Statistical **I**nference and **M**achine **L**earning in **E**arth **S**cience
SIMLES

Module 1
Probability Models

Lecture 3
Moments
Sampling Distributions

Cristi Proistosescu

Probability Models

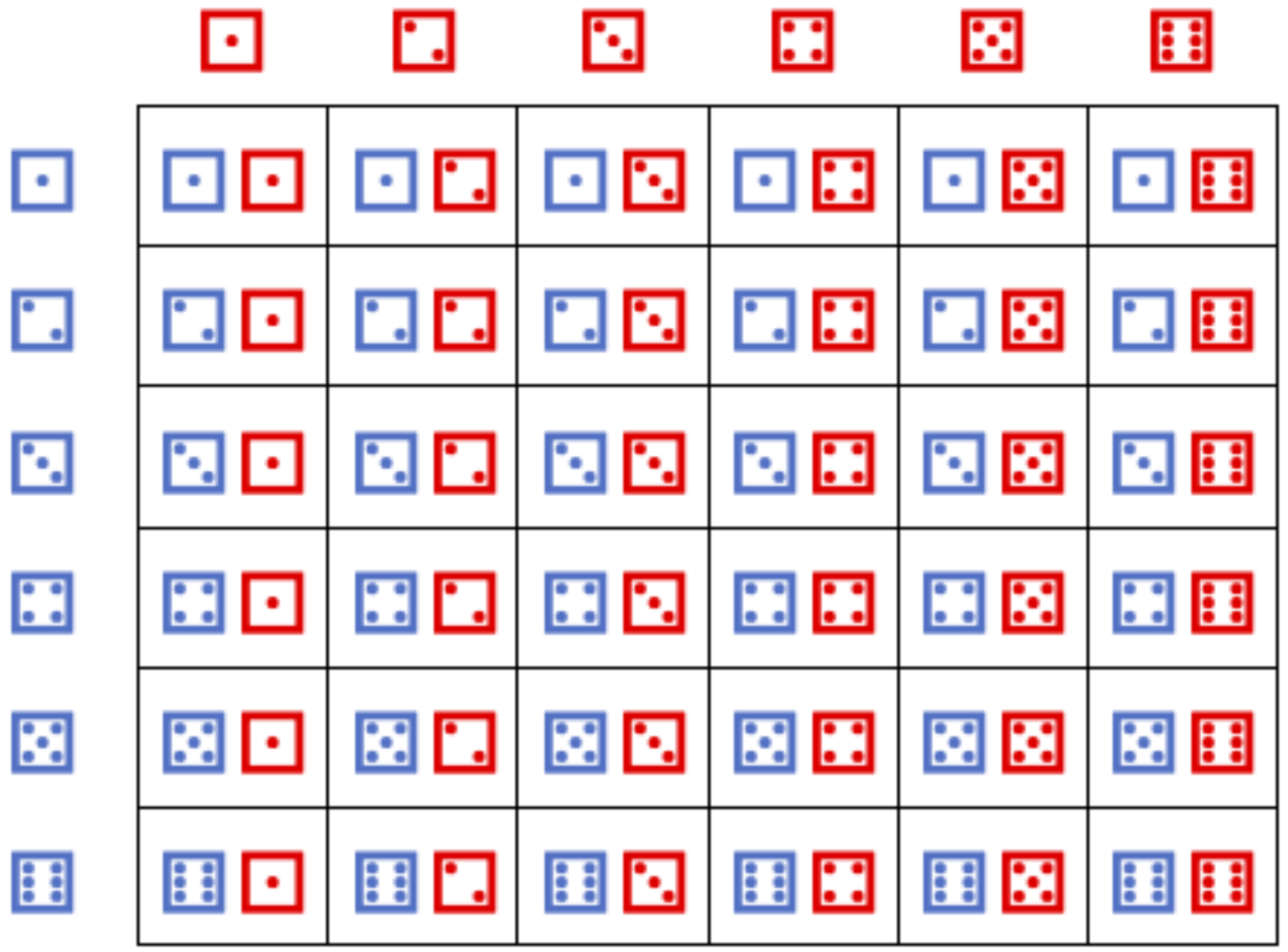
Probability models: allow us to model a random process

Probability Models: Random Variable

Probability models: allow us to model a random process

Random variables allow us to simulate the outcome of random processes.

- dice roll



```
X=stats.randint.rvs(low=1, high=7, size=1)
Y=stats.randint.rvs(low=1, high=7, size=1)
S=X+Y
print(S)
```

```
[6]
```

Probability Models: Random Variable

Probability models: allow us to model a random process

Random variables allow us to simulate the outcome of random processes.

- dice roll
- coin flip

0 1

$X \sim \text{Bern}(p)$

```
# Rerun this cell to generate Bernoulli random variables with $p=0.9$
X=stats.bernoulli.rvs(p=0.9, size=50)
print(X)
```

```
[1 1 1 1 0 1 1 1 1 0 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 0 1 1 1 1 0
 1 1 1 1 1 1 1 1 1 1 0 1 1]
```

Probability Models: Random Variable

Probability models: allow us to model a random process

Random variables allow us to simulate the outcome of random processes.

- dice roll
- coin flip

0 1

$X \sim \text{Bern}(p)$

```
# rerun this cell to simulate more realizations of a bernoulli random variable
X=stats.randint.rvs(low=1, high=7, size=1)
if X>3:
    Y=1
else:
    Y=0
print(Y)
```

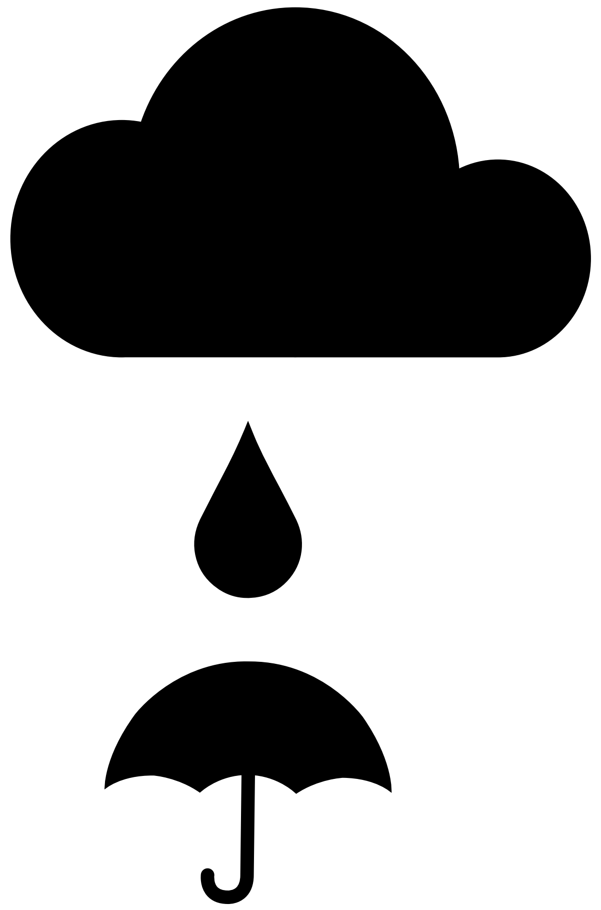
```
1
```

Probability Models: Random Variable

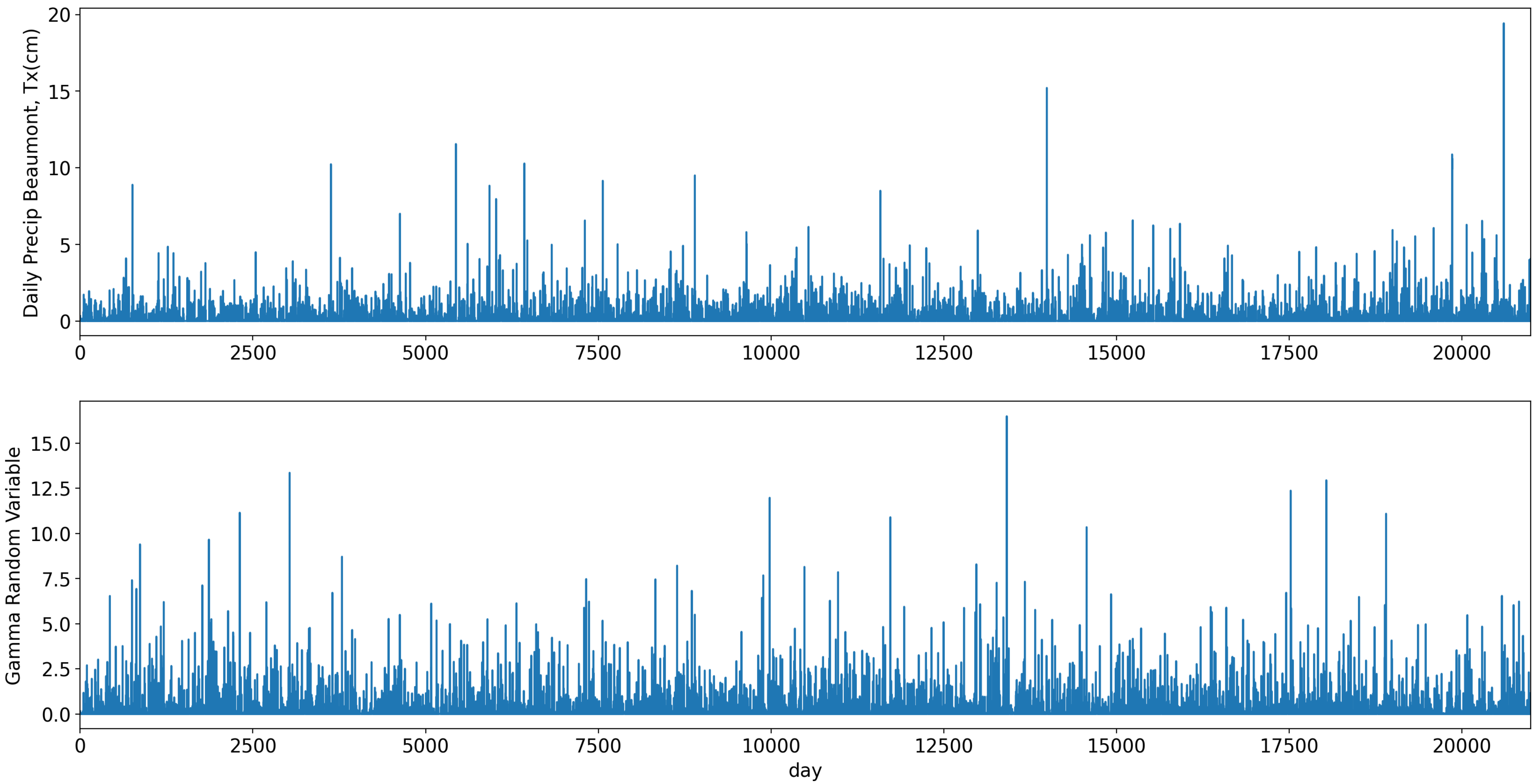
Probability models: allow us to model a random process

Random variables allow us to simulate the outcome of random processes.

- dice roll
- coin flip
- precipitation



$$X \sim \Gamma(\alpha, \beta)$$

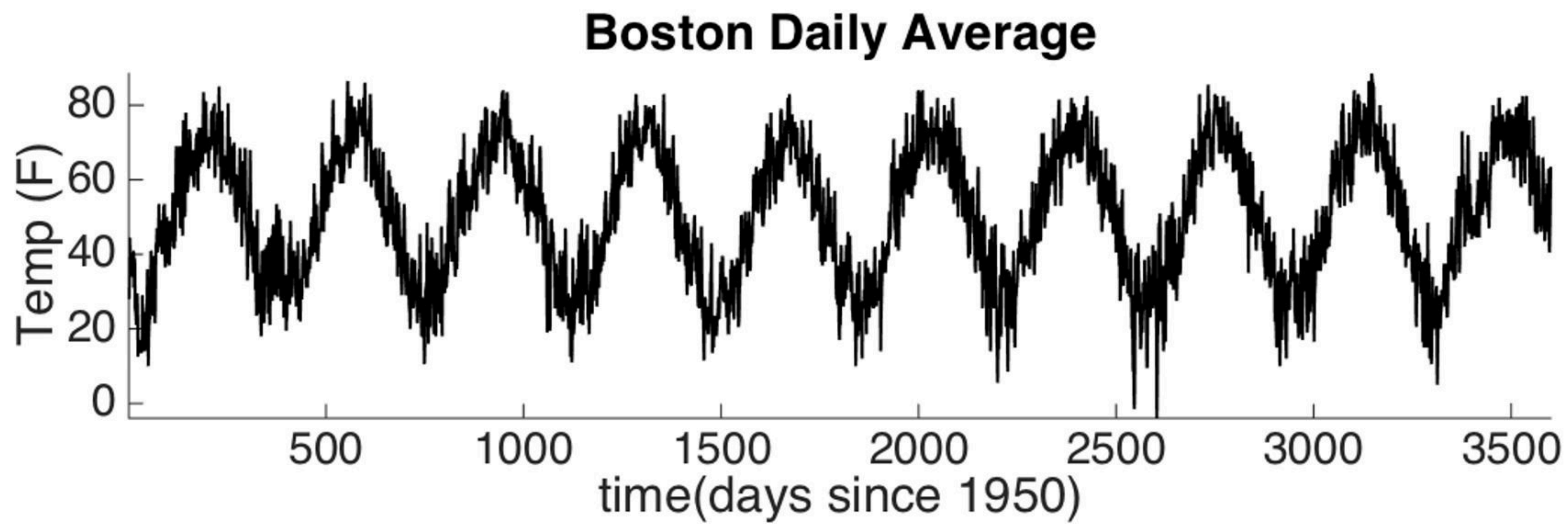


Probability Models: Random Variable

Probability models: allow us to model a random process

Random variables allow us to simulate the outcome of random processes.

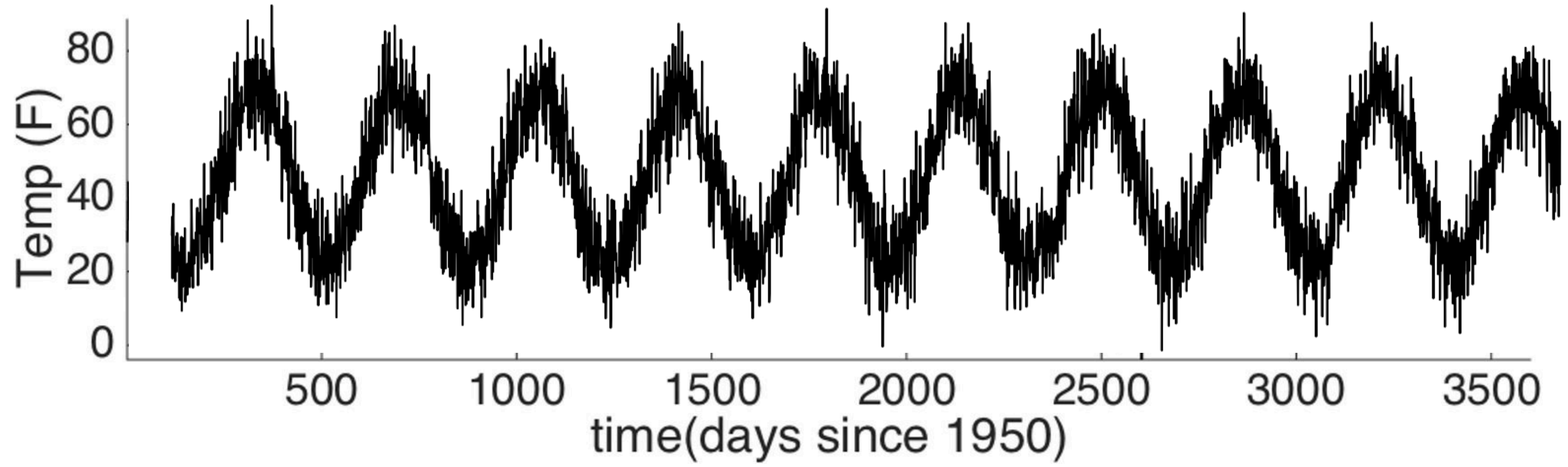
- dice roll
- coin flip
- precipitation
- temperature



$$T \sim A \sin\left(\frac{2\pi}{T}t + \phi\right) + X$$

$$X \sim \mathcal{N}(0,\sigma)$$

Probabilistic model



Probability Models: Random Variable

Probability models: allow us to model a random process

Random variables allow us to simulate the outcome of random processes.

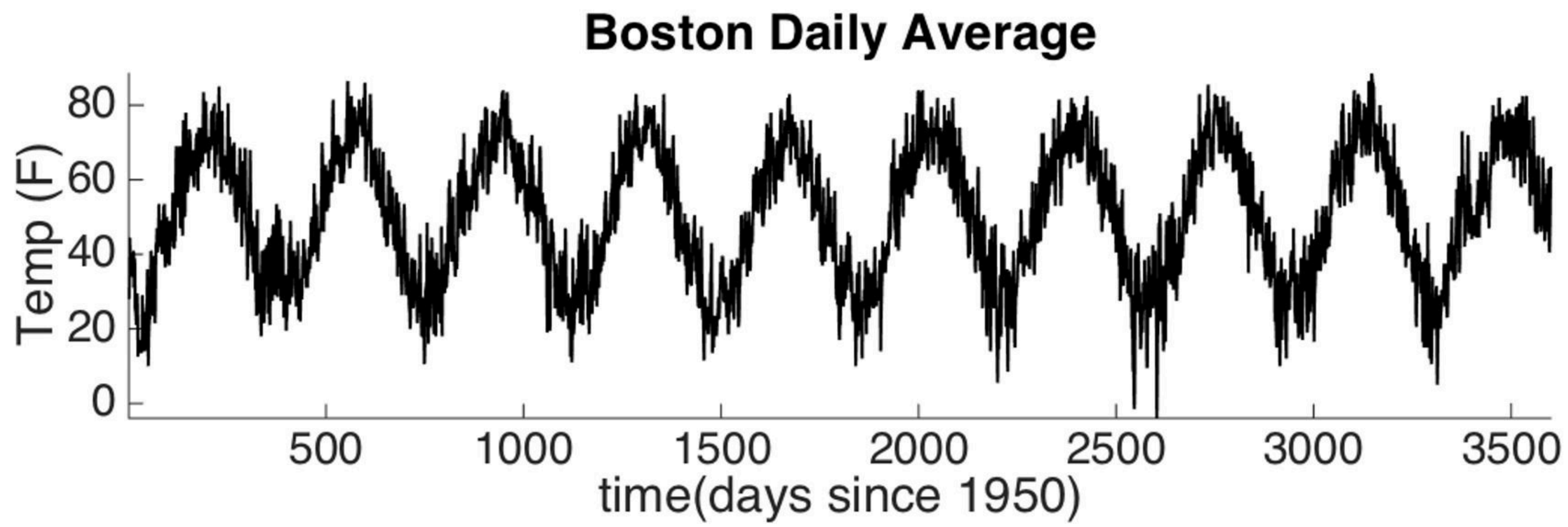
- dice roll
- coin flip
- precipitation
- temperature

$$T \sim A \sin\left(\frac{2\pi}{T}t + \phi\right) + X$$

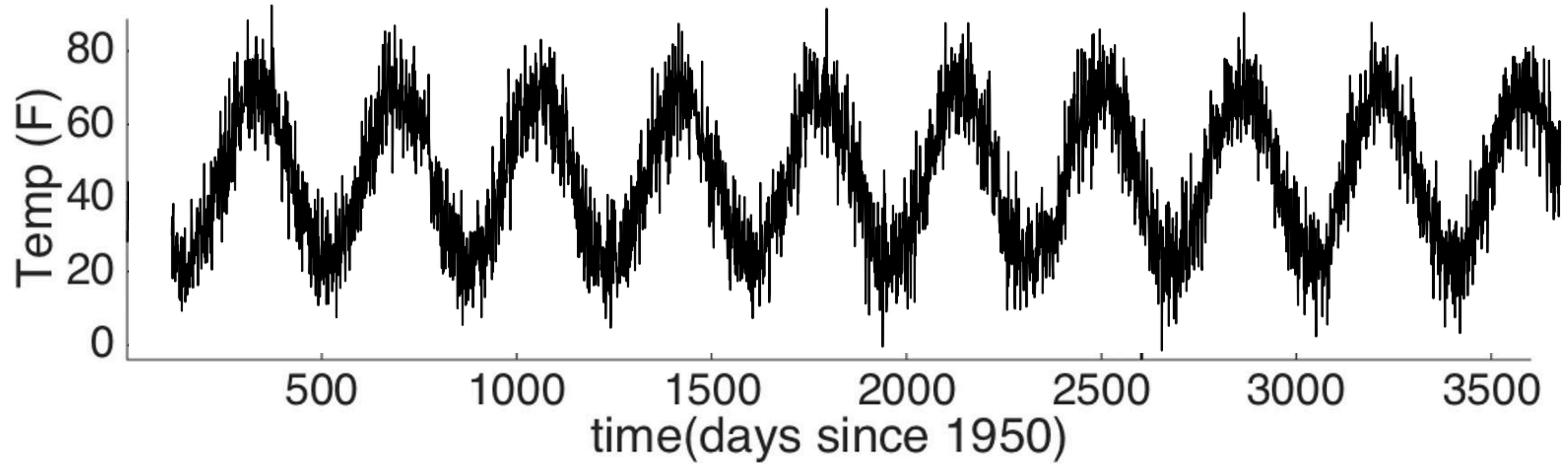
$$T \sim \mathcal{N}(0,\sigma)$$

```
Ndays=3600        #number of days to simulat
T = 365.25        #period of the annual cycle (days)
A = 30            #amplitude of annual cycle
phase = 130       # phase shift (to the sine to start at the right poin
mu     = 50       # mean temperature
sig= 30           # spread of weather

T=A*np.sin(2*np.pi*(t-phase)/T) + stats.norm.rvs(loc=mu,scale=sig,size
```



Probabilistic model



Probability Models: Distributions

Probability models: allow us to model a random process

Random variables allow us to simulate the outcome of random processes.

Distributions: summarize the outcome of a random process.

- **pmfs, pdfs** describe the underlying *process*.
- **Histograms** summarize a finite sample.

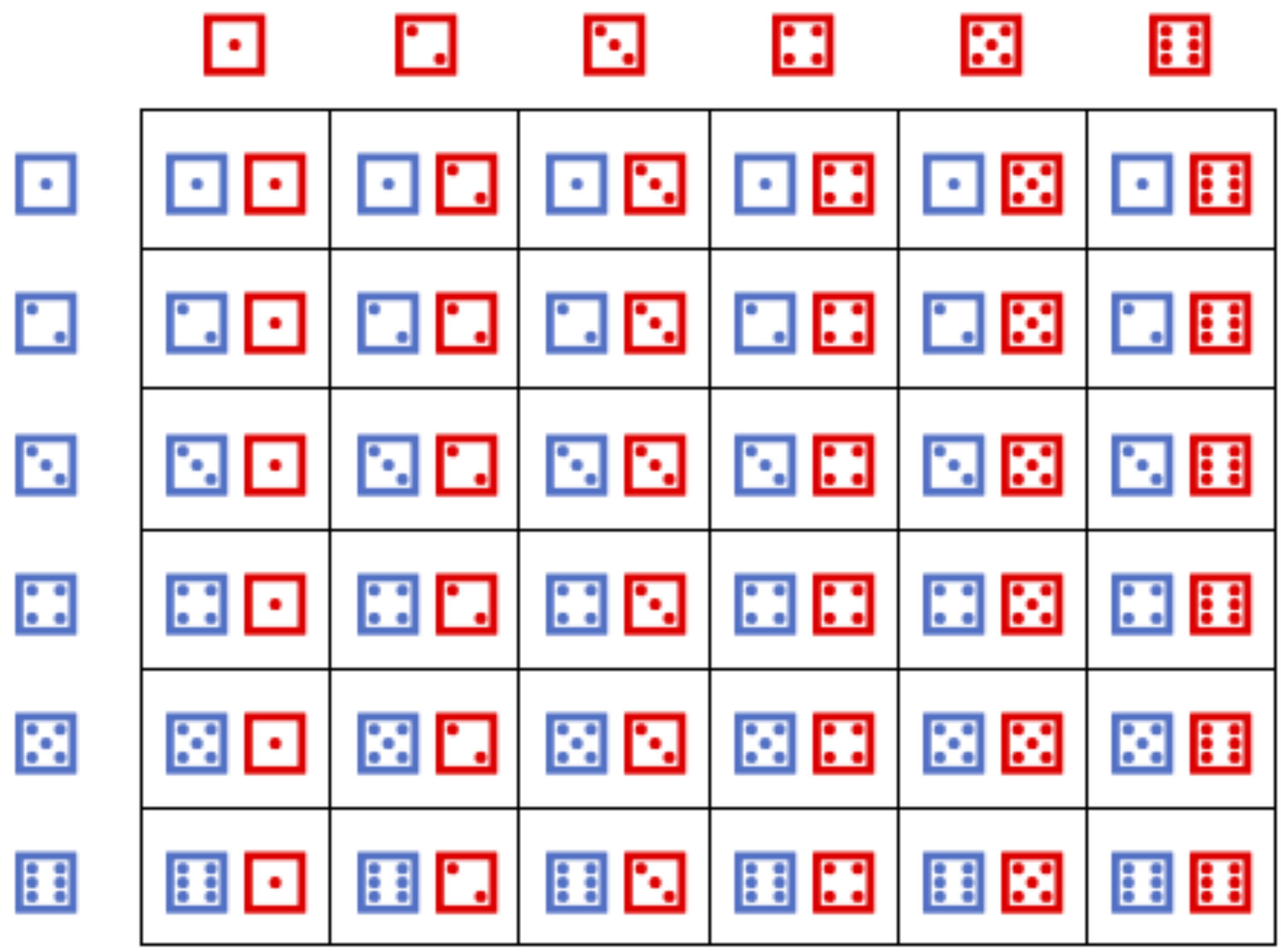
Probability Models: Distributions

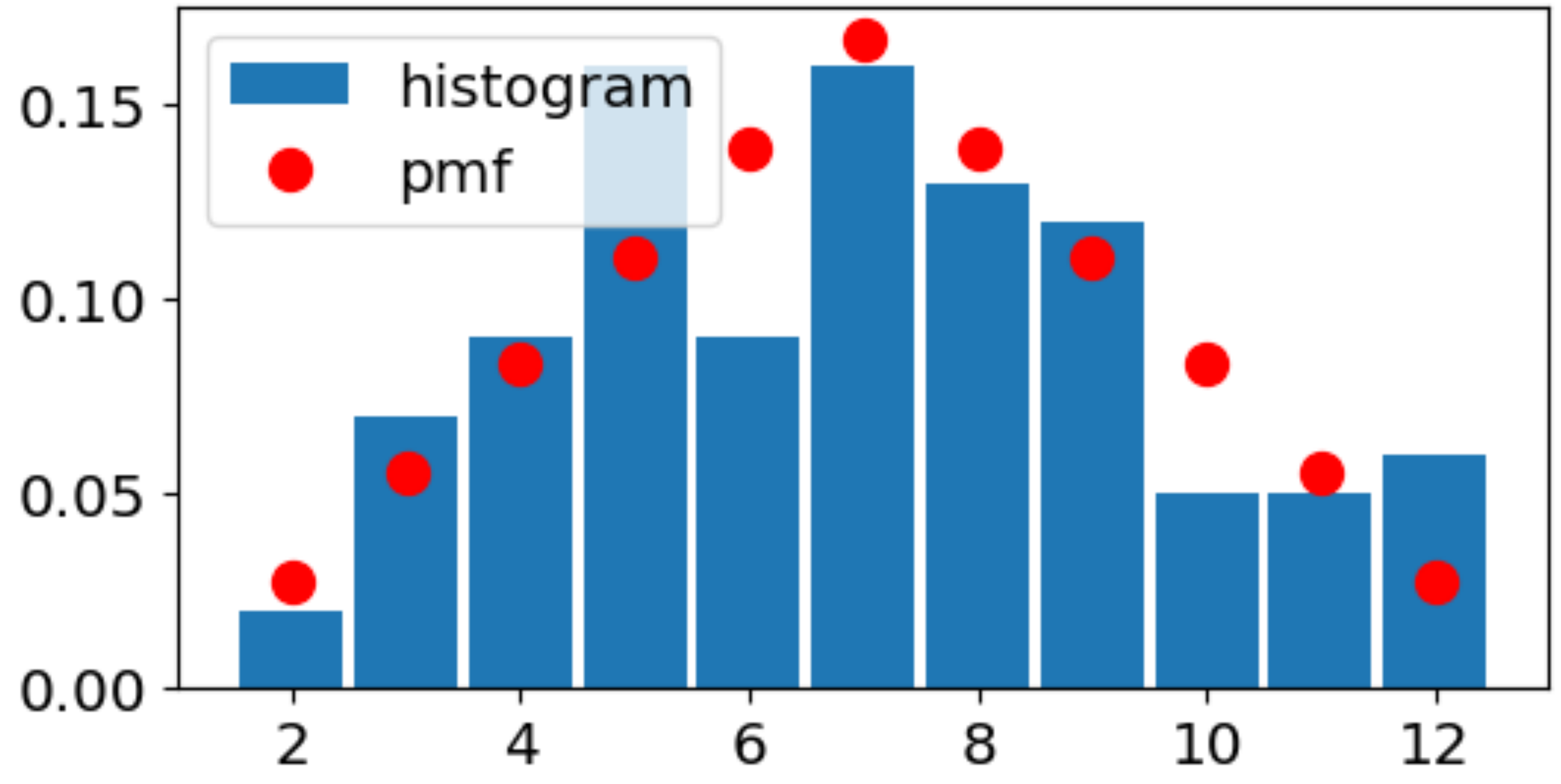
Probability models: allow us to model a random process

Random variables allow us to simulate the outcome of random processes.

Distributions: summarize the outcome of a random process.

- **pmfs, pdfs** describe the underlying *process*.
- **Histograms** summarize a finite sample.





$$N_{draws} = 100$$

Probability Models: Moments

Probability models: allow us to model a random process

Random variables allow us to simulate the outcome of random processes.

Distributions: summarize the outcome of a random process.

- **pmfs, pdfs** describe the underlying *process*.
- **Histograms** summarize a finite sample.

Moments: another useful summary of the outcome of a random process.

Moments: Expected Value

Expected value

$$E(X) = \sum_{i=1}^{N} x \cdot P(X = x_i)$$

$$E(X) = \int_{\mathbb{R}} x \cdot p(x) dx$$

The expected value of a random variable is the average value we would expect to get if we could sample a random variable an infinite number of times.

It is a average of all the possible outcomes, weighted by how probably they are.

Moments: mean

Key property: linearity

$$E(aX + bY) = aE(x) + bE(y)$$

Moments: mean

Key property: linearity

$$E(aX + bY) = aE(x) + bE(y)$$

We can also define the expected value, or mean, of any function of a random variable:

$$E(g(X)) = \begin{cases} \sum_{i=1}^{N} g(x_i)P(X = x_i) & \text{if } X \text{ is discrete} \\ \int_{\mathbb{R}} g(x)f(x)dx & \text{if } X \text{ is continuous} \end{cases}$$

Moments: variance

A closely related notion to the second order moment is the **variance** or centered second moment, defined as:

$$V(X) = E([X - E(x)]^2) = E([X - \mu]^2) = \int_{\mathbb{R}} (x - \mu)^2 f(x) dx$$

Moments: variance

A closely related notion to the second order moment is the **variance** or centered second moment, defined as:

$$V(X) = E([X - E(x)]^2) = E([X - \mu]^2) = \int_{\mathbb{R}} (x - \mu)^2 f(x) dx$$

Expanding the square and using the linearity:

$$V(X) = E(X^2) - (E(X))^2 = E(X^2) - \mu^2$$

Moments: variance

A closely related notion to the second order moment is the **variance** or centered second moment, defined as:

$$V(X) = E([X - E(x)]^2) = E([X - \mu]^2) = \int_{\mathbb{R}} (x - \mu)^2 f(x) dx$$

Expanding the square and using the linearity:

$$V(X) = E(X^2) - (E(X))^2 = E(X^2) - \mu^2$$

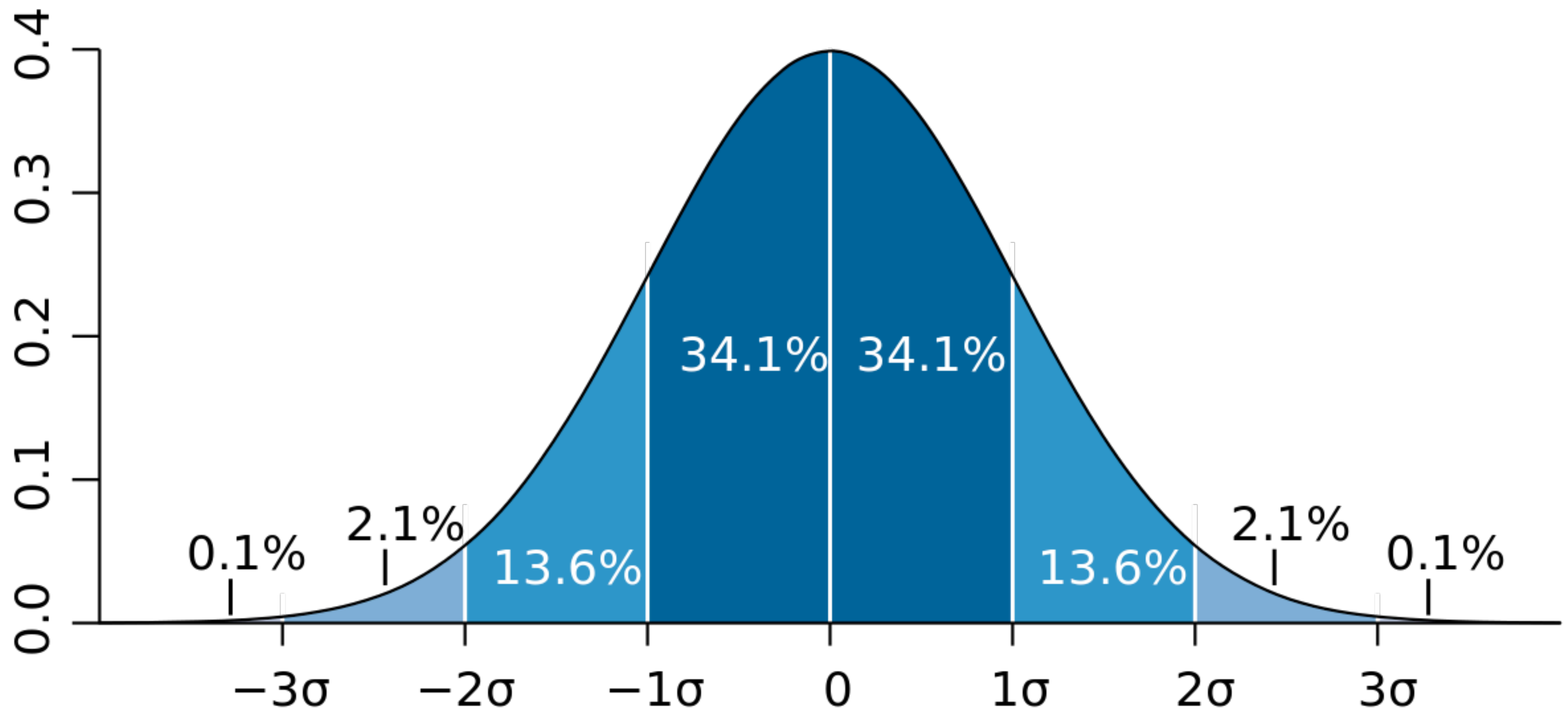
Another closely related measure is standard deviation,

$$\text{std} = \sqrt{V(X)} = \sqrt{E([X - \mu]^2)}$$

Normal distribution

X= normal with location parameter μ and scale parameter σ

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$



$$E(X) = \mu$$
$$V(X) = \sigma^2$$
$$\text{std} = \sigma$$

Normal distribution

DEMO: mean and variance of normal distribution

Moments

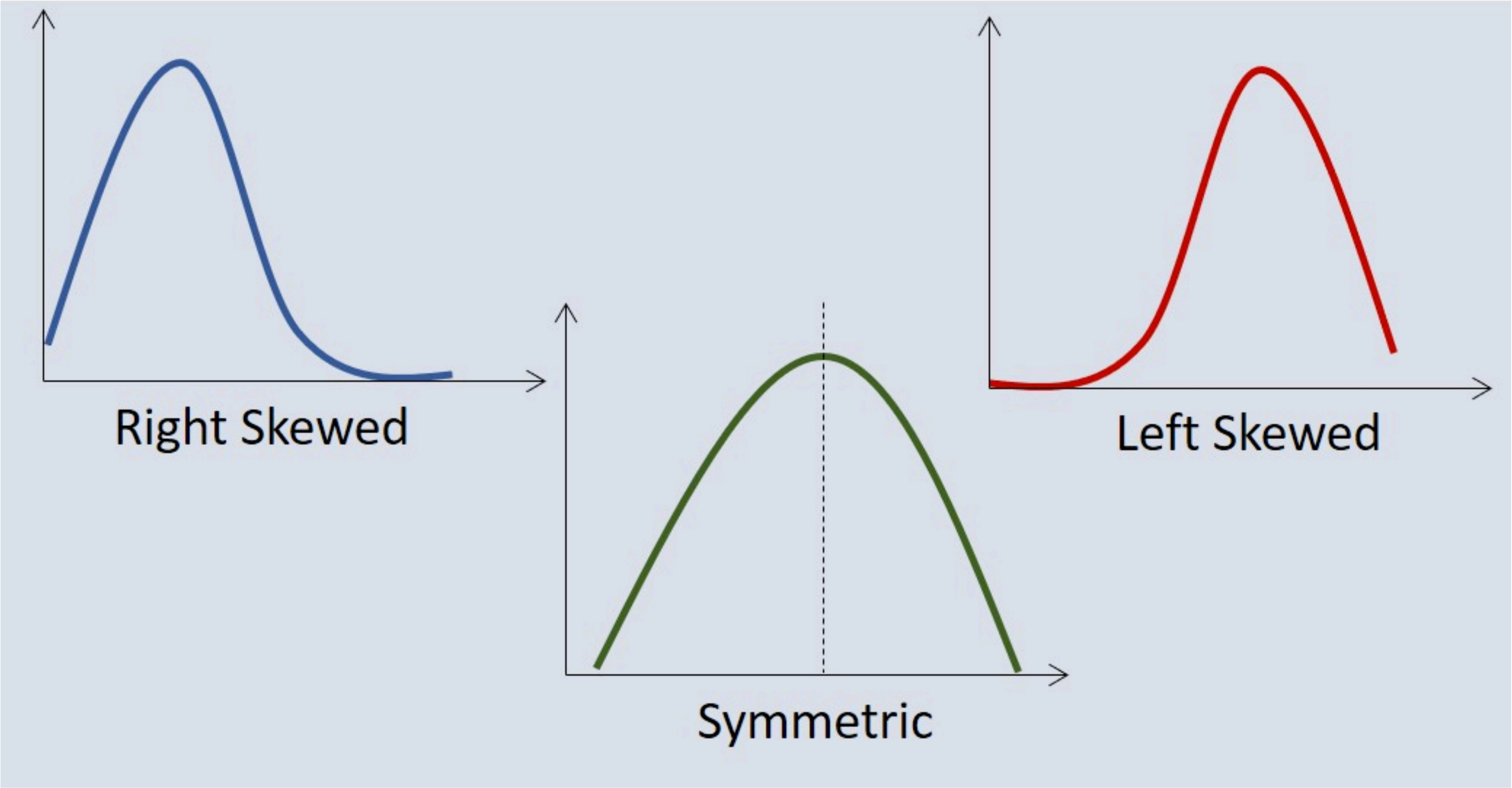
Moments of a distribution

$$m(X,n) = E(X^n) = \sum_{i=1}^{N} x_i^n \cdot P(X = x_i)$$

$$m(X,n) = E(X^n) = \int_{\mathbb{R}} x^n \cdot p(x)dx$$

Higher order moments: Skewness

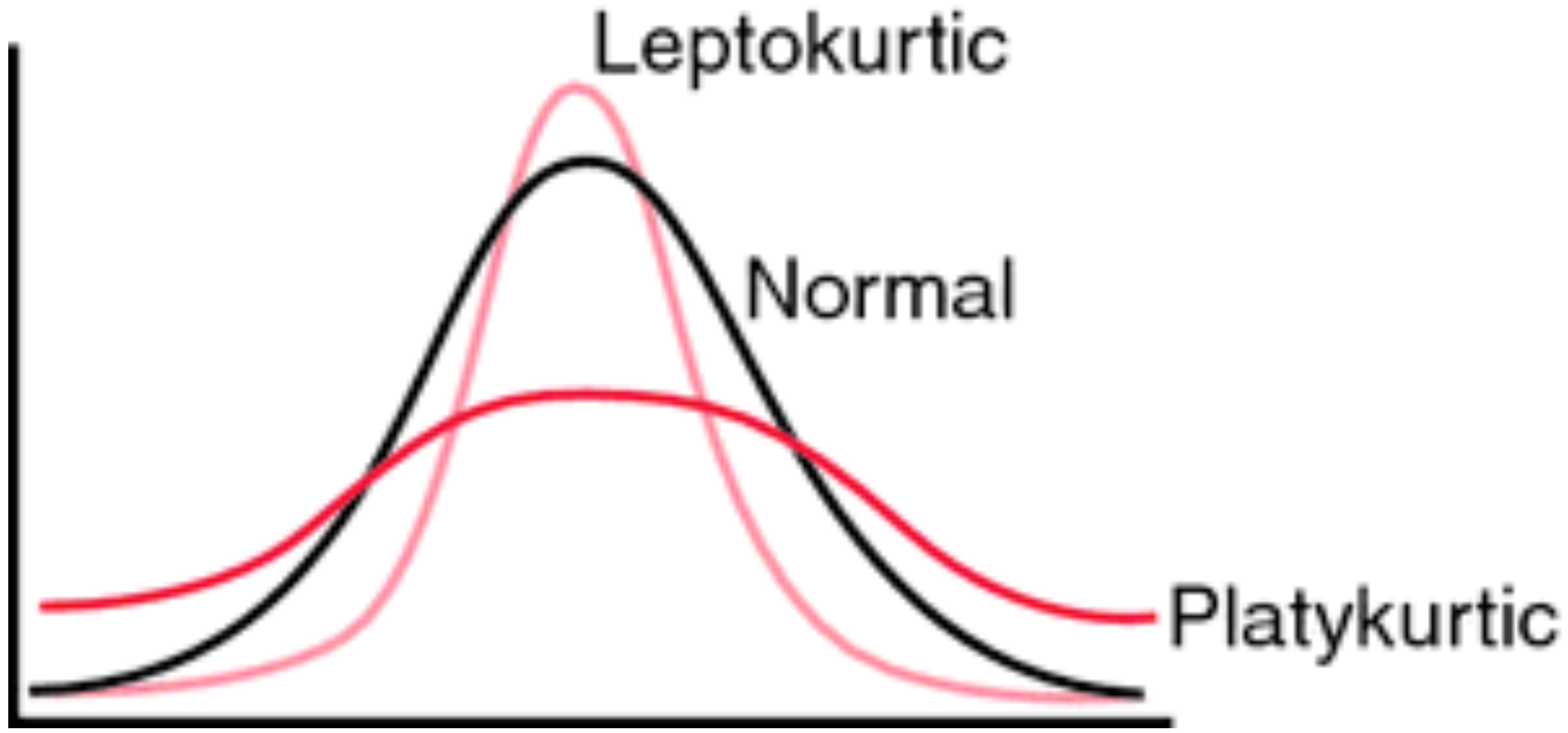
$$S = E\left(\left[\frac{X-\mu}{\sigma}\right]^3\right)$$



Higher order moments: Kurtosis

$$K = E\left(\left[\frac{X-\mu}{\sigma}\right]^4\right)$$

$$K_{\text{normal}} = 3$$



Bivariate distributions

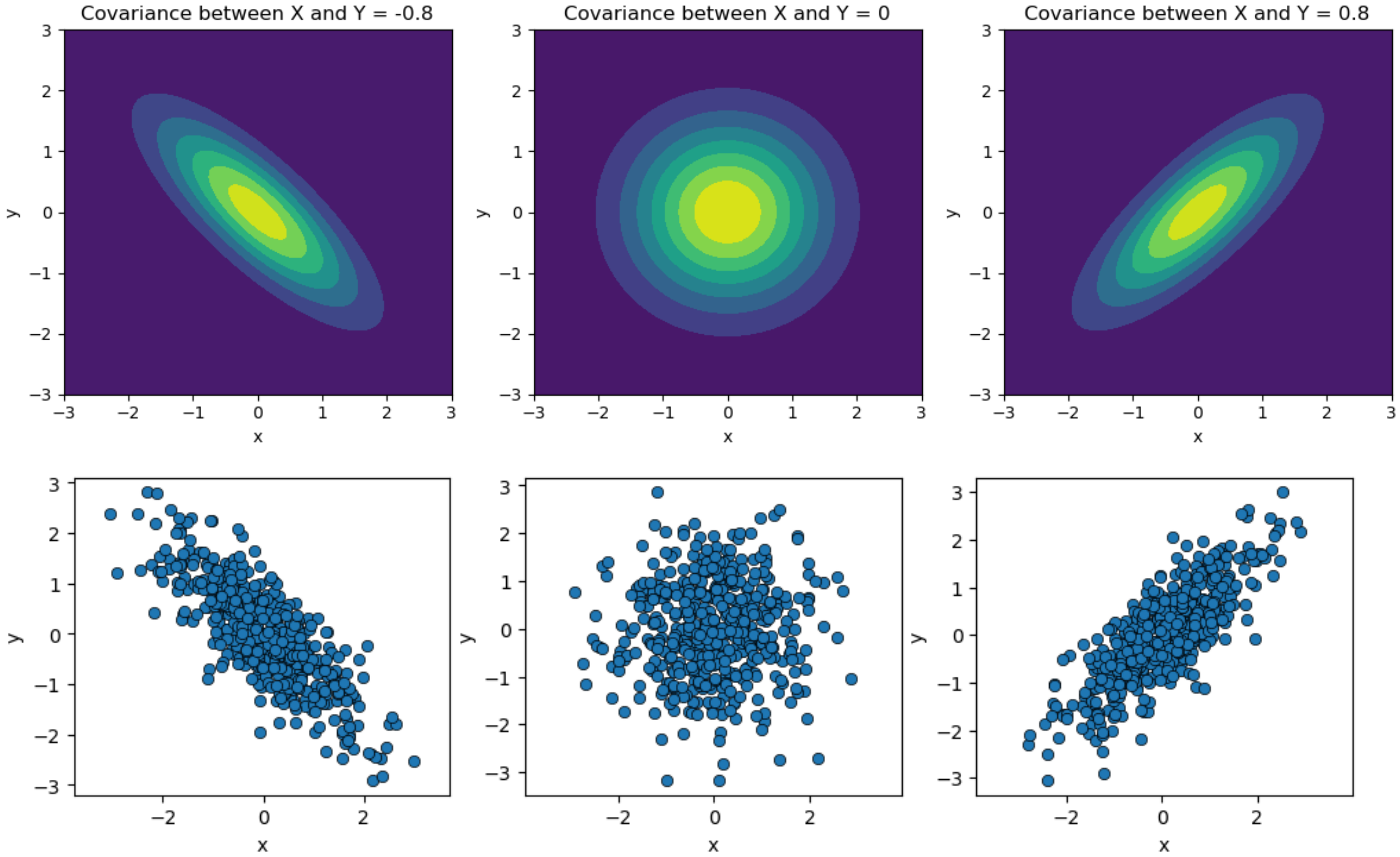
Joint Distributions

$$p(x, y) = P(X \in [x, x + dx] \text{ AND } Y \in [y, y + dy])$$

Bivariate pdf

Bivariate histogram

$N = 500$



Covariance & Correlation

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = \left\langle \left(X - \langle X \rangle\right) \left(Y - \langle Y \rangle\right) \right\rangle$$

Covariance & Correlation

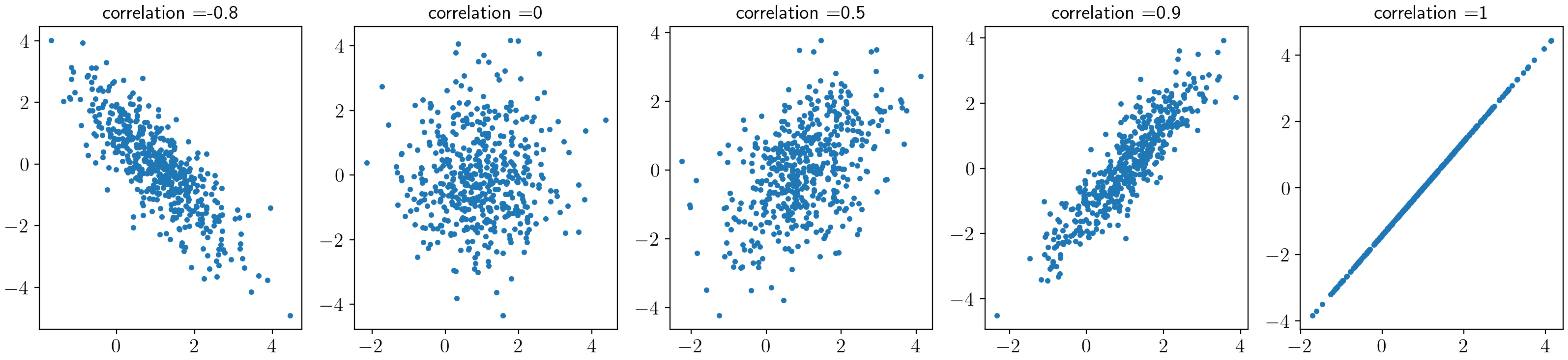
$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = \Big\langle \big(X - \langle X \rangle\big)\,\big(Y - \langle Y \rangle\big) \Big\rangle$$

$$Corr(X, Y) = \rho_{XY} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

Covariance & Correlation

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = \Big\langle \big(X - \langle X\rangle\big)\,\big(Y-\langle Y\rangle\big)\Big\rangle$$

$$Corr(X,Y) = \rho_{XY} = \frac{Cov(X,Y)}{\sigma_X\sigma_Y}$$



Covariance & Correlation

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = \left\langle \left(X - \langle X \rangle\right) \left(Y - \langle Y \rangle\right) \right\rangle$$

$$Corr(X,Y) = \rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

Important

Key Properties of Covariance

$$Cov(X,Y) = Cov(Y,x)$$

$$Cov(aX,Y) = a \cdot Cov(X,y)$$

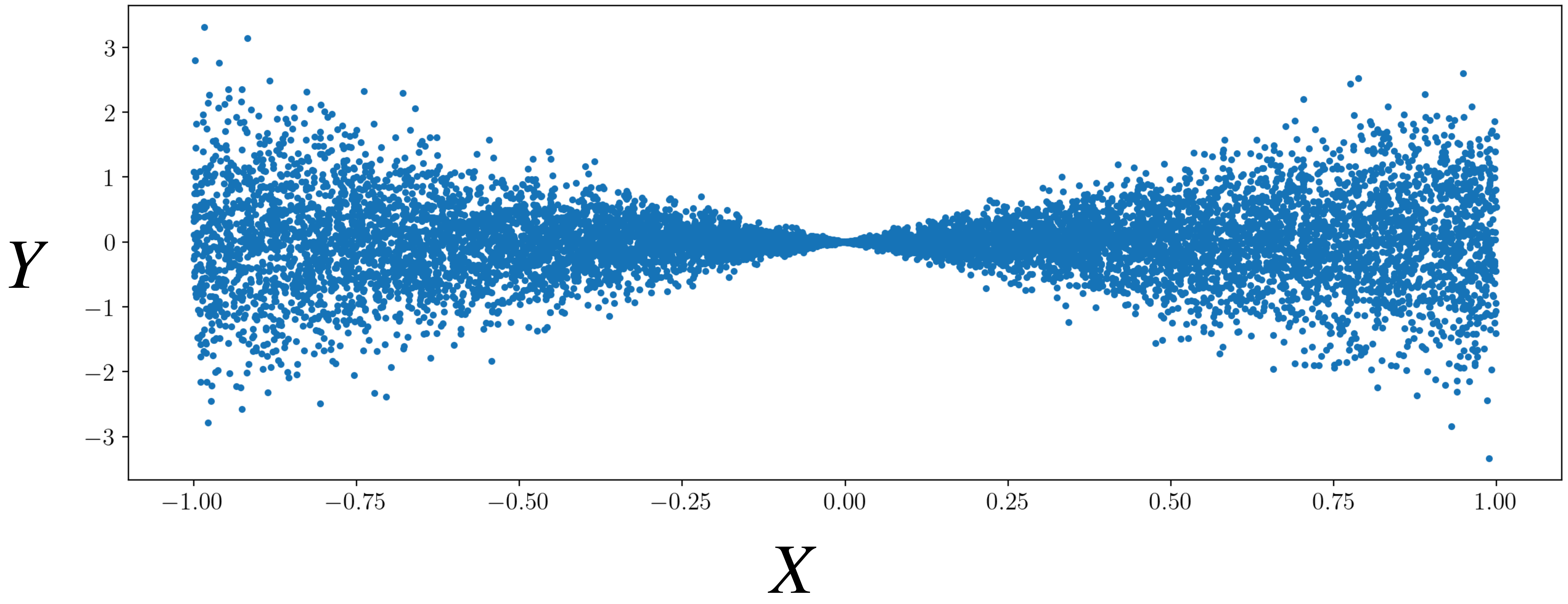
$$Cov(X,Y+Z) = Cov(X,Y) + COV(Y,Z)$$

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$

Correlation vs independence

$$X \sim U[-1,1]$$

$$Y|X \sim \mathcal{N}\left(0, X^2\right)$$



Probability Models: Moments

Probability models: allow us to model a random process

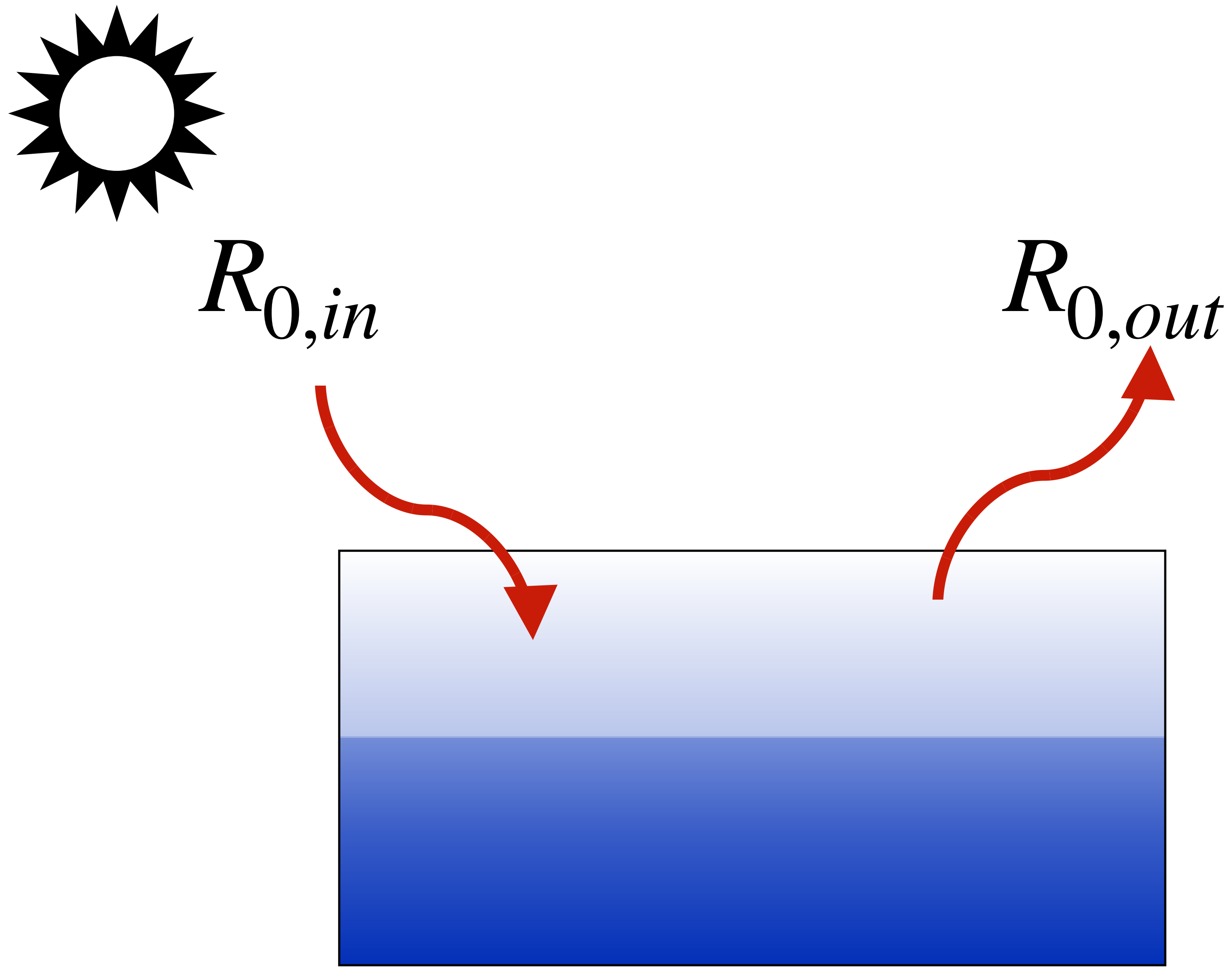
Random variables allow us to simulate the outcome of random processes.

Distributions: summarize the outcome of a random process.

- **pmfs, pdfs** describe the underlying *process*.
- **Histograms** summarize a finite sample.

Moments: another useful summary of the outcome of a random process.

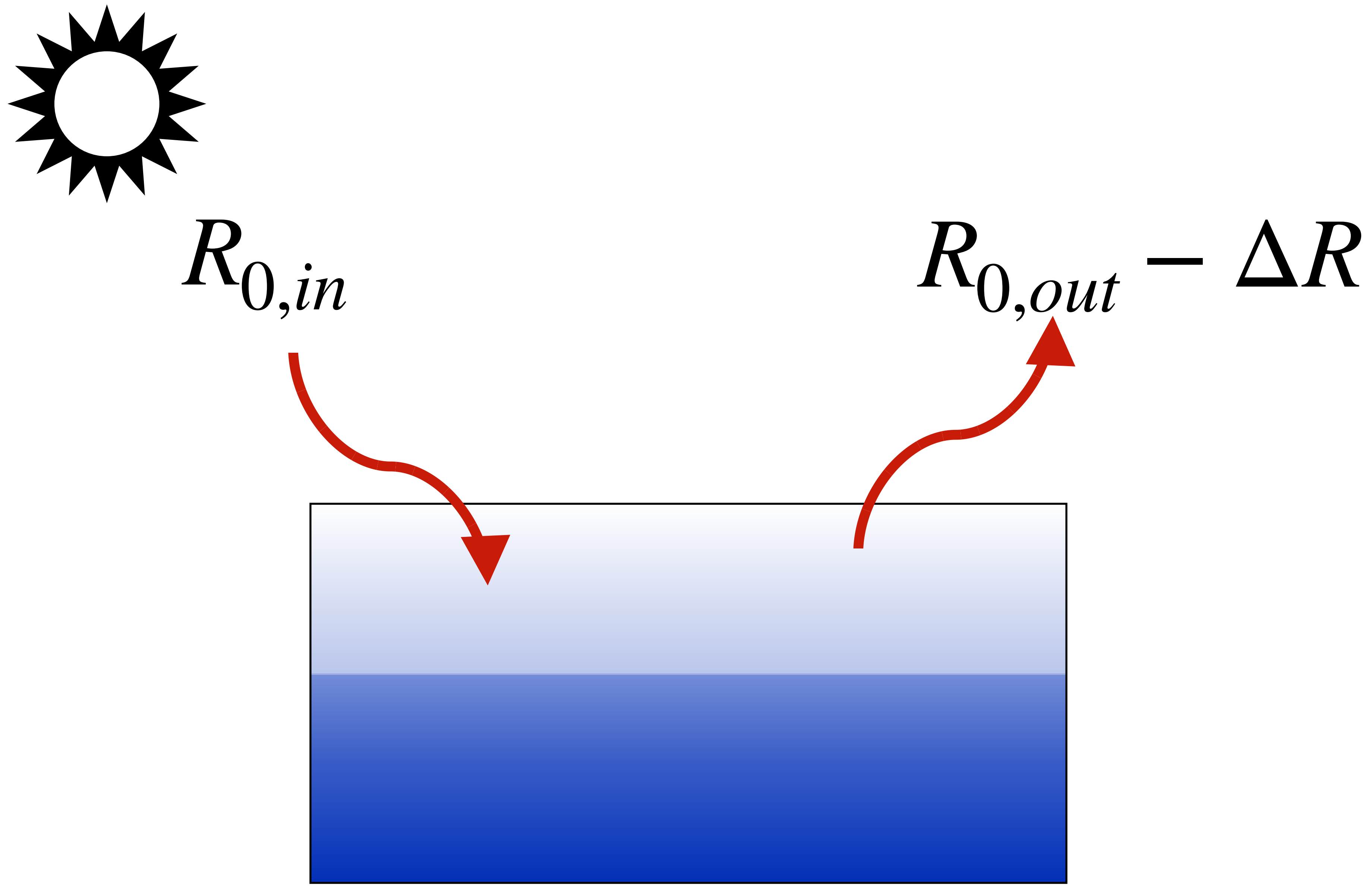
Earth's Energy Budget



Constant temperature (equilibrium)
Requires balanced energy budget

$$R_{in} + R_{out} = 0$$

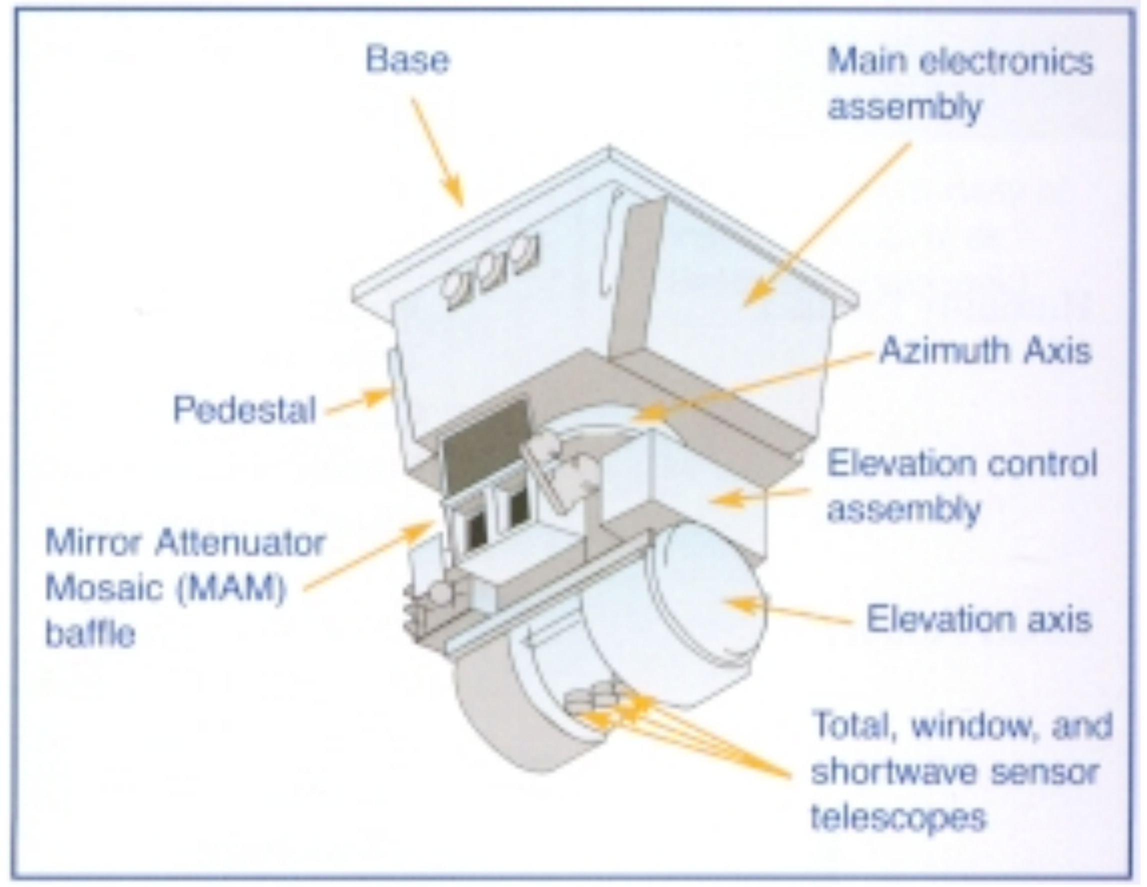
Earth's Energy Budget



Global warming:

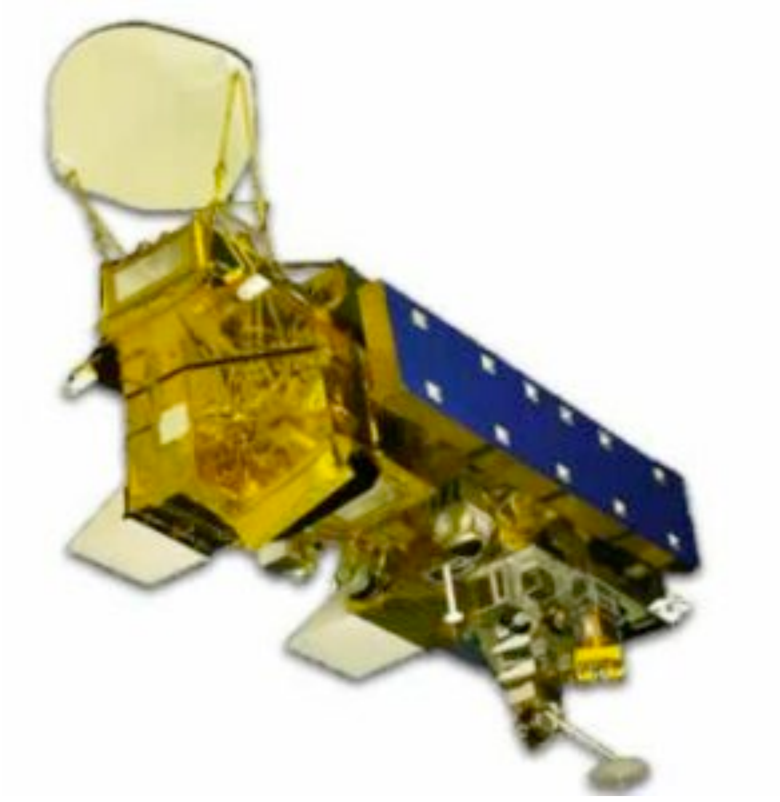
$$R_{in} + R_{out} = \Delta R > 0$$

Earth's Energy Imbalance



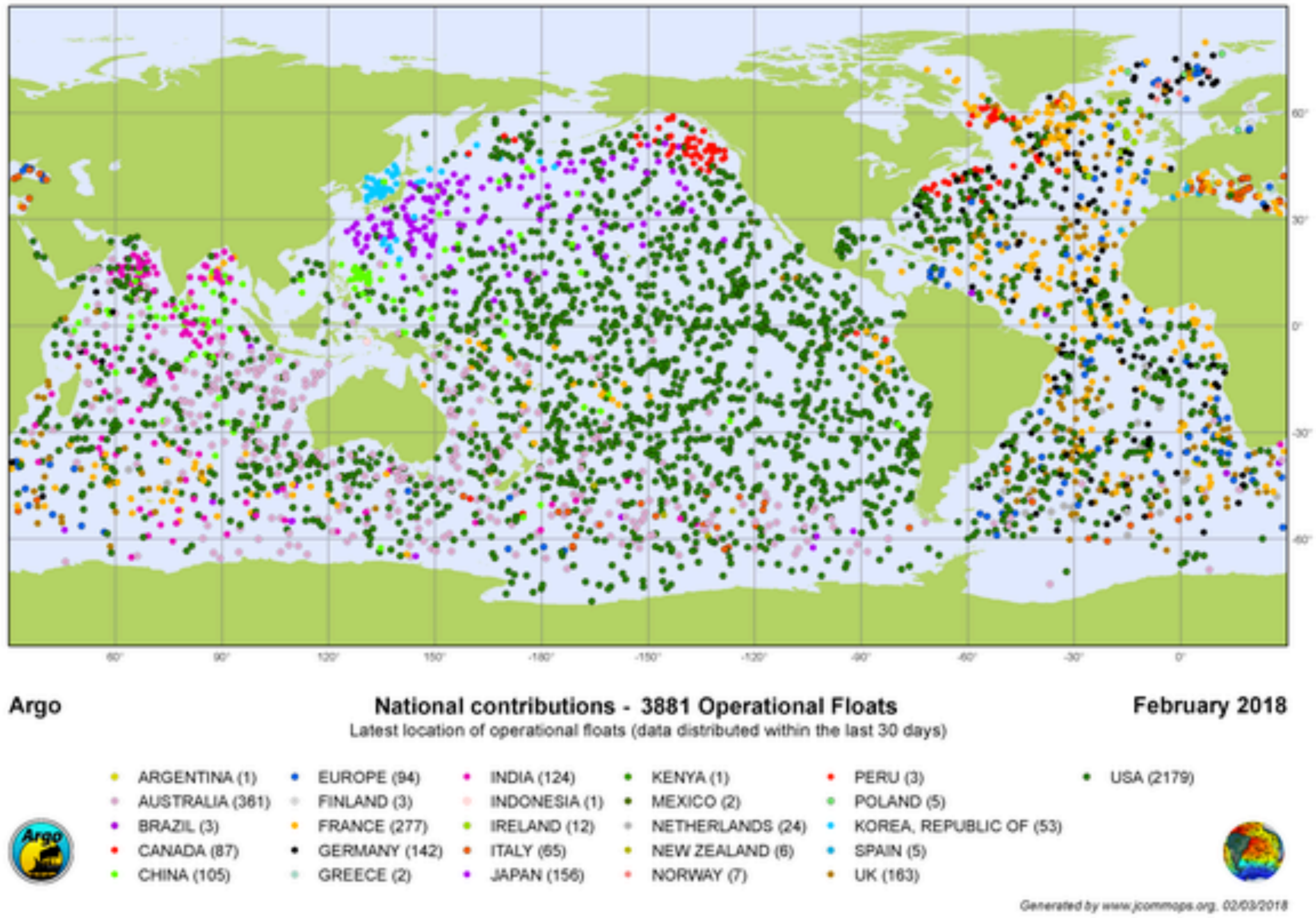
CERES instrument

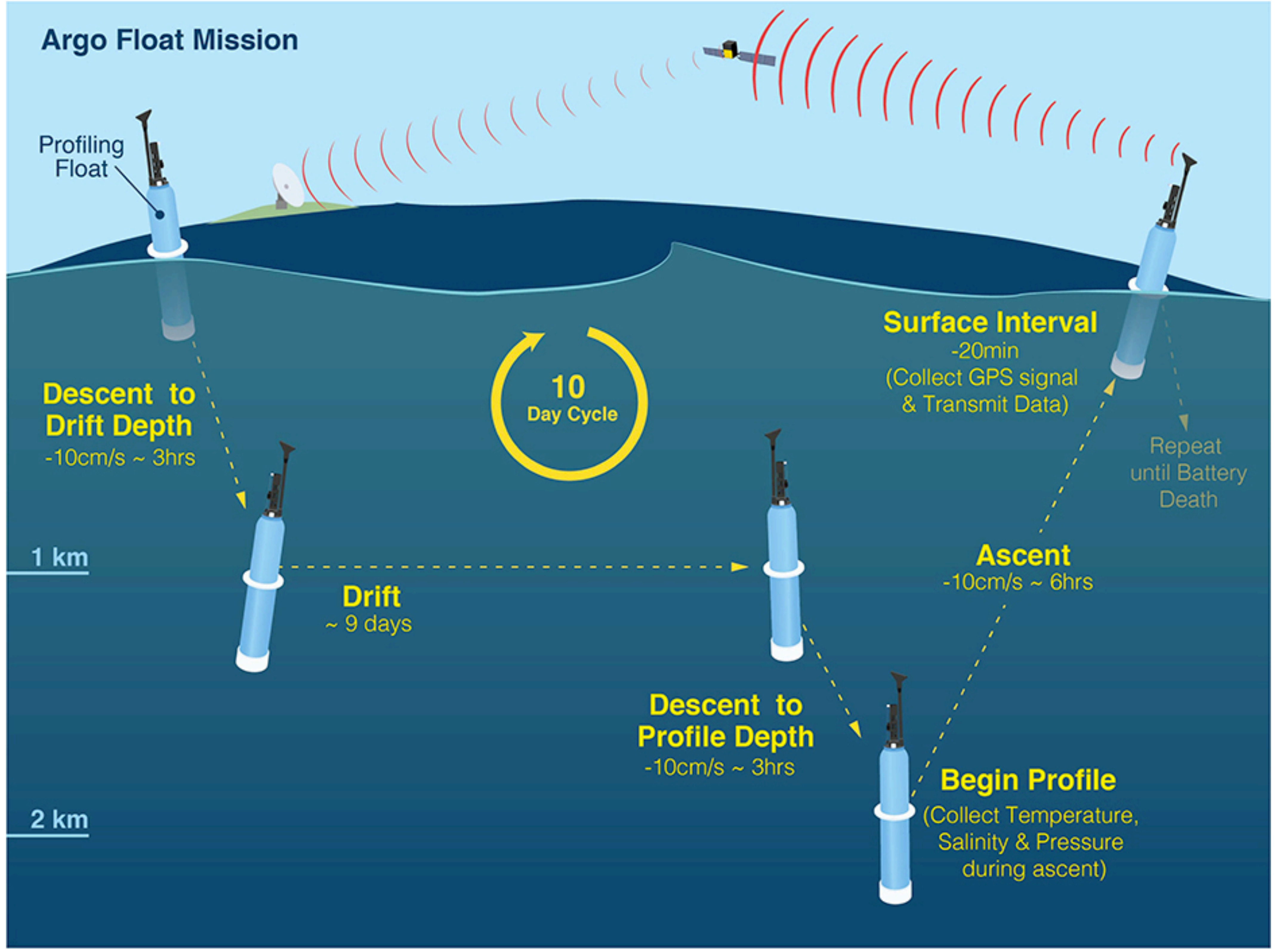
Terra satellite



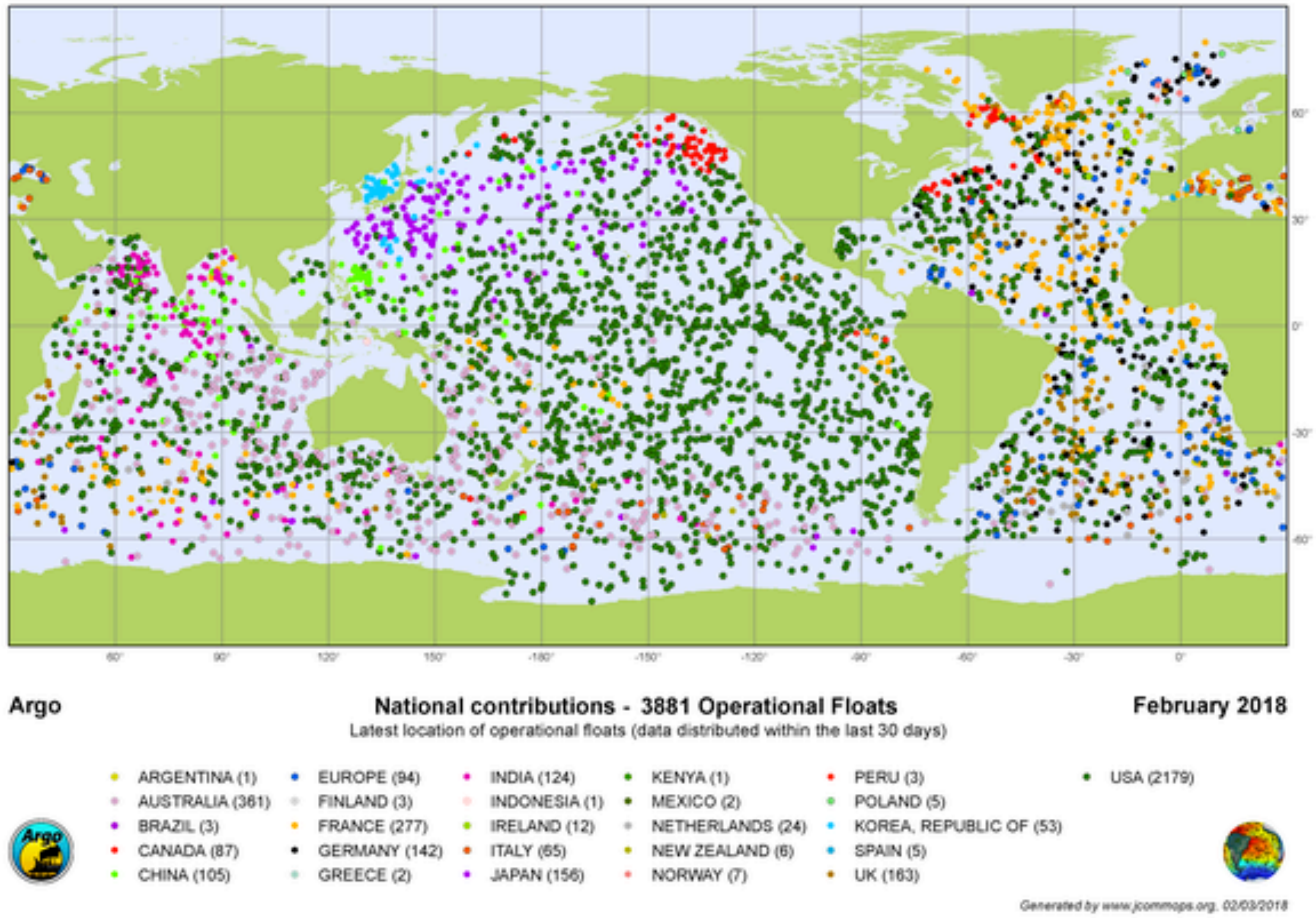
Aqua satellite

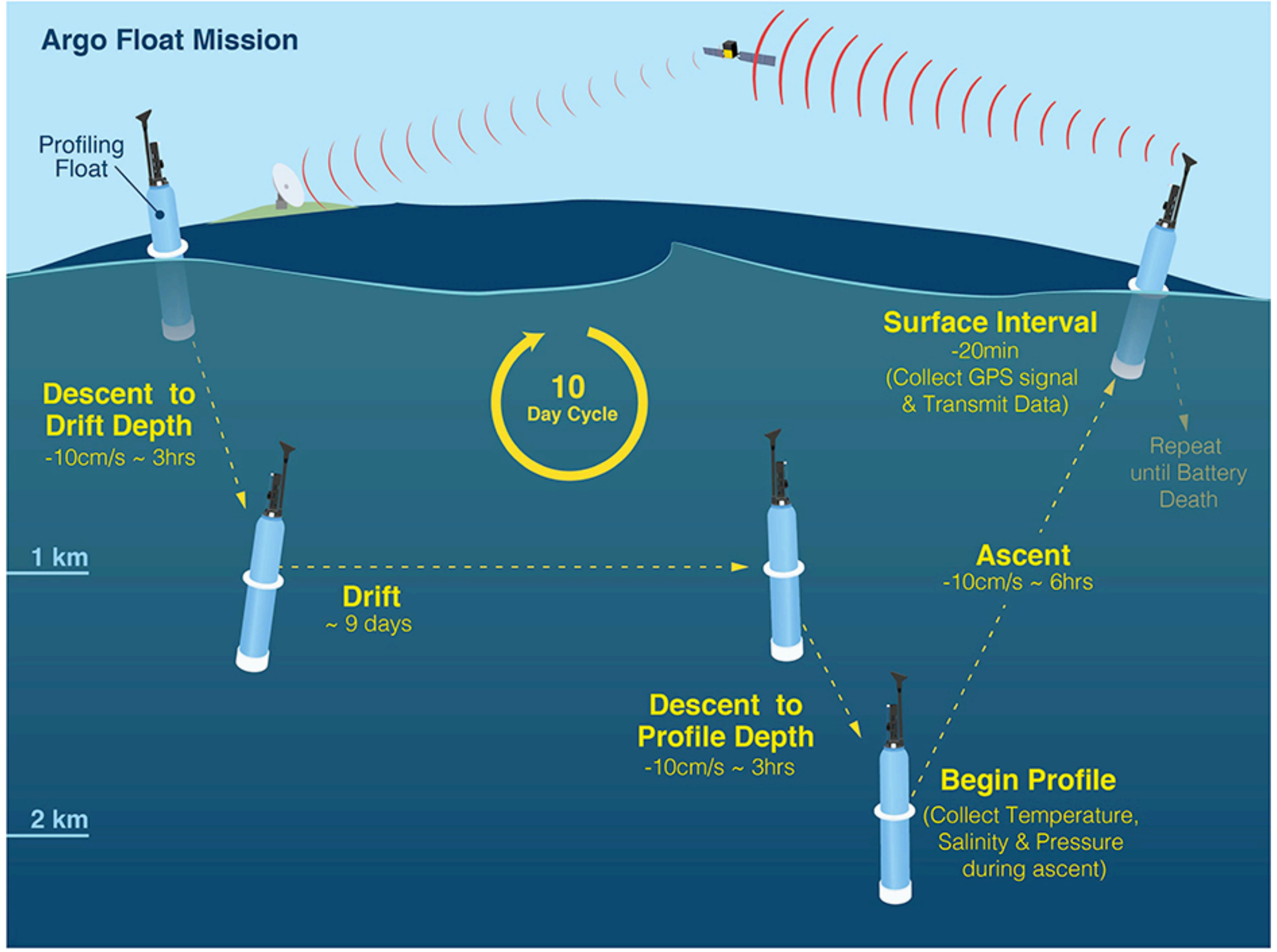
Earth's Energy Imbalance



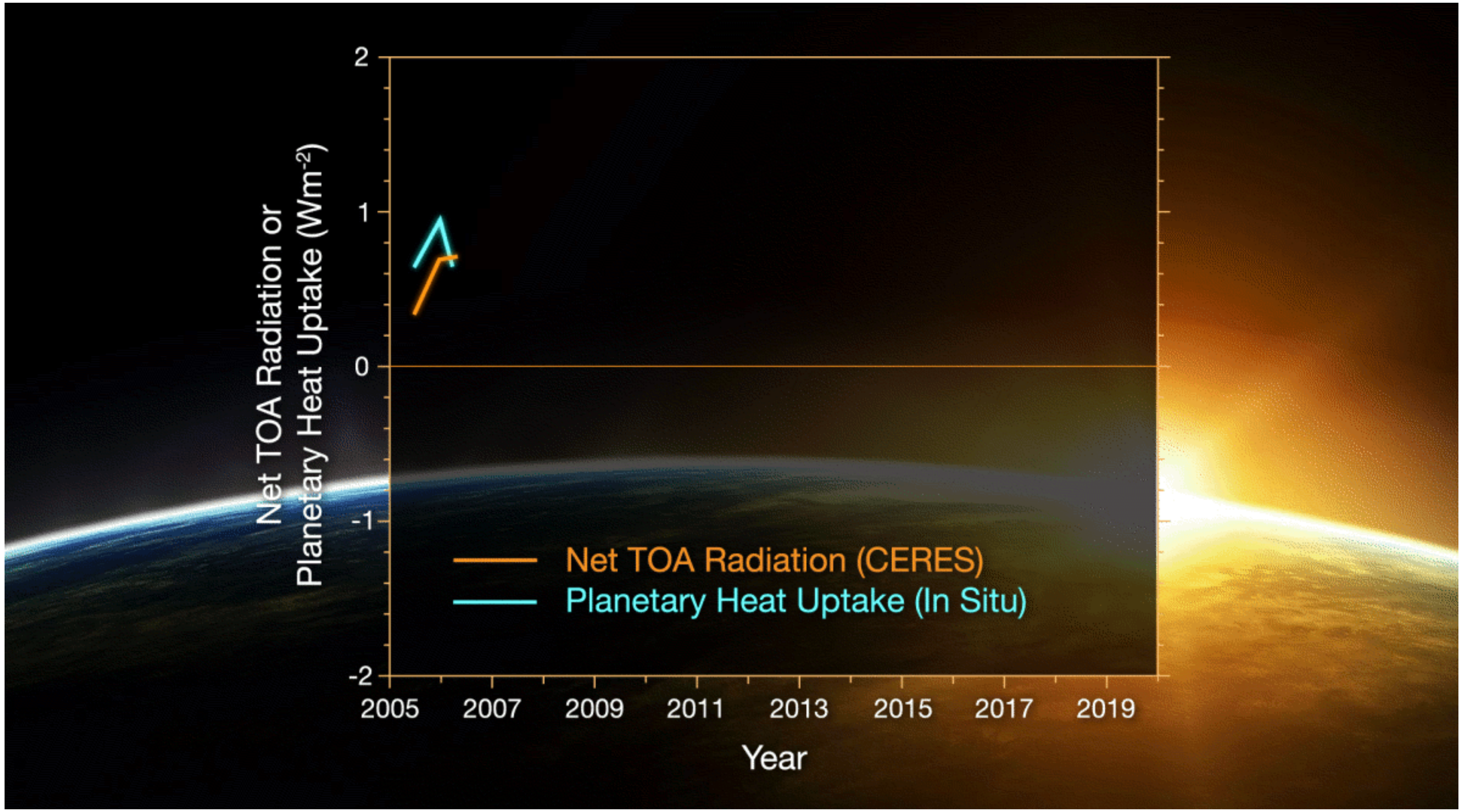


Earth's Energy Imbalance

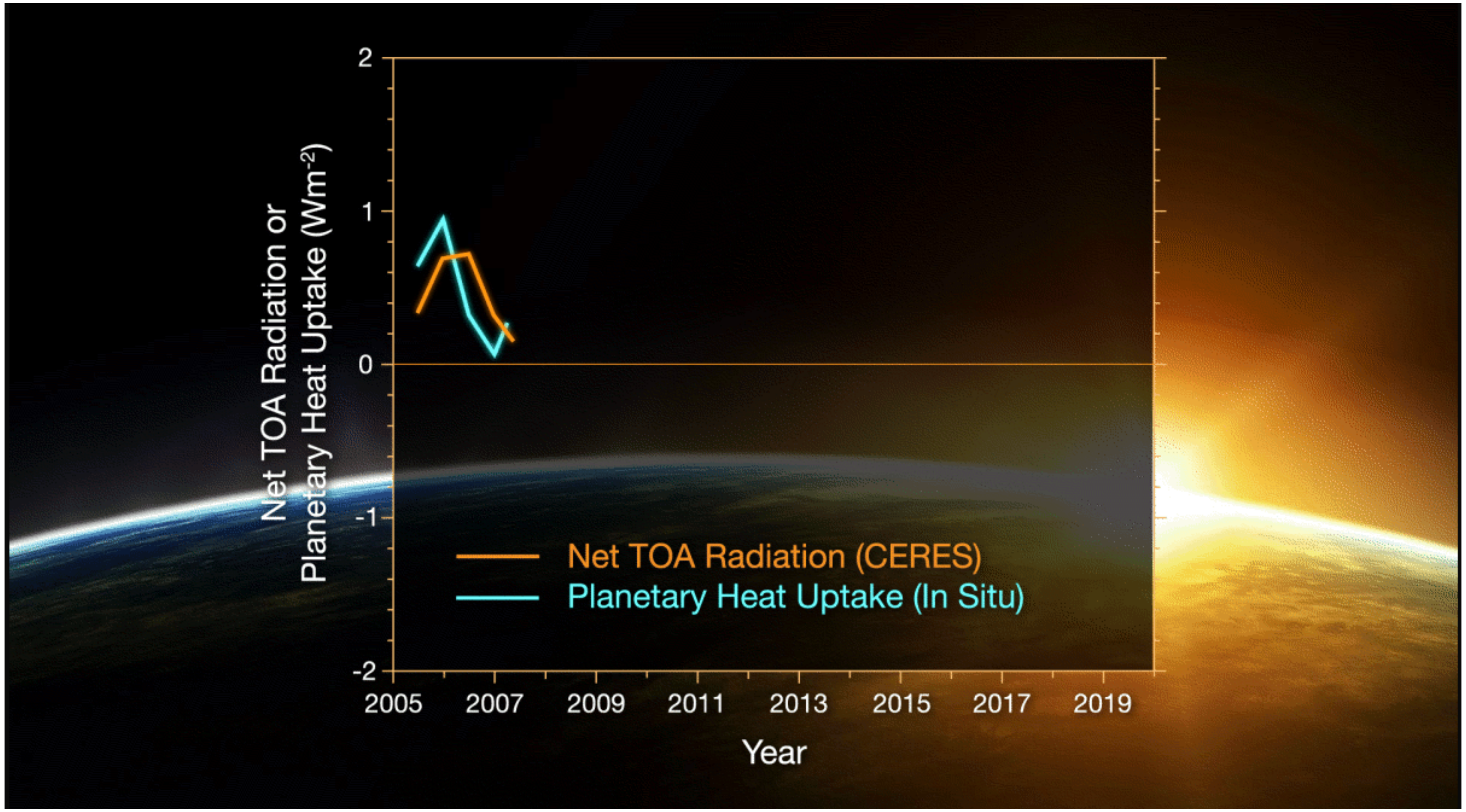




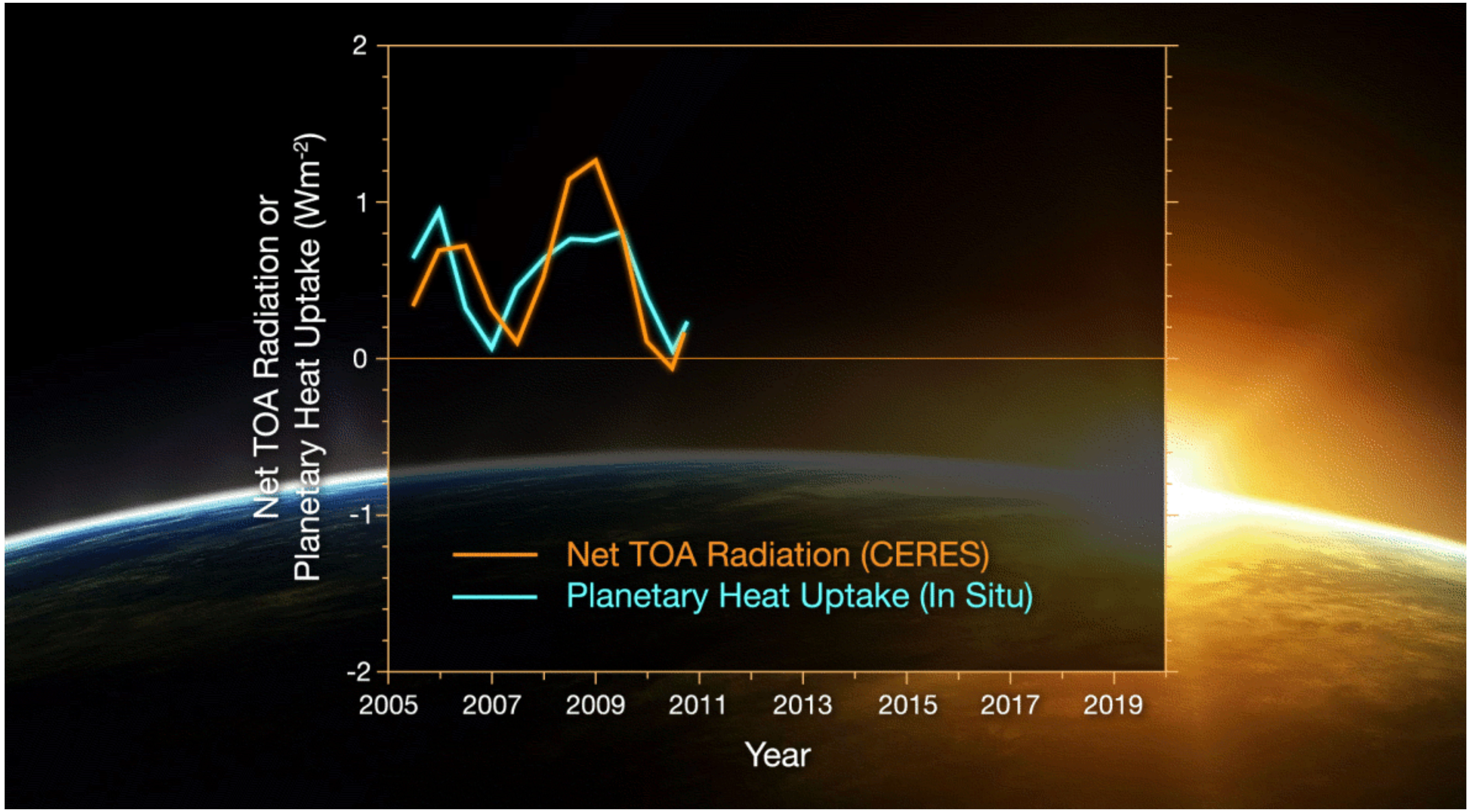
Earth's Energy Imbalance



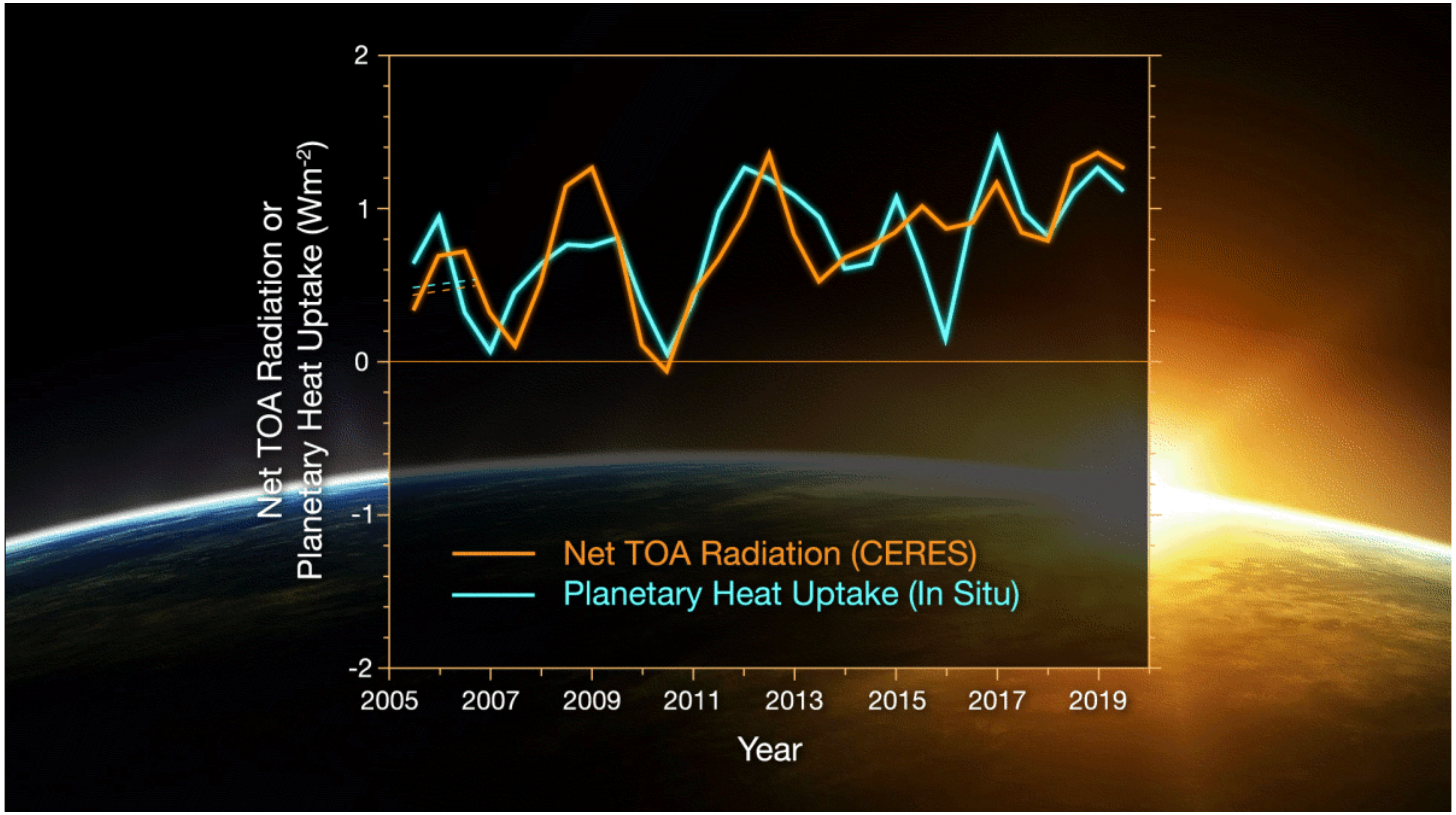
Earth's Energy Imbalance



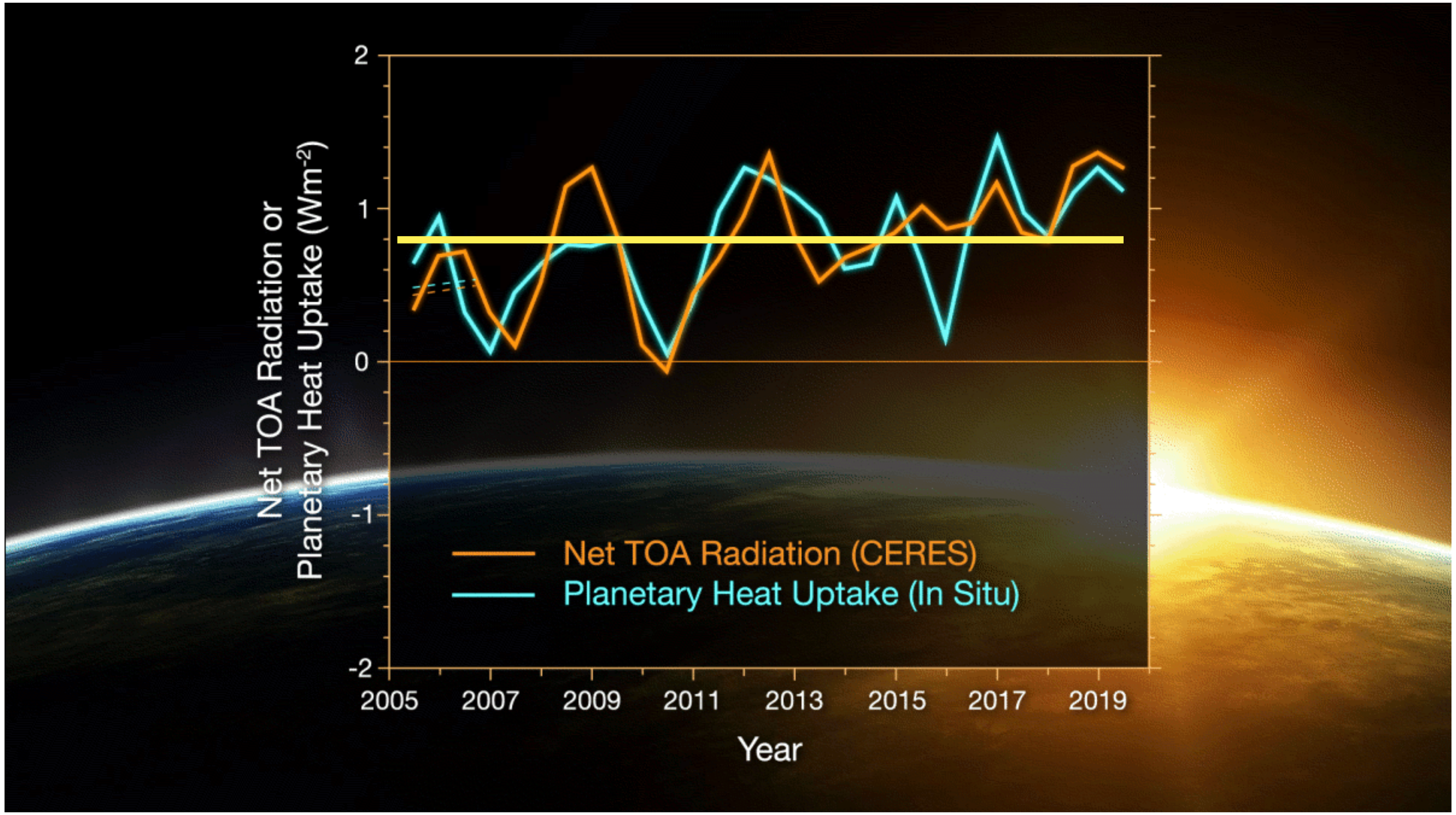
Earth's Energy Imbalance



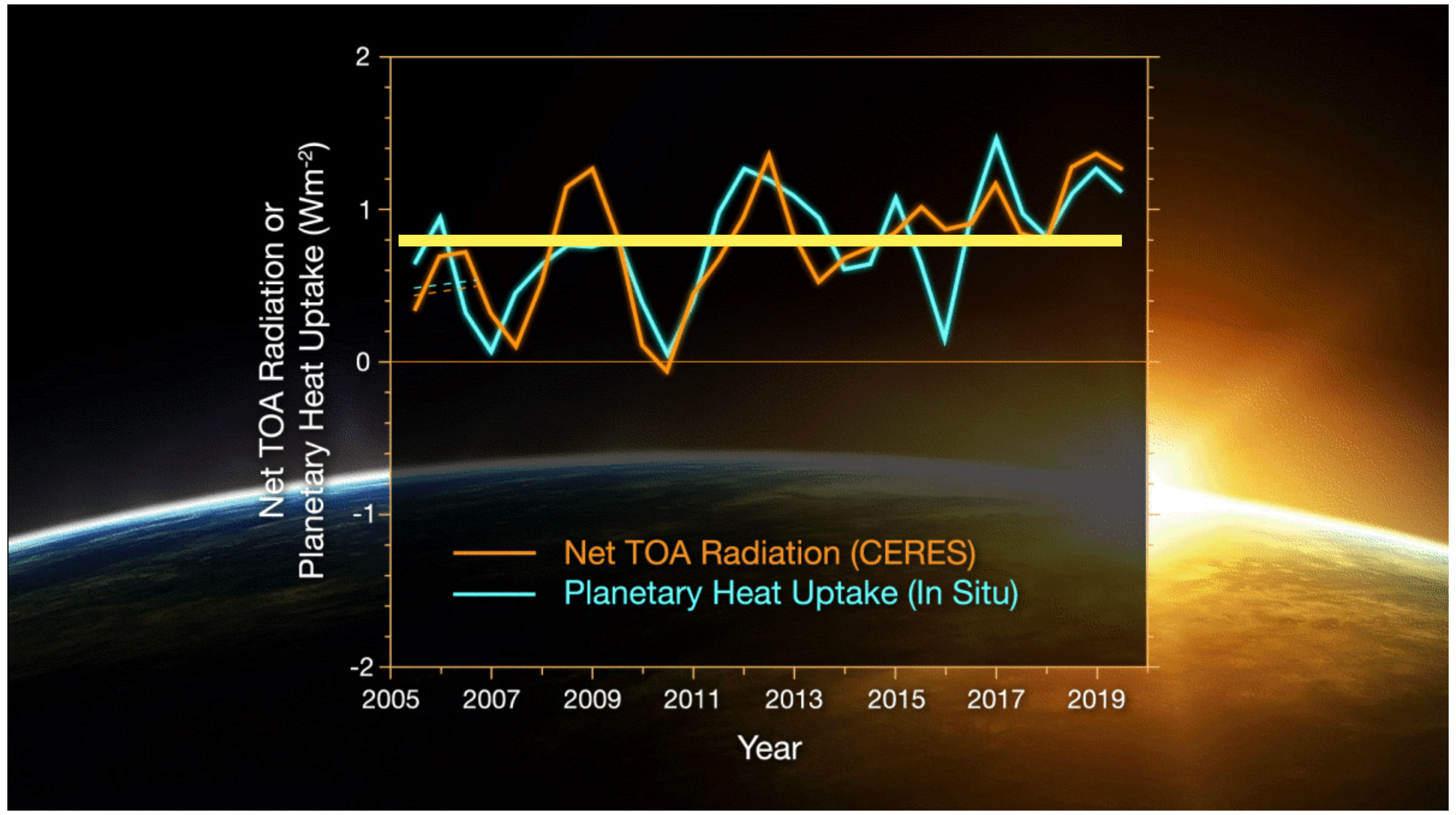
Earth's Energy Imbalance



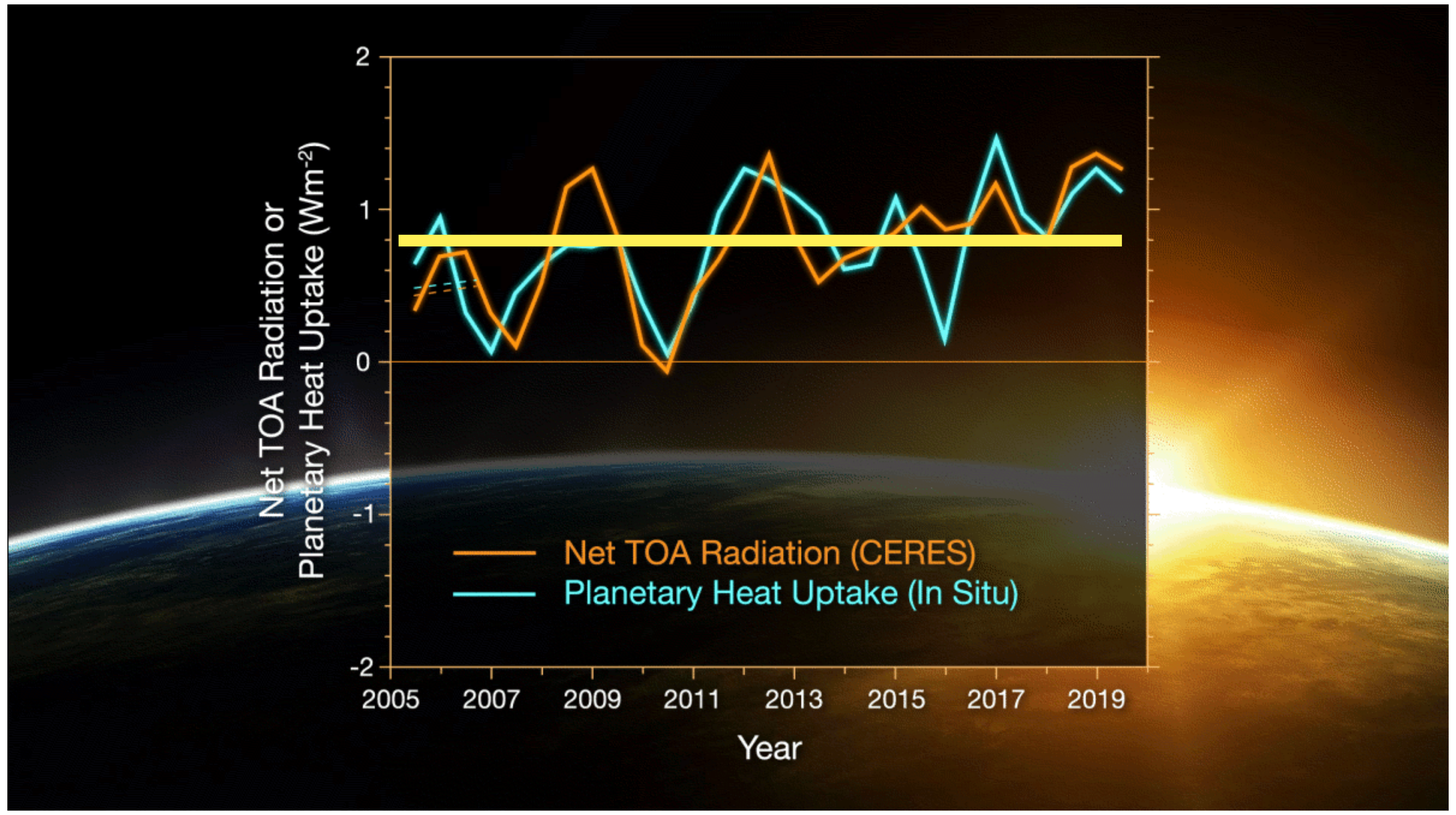
Earth's Energy Imbalance



Model the data!



Model the data!



A probabilistic model for the measurements:

$$X = \mu + \varepsilon$$

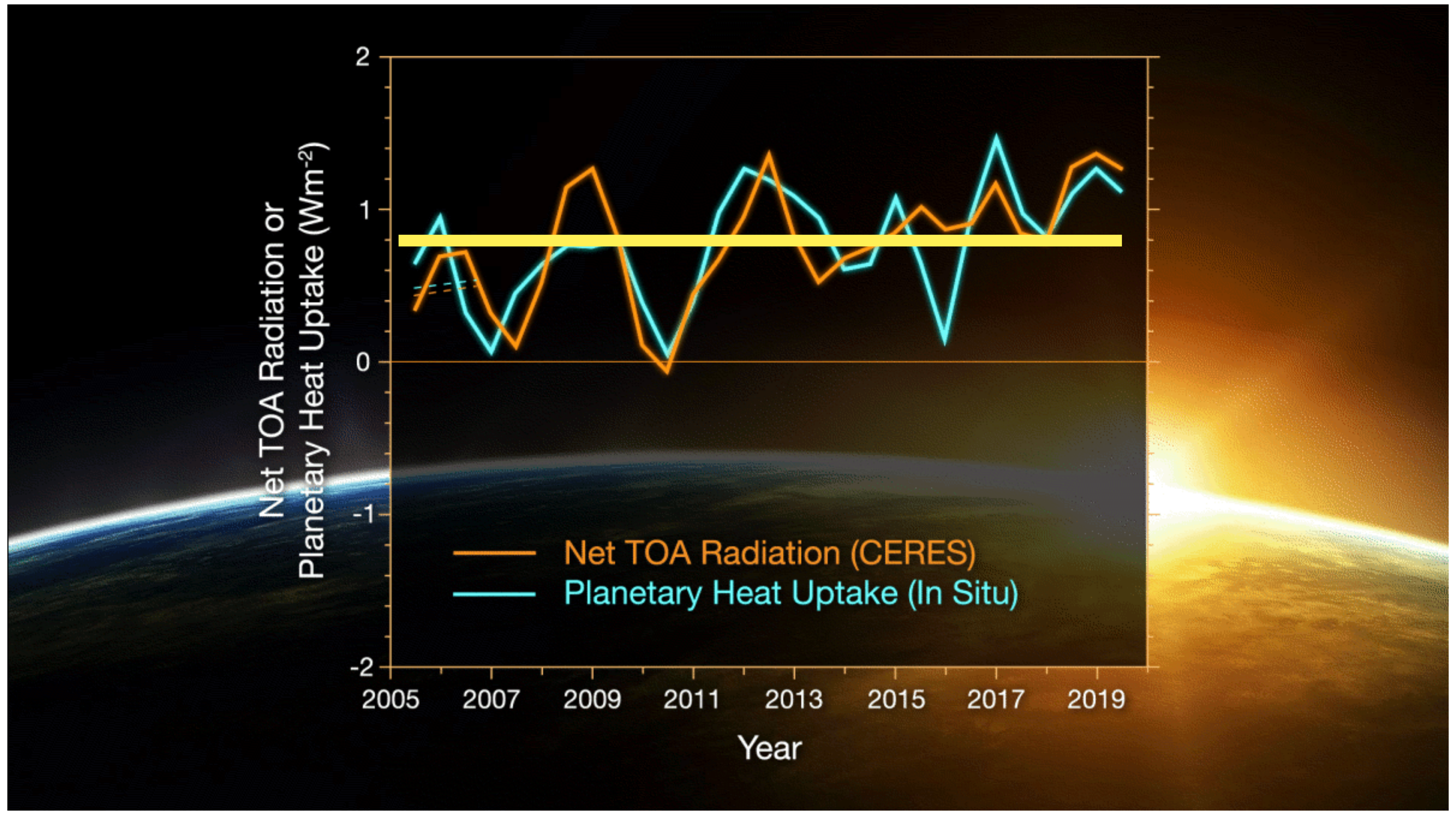
Quantity of interest (climate change component)

$$\mu$$

Random errors (weather, instrument error):

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

Estimate model parameters



A probabilistic model for the measurements:

$$X = \mu + \varepsilon$$

Quantity of interest (climate change component)

$$\mu$$

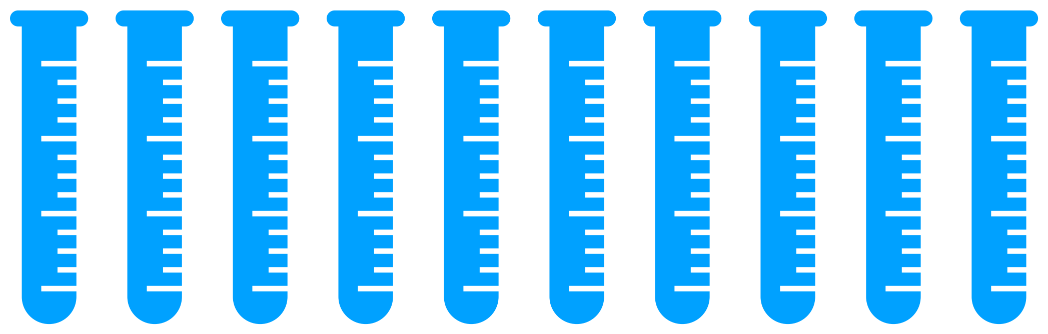
Random errors (weather, instrument error):

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

Estimator: Sample mean

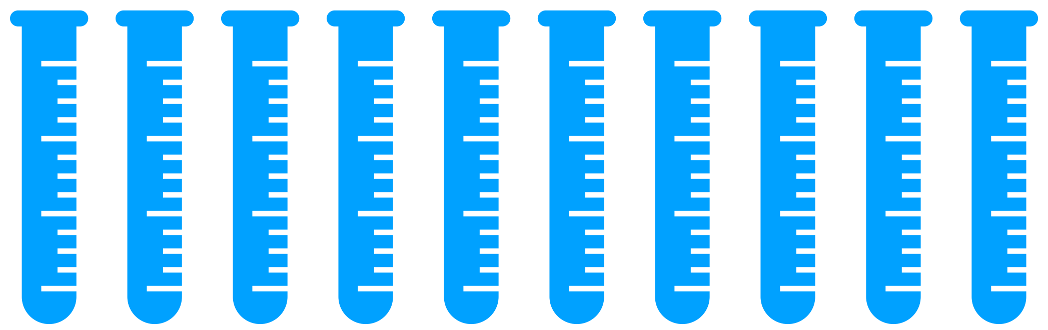
$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

A sample



N measurements

A sample



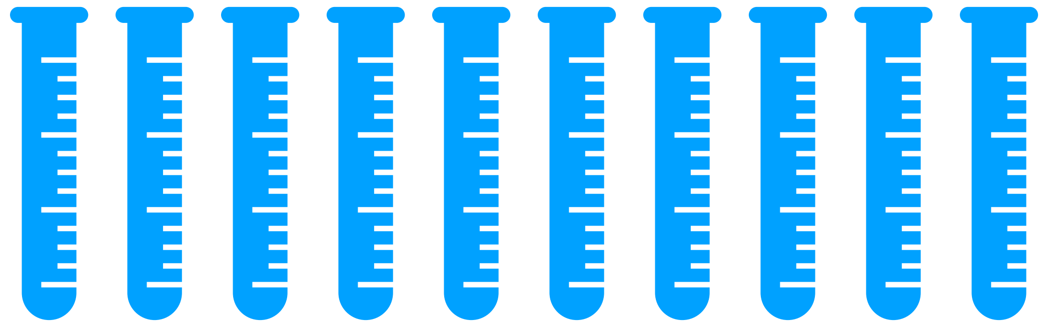
$x_1 \quad x_2 \qquad\qquad\qquad x_n$

N measurements

Sample mean

$$z = \bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Many samples

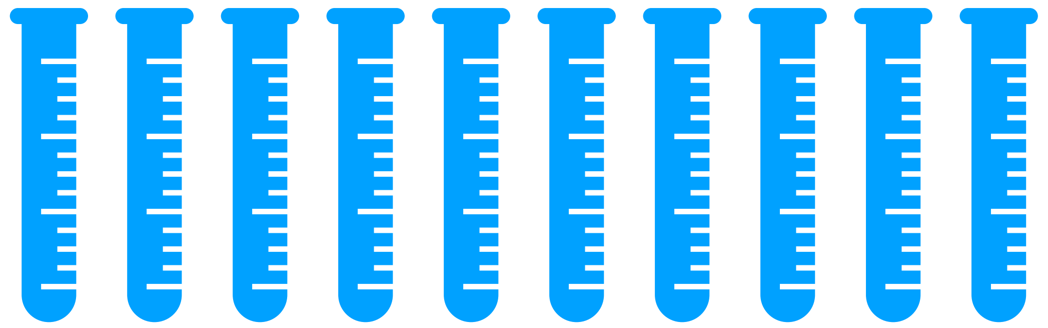


N measurements

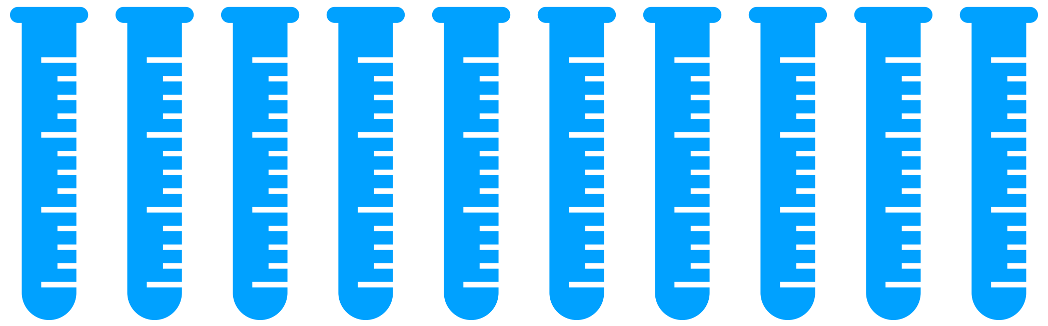
x_{11} x_{21} x_{n1}

Sample mean

$$z_1 = \bar{x}_{n1} = \frac{1}{n}\sum_{i=1}^{n} x_{i1}$$

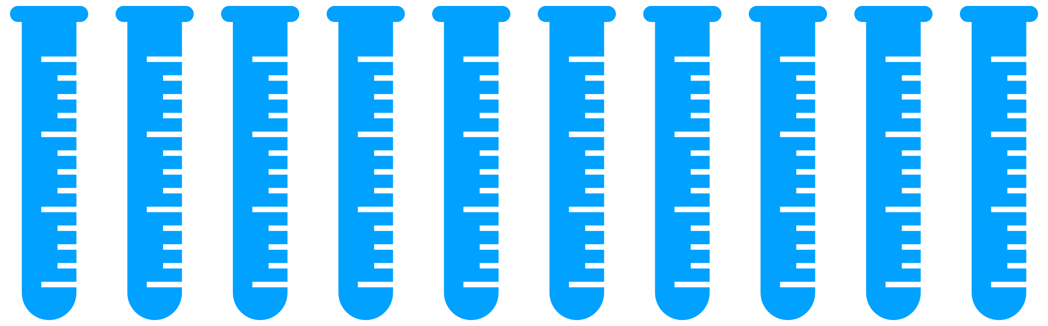


$$z_2 = \bar{x}_{n2} = \frac{1}{n}\sum_{i=1}^{n} x_{i2}$$



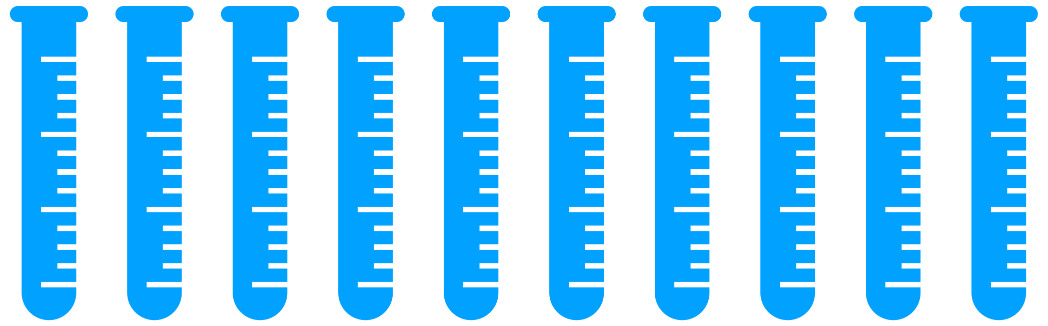
$$z_3 = \bar{x}_{n3} = \frac{1}{n}\sum_{i=1}^{n} x_{i3}$$

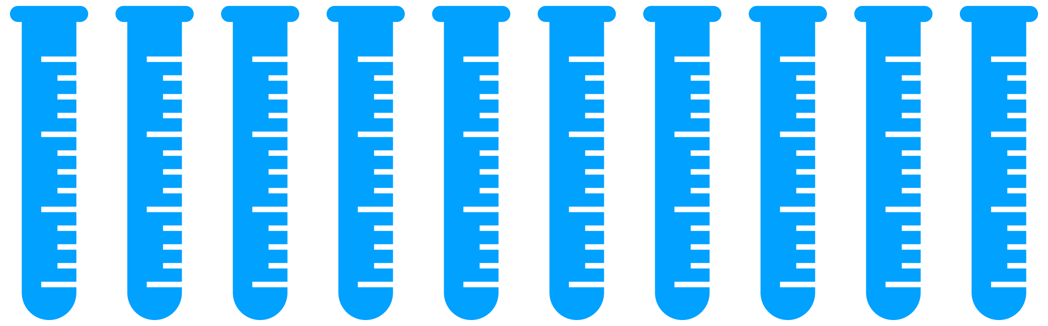
Sample mean is a random variable



X_1 X_2 X_n

N measurements





Sample mean is a random variable

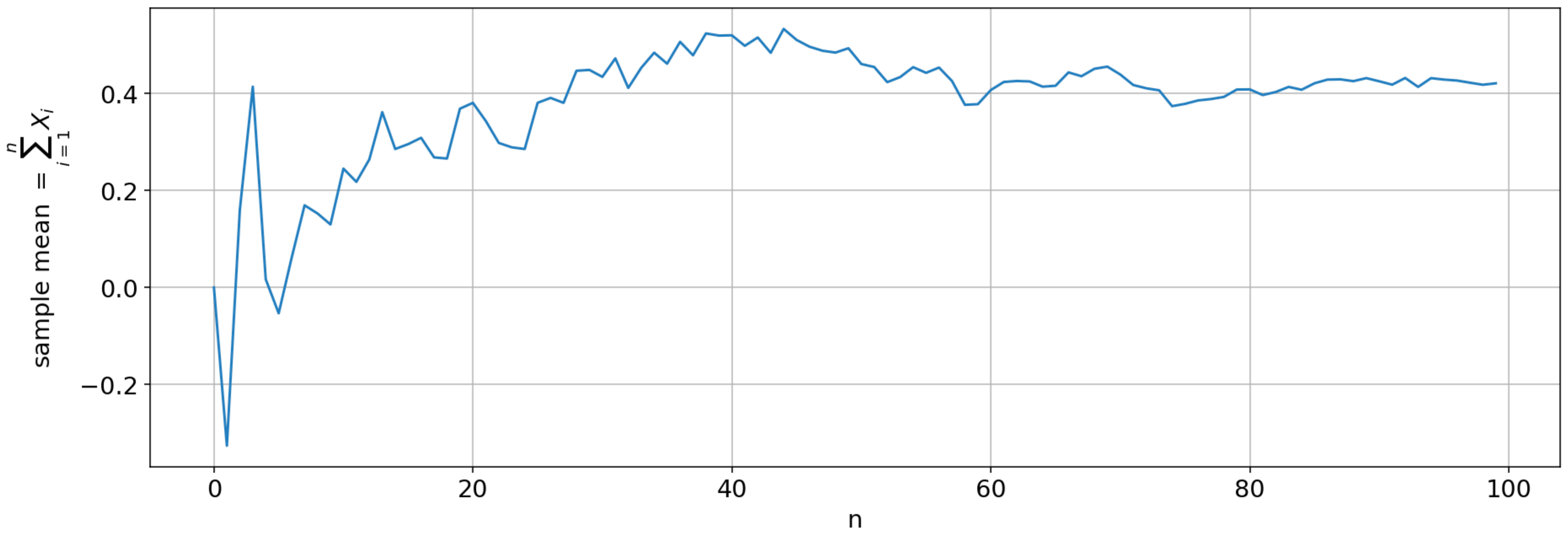
$$Z = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_n$$

$$z_j = \overline{x}_{nj} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

Law of large numbers

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

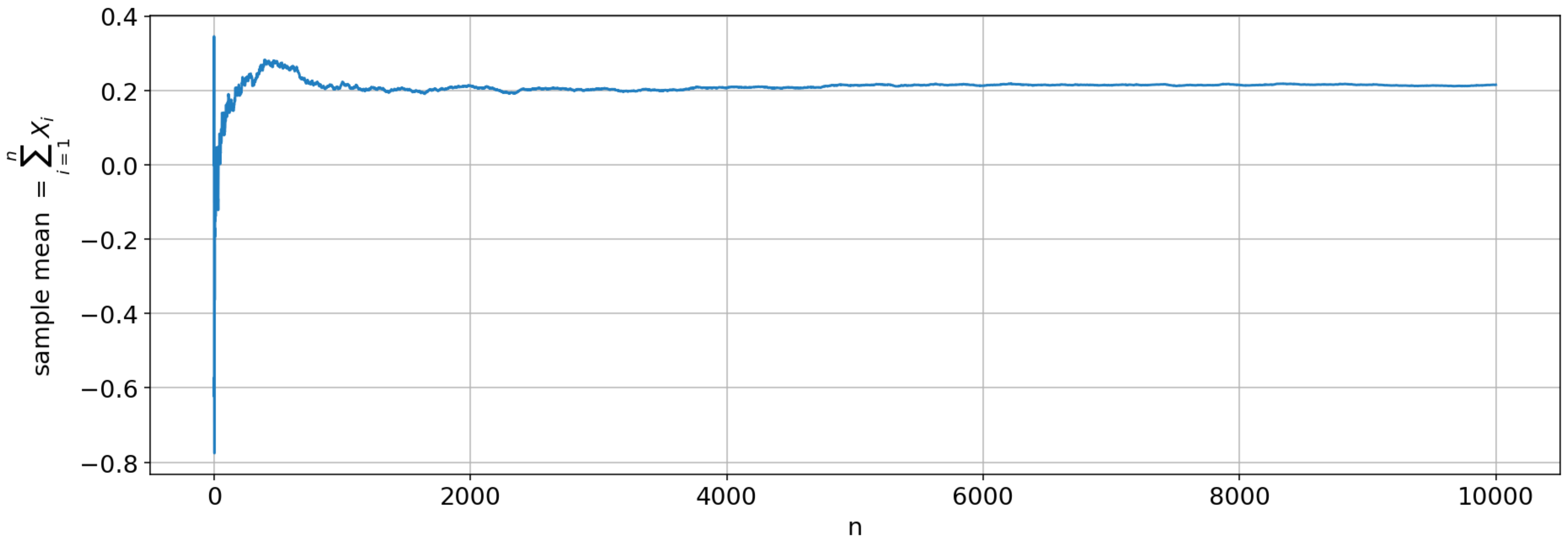
$$X = \mu + \varepsilon \sim \mathcal{N}(\mu, \sigma^2)$$



Law of large numbers

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$X = \mu + \varepsilon \sim \mathcal{N}(\mu, \sigma^2)$$



Central Limit Theorem

What is the distribution of $\overline{X}$?

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$X = \mu + \varepsilon \sim \mathcal{N}(\mu, \sigma^2)$$

Central Limit Theorem

$$\overline{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$$

Central Limit Theorem

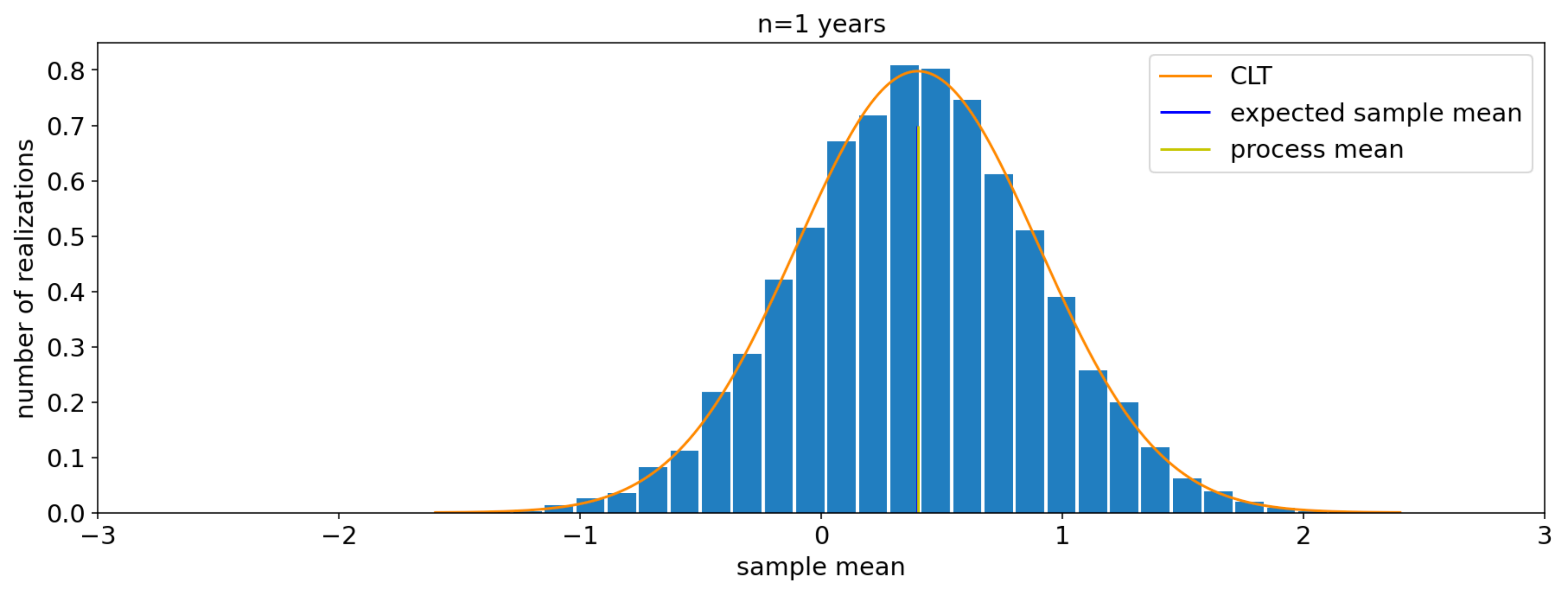
What is the distribution of $\overline{X}$?

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$X = \mu + \varepsilon \sim \mathcal{N}(\mu, \sigma^2)$$

Central Limit Theorem

$$\overline{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$$



Central Limit Theorem

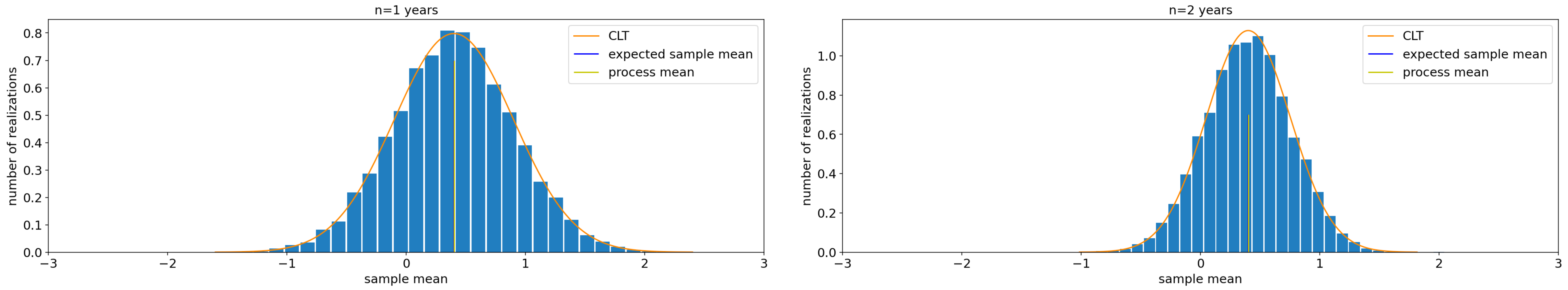
What is the distribution of $\overline{X}$?

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$X = \mu + \varepsilon \sim \mathcal{N}(\mu, \sigma^2)$$

Central Limit Theorem

$$\overline{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$$



Central Limit Theorem

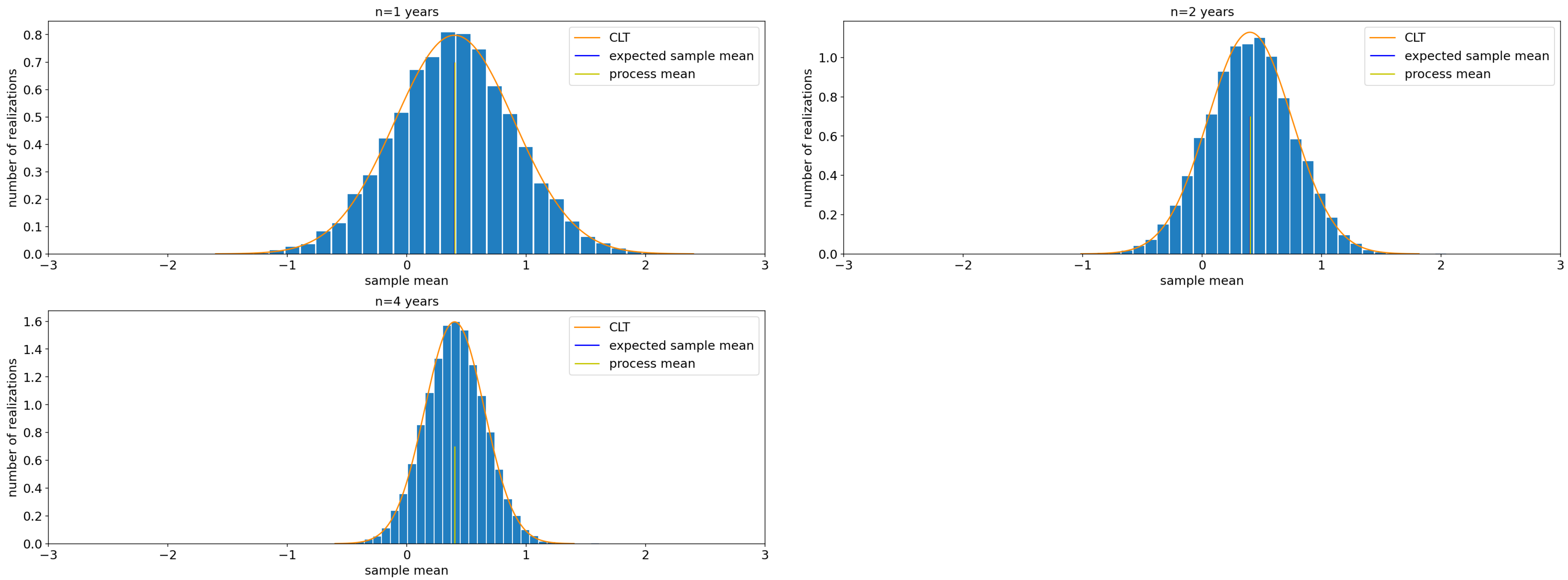
What is the distribution of $\bar{X}$?

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$X = \mu + \varepsilon \sim \mathcal{N}(\mu, \sigma^2)$$

Central Limit Theorem

$$\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$$



Central Limit Theorem

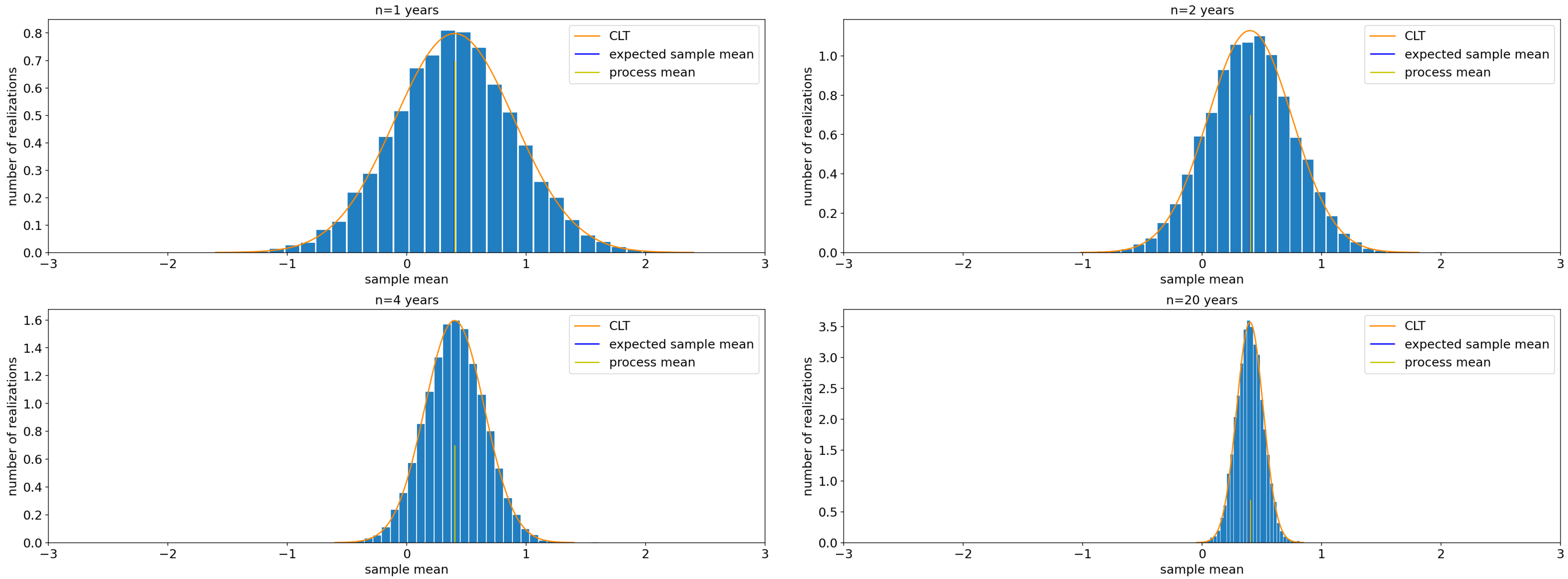
What is the distribution of $\bar{X}$?

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$X = \mu + \varepsilon \sim \mathcal{N}(\mu, \sigma^2)$$

Central Limit Theorem

$$\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$$



Central Limit Theorem

What is the distribution of $\overline{X}$?

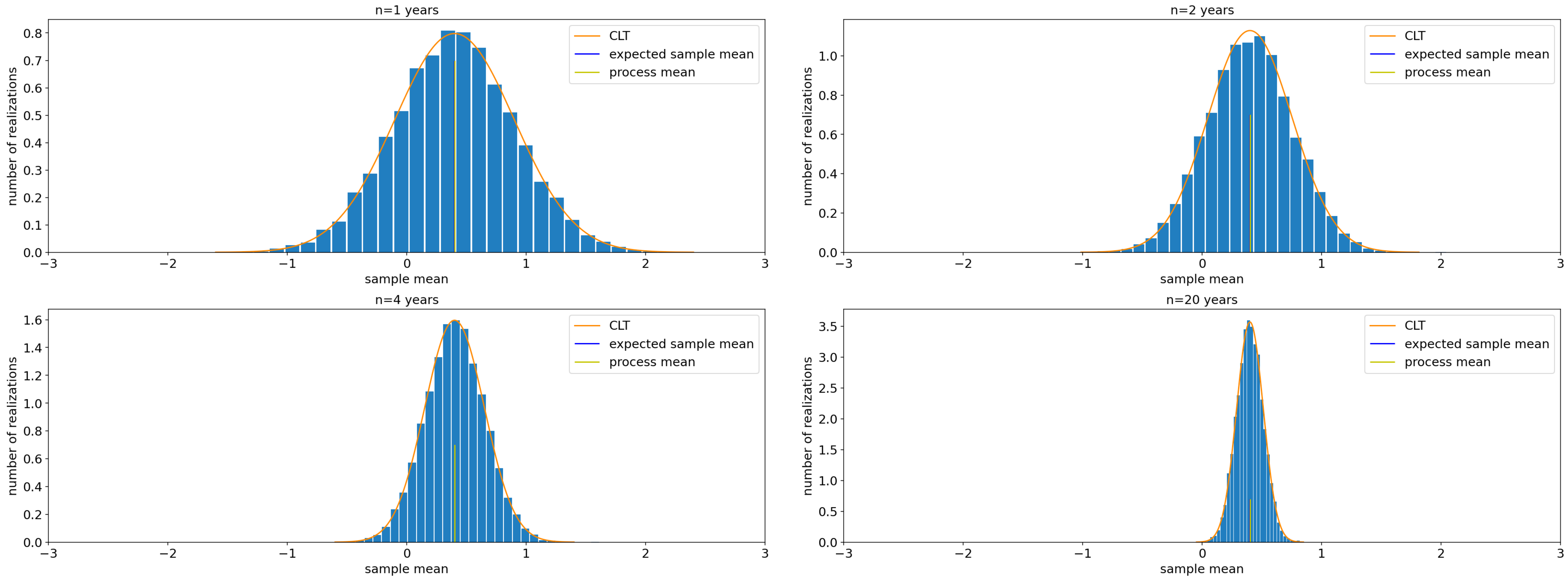
$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$X = \mu + \varepsilon \sim \mathcal{N}(\mu, \sigma^2)$$

Central Limit Theorem

$$\overline{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$$

$$\sigma_{\overline{X}_{20}} = \sigma_X/\sqrt{(20)} \approx \sigma_X/4.5$$



But what is σ?!

What is the distribution of $\overline{X}$?

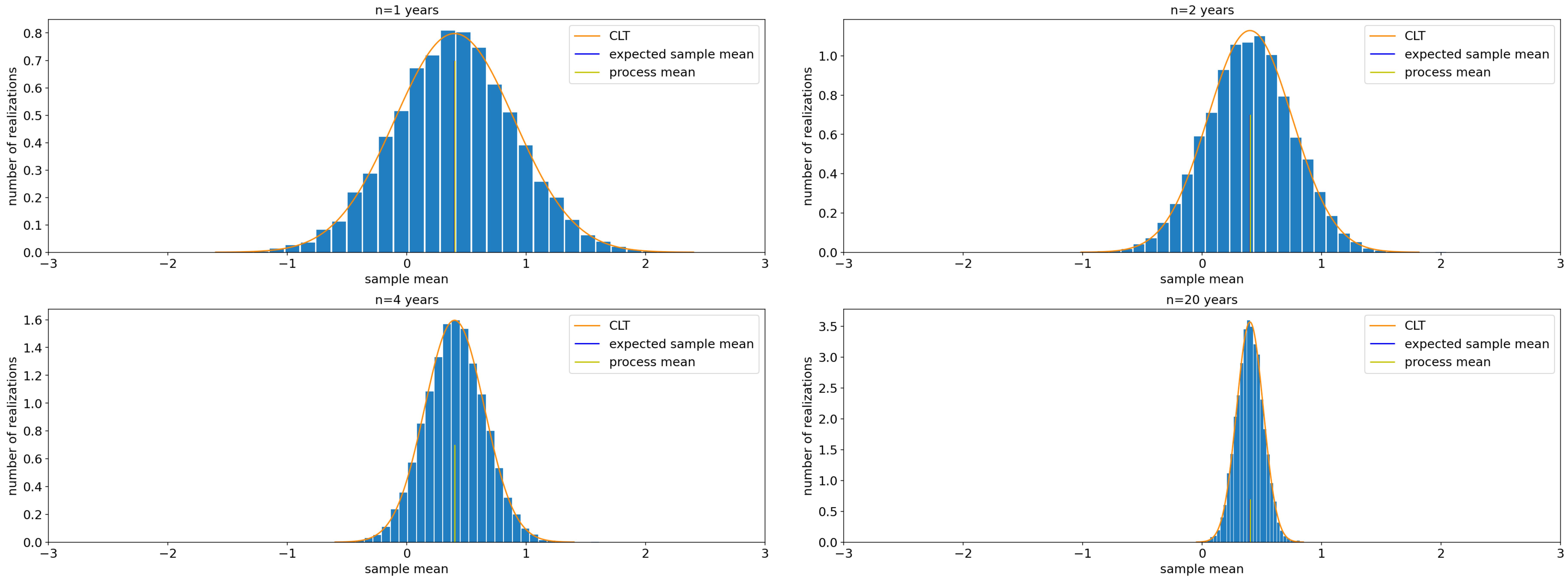
$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$X = \mu + \varepsilon \sim \mathcal{N}(\mu, \sigma^2)$$

Central Limit Theorem

$$\overline{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$$

$$\sigma_{\overline{X}_{20}} = \sigma_X/\sqrt{(20)} \approx \sigma_X/4.5$$



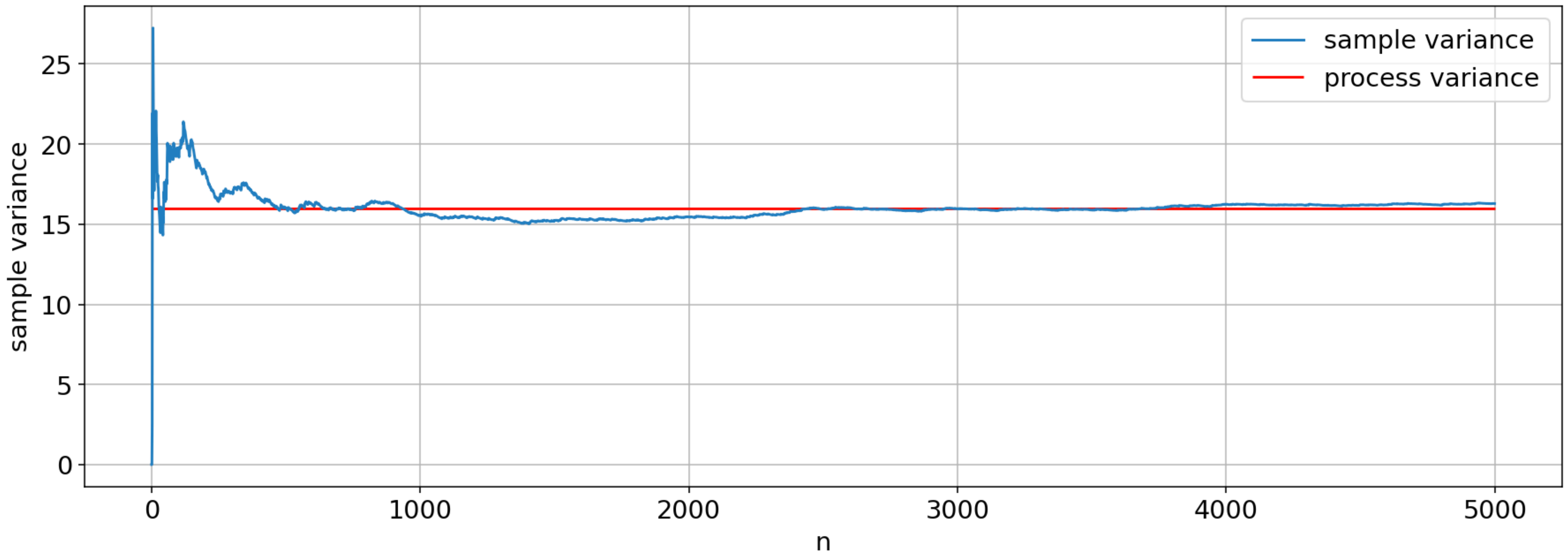
Sample Variance

$$s_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

Sample Variance

$$s_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

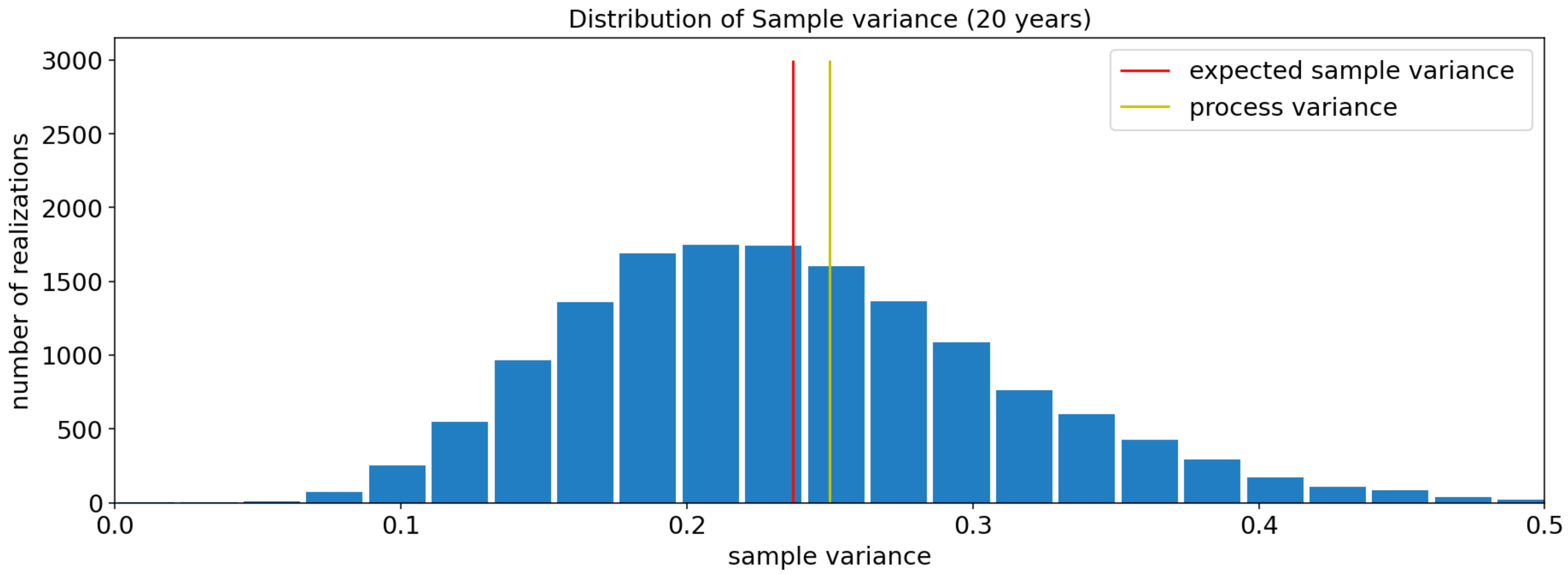
Consistent



Sample Variance

$$s_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

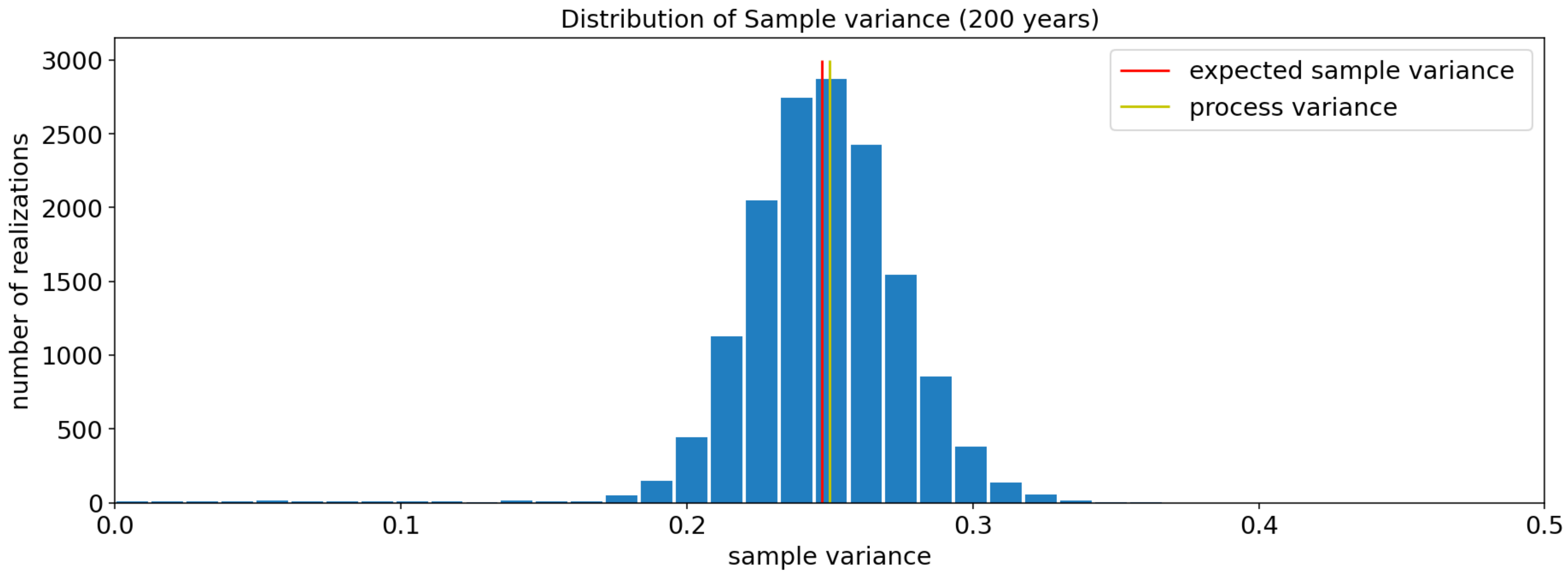
Biased



Sample Variance

$$s_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

Biased

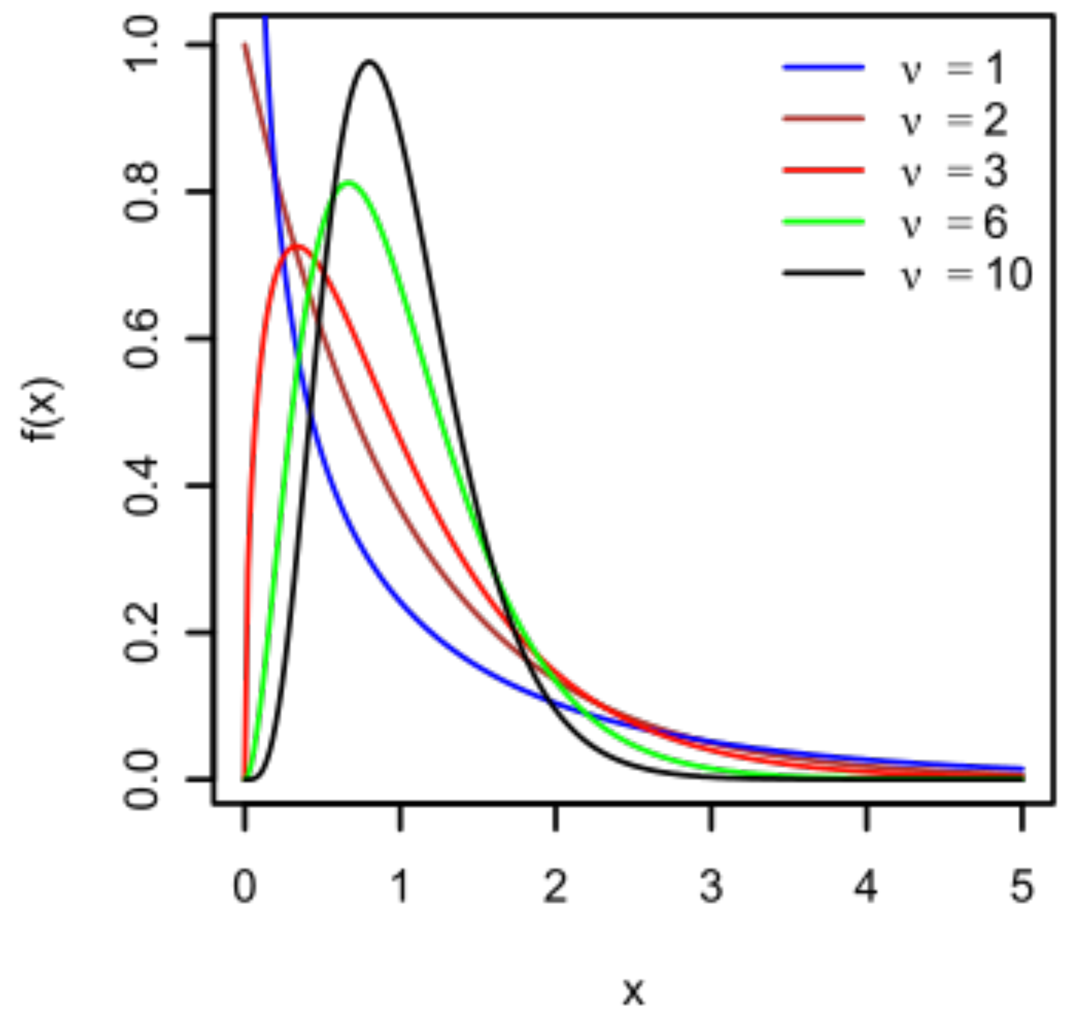


Sample Distributions:

Sample mean: $\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma_X}{n}\right)$

$$\bar{X}_n \sim \frac{\sigma_X}{\sqrt{(n)}}\mathcal{N}(\mu,1)$$

Sample variance: $s_{n-1} \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}$



Sample mean with unknown variance

$$\bar{X}_n \sim \frac{s_{n-1}}{\sqrt{(n)}}\mathcal{N}(\mu,1/n)$$