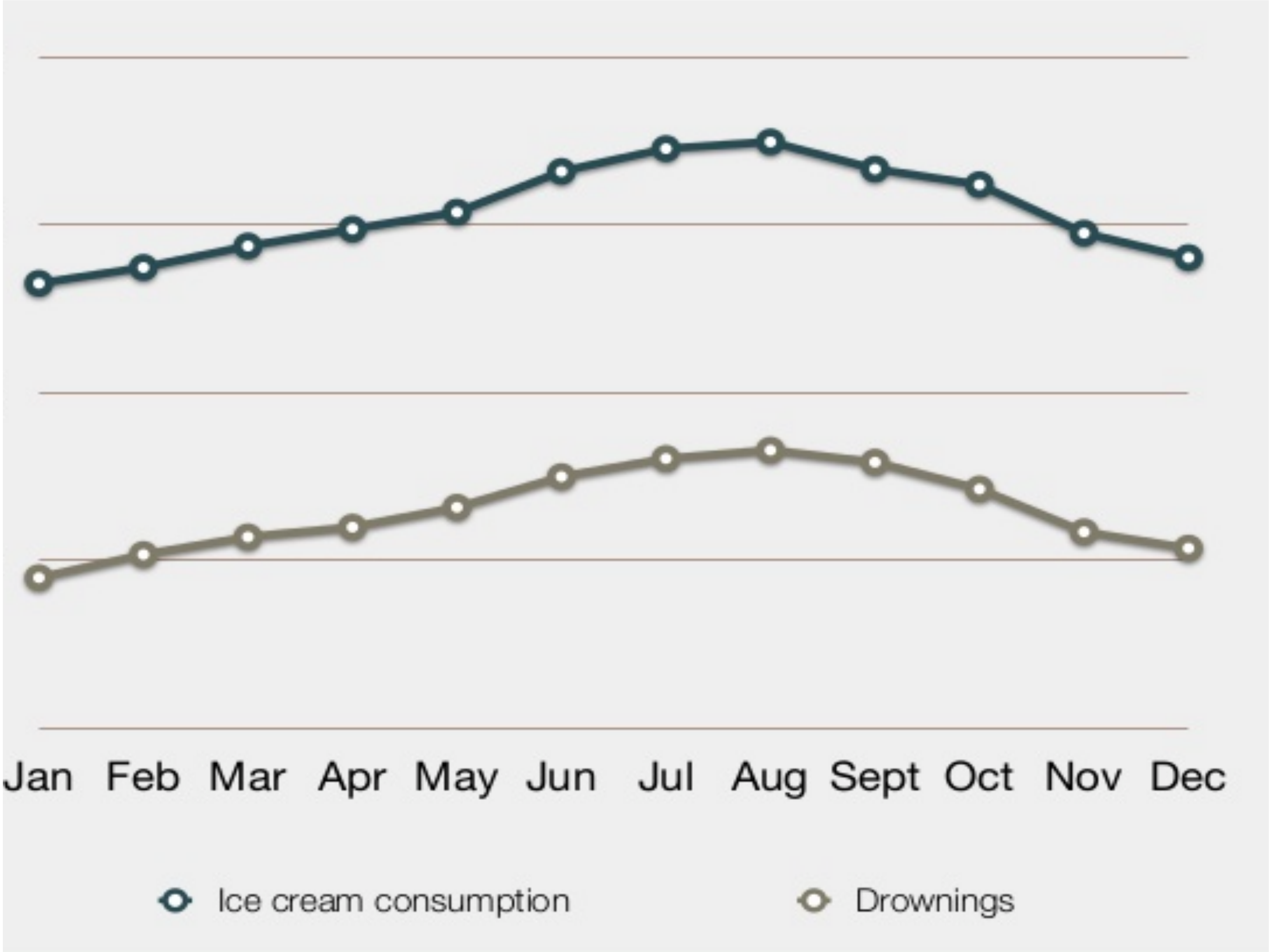
Correlation and Causality



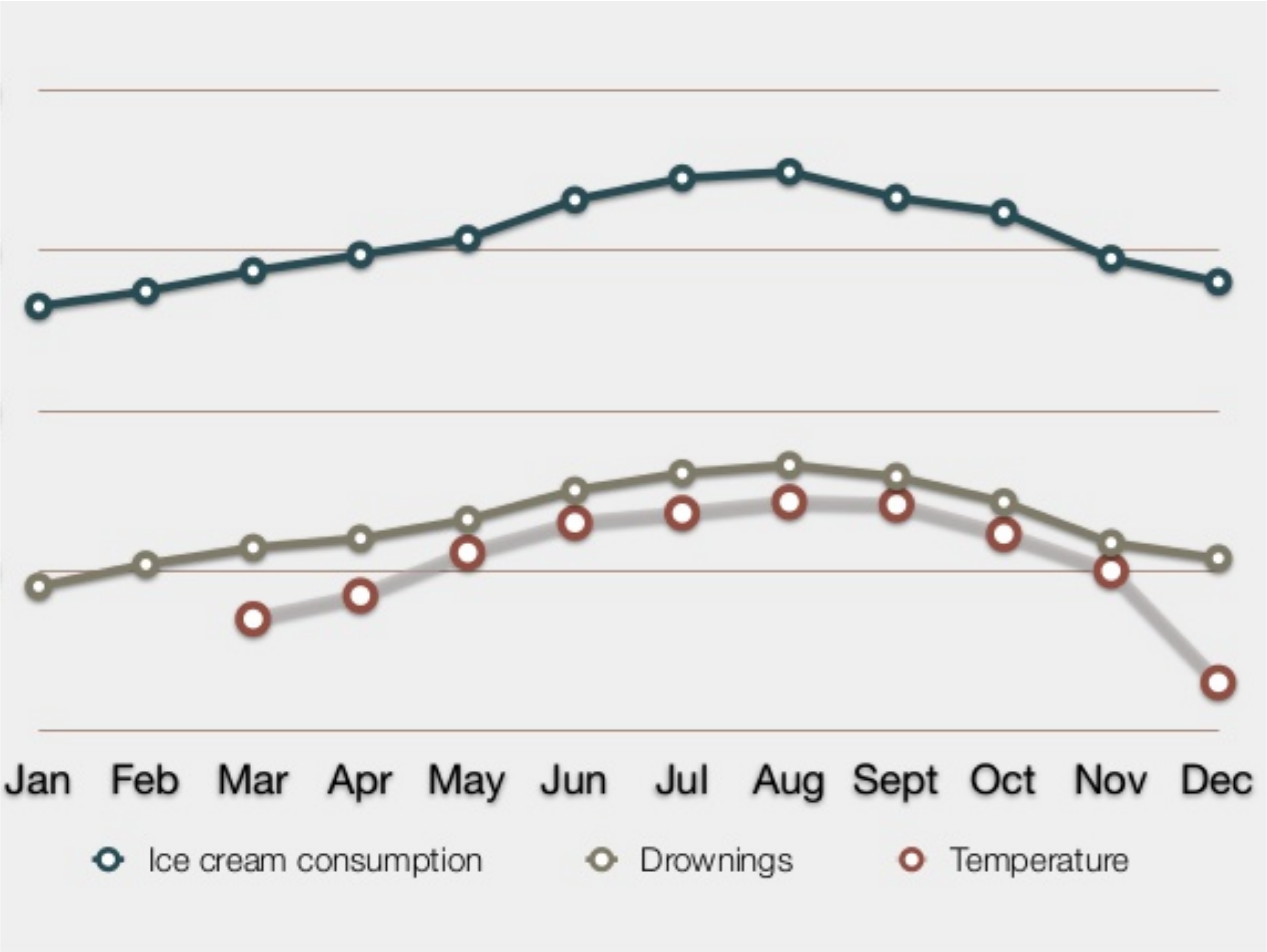
monthly data

Ice Cream vs Drowning

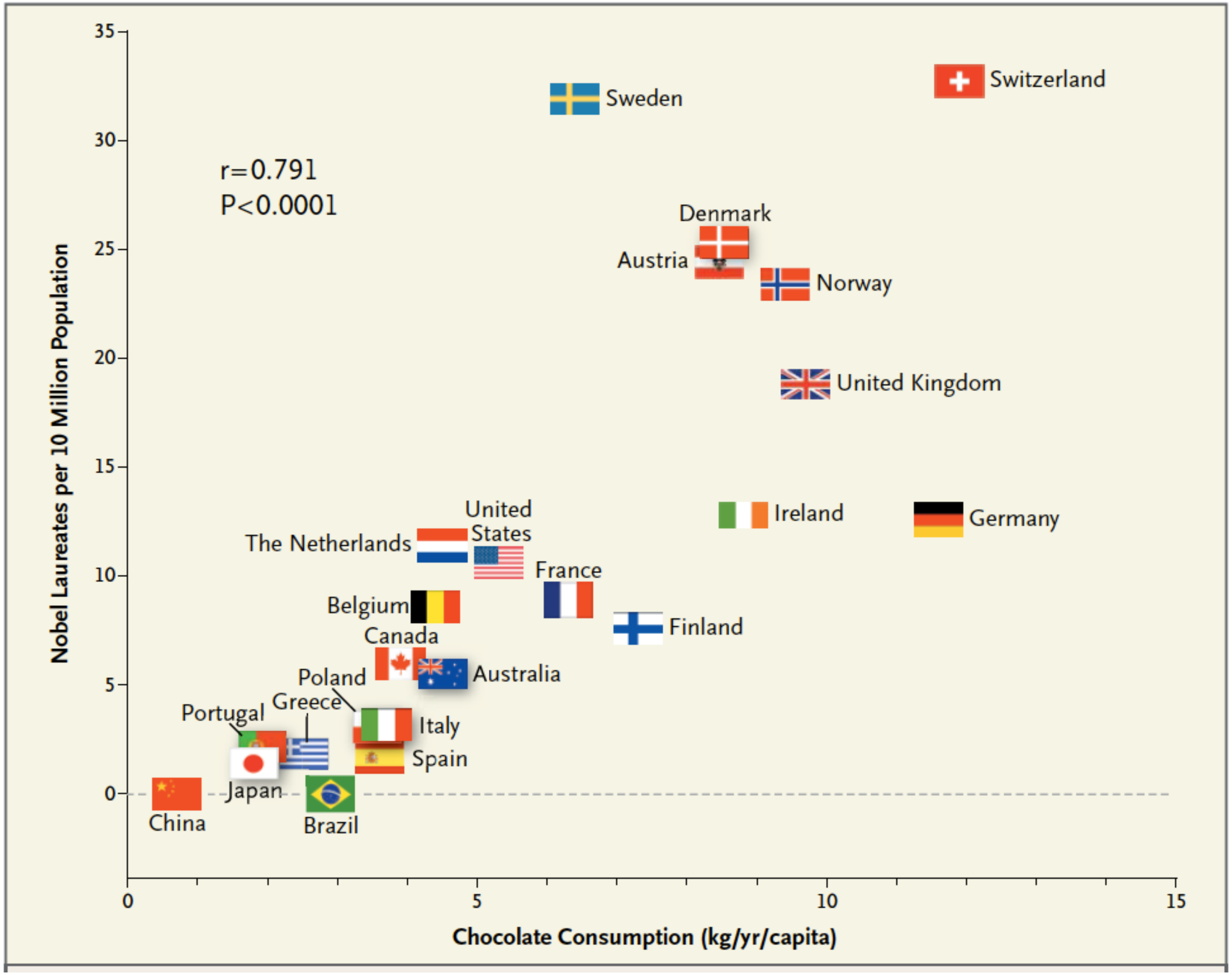
Ice Cream vs Drowning



Ice Cream vs Drowning

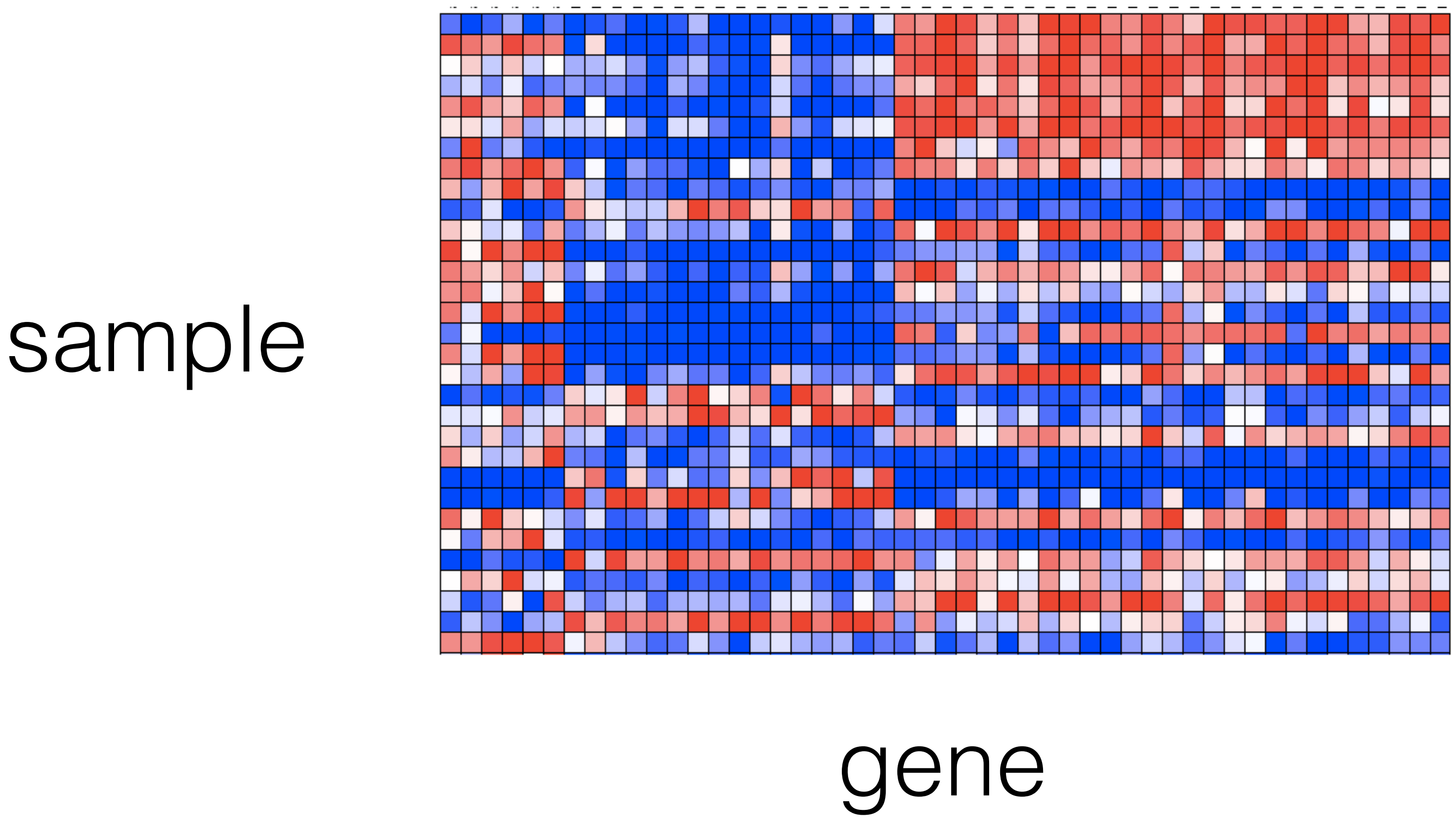


Chocolate vs Nobel Prizes



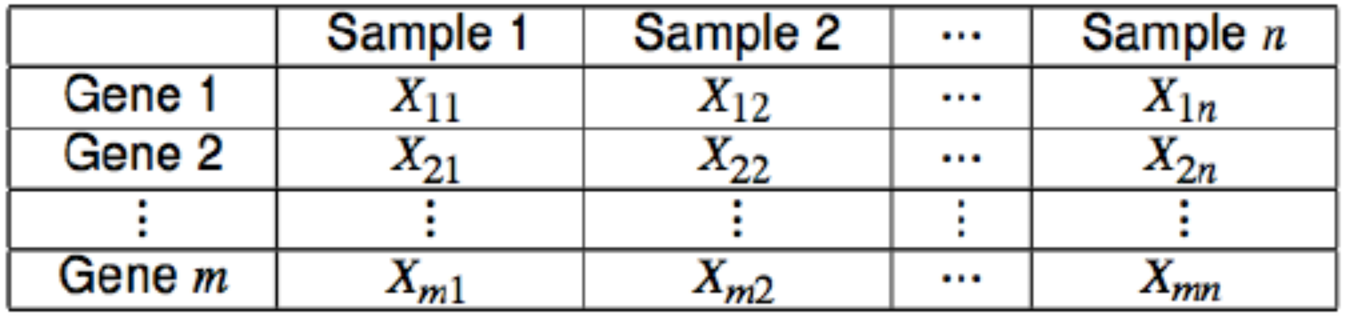
credit: NEJM, 2012

Gene expression analysis



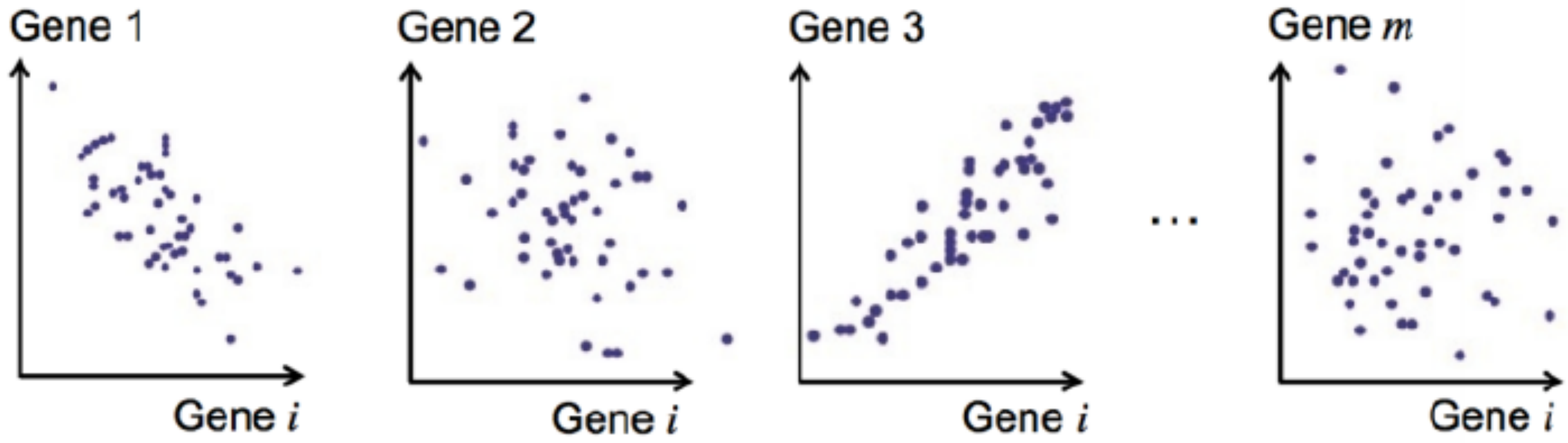
Correlation of genes across experimental conditions → coregulation of genes

Correlation analysis



	Sample 1	Sample 2	⋯	Sample n
Gene 1	X_{11}	X_{12}	⋯	X_{1n}
Gene 2	X_{21}	X_{22}	⋯	X_{2n}
⋮	⋮	⋮	⋮	⋮
Gene m	X_{m1}	X_{m2}	⋯	X_{mn}

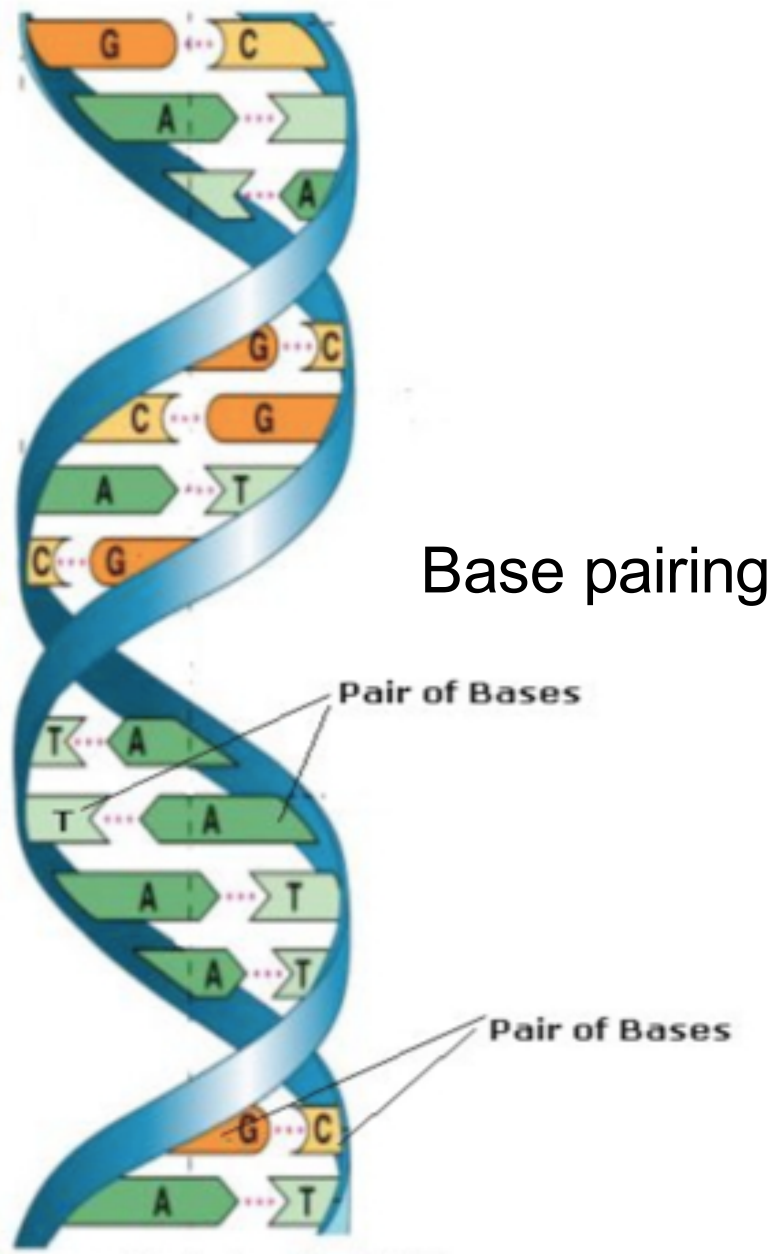
$$r = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2}\sqrt{\sum(Y-\overline{Y})^2}}$$



r=-0.8 r=-0.2 r=0.85 r=-0.15

Biological sequences

DNA



Base Pairing in DNA
Double Helix

Base pairing property

=

The DNA Molecule

5’

```
G -- C
A -- T
T -- A
G -- C
C -- G
G -- C
T -- A
G -- C
T -- A
T -- A
A -- T
A -- T
C -- G
T -- A
```

3’

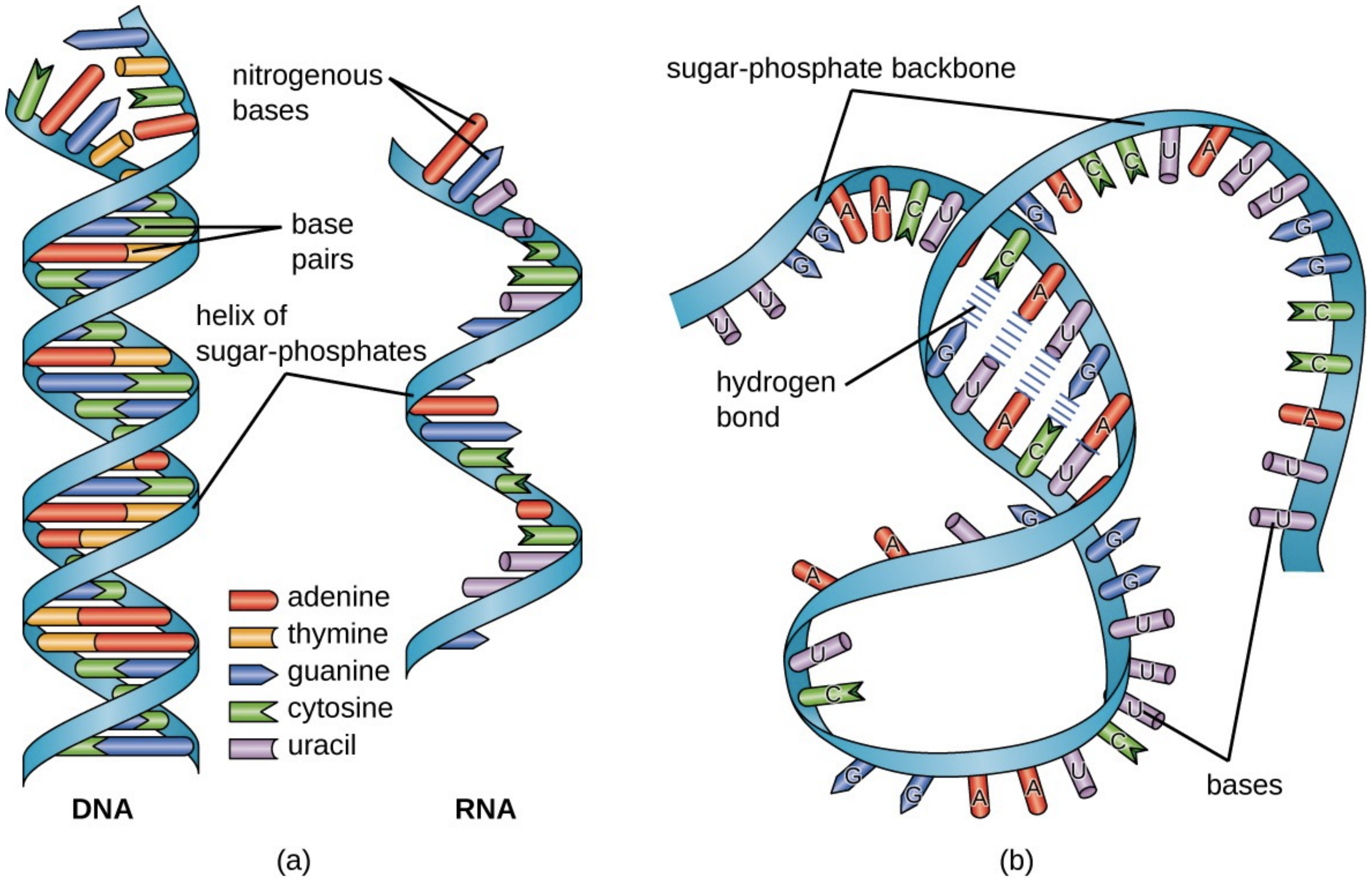
DNA to chromosome

SOURCE: http://www.microbe.org/espanol/news/human_genome.asp

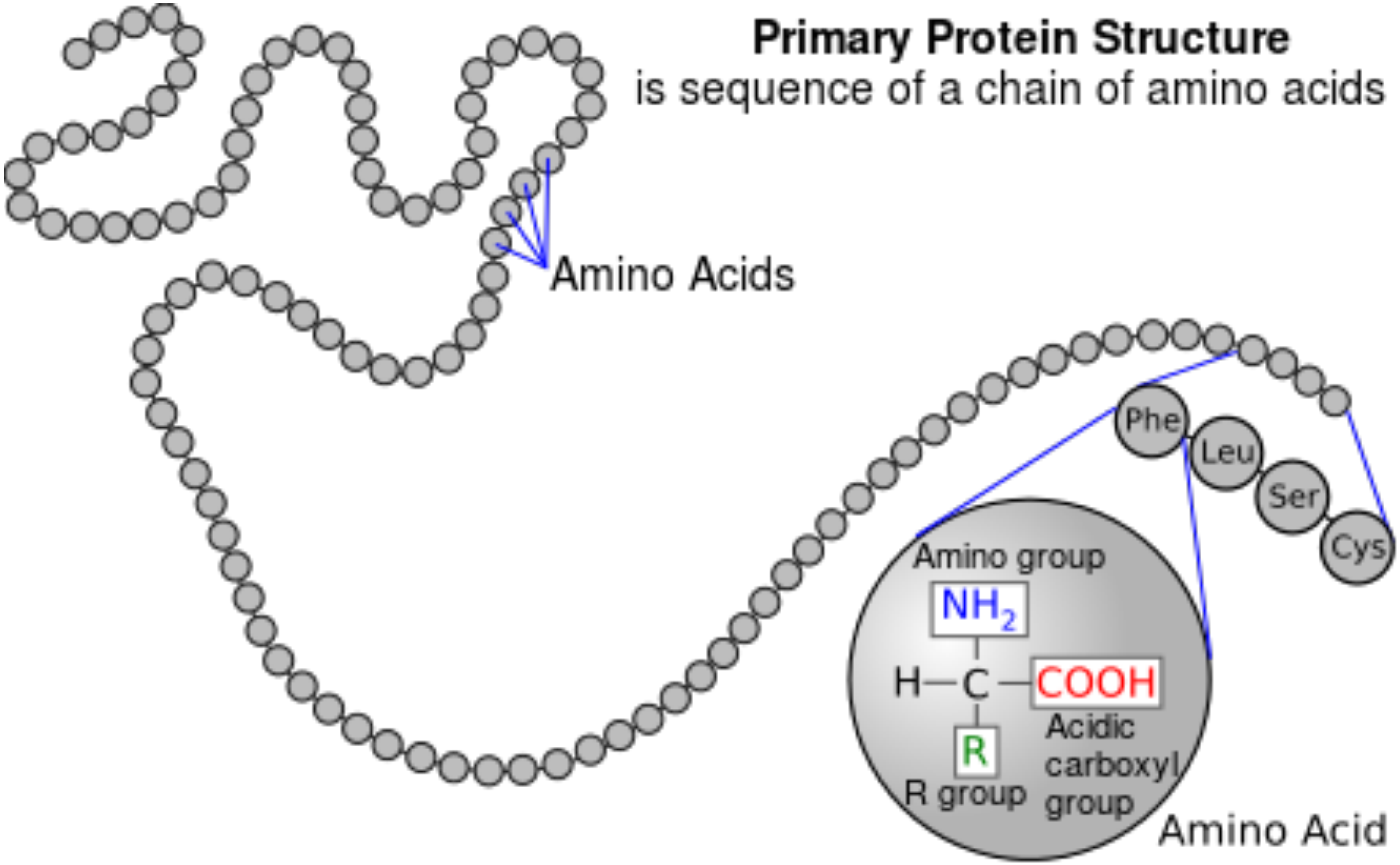
What is RNA?

RNA = ribonucleic acid

- “U” instead of “T”
- Usually single stranded
- Has base-pairing capability
  - Can form simple non-linear structures
- Life may have started with RNA



Protein sequence



A short summary: string transformation

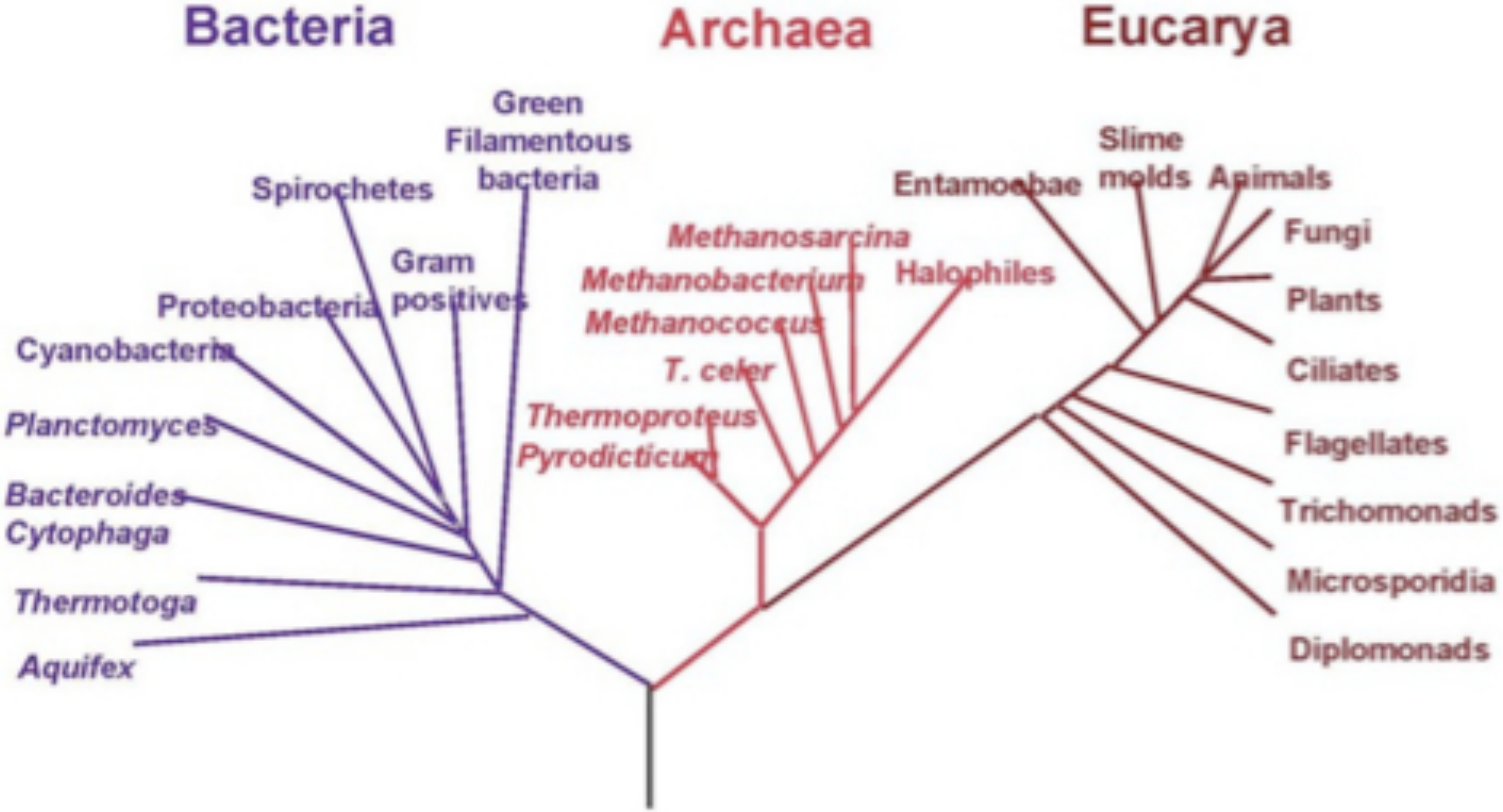
- DNA = nucleotide sequence
  - Alphabet size = 4 (A,C,G,T)

- DNA to mRNA (single stranded)
  - Alphabet size = 4 (A,C,G,U)

- mRNA to amino acid sequence
  - Alphabet size = 20

- Amino acid sequence “folds” into 3-dimensional protein

Phylogenetic Tree of Life



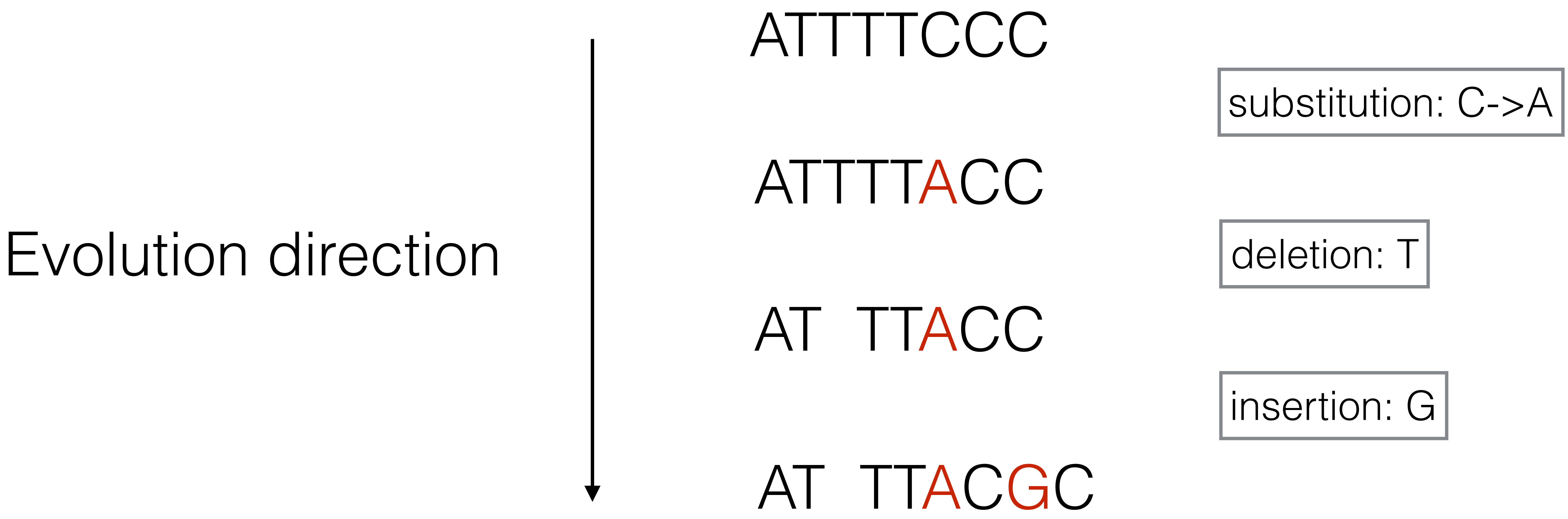
determined by DNA sequences

Evolution theory

- All organisms share the genetic code
- Similar genes across species
- Probably had a common ancestor
- Genomes are a wonderful resource to trace back the history of life

Evolutionary process of sequences

- Substitutions
- Insertions and Deletions
- Representing an alignment; “gaps”



Evolution direction

ATTTTCCC

substitution: C->A

ATTTTACC

deletion: T

AT TTACC

insertion: G

AT TTACGC

Sequence alignment

Correspondence between bases of two DNA sequences, or between amino acids of two protein sequences

Alignment : 2 x k matrix (k ≥ m, n)

V = ACCTGGTAAA n = 10

W = ACTGCGTATA m = 10

8	matches
1	mismatches
1	deletions
1	insertions

V	A	C	C	T	G	—	G	T	A	A	A
W	A	C	—	T	G	C	G	T	A	T	A