Probability and statistical thinking

Biofísica Guatemala 2025

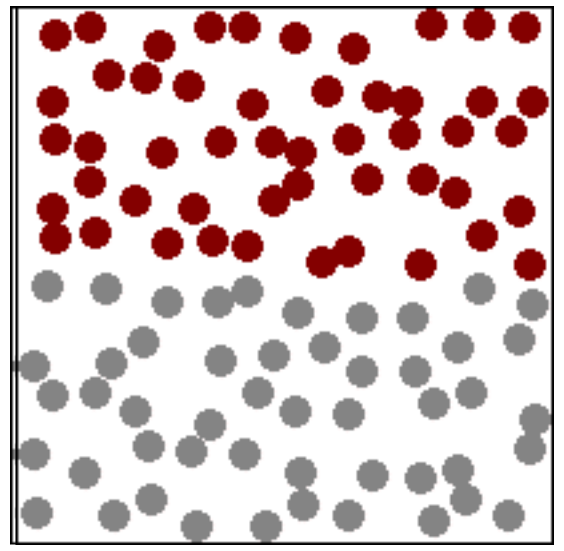
Javier Aguilar

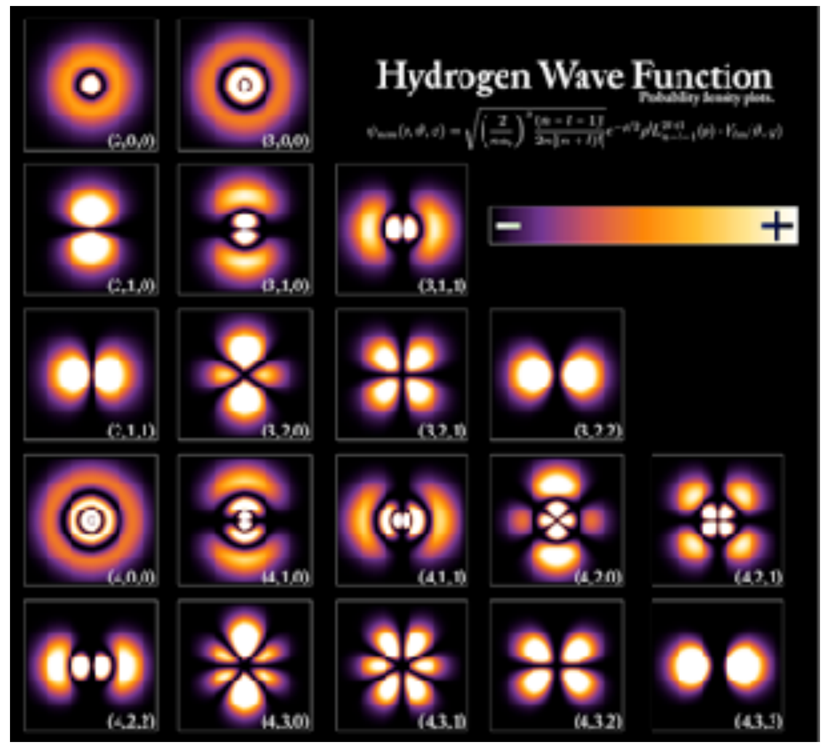
Aim and success

Probability is the logic of uncertainty

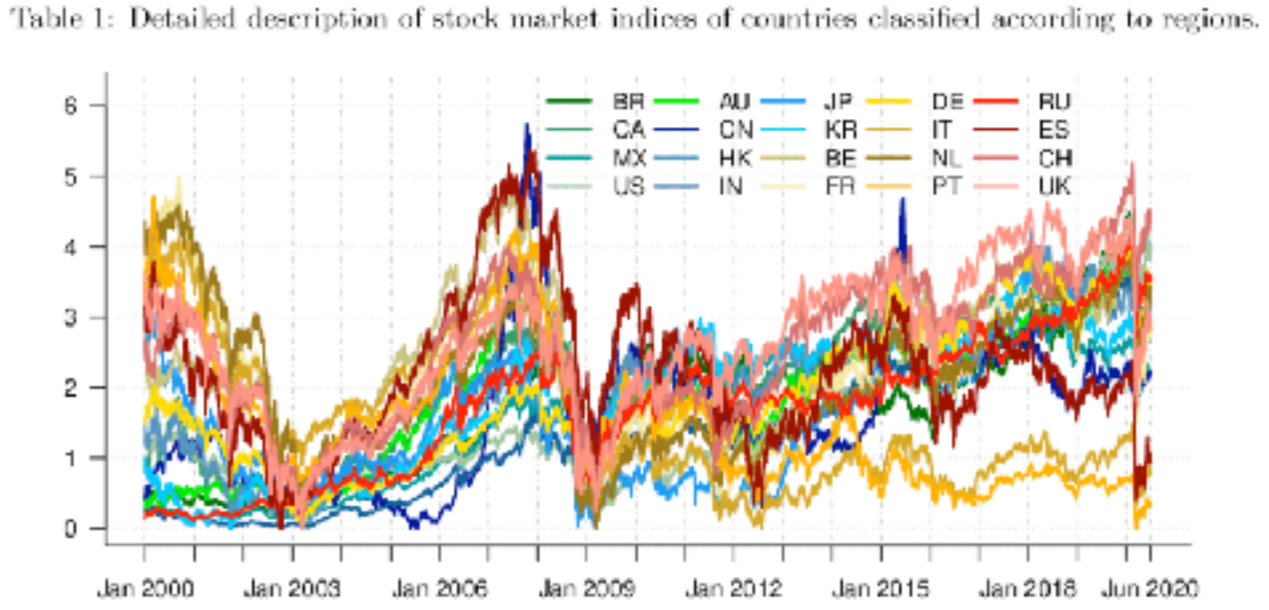
Luck. Coincidence. Randomness. Risk. Doubt. Fortune. Chance.

Physics

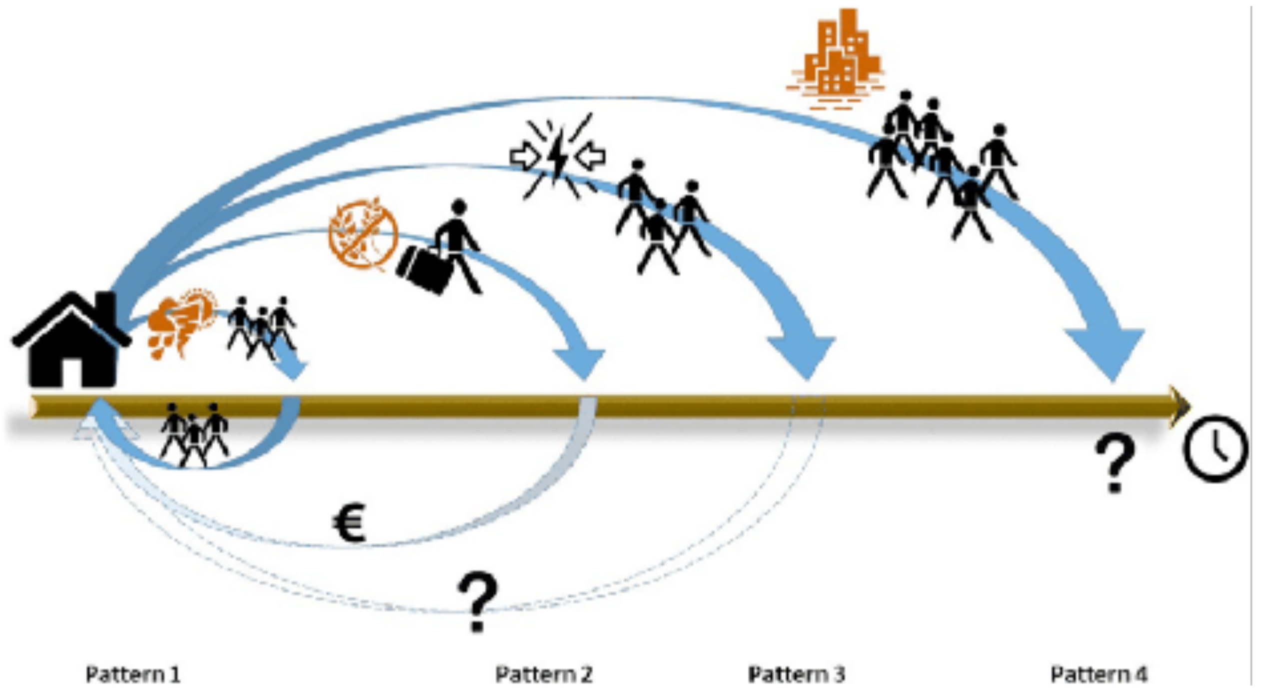




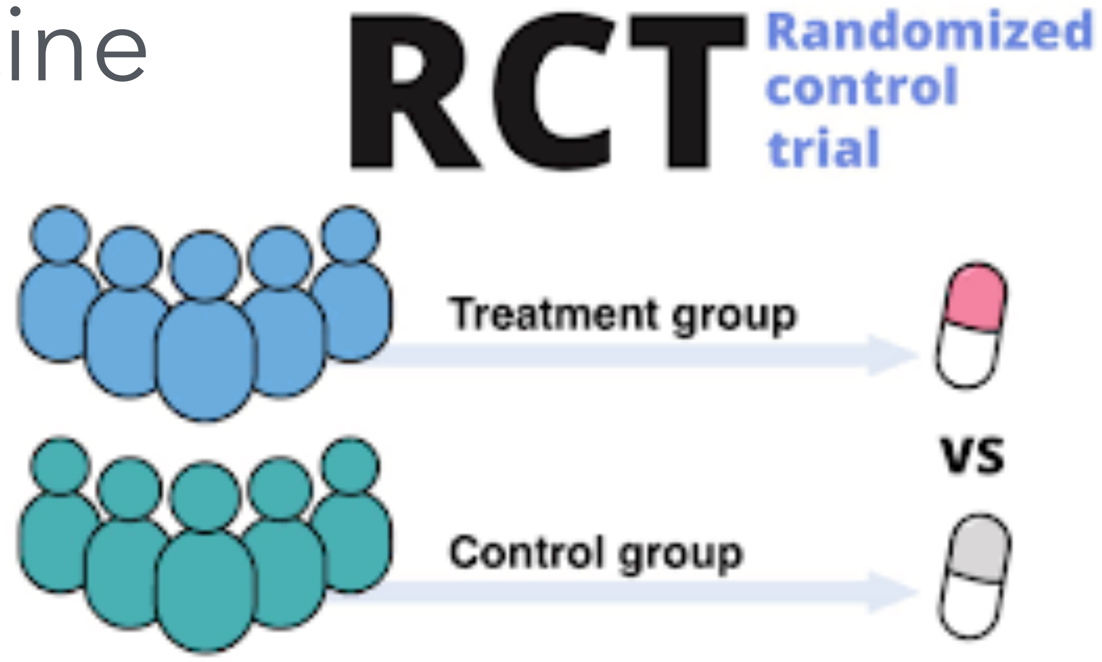
Finance



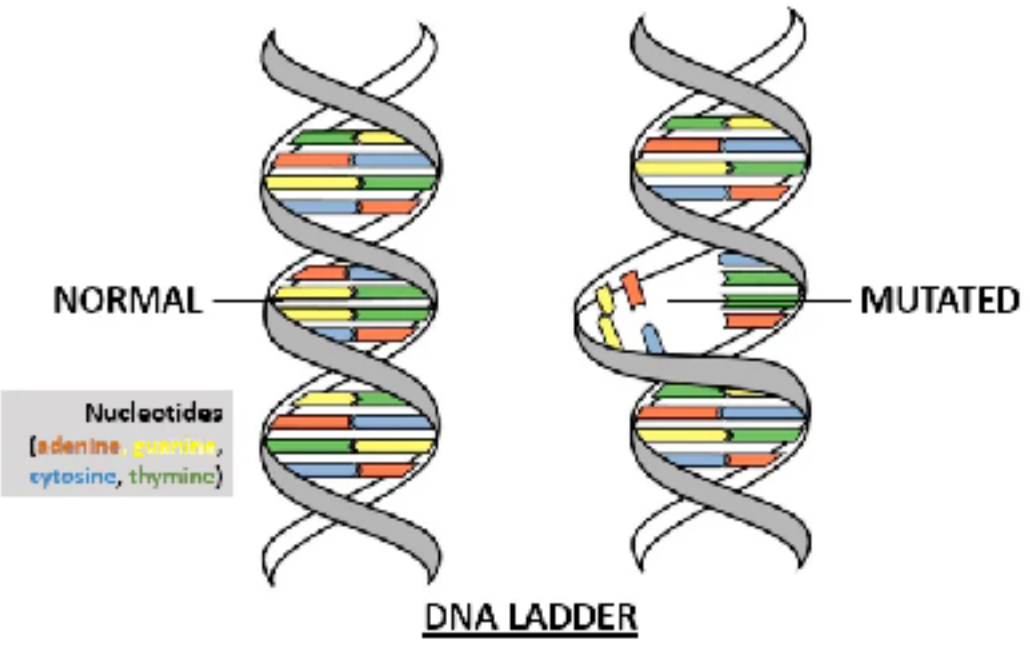
Social sciences



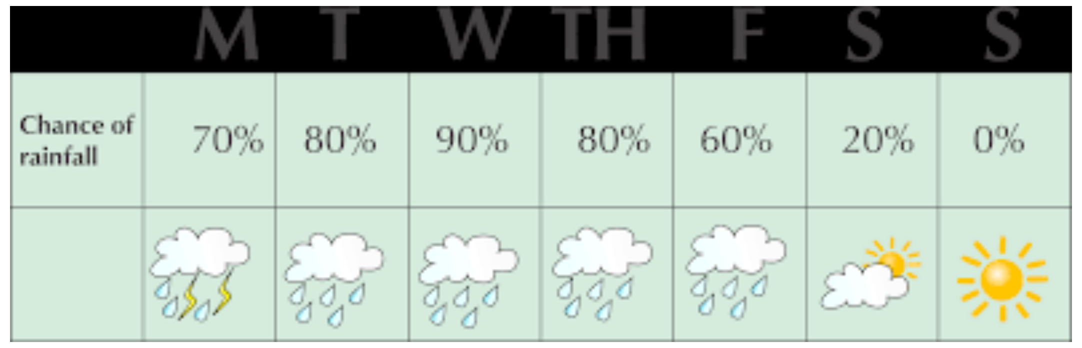
Medicine



Biology



Weather forecast



	M	T	W	TH	F	S	S
Chance of rainfall	70%	80%	90%	80%	60%	20%	0%

History

“Since traveling was onerous (and expensive), and eating, hunting and wenching generally did not fill the 17" century gentleman's day, two possibilities remained to occupy the empty hours, praying and **gambling**; **many preferred the latter**.”

Chris Van den Broeck

The rising of probability and statistical thinking

XVI XVII XVIII XIX XX XXI

Gombaud

In 1654 a French nobleman poses a gambling question: Is it better to bet on a double six in 24 throws of two dices or one six in 4 throws?

Law of large numbers

Bayes theorem

Notion of standard deviation

Systematic use of statistics

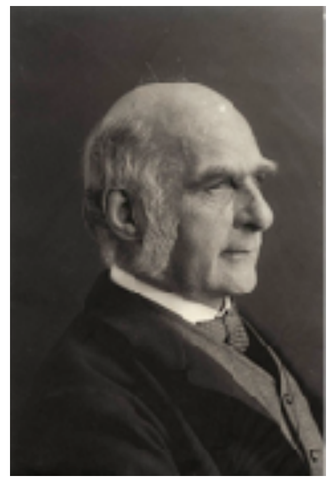
First book on probabilities: Generating functions

Probability postulates

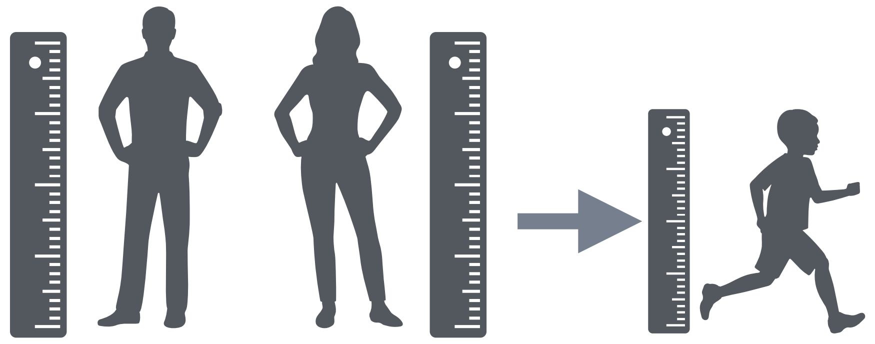
Statistics

Characterizing observed randomness

Sir Francis Galton
(1822-1911)



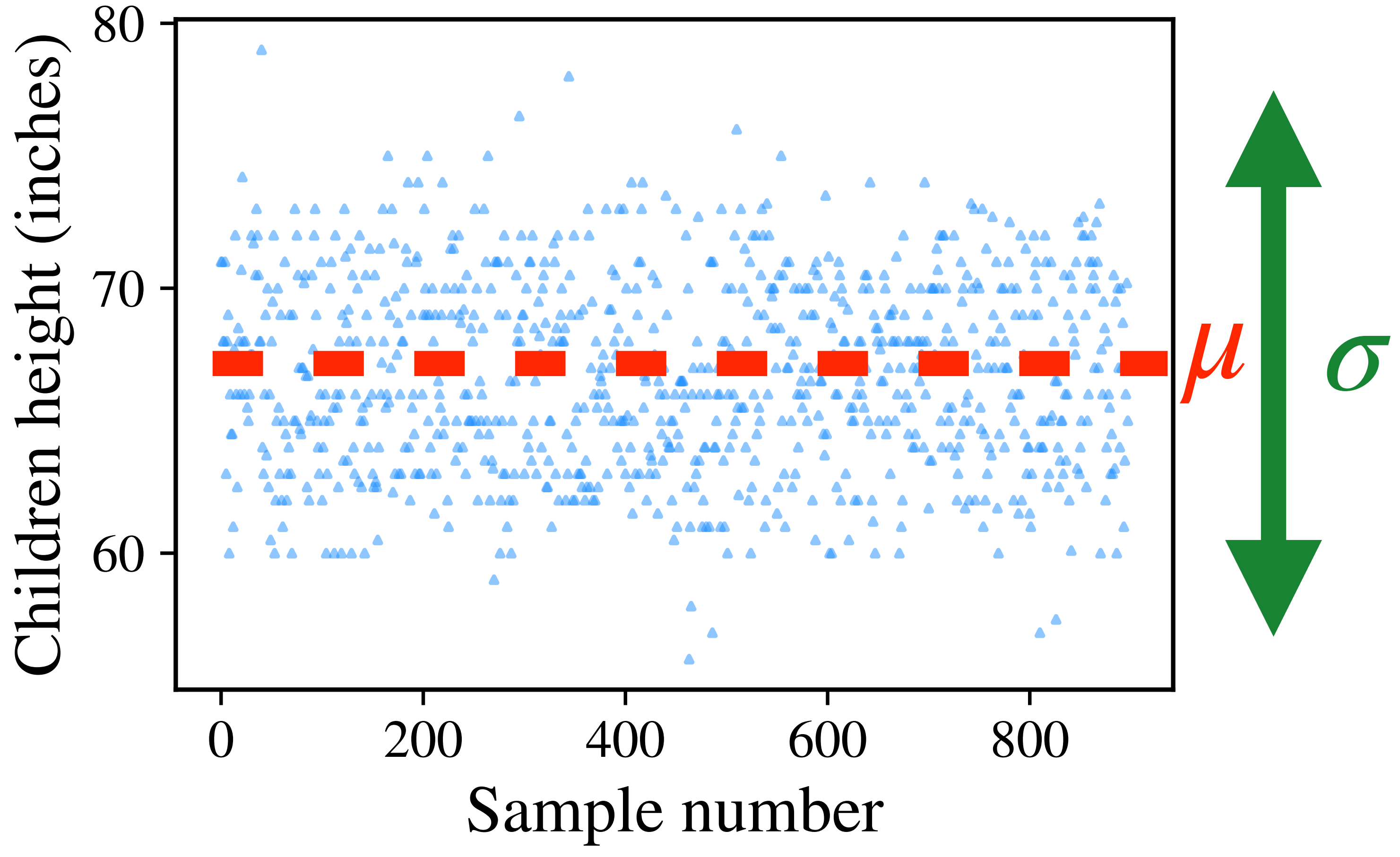
Case of study:
Human height heritage
Mean, variance, histogram, correlations



Data

	family	father	mother	gender	height	kids	male	female
0	1	78.5	67.0	M	73.2	4	1.0	0.0
1	1	78.5	67.0	F	69.2	4	0.0	1.0
2	1	78.5	67.0	F	69.0	4	0.0	1.0
3	1	78.5	67.0	F	69.0	4	0.0	1.0
4	2	75.5	66.5	M	73.5	4	1.0	0.0
...	...	...	...	...	...	...	...	...
893	136A	68.5	65.0	M	68.5	8	1.0	0.0
894	136A	68.5	65.0	M	67.7	8	1.0	0.0
895	136A	68.5	65.0	F	64.0	8	0.0	1.0
896	136A	68.5	65.0	F	63.5	8	0.0	1.0
897	136A	68.5	65.0	F	63.0	8	0.0	1.0

898 rows × 8 columns



Summary statistics

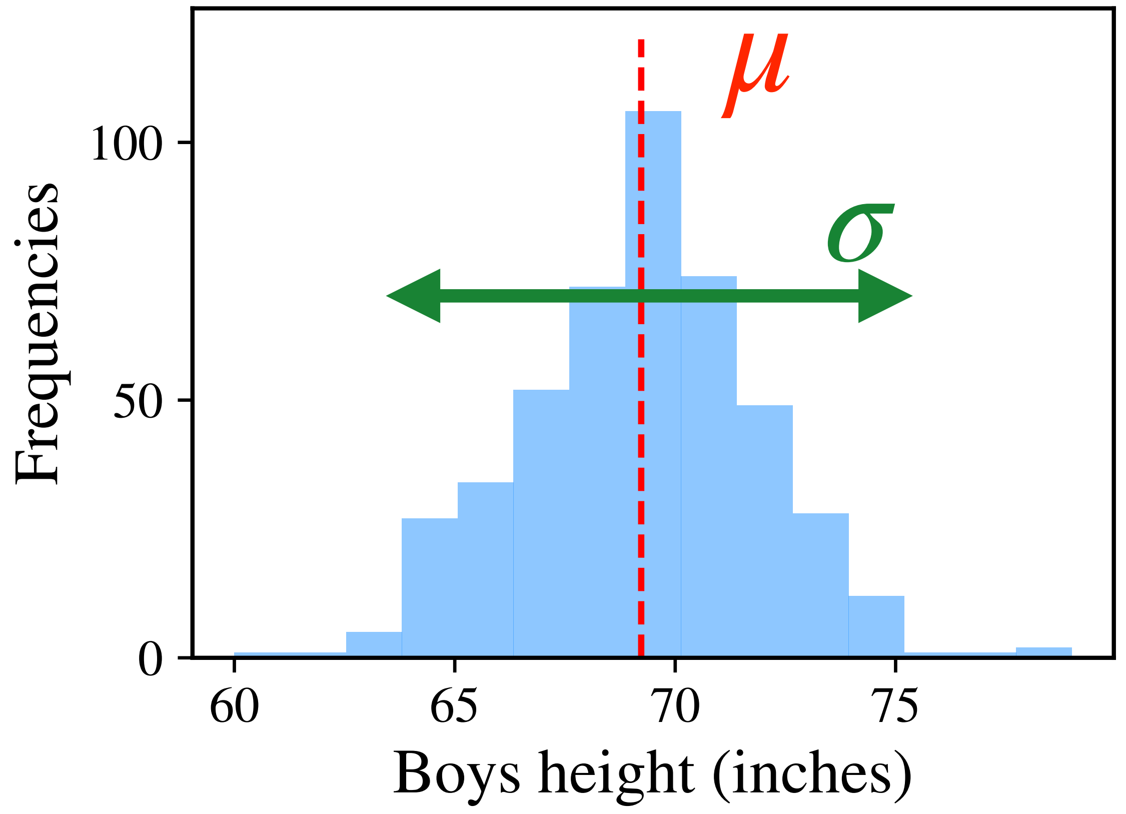
measures: N

measures: $x_1, x_2, \ldots, x_N$

Sample Mean: $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$

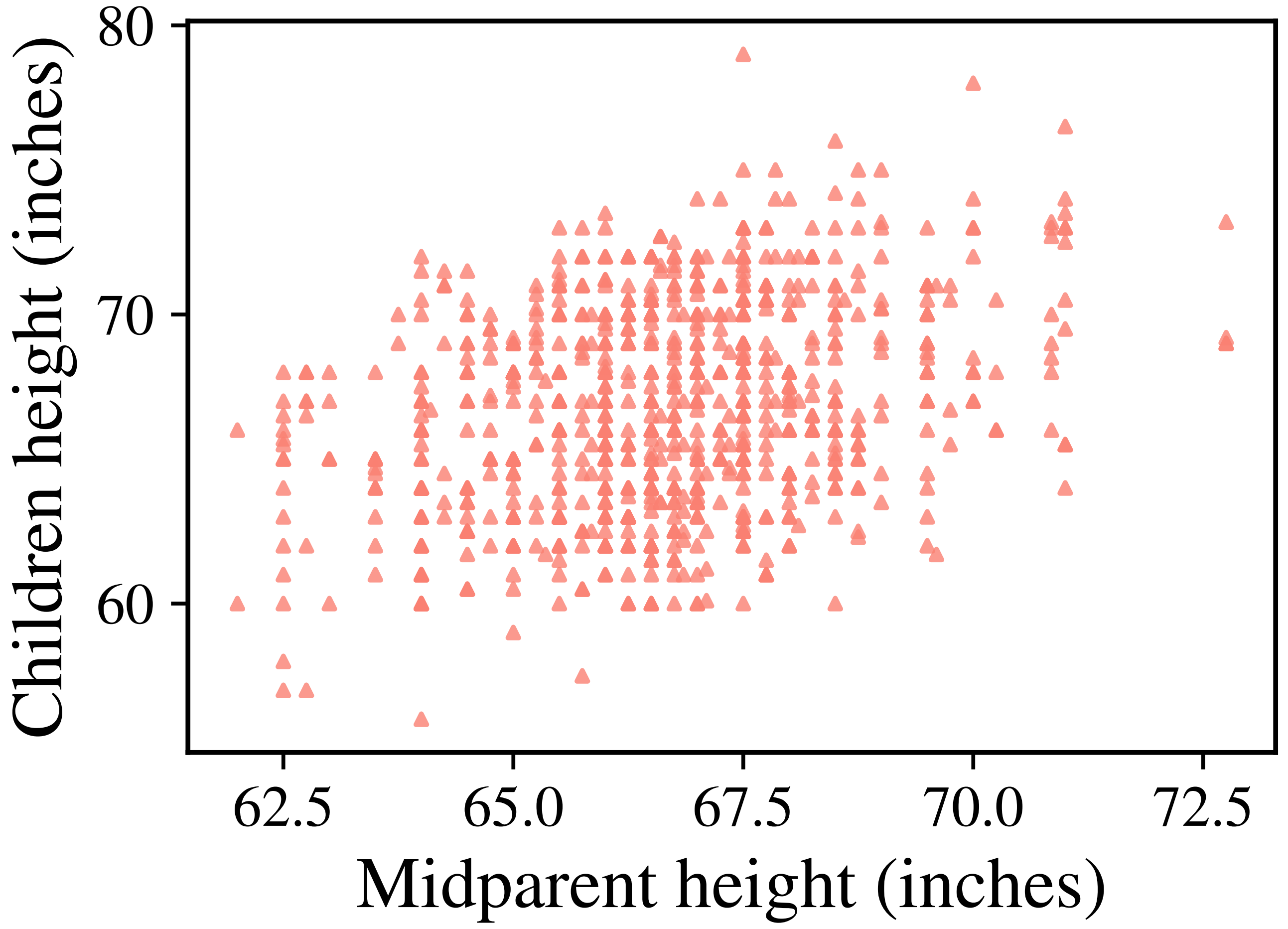
Sample Variance: $\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i^2 - \hat{\mu}\right)^2$

Frequency: (histogram) $f(x) = \sum_{i=1}^{N} I(x_i \in [x, x+dx])$



Correlations

$\hat{\rho} \sim 0.3$



"Midparent" height : $h_m = \frac{1}{2}\left(h_{father} + h_{mother}\right)$

Sample Covariance: $\hat{C} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu}_x)(y_i - \hat{\mu}_y)$

Pearson's coefficient: $\hat{\rho} = \frac{\hat{C}}{\hat{\sigma}_x \hat{\sigma}_y} \in [-1,1]$

$\hat{\rho} > 0$ The two variables tend to increase together

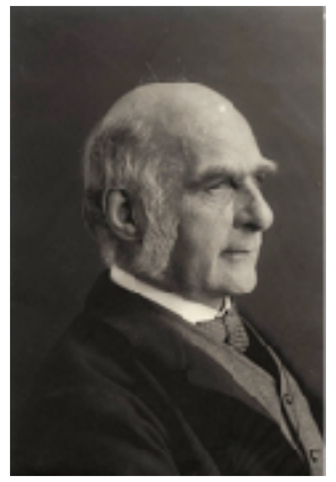
$\hat{\rho} < 0$ If one increases the other tend to decrease

$\hat{\rho} \approx 0$ Decorrelated variables

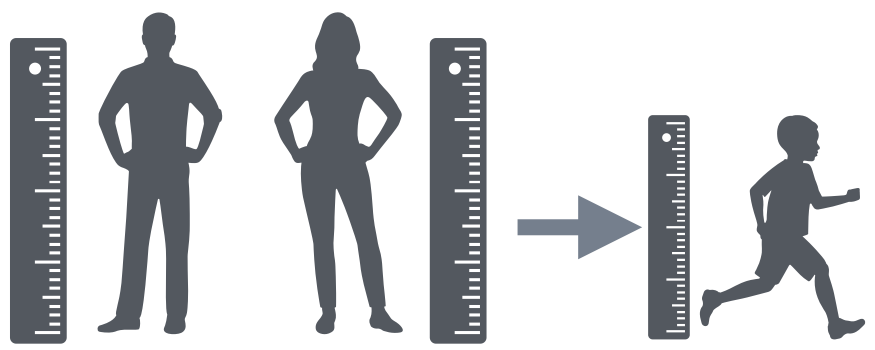
Statistics

Characterizing observed randomness

Sir Francis Galton
(1822-1911)



Case of study:
Human height heritage
Mean, variance, histogram, correlations



Linear regression

Best straight line fit to data

$$y = \alpha x + \beta$$

$$\alpha = \frac{\hat{C}}{\hat{\sigma}_x^2} = \rho \frac{\hat{\sigma}_y}{\hat{\sigma}_x}$$

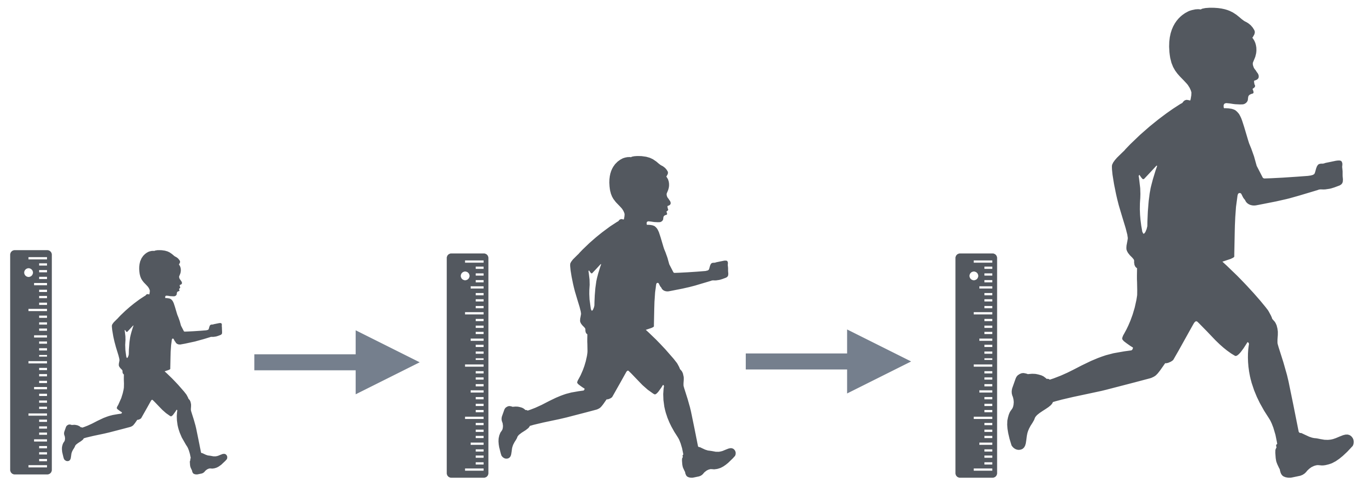
$$\beta = \hat{\mu}_y - \alpha \hat{\mu}_x$$

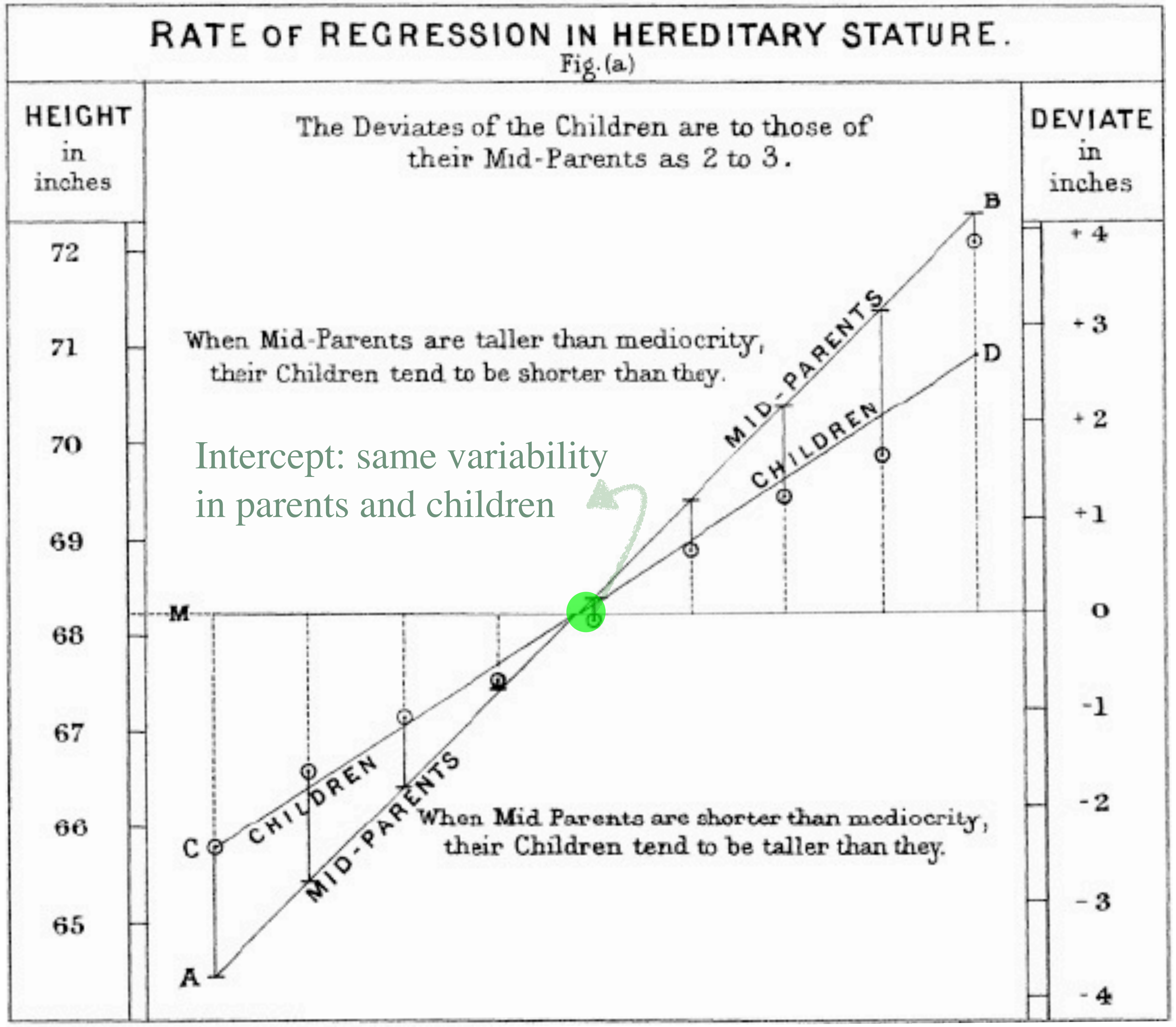
$$\boxed{\alpha = \frac{2}{3}}$$

Variability controlled through generations

Regression to mediocrity

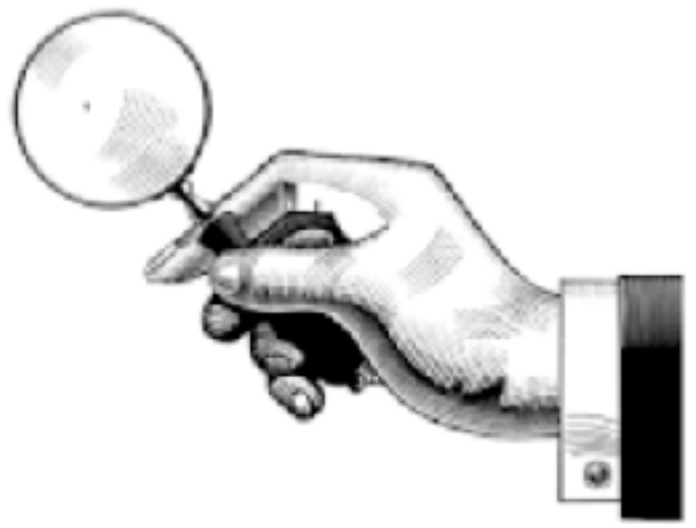
$\alpha = 1 \rightarrow$





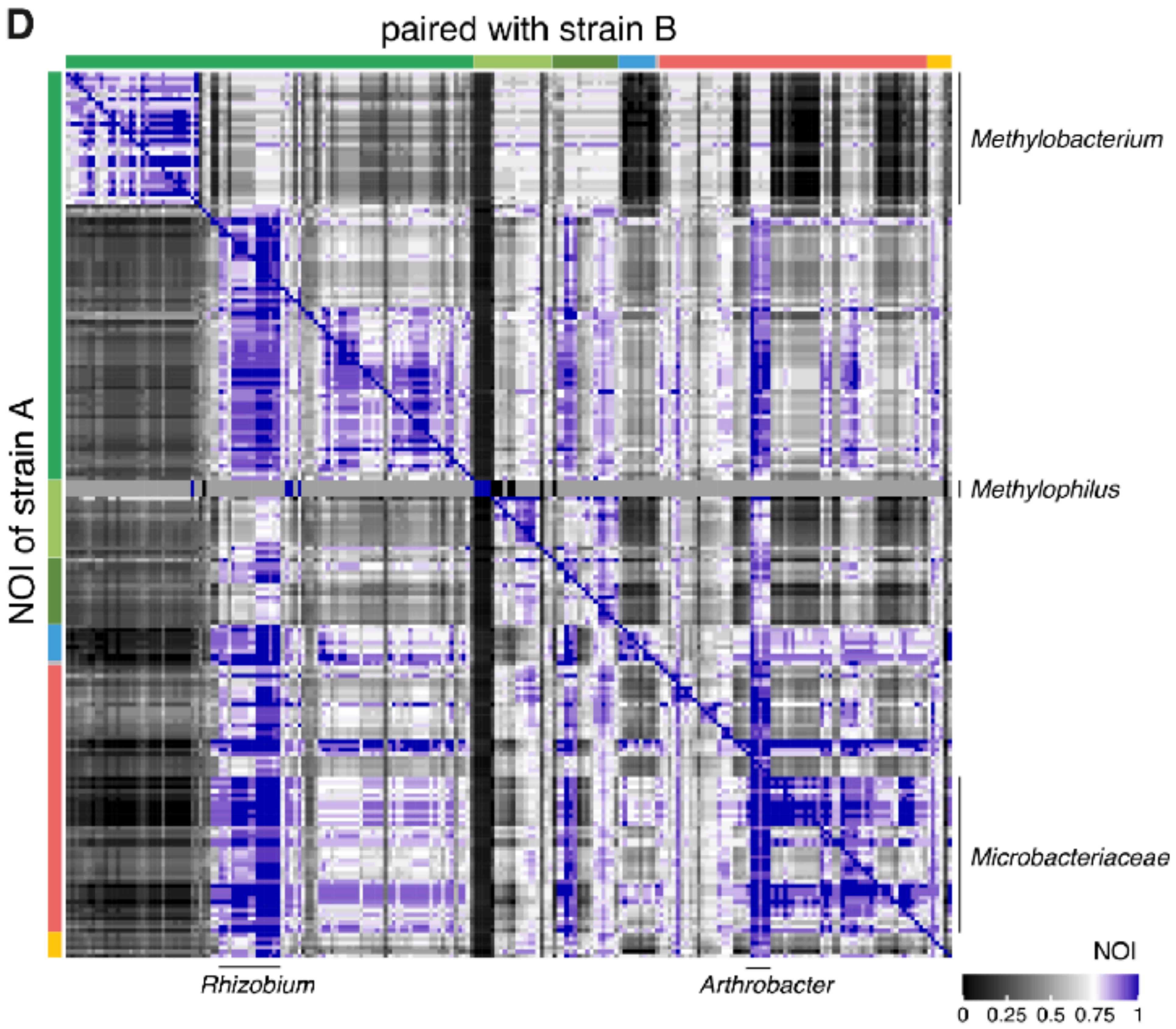
Statistics

Characterizing observed randomness



THE DEVIL IS
IN THE DETAIL

How to visualize complex data?

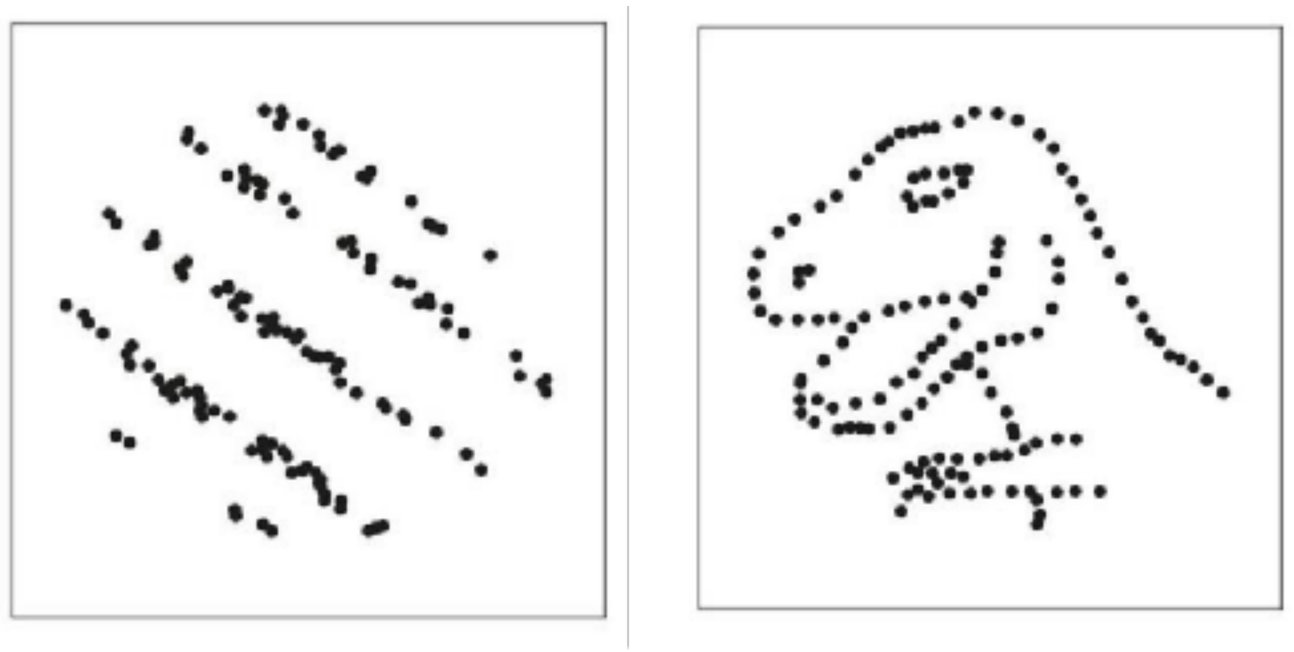


How to interpret summary statistics?

What measures the central tendency?
Mean, mode or median

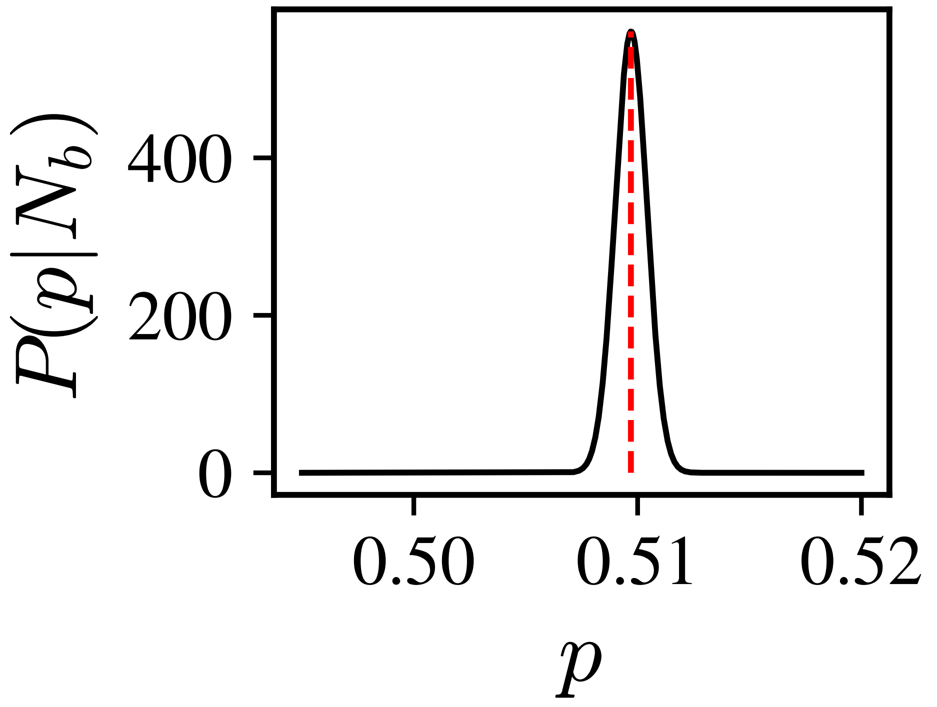
How many sigmas?
Chebyshev's theorem

Is it correlation or causation?
Modeling

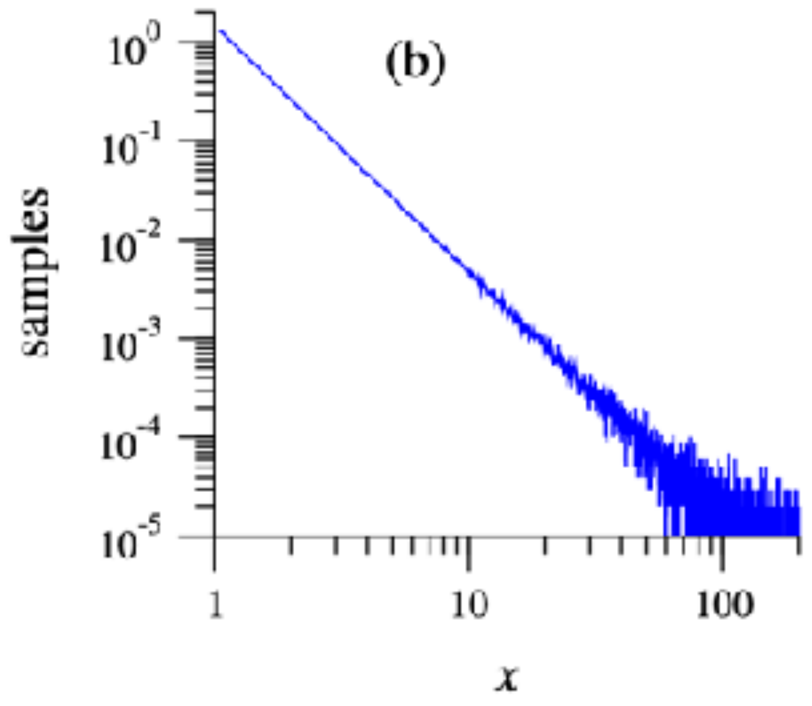


How to deal with fluctuations?

Confidence intervals



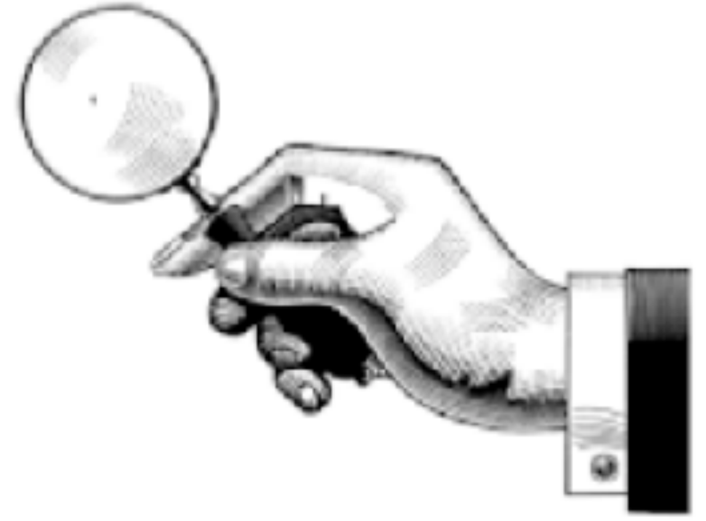
Logarithmic vs. linear binning



Statistics

Characterizing observed randomness

Probability estimators: normalized histograms



THE DEVIL IS
IN THE DETAIL

Discrete data

$\hat{x}_i \in \{a, b, c, \dots\}$

 $\rightarrow$ $\hat{x}_i \in \{1,2,3,4,5,6\}$

Frequency estimate

$$\hat{f}(x) = \sum_{i=1}^{N} I(\hat{x}_i = x) \in [1,N]$$

Probability estimate

$$\hat{p}(x) = \frac{1}{N}\hat{f}(x) \in [0,1]$$

Example:

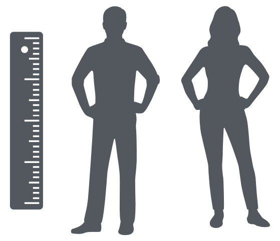
 $\rightarrow$ $\hat{x}_1 = 4 \quad \hat{x}_2 = 6$
$\hat{x}_3 = 1 \quad \hat{x}_4 = 4$

$$\hat{f}(1) = 1 \quad \hat{f}(4) = 2 \quad \hat{f}(3) = 0$$

$$\hat{p}(1) = \frac{1}{4} \quad \hat{f}(4) = \frac{1}{2} \quad \hat{p}(3) = 0$$

Continuous data

$\hat{x}_i \in [a, b]$

 $\rightarrow$ $\hat{x}_i \in [0\text{ m}, 4\text{ m}]$

$$\boxed{I(\hat{x}_i = x) = 0}$$

Two real numbers are never equal:
real numbers have zero measure

$$\hat{f}(x, \Delta x) = \sum_{i=1}^{N} I(\hat{x}_i \in [x, x + \Delta x])$$ Frequency estimate

$$\hat{p}(x, \Delta x) = \frac{1}{N}\sum_{i=1}^{N} \hat{f}(x, \Delta x)$$ Probability estimate

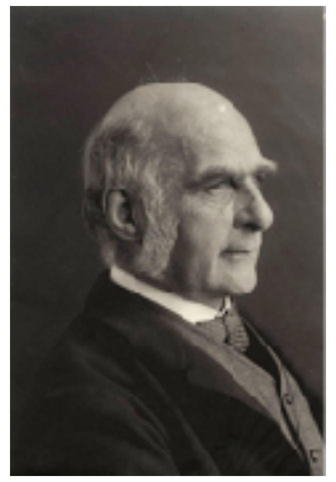
Example:

$$\hat{f}(x = 55\text{ in.}, \Delta x = 30\text{ in.}) = 896$$
$$\hat{p}(x = 55\text{ in.}, \Delta x = 30\text{ in.}) = 1$$

$$\hat{f}(x = 55\text{ in.}, \Delta x = 5\text{ in.}) = 31$$
$$\hat{p}(x = 55\text{ in.}, \Delta x = 5\text{ in.}) \approx 0.03$$

$$\hat{f}(x = 55\text{ in.}, \Delta x = 0\text{ in.}) = 0$$
$$\hat{p}(x = 55\text{ in.}, \Delta x = 5\text{ in.}) \approx 0$$

Galton's data



Getting rid of bins

Probability density function (estimate)

$$\hat{\rho}(x) = \frac{\hat{\rho}(x, \Delta x)}{\Delta x} \in [0,\infty]$$

Cumulative distribution function (estimate)

$$\hat{F}_{<}(x) = \frac{1}{N}\sum_{i=1}^{N} I(\hat{x}_i < x) \in [0,1]$$

$$\hat{F}_{>}(x) = \frac{1}{N}\sum_{i=1}^{N} I(\hat{x}_i > x) \in [0,1]$$

$$\hat{F}_{<}(x) + \hat{F}_{>}(x) = 1$$

Probability

Modeling randomness

Jacob Bernuilli
(1654 - 1705)

Case of study:

Flipping coins
Bernoulli distribution
Law of large numbers

Kolgomorov's receipt for probability theory

Andréi Kolmogórov
(1903 - 1987)

Ingredient 1: Sample space
(all possible outcomes)

H: heads

T: tails

Ingredient 2: Random variable
(Each outcome has a number)

X(H)=1 X(T)=0

Ingredient 3: Define probabilities
(each outcome has a probability)

$P(H) \in [0,1]$

$P(T) \in [0,1]$

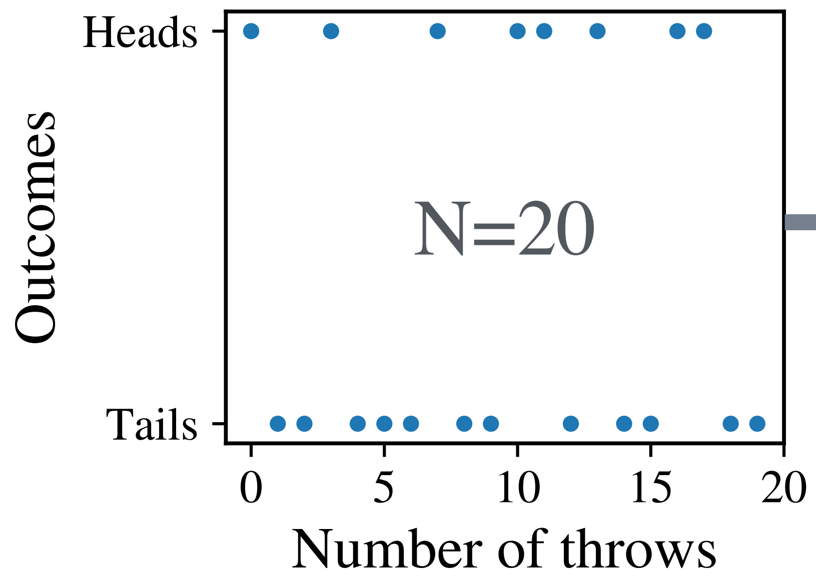
$P(H) + P(T) = 1$

How to assign probabilities?

Method I: Maximum ignorance
Symmetric guess

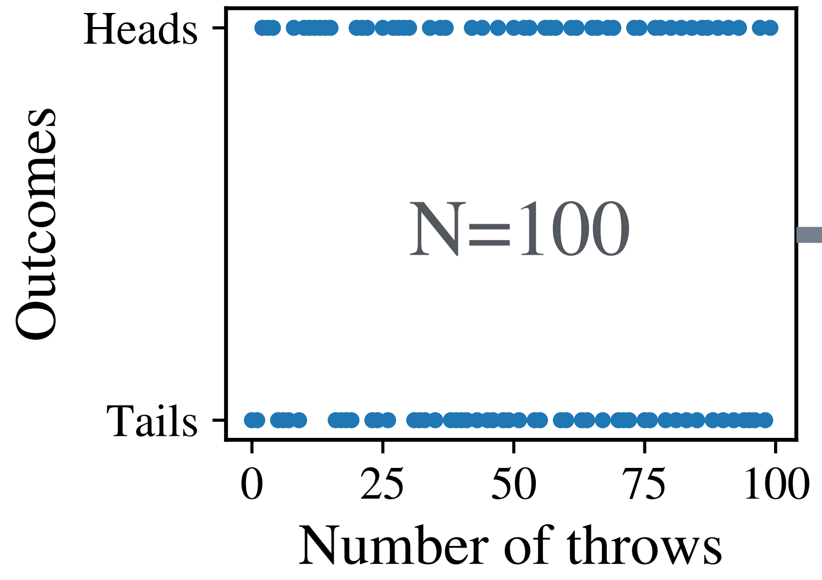
$$P(H) = P(T) = \frac{1}{2}$$

Moments



$\hat{\mu} = 0.40$

$\hat{\sigma} = 0.24$



$\hat{\mu} = 0.48$

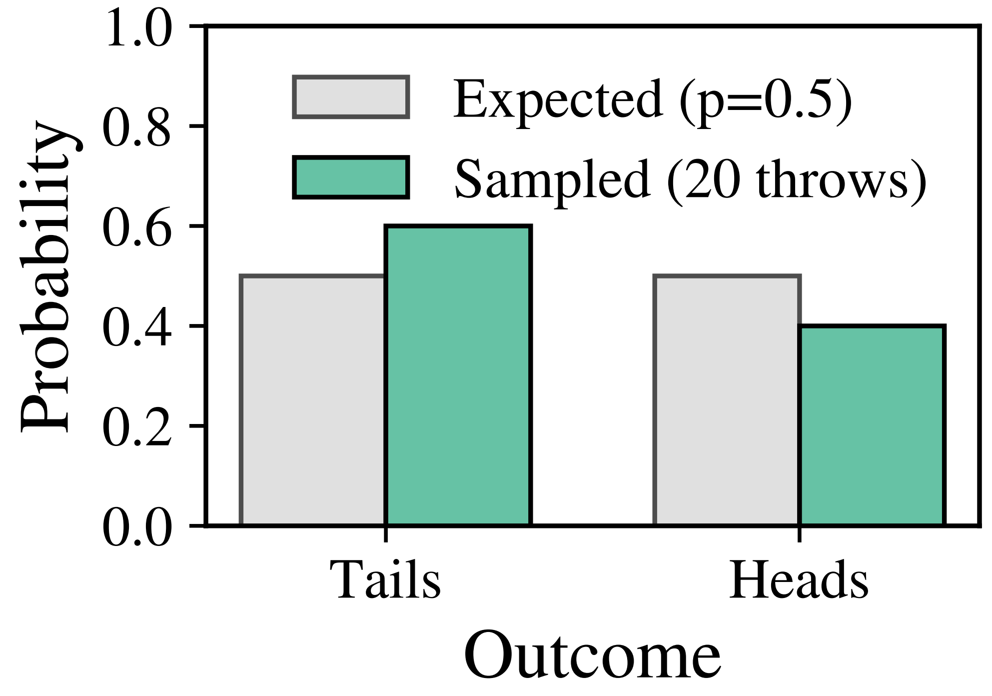
$\hat{\sigma} = 0.2496$

$$\langle X \rangle = E(X) = \mu = \sum_{x \in \Omega} xP(X = x) = 0.5 \longrightarrow \hat{\mu} \approx \mu$$

$$\langle (X - \mu)^2 \rangle = \mathrm{var}(X) = \sigma^2 = E(X^2) - \mu^2 = 0.25 \longrightarrow \hat{\sigma} \approx \sigma$$

Probabilities

$\hat{P} \approx P$



Law of large numbers

$$\lim_{N\to\infty} \hat{P}(X \in A) = P(X \in A)$$

$$Z = z(X)$$

$$\downarrow$$

$$\lim_{N\to\infty} \hat{Z} = E(Z)$$

Probability
Modeling randomness

Antoine Gombaud (1607-1684)

Blaise Pascal (1623-1662)

Pierre Fermat (1607-1665)

Case of study:

Throwing dices
Uniform distribution
Binomial distribution

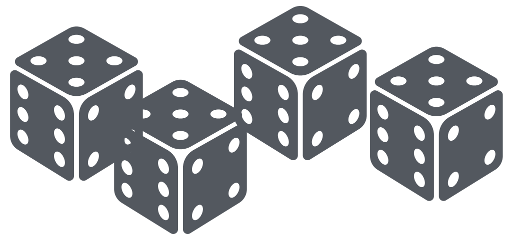
How to assign probabilities of one die?

Method I: Maximum entropy/ignorance
Symmetric guess

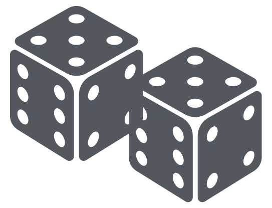
$$P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}$$

Which is more likely?

Game A:
Rolling, at least, one six in 4 throws of one die



Game B:
Rolling, at least, one double six in 24 throws of two dice



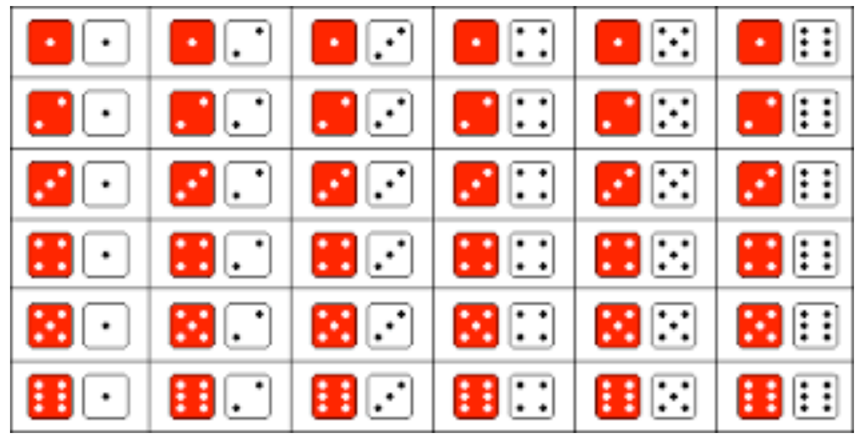
$\times 24$

How to assign probabilities of two dice?

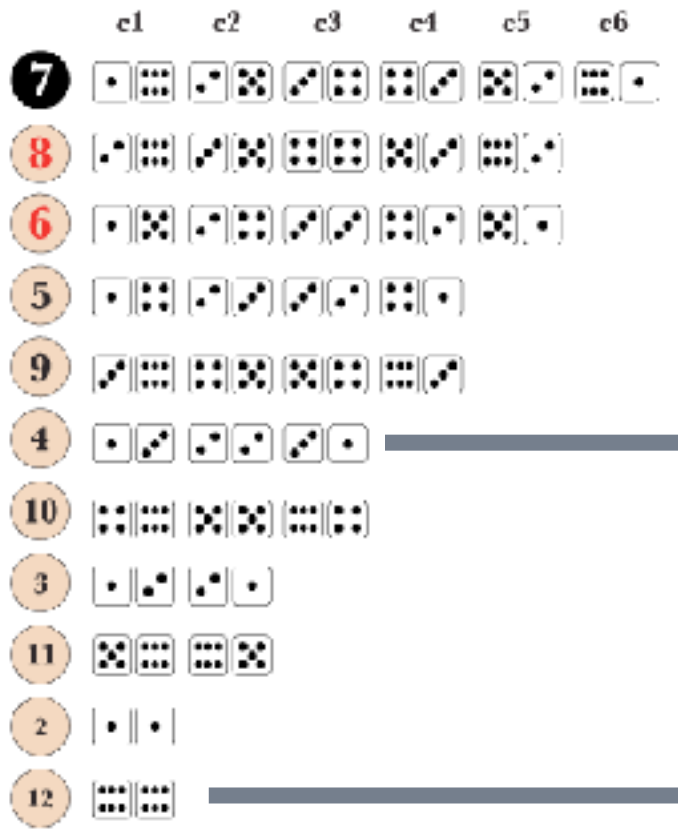
Method II: Intuitive definition of probability

$$P(A) = \frac{\#A}{\#\Omega}$$

$$\#\Omega = 6^2 = 36$$



$\#A$



$$P(7) = \frac{6}{36} = \frac{1}{6}$$

$$P(4) = \frac{3}{36} = \frac{1}{12}$$

$$P(12) = \frac{1}{36}$$

Computing probabilities

$P(V) + P(V^C) = 1 \rightarrow P(V) = 1 - P(V^C)$ (Either it rains or it doesn't)

$V_1, \ldots, V_n$ Independent events $\rightarrow P(V_1, \ldots, V_n) = P(V_1)P(V_2)\ldots P(V_n)$

$$\boxed{P(A\text{ wins}) = 1 - P(A\text{ loses}) = 1 - \left(\frac{5}{6}\right)^4 \approx 0.52}$$

$$P(B\text{ wins}) = 1 - P(B\text{ loses}) = 1 - \left(\frac{35}{36}\right)^{24} \approx 0.49$$

Game B loses on average, game A wins on average

Probability

Modeling randomness

Antoine Gombaud (1607-1684)

Blaise Pascal (1623-1662)

Pierre Fermat (1607-1665)

Case of study:

Throwing dices

Uniform distribution

Binomial distribution

Central limit theorem:

For any R.V. X with $\mu = E(X), \sigma^2 = Var(X) < \infty$

Let us define $Y := \frac{1}{N}\sum_{i=1}^{N} X^{(i)}$, Then

$E(Y) \approx \mu \quad Var(Y) \approx \sigma^2/N \quad \rho(y) \approx G\left(x; \mu, \frac{\sigma}{\sqrt{N}}\right)$

$$G(x; \mu, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

Probability

Modeling randomness

Antoine Gombaud (1607-1684)

Blaise Pascal (1623-1662)

Pierre Fermat (1607-1665)

Case of study:

Throwing dices

Uniform distribution

Binomial distribution

Central limit theorem:

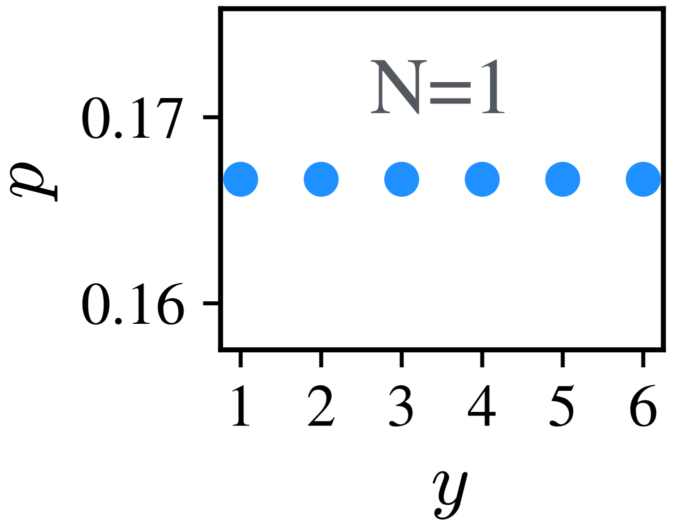
For any R.V. X with $\mu = E(X), \sigma^2 = Var(X) < \infty$

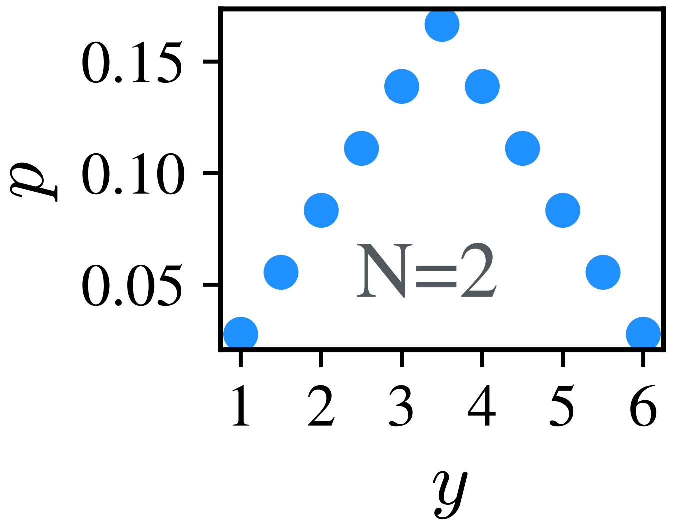
Let us define $Y := \frac{1}{N}\sum_{i=1}^{N} X^{(i)}$, Then

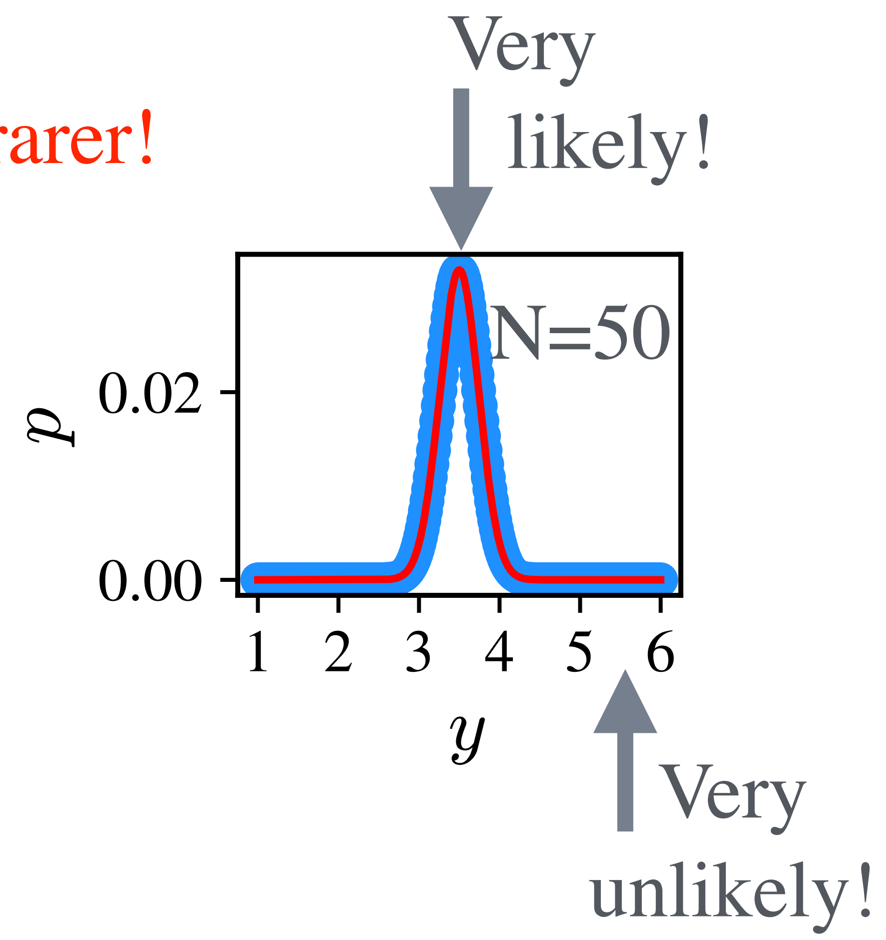
$$E(Y) \approx \mu \quad Var(Y) \approx \sigma^2/N \quad \rho(y) \approx G\left(x; \mu, \frac{\sigma}{\sqrt{N}}\right)$$

$$G(x; \mu, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

Deviations from the mean become rarer!







Probability

Modeling randomness



Antoine Gombaud (1607-1684) Blaise Pascal (1623-1662) Pierre Fermat (1607-1665)

Case of study:

Throwing dices

Uniform distribution

Binomial distribution

Central limit theorem:

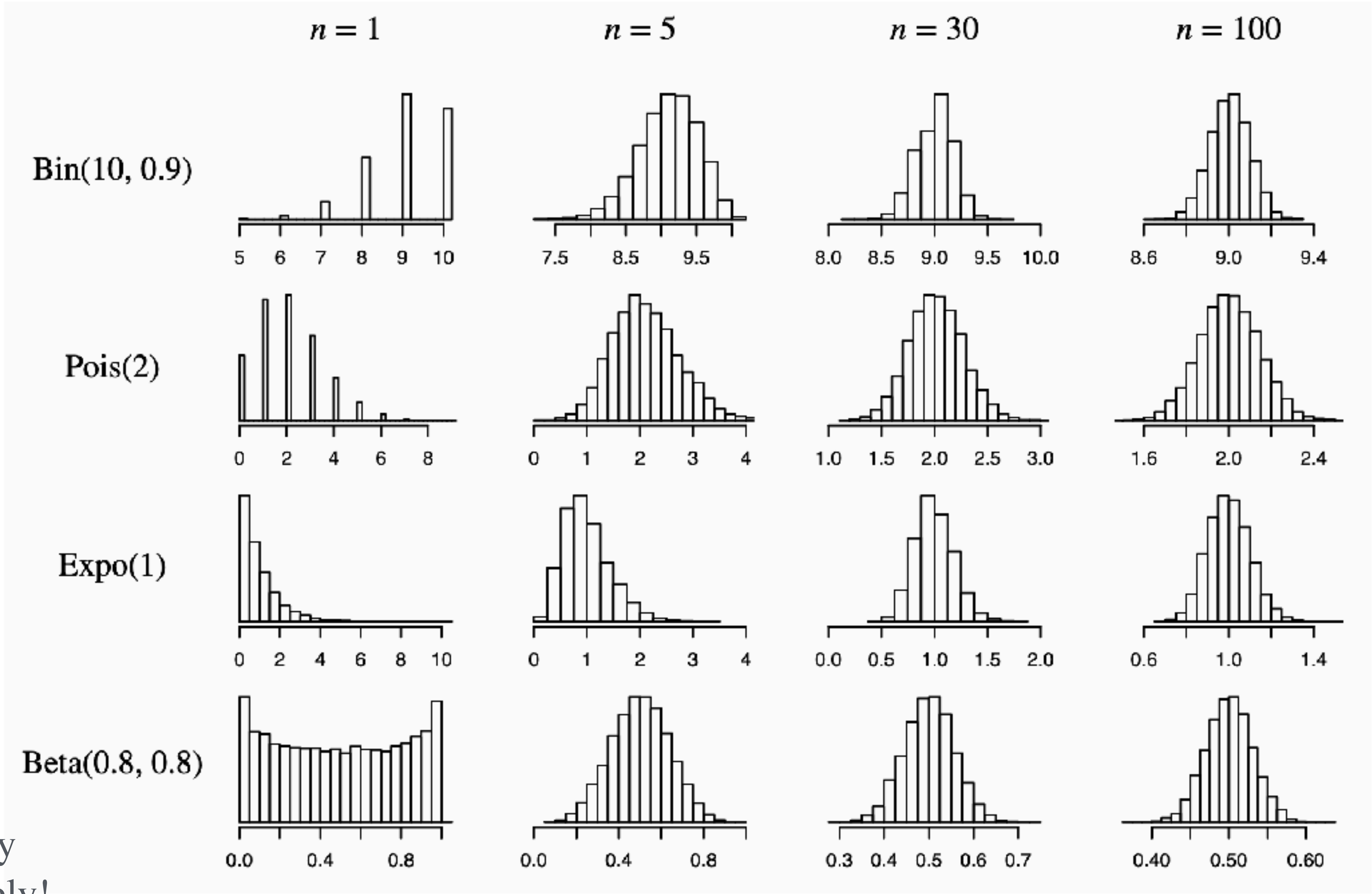
For any R.V. X with $\mu = E(X), \sigma^2 = Var(X) < \infty$

Let us define $Y := \frac{1}{N}\sum_{i=1}^{N} X^{(i)}$, Then

$$E(Y) \approx \mu \quad Var(Y) \approx \sigma^2/N \quad \rho(y) \approx G\left(x; \mu, \frac{\sigma}{\sqrt{N}}\right)$$

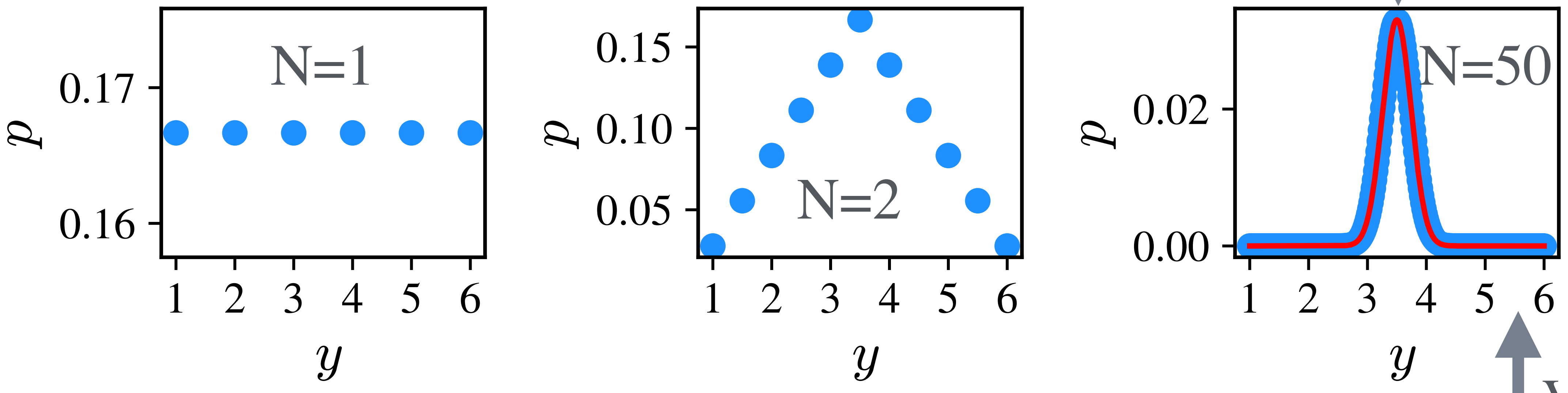
$$G(x; \mu, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

Robustness of statistical patterns



Deviations from the mean become rarer!

Very likely!



Very unlikely!

Probability

Modeling randomness

Sally Clarck
(1964-2007)

Case of study:

Prosecutor’s fallacy

Bayes theorem

Conditioned probabilities:

Joint probability

$$\boxed{P(A\,|\,B)} = \frac{\boxed{P(A,B)}}{P(B)}$$

Prob. of A **GIVEN** B

Probability

Modeling randomness

Sally Clarck
(1964-2007)

Case of study:

Prosecutor's fallacy

Bayes theorem

Conditioned probabilities:

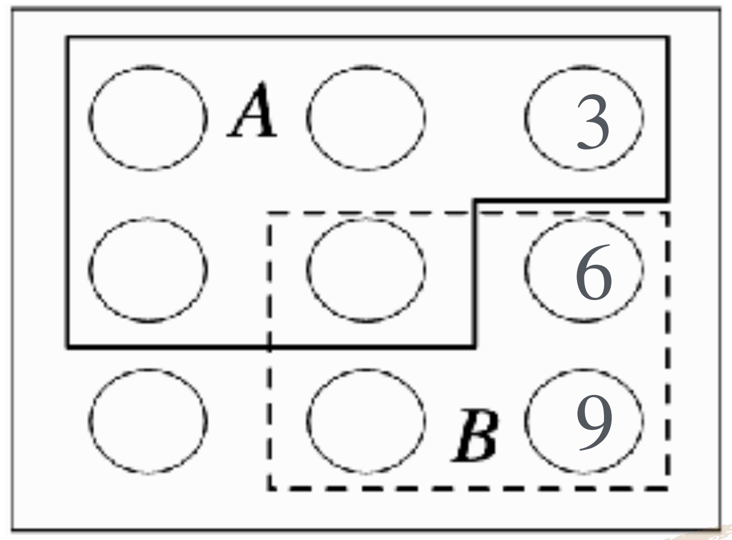
Joint probability

$$\boxed{P(A\,|\,B)} = \frac{\boxed{P(A,B)}}{P(B)}$$

Prob. of A **GIVEN** B

Example: Uniform discrete distribution

$$X \in \{1,2,\ldots,9\} \quad A = \{1,2,3,4,5\} \quad B = \{5,6,8,9\}$$



$$P(A) = \frac{5}{9} \qquad P(B) = \frac{4}{9}$$

$$P(A,B) = \frac{1}{9}$$

$$P(A\,|\,B) = \frac{1}{4} \qquad P(B\,|\,A) = \frac{1}{5}$$

Conditioning is **not** symmetric

Probability

Modeling randomness

Conditioned probabilities:

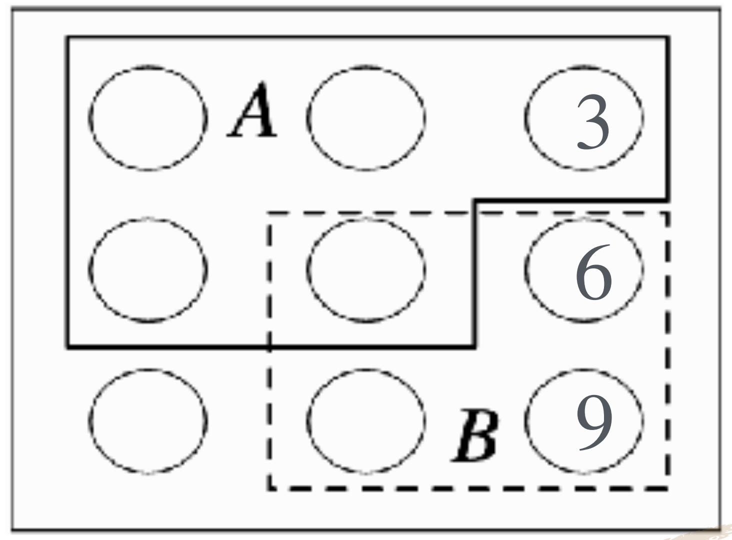
Joint probability

$$\boxed{P(A\,|\,B)} = \frac{\boxed{P(A,B)}}{P(B)}$$

Prob. of A **GIVEN** B

Example: Uniform discrete distribution

$X \in \{1,2,\ldots,9\}$ $\quad A = \{1,2,3,4,5\}$ $\quad B = \{5,6,8,9\}$



$$P(A) = \frac{5}{9} \qquad P(B) = \frac{4}{9}$$

$$P(A,B) = \frac{1}{9}$$

$$P(A\,|\,B) = \frac{1}{4} \qquad P(B\,|\,A) = \frac{1}{5}$$

Conditioning is **not** symmetric

Sally Clarck (1964-2007)

Case of study:

Prosecutor's fallacy

Bayes theorem

Bayes theorem: $\quad P(A\,|\,B) = P(B\,|\,A)\dfrac{P(A)}{P(B)}$

$E =$ Evidence: two babies sadly die at very young age in one family

$I =$ Sally is innocent of the crime $\qquad G =$ Sally is guilt of the crime

Knowledge : If Sally is innocent, the event is very unlikely $\quad P(E\,|\,I) \approx \dfrac{1}{73 \cdot 10^6}$

Fallacy: $\quad P(I\,|\,E) \approx P(E\,|\,I) \sim 0$

Solution : The probability that Sally was a murderer was ALSO very small

$$P(I\,|\,E) = P(E\,|\,I)\frac{P(I)}{P(E)} \qquad P(E) = P(E\,|\,I)P(I) + P(E\,|\,G)P(G)$$

Knowledge : Two infanticides are extremely unlikely, $\quad P(G) \approx \dfrac{1}{2 \cdot 10^9}$

Truth: $\quad P(I\,|\,E) \approx 0.9 \longrightarrow$ It was more likely that Sara was innocent

ALL PROBABILITIES ARE CONDITIONED

PROBABILITIES ARE SERIOUS: USED TO TAKE DECISIONS

Probability

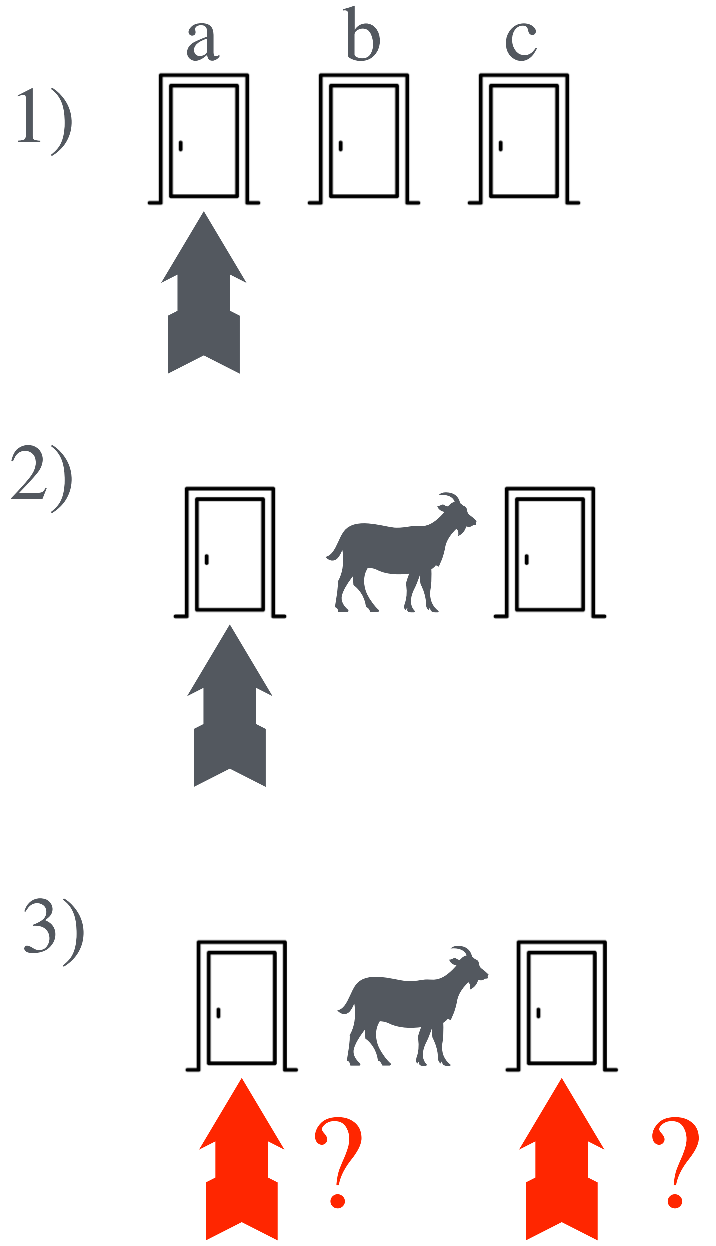
Modeling randomness

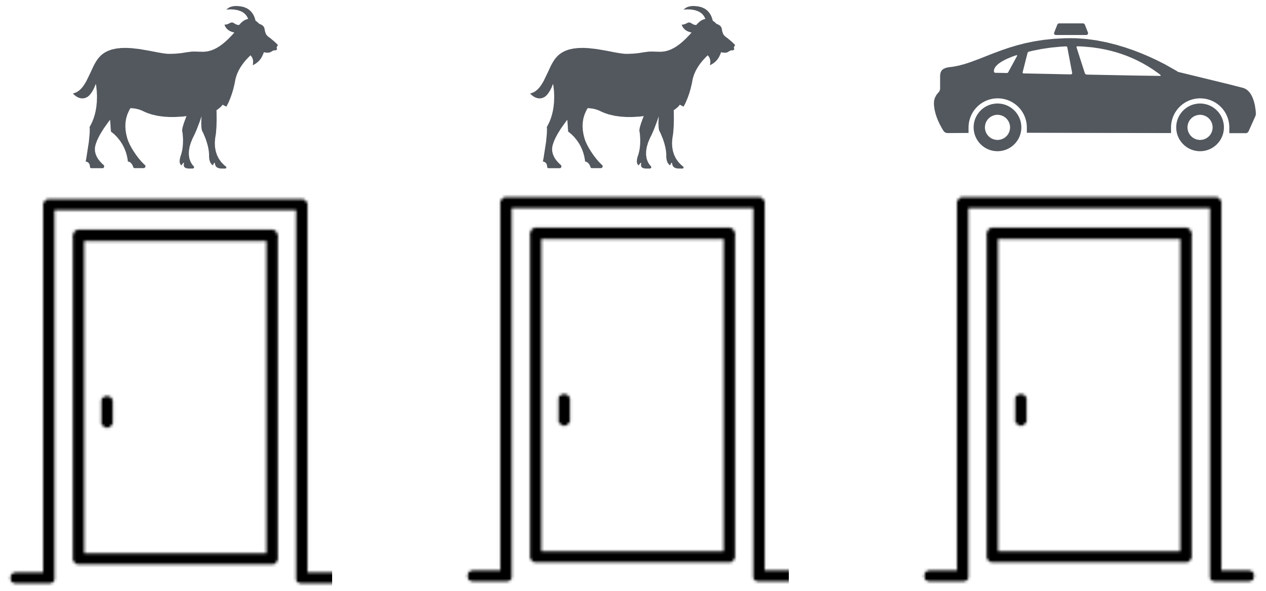
Marilyn vos Savant
(1946 -)

Case of study:

Monty Hall paradox

Conditioned probabilities





Rules:

1) Choose one door.
2) M.H. opens another door with no prize.
3) Stay with your initial choice or change.

Probability

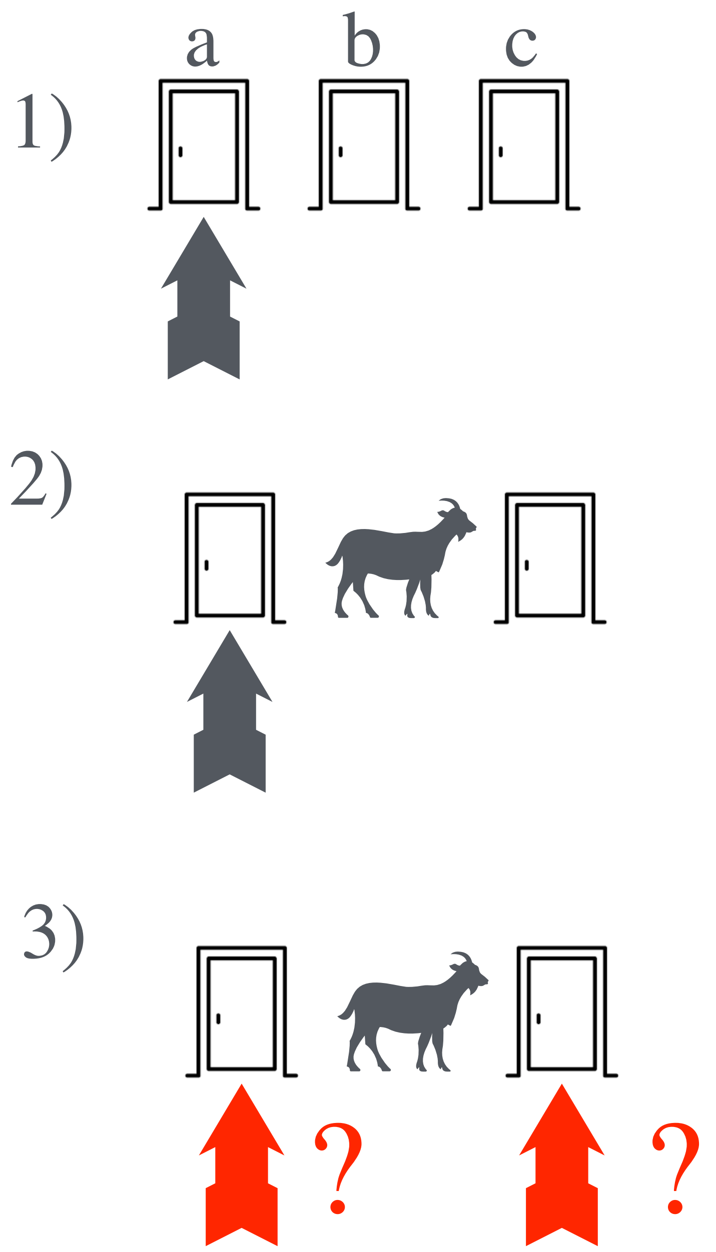
Modeling randomness

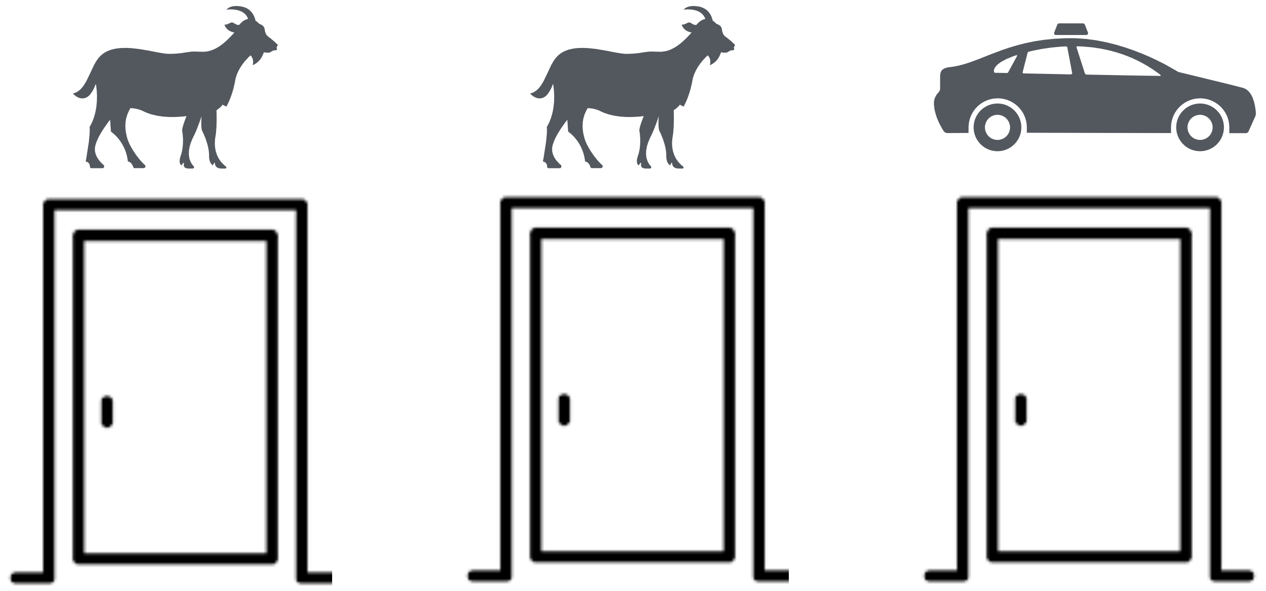
Marilyn vos Savant
(1946 -)

Case of study:

Monty Hall paradox

Conditioned probabilities





Rules:

1) Choose one door.
2) M.H. opens another door with no prize.
3) Stay with your initial choice or change.

Q: Is there a winner strategy?

- It doesn't matter if we stay or change.
- It is better to stay.
- It is better to change.

Reasoning 1:

M.H. provides information that
favors one choice.
P(win) > 1/2.

Reasoning 2:

M.H. is just goofing around.
P(win) =1/2

Probability

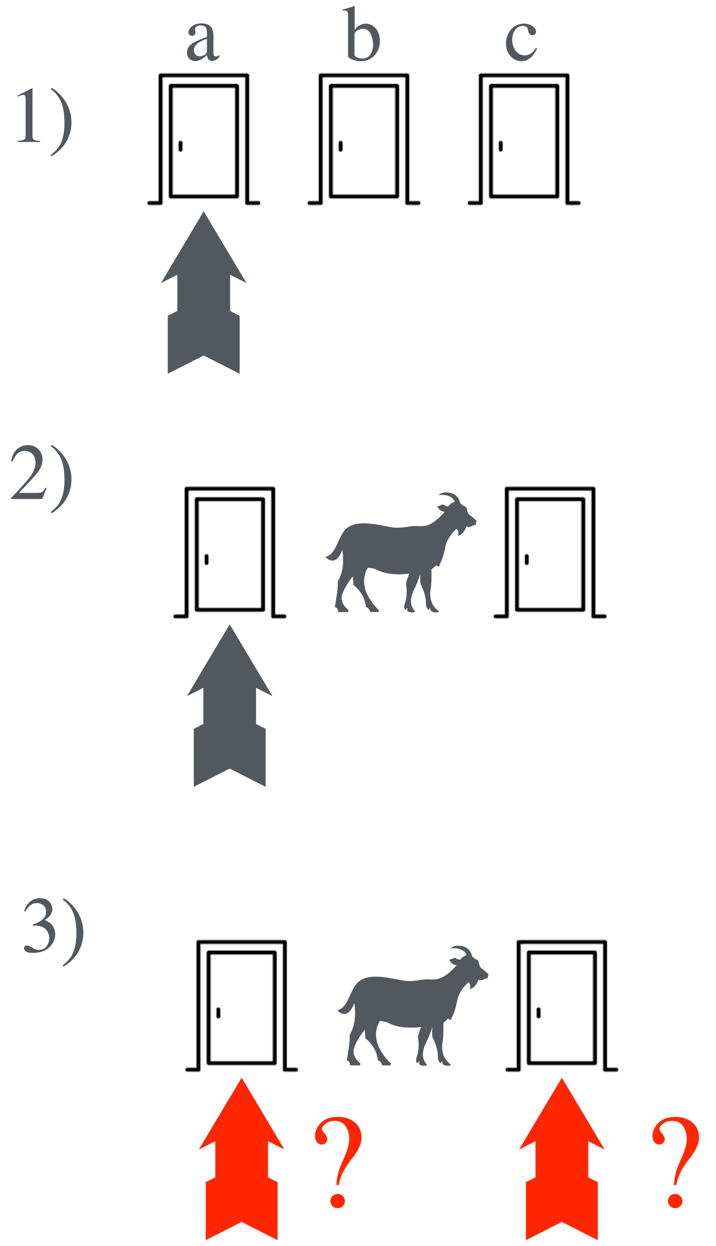
Modeling randomness

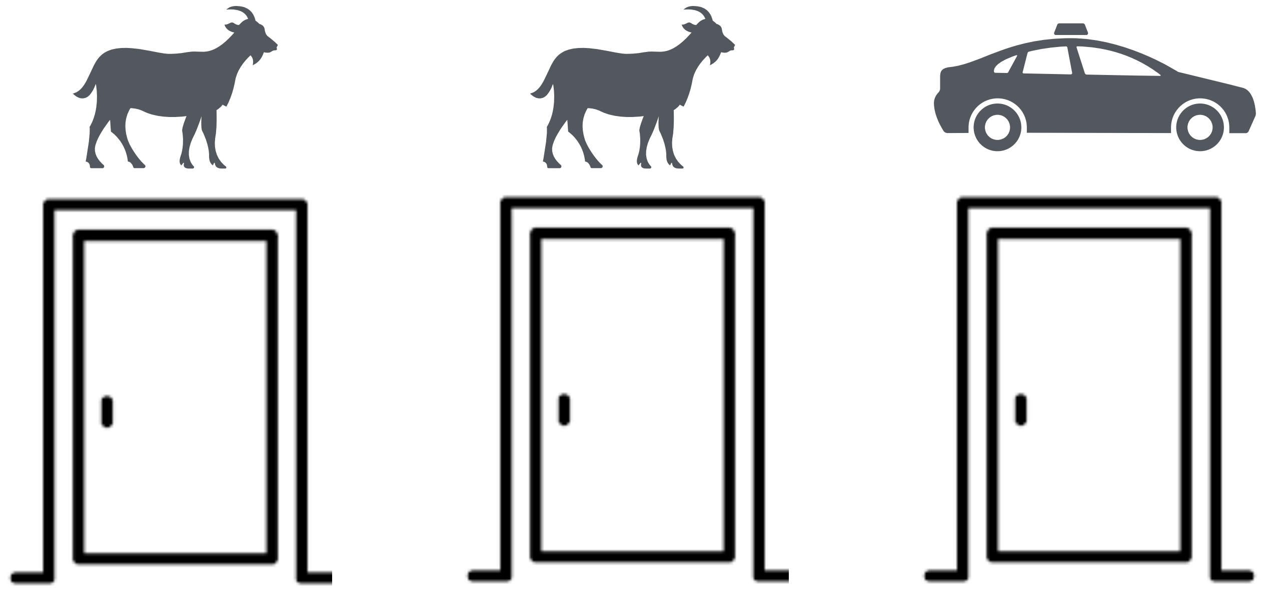
Marilyn vos Savant
(1946 -)

Case of study:

Monty Hall paradox

Conditioned probabilities





Rules:

1) Choose one door.
2) M.H. opens another door with no prize.
3) Stay with your initial choice or change.

Q: Is there a winner strategy?

- It doesn't matter if we stay or change.
- It is better to stay.
- It is better to change.

Reasoning 1:

M.H. provides information that favors one choice.
P(win) > 1/2.

Reasoning 2:

M.H. is just goofing around.
P(win) =1/2

C: Where is the car M: Where is M.H $C, M \in \{a, b, c\}$

Say we choose the door "a" and M.H. shows a goat in door "b"

$P(C = a | M = b) = P(\text{win staying}) = ?$

$P(C = b | M = b) = 0$

$P(C = c | M = b) = P(\text{win changing}) = ?$

What do we know? $P(M = b | C = a) = 1$ $P(M = b | C = B) = 0$ $P(M = b | C = B) = 1/2$

Bayes: $P(C = a | M = b) = \dfrac{P(M = b | C = a)P(C = a)}{P(M = B)} = \dfrac{2}{3} \longrightarrow P(C = a | M = b) = \dfrac{1}{3}$

A: It is more likely to win if we change!!

Inference
Connecting experiments and models

Karl Pearson
(1857-1936)

Case of study:

German tank problem

Method of moment

Maximum likelihood estimator (MLE)

Data:

In WWII Allies captured n serial numbers of German tanks

$0 > s_1 > s_2 > \ldots > s_n$

Q: What was the total number of tanks?

A (Naive): using the uniform distribution

H: It is equally likely to "sample" any serial number.

H: Assuming continuous range.

H: Counts with replacement.

$$s_i \sim U_{[0,N]}$$

$$P(s_i \in [s, s+ds]) = \frac{ds}{N}, \quad s \in [0,N] .$$

Inference

Connecting experiments and models

Karl Pearson
(1857-1936)

Case of study:

German tank problem

Method of moment

Maximum likelihood estimator (MLE)

Data:

In WWII Allies captured n serial numbers of German tanks

$$0 > s_1 > s_2 > \ldots > s_n$$

Q: What was the total number of tanks?

A (Naive): using the uniform distribution

H: It is equally likely to "sample" any serial number.

H: Assuming continuous range.

H: Counts with replacement.

$$s_i \sim U_{[0,N]}$$

$$P(s_i \in [s, s+ds]) = \frac{ds}{N}, \quad s \in [0,N]\,.$$

Method of moments:

$$E(s_i) = \frac{N}{2}$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} s_i \longrightarrow \hat{N}_m = 2\hat{\mu}$$

Maximum likelihood estimator:

$$L\left(s_1, s_2, \ldots, s_n \,|\, N\right) = N^{-n}, \quad N \geq s_n, \quad 0 > s_1$$

$$\ell = \log(L) = -n\log(N)$$

$$\partial_N \ell = -\frac{n}{N} < 0 \longrightarrow \hat{N}_{MLE} = s_n$$

Inference

Connecting experiments and models

Karl Pearson
(1857-1936)

Case of study:

German tank problem

Method of moment

Maximum likelihood estimator (MLE)

Data:

In WWII Allies captured n serial numbers of German tanks

$$0 > s_1 > s_2 > \ldots > s_n$$

Q: What was the total number of tanks?

A (Naive): using the uniform distribution

H: It is equally likely to "sample" any serial number.

H: Assuming continuous range.

H: Counts with replacement.

$$s_i \sim U_{[0,N]}$$

$$P(s_i \in [s, s+ds]) = \frac{ds}{N}, \quad s \in [0,N]\,.$$

Method of moments:

$$E(s_i) = \frac{N}{2}$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} s_i \longrightarrow \hat{N}_m = 2\hat{\mu}$$

Maximum likelihood estimator:

$$L\left(s_1, s_2, \ldots, s_n \,|\, N\right) = N^{-n}, \quad N \geq s_n, \quad 0 > s_1$$

$$\ell = \log(L) = -n\log(N)$$

$$\partial_N \ell = -\frac{n}{N} < 0 \longrightarrow \hat{N}_{MLE} = s_n$$

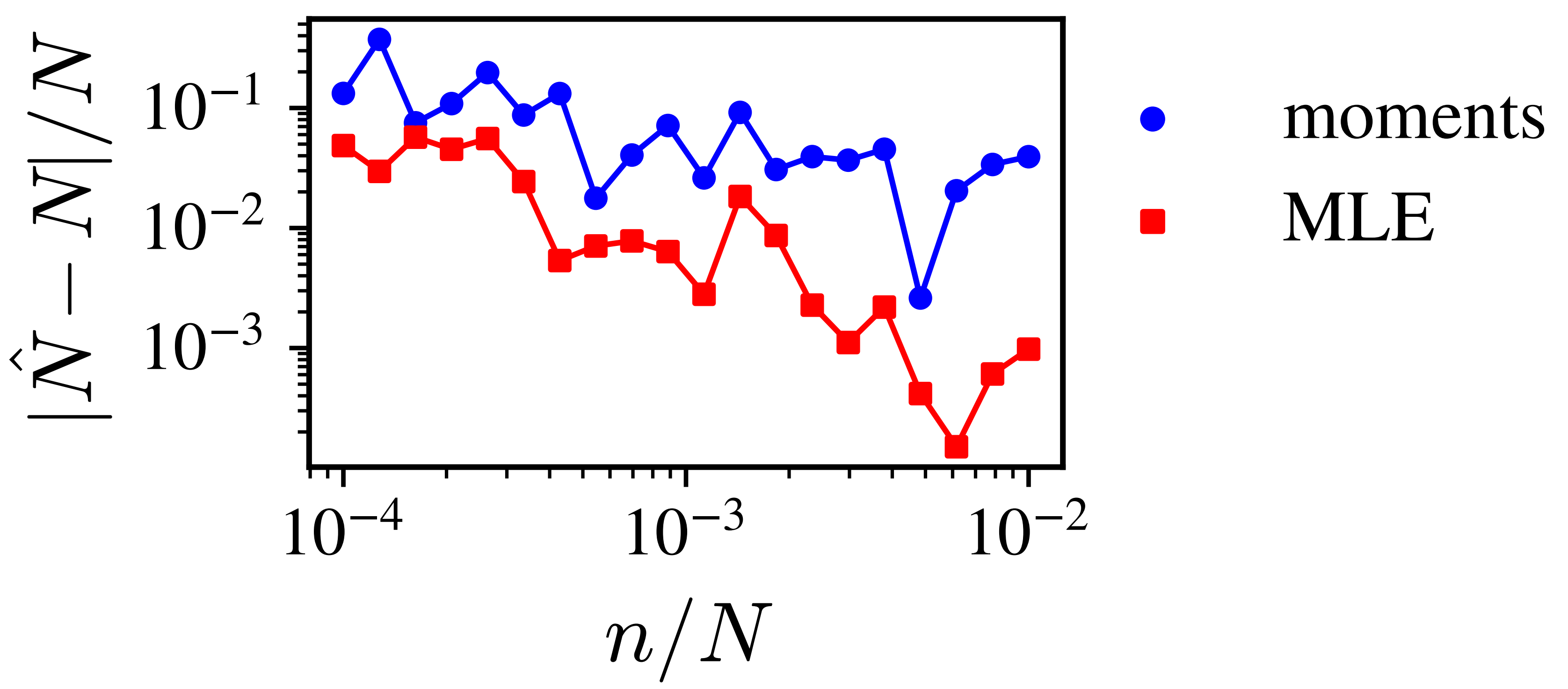
Comparing the methods:

Method of moments is unbiased

$$E(\hat{N}_m) = N$$

MLE is consistent

$$E(\hat{N}_{MLE}) = \frac{n}{n+1}N$$

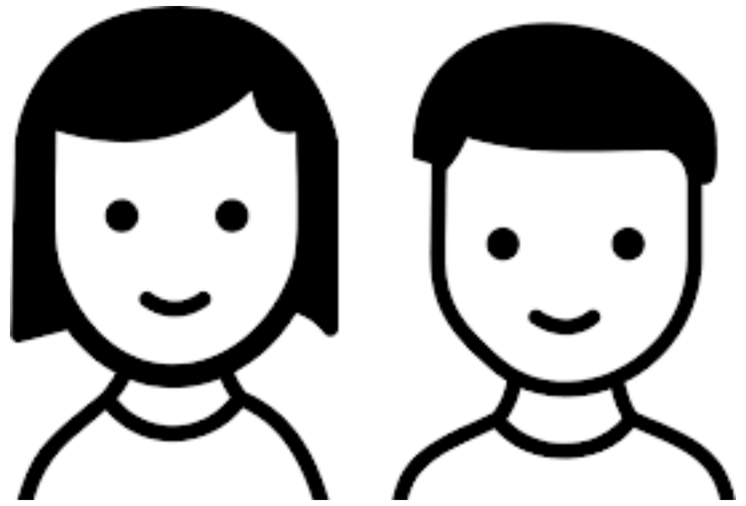


Inference
Connecting experiments and models

Pierre-Simon
Laplace
(1749-1827)

Case of study:

Asymmetry in birth rates
Bayesian inference
Beta distribution



Data:

There are more male than female newborns:
Laplace data: 251527 boys and 241945 girls

Q: How **robust** is this observation?

Inference

Connecting experiments and models

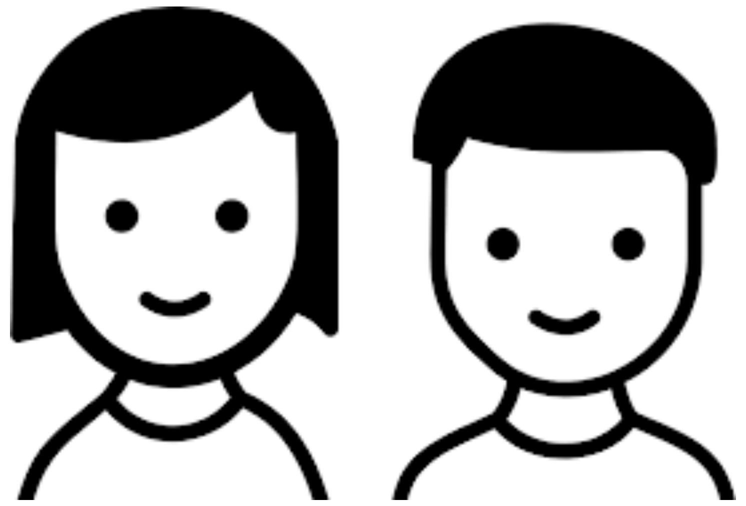
Pierre-Simon Laplace (1749-1827)

Case of study:

Asymmetry in birth rates

Bayesian inference

Beta distribution



Data:

There are more male than female newborns:

Laplace data: 251527 boys and 241945 girls

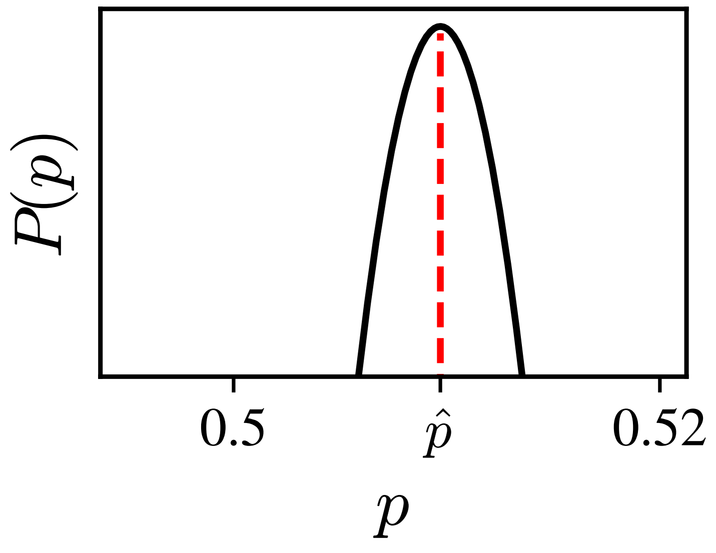
Q: How **robust** is this observation?

$$N = N_b + N_g = 493472$$

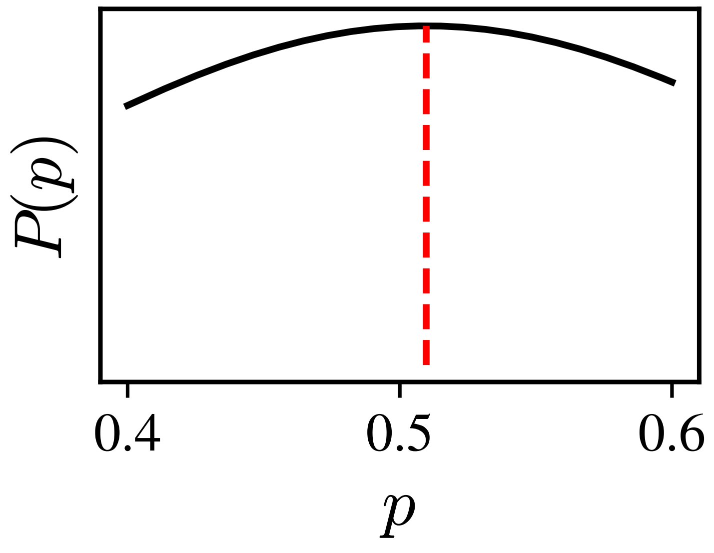
$$N_b \sim B(N, p)$$

$$E(N_b) = Np \rightarrow \hat{p} = N_b/N = 0.5097$$

Real feature



Due to fluctuations



Inference
Connecting experiments and models

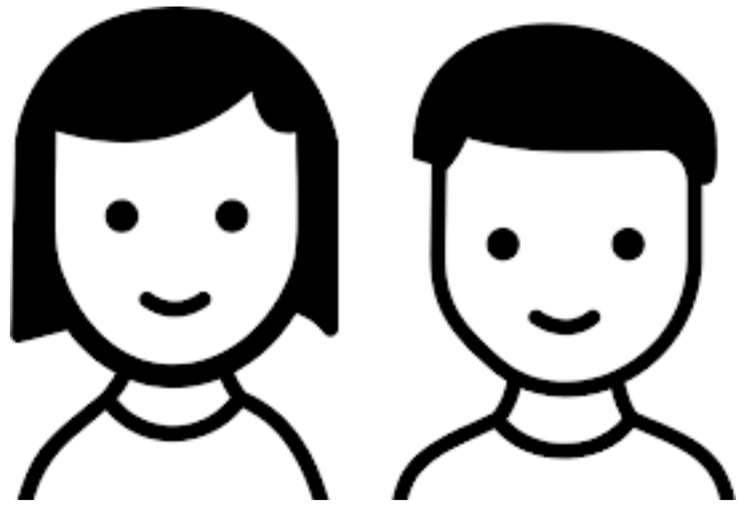
Pierre-Simon Laplace (1749-1827)

Case of study:

Asymmetry in birth rates

Bayesian inference

Beta distribution



Data:
There are more male than female newborns:
Laplace data: 251527 boys and 241945 girls

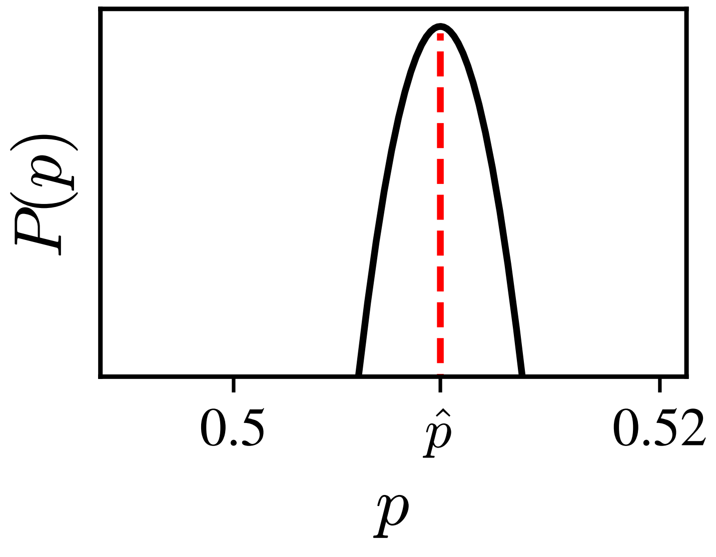
Q: How **robust** is this observation?

$N = N_b + N_g = 493472$

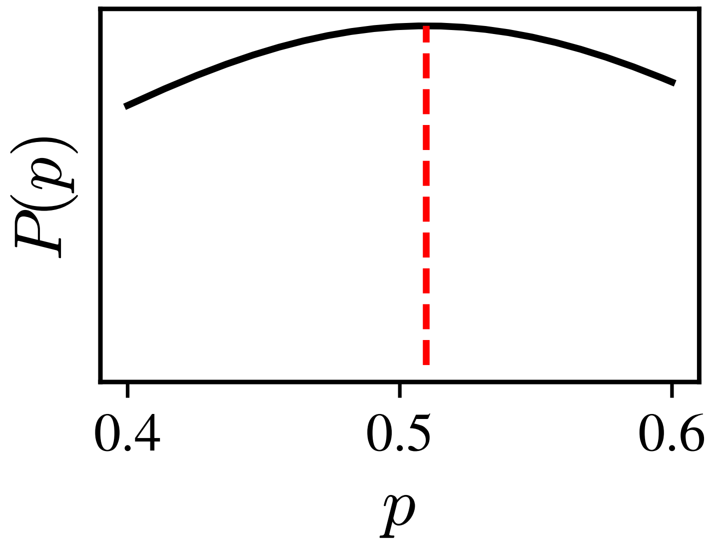
$N_b \sim B(N, p)$

$E(N_b) = Np \rightarrow \hat{p} = N_b/N = 0.5097$

Real feature



Due to fluctuations



Bayes method

Posterior, Likelihood, Prior, Normalization

$$\boxed{P\left(p\,|\,N_b\right)} = \boxed{P\left(N_b\,|\,p\right)} \frac{\boxed{P\left(p\right)}}{\boxed{P\left(N_b\right)}}$$

Prior	$P\left(p\right) = 1, \quad p \in [0,1]$	**Assumed:** based on experimenter's belief.
Likelihood	$P\left(N_b\,|\,p\right) = p^{N_b}(1-p)^{N-N_b}\binom{N}{N_b}$	**Known:** model evaluated on data.
Normalization	$P\left(N_b\right) = \int dp\, P(p)\, P(N_b\,|\,p)$	**Computed:** independent of p.

Inference

Connecting experiments and models

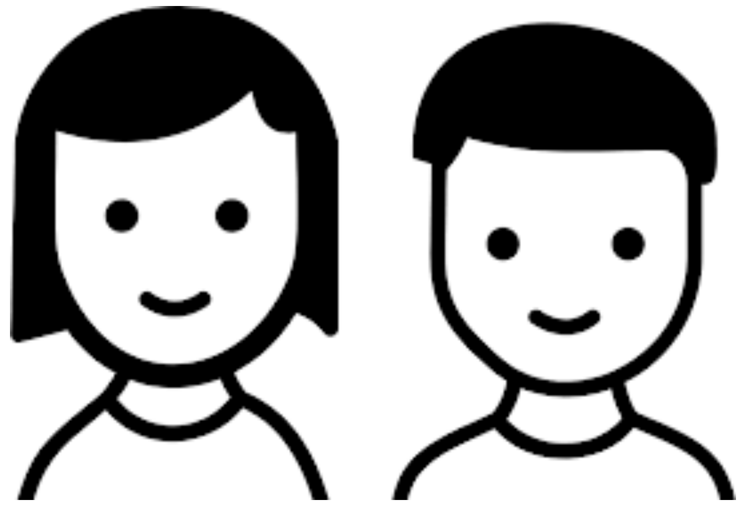
Pierre-Simon Laplace (1749-1827)

Case of study:

Asymmetry in birth rates

Bayesian inference

Beta distribution



Data:

There are more male than female newborns:

Laplace data: 251527 boys and 241945 girls

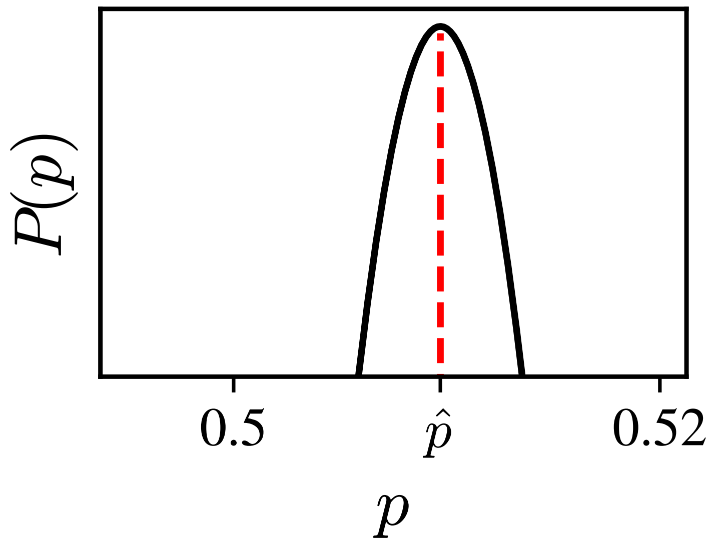
Q: How **robust** is this observation?

$N = N_b + N_g = 493472$

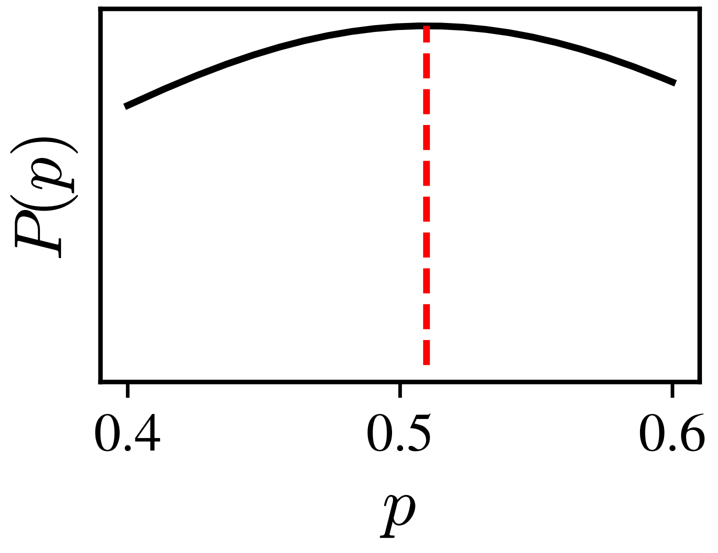
$N_b \sim B(N, p)$

$E(N_b) = Np \rightarrow \hat{p} = N_b/N = 0.5097$

Real feature



Due to fluctuations



Bayes method

Posterior Likelihood Prior Normalization

$$\boxed{P(p|N_b)} = \boxed{P(N_b|p)} \frac{\boxed{P(p)}}{\boxed{P(N_b)}}$$

Prior $P(p) = 1, \quad p \in [0,1]$ **Assumed:** based on experimenter's belief.

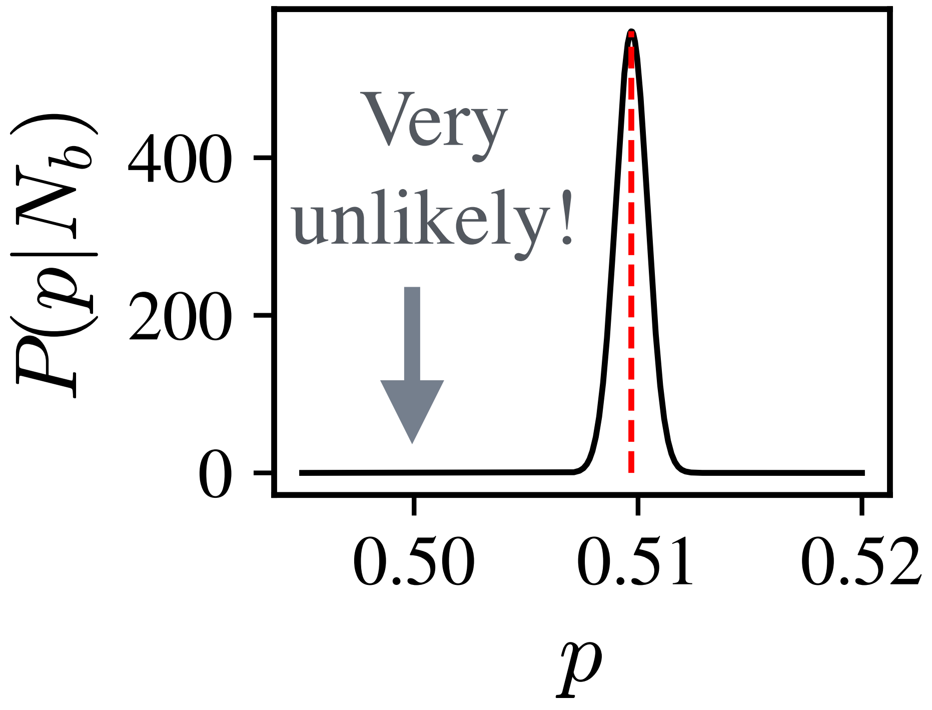
Likelihood $P(N_b|p) = p^{N_b}(1-p)^{N-N_b}\binom{N}{N_b}$ **Known:** model evaluated on data.

Normalization $P(N_b) = \int dp\, P(p)\, P(N_b|p)$ **Computed:** independent of p.

Uniform prior (no prior information) $\rightarrow$ $P(p|N_b) = Z^{-1} p^{N_b}(1-p)^{N-N_b}$ (Beta distribution)

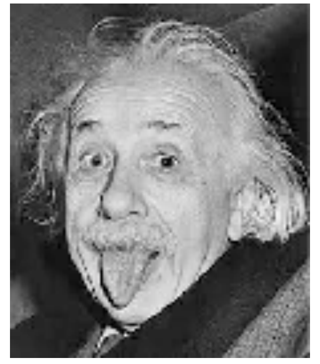
We can trust the result!!

$P(p <= 0.5) \sim 10^{-42}$



Stochastic processes

Marian Smoluchowski (1872-1917)



Albert Einstein (1879-1955)

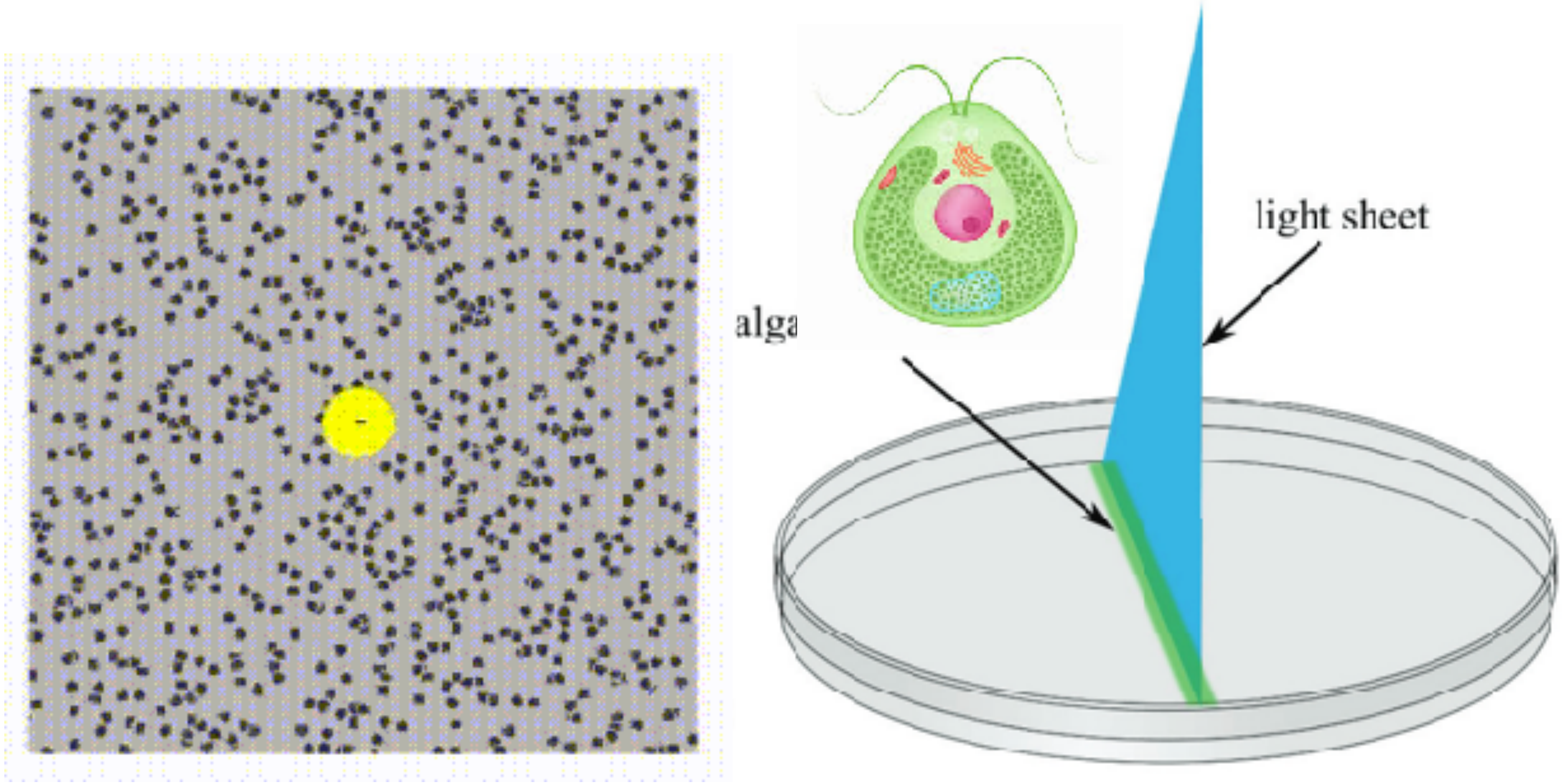
Robert Brown (1857-1936)

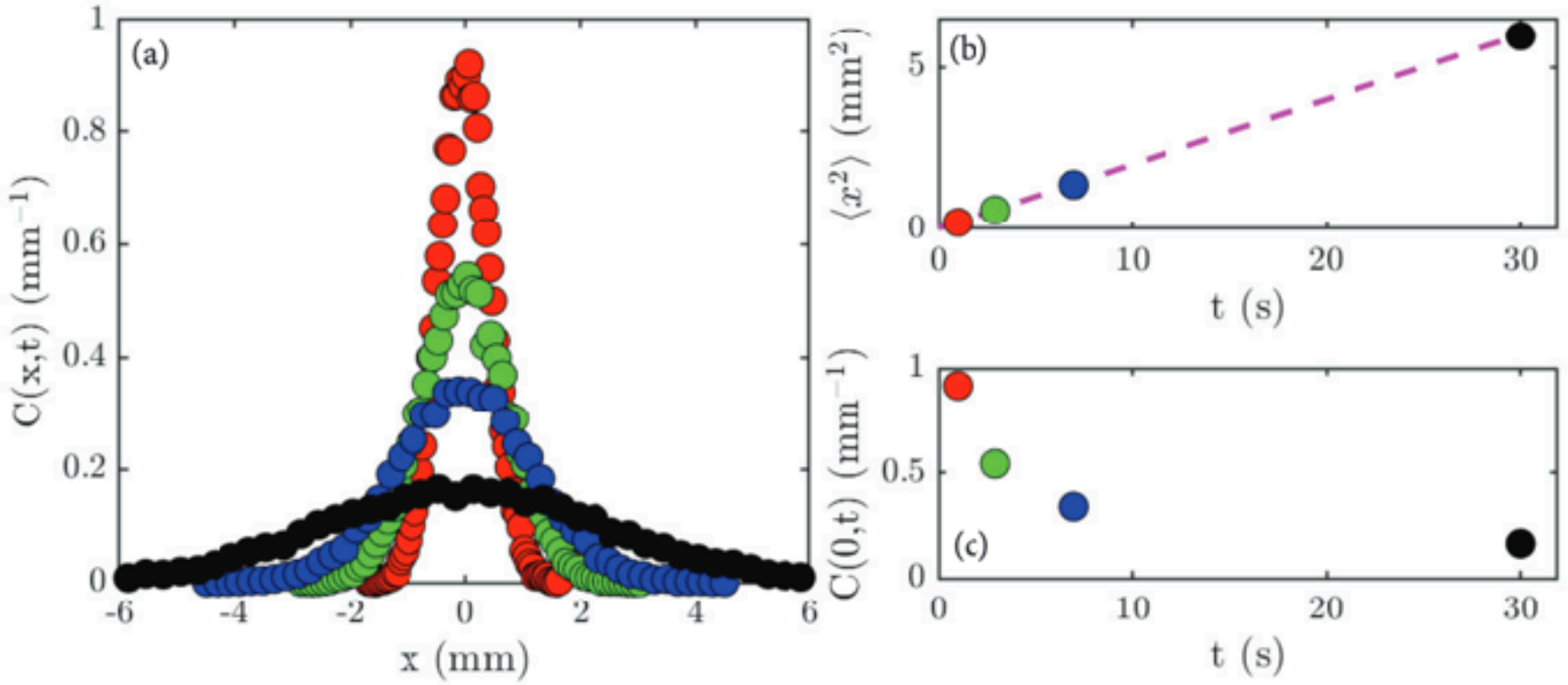
Case of study:

Diffusion, random walks

Markov chains, Fokker-Planck eq.

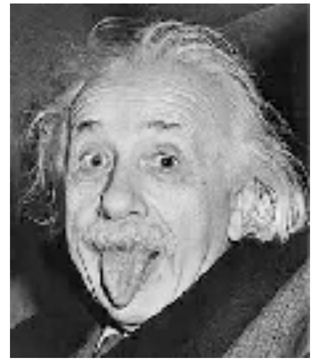
Q: Origin of statistical patterns in erratic movements?





Stochastic processes

Marian Smoluchowski (1872-1917)



Albert Einstein (1879-1955)

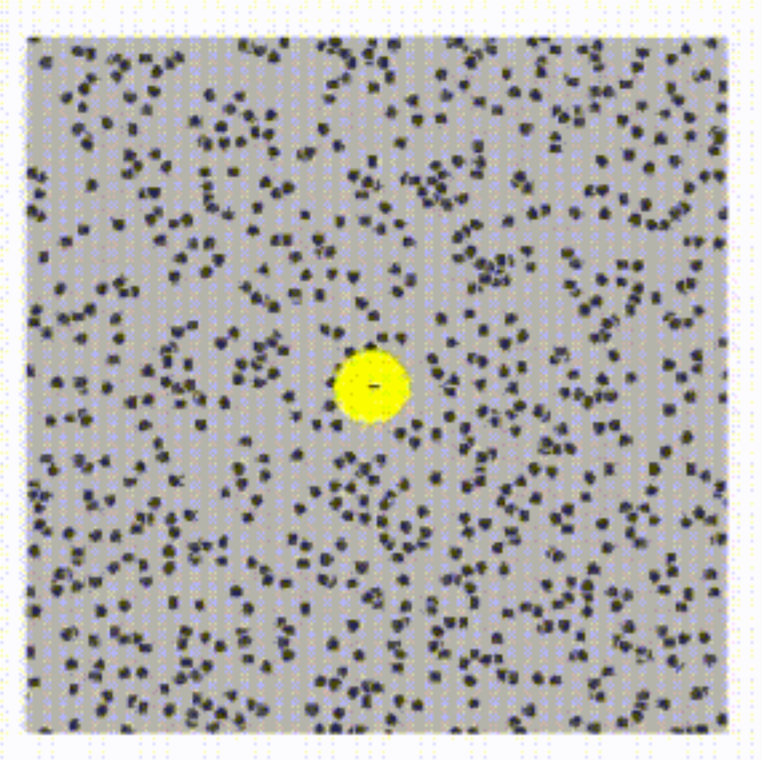
Robert Brown (1857-1936)

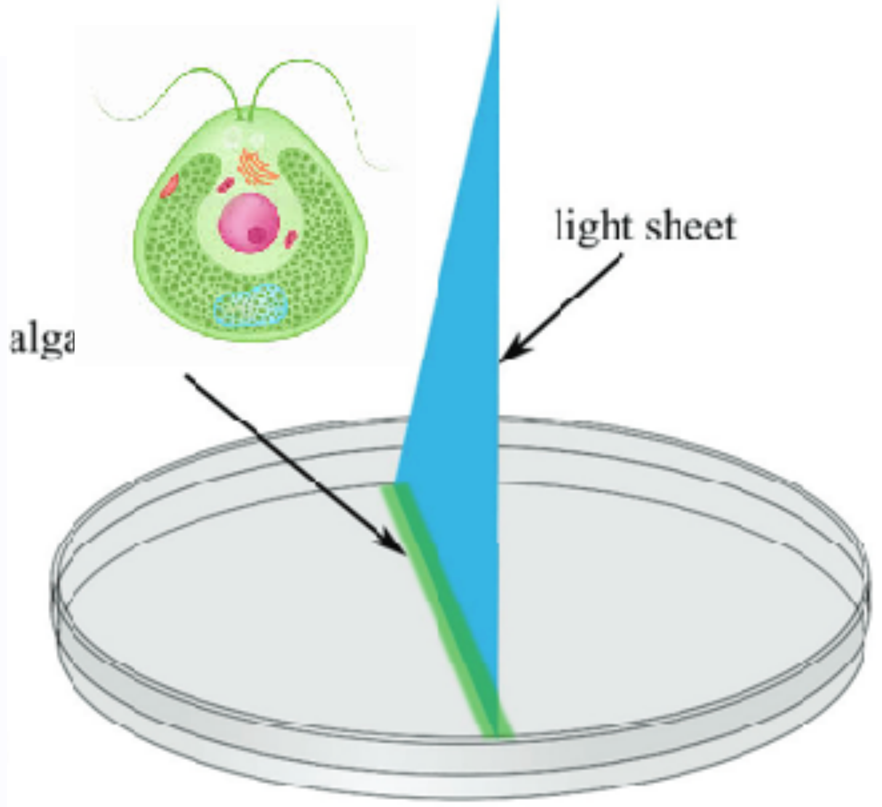
Case of study:

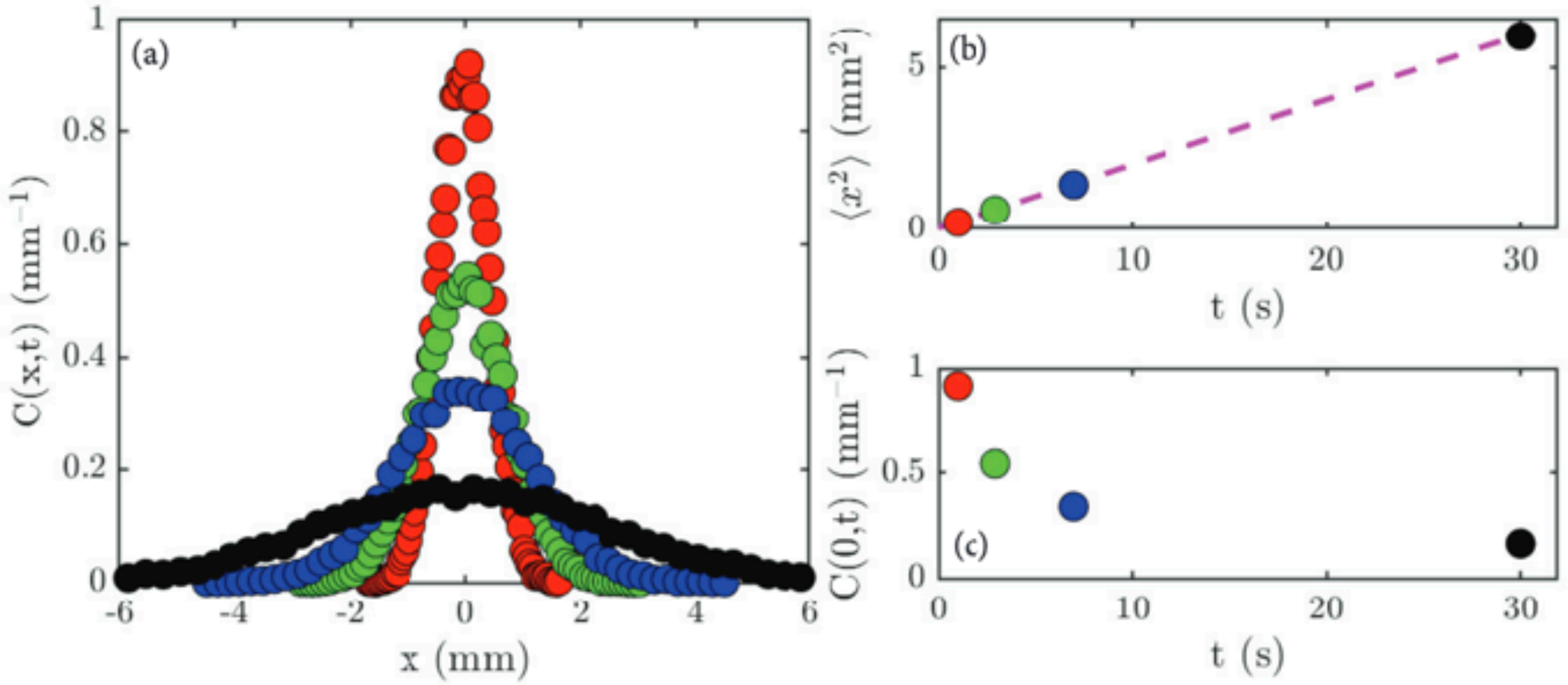
Diffusion, random walks

Markov chains, Fokker-Planck eq.

Q: Origin of statistical patterns in erratic movements?







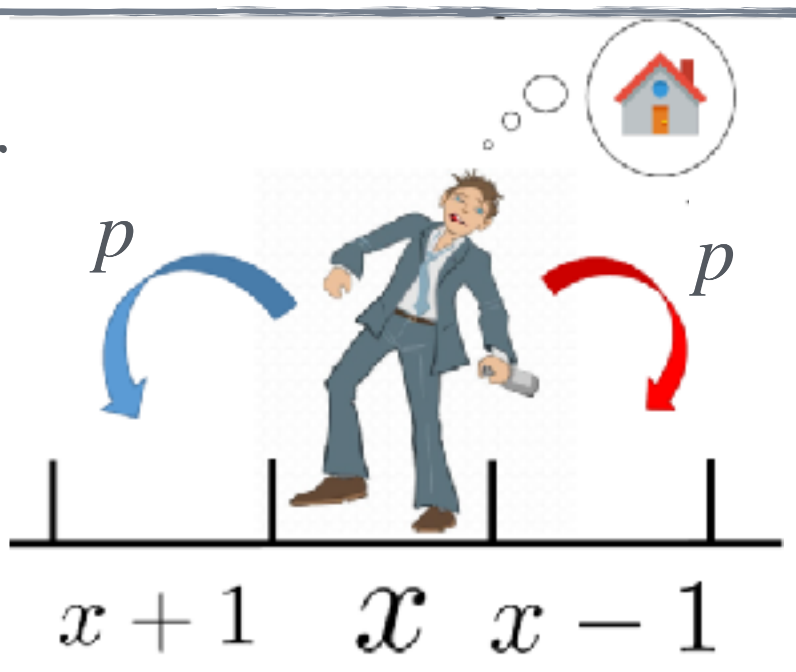
Random walk

$X_t \longrightarrow$ Position of the (drunk) walker

Transition probabilities only depend on current state: **Markov process**

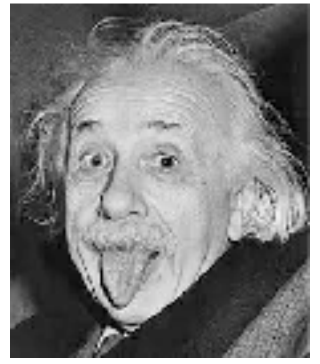
$p = P(X_{t+\Delta t} = x \pm \Delta x \,|\, X_t = x)$ $\qquad 1 - p = P(X_{t+\Delta t} = x \,|\, X_t = x)$

$P_t(x) = P(X_t = x \,|\, X_0 = x_0)$



Stochastic processes

Marian Smoluchowski (1872-1917)



Albert Einstein (1879-1955)

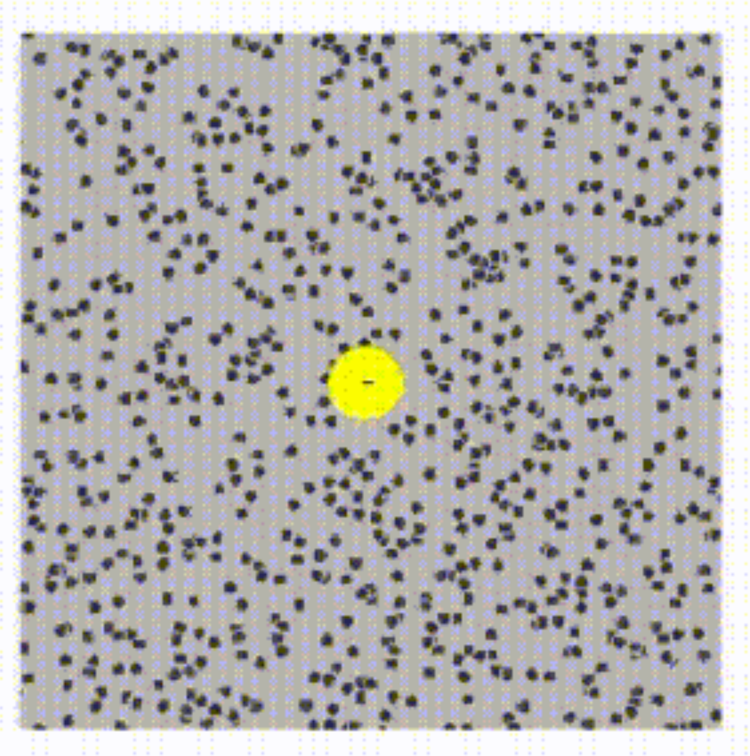
Robert Brown (1857-1936)

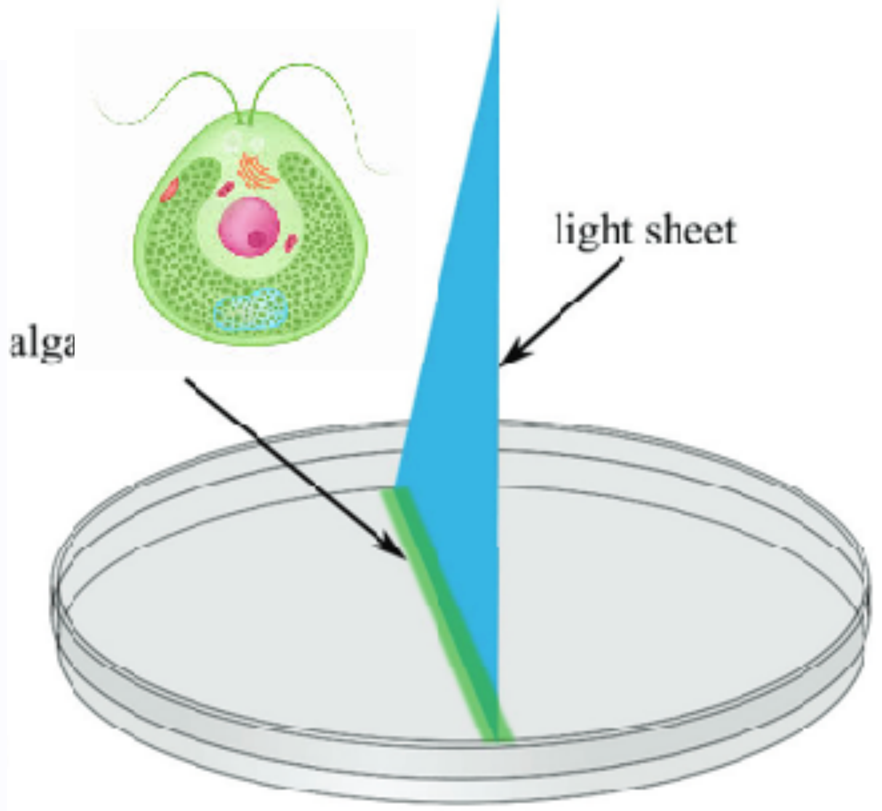
Case of study:

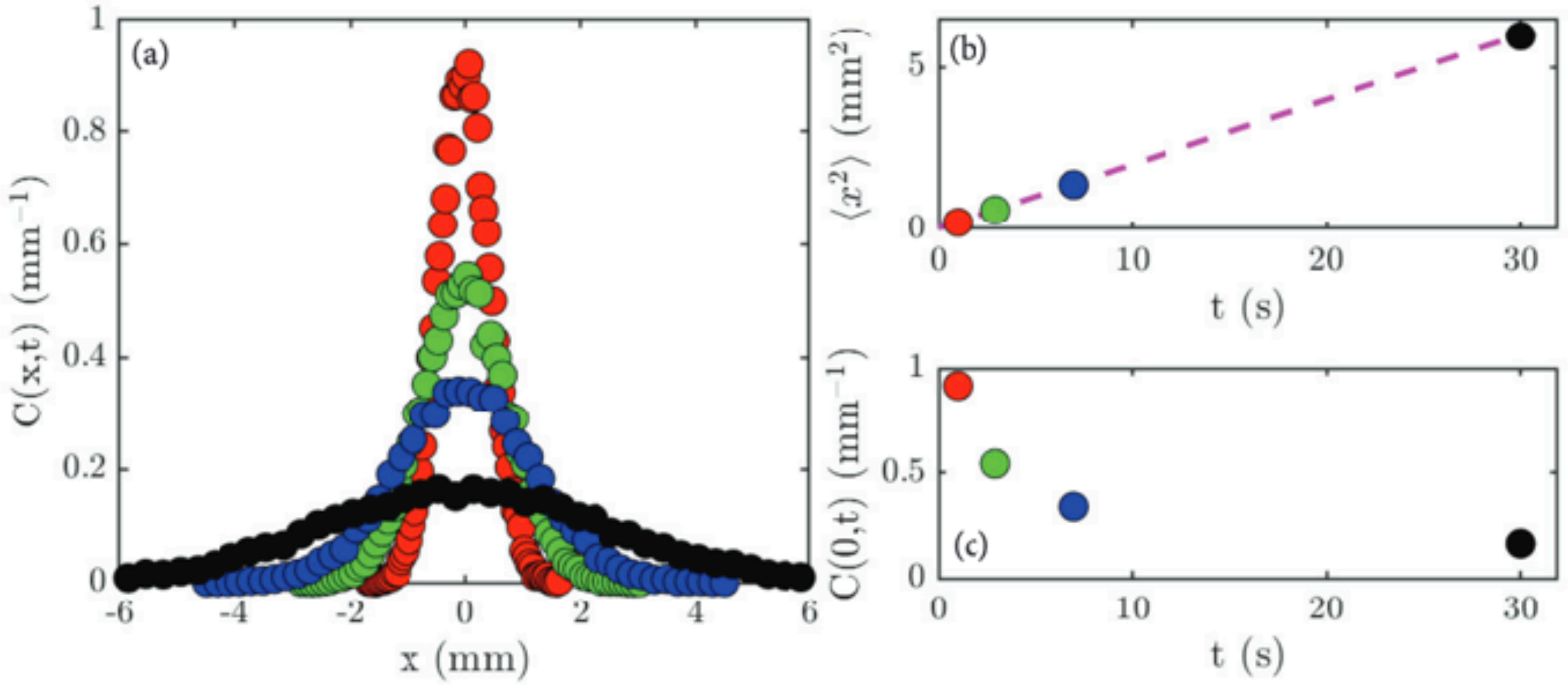
Diffusion, random walks

Markov chains, Fokker-Planck eq.

Q: Origin of statistical patterns in erratic movements?

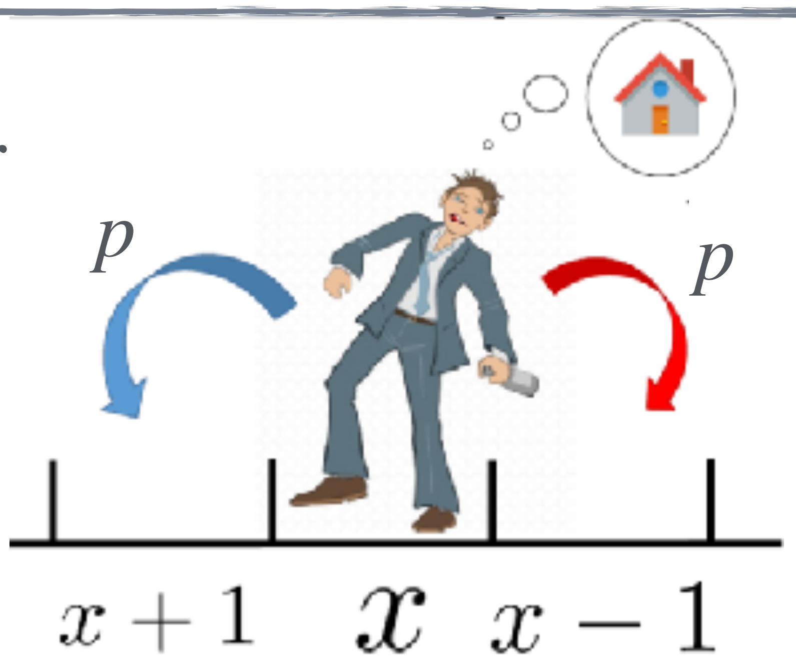






Random walk

$X_t \longrightarrow$ Position of the (drunk) walker



Transition probabilities only depend on current state: **Markov process**

$p = P(X_{t+\Delta t} = x \pm \Delta x \,|\, X_t = x) \qquad 1 - p = P(X_{t+\Delta t} = x \,|\, X_t = x)$

$P_t(x) = P(X_t = x \,|\, X_0 = x_0)$

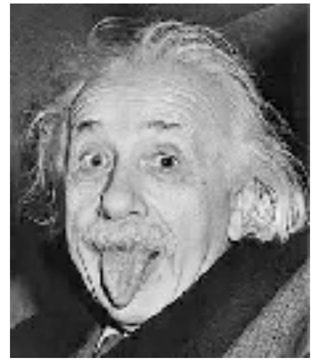
$$P_{t+\Delta t}(x) = \sum_y P_t(y) P(X_{t+\Delta t} = x \,|\, X_t = y) = (1-p)P_t(x) + pP_t(x+\Delta x) + qP_t(x - \Delta x)$$

Diffusion eq. Jumping prob. Per unit of time

Continuum limit $\longrightarrow$ $\boxed{\partial_t P_t(x) = \frac{D^2}{2}\partial_x^2 P_t(x)}$ $\quad D^2 = \boxed{\frac{p}{\Delta t}}\boxed{\Delta x^2}$ Size of jumps squared

Stochastic processes

Marian Smoluchowski (1872-1917)



Albert Einstein (1879-1955)

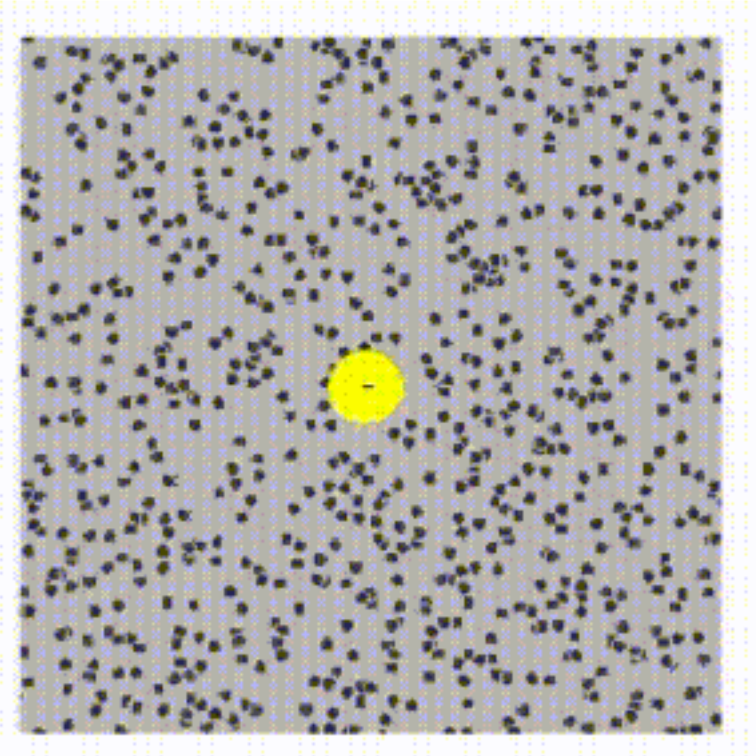
Robert Brown (1857-1936)

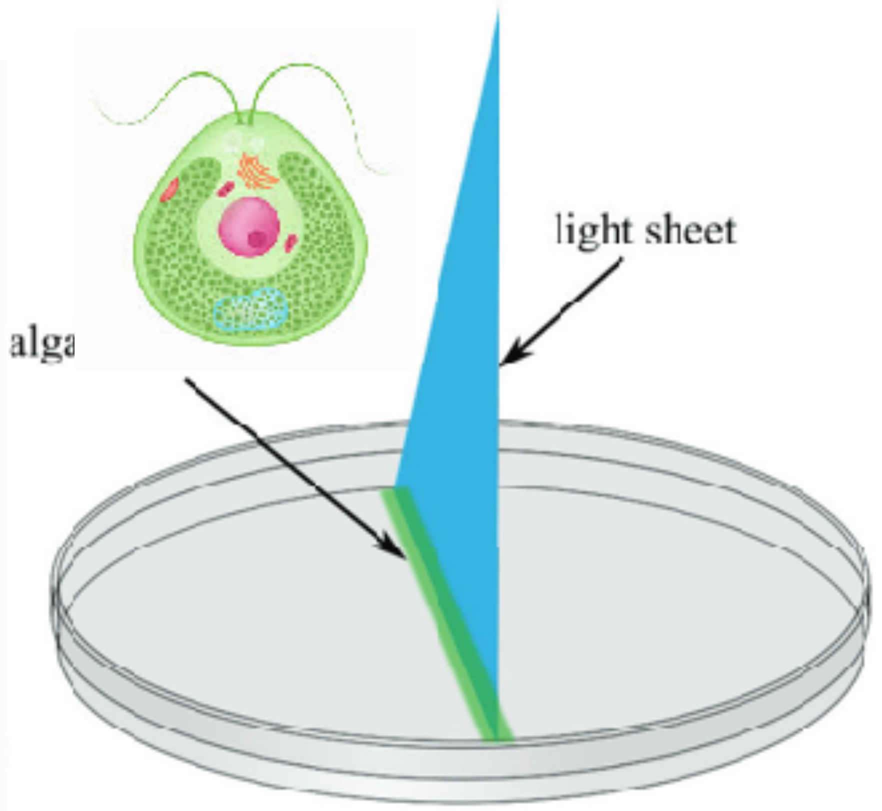
Case of study:

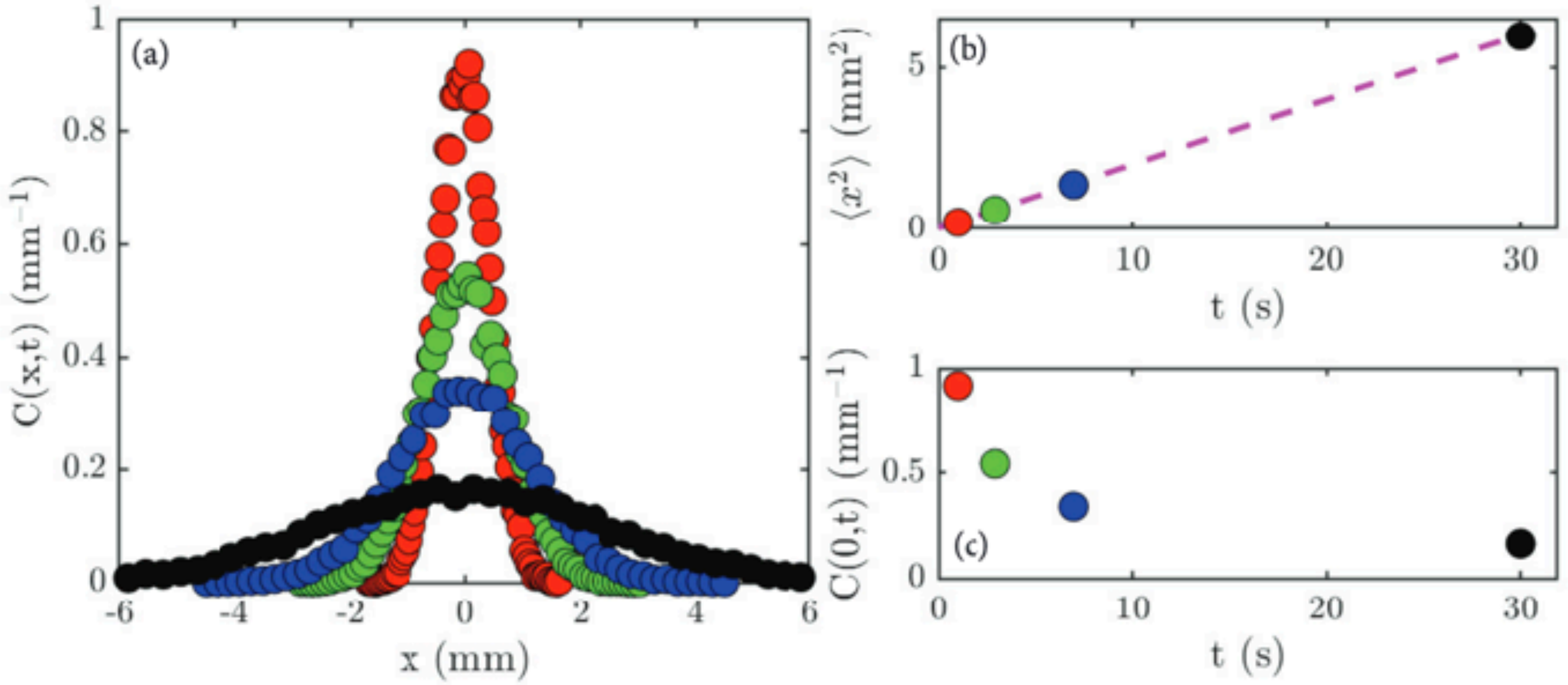
Diffusion, random walks

Markov chains, Fokker-Planck eq.

Q: Origin of statistical patterns in erratic movements?

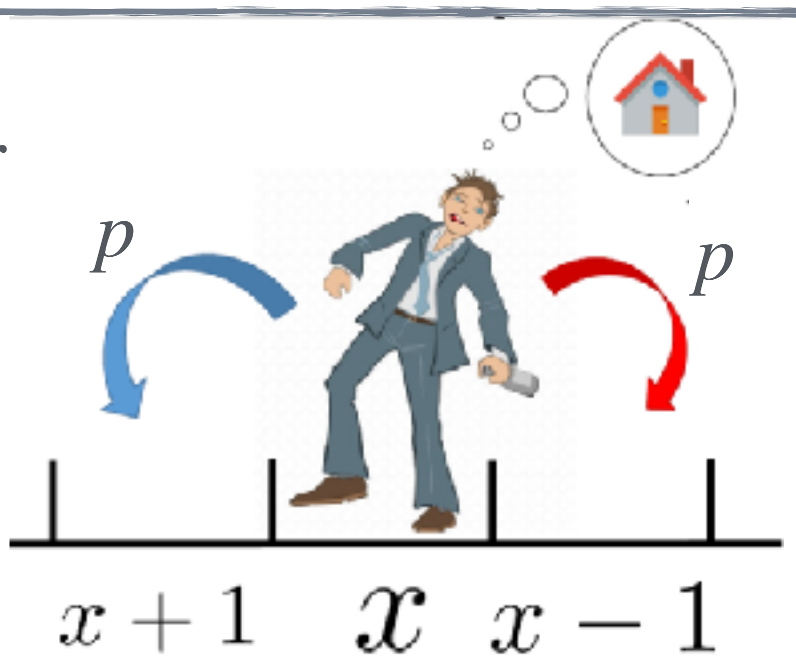






Random walk

$X_t \longrightarrow$ Position of the (drunk) walker



Transition probabilities only depend on current state: **Markov process**

$p = P(X_{t+\Delta t} = x \pm \Delta x \,|\, X_t = x) \qquad 1 - p = P(X_{t+\Delta t} = x \,|\, X_t = x)$

$P_t(x) = P(X_t = x \,|\, X_0 = x_0)$

$$P_{t+\Delta t}(x) = \sum_y P_t(y) P(X_{t+\Delta t} = x \,|\, X_t = y) = (1-p)P_t(x) + pP_t(x+\Delta x) + qP_t(x - \Delta x)$$

Diffusion eq.

Jumping prob. Per unit of time

Continuum limit $\longrightarrow$ $\boxed{\partial_t P_t(x) = \dfrac{D^2}{2}\partial_x^2 P_t(x)}$ $\quad D^2 = \boxed{\dfrac{p}{\Delta t}}\boxed{\Delta x^2}$

Size of jumps squared

$$P_t(x) = \frac{1}{\sqrt{2\pi D^2 t}} e^{\frac{(x-x_0)^2}{2D^2 t}} \longrightarrow \begin{array}{l} \langle X_t \rangle = x_0 \\ \sigma^2(X_t) = D^2 t \\ \max\limits_x \{P_t(x)\} = t^{-\frac{1}{2}} \end{array} \longrightarrow \boxed{\Delta l \sim D\sqrt{t}}$$

Signature of Diffusion

Stochastic processes

Norbert Wiener (1894-1964)

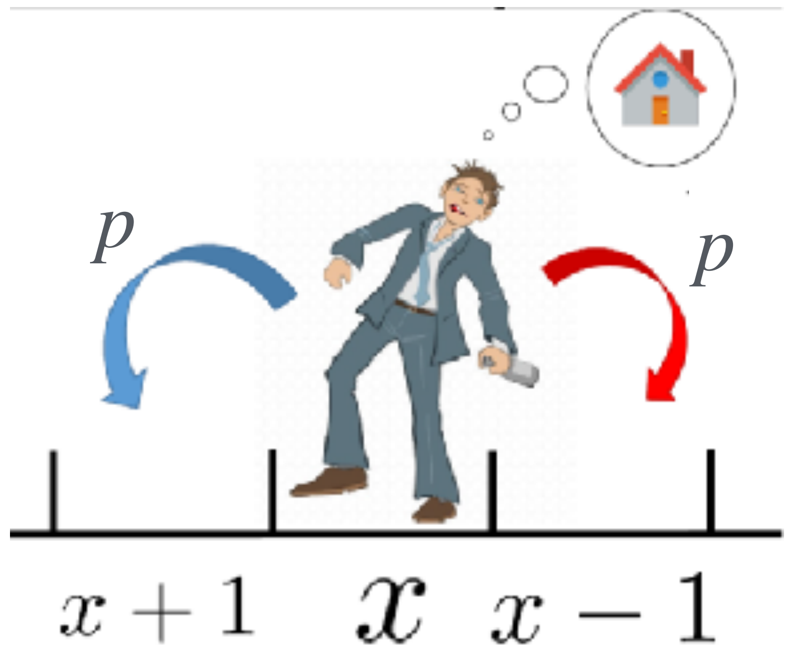
Paul Langevin (1872-1946)

Case of study:

Diffusion, random walks

Markov chains, Fokker-Planck eq.

Discrete space and time



Continuous space and time

$$\partial_t P_t(x) = \frac{D^2}{2}\partial_x^2 P_t(x)$$

Trajectories??

$$\frac{d}{dt}X_t = ??$$

Stochastic processes

Norbert Wiener (1894-1964)

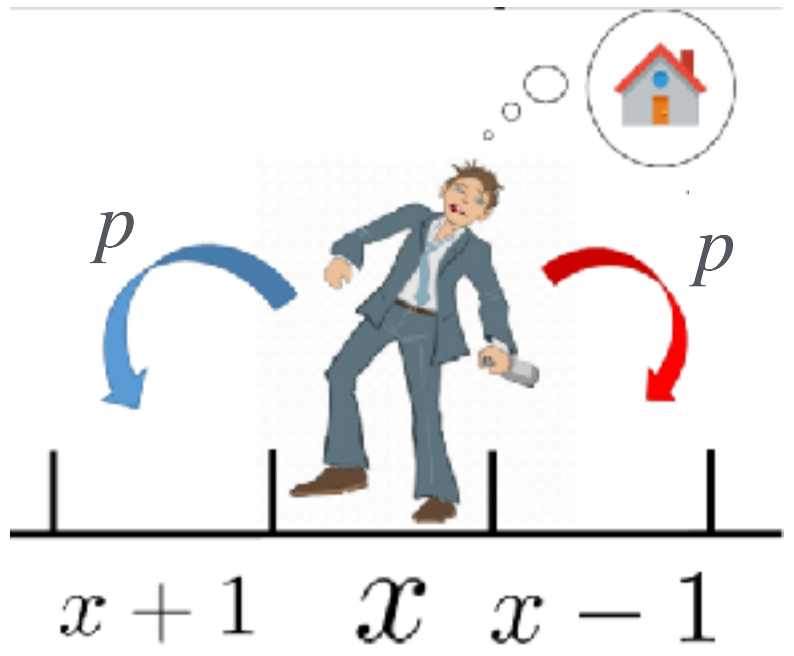
Paul Langevin (1872-1946)

Case of study:

Diffusion, random walks

Markov chains, Fokker-Planck eq.

Discrete space and time



Continuous space and time

$$\partial_t P_t(x) = \frac{D^2}{2}\partial_x^2 P_t(x)$$

Trajectories??

$$\frac{d}{dt}X_t = ??$$

Deterministic dynamics:

ODE

$$\frac{d}{dt}X_t = A(X_t)$$

$$X_{t+\Delta t} = X_t + A(X_t)\Delta t$$

Euler Method

Stochastic processes

Norbert Wiener (1894-1964)

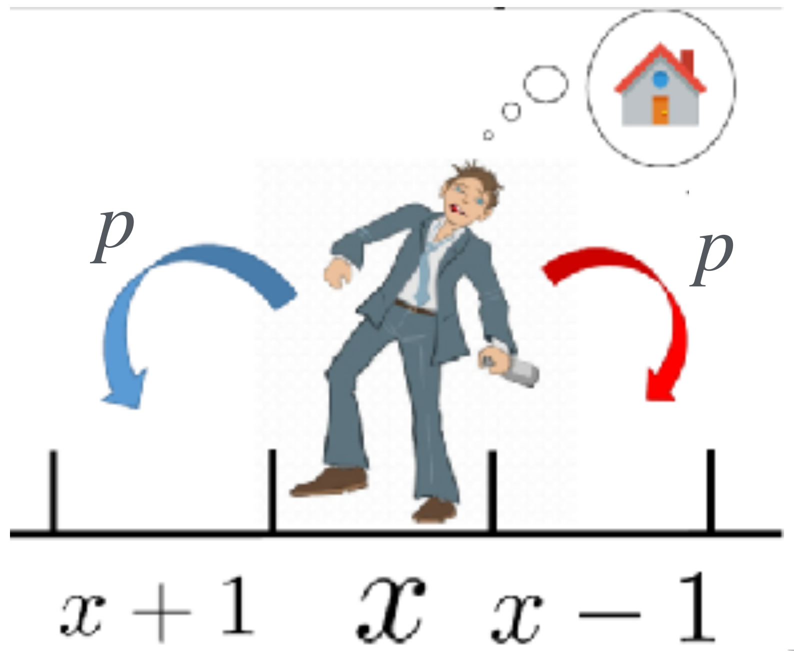
Paul Langevin (1872-1946)

Case of study:

Diffusion, random walks

Markov chains, Fokker-Planck eq.

Discrete space and time



Continuous space and time

$$\partial_t P_t(x) = \frac{D^2}{2}\partial_x^2 P_t(x)$$

Trajectories??

$$\frac{d}{dt}X_t = ??$$

Deterministic dynamics:

ODE

$$\frac{d}{dt}X_t = A(X_t)$$

$$X_{t+\Delta t} = X_t + A(X_t)\Delta t$$

Euler Method

Stochastic dynamics:

Euler-Maruyama Method

$$\frac{d}{dt}X_t = A(X_t) + B(X_t)\xi(t) \quad \rightarrow \quad X_{t+\Delta t} = X_t + A(X_t)\Delta t + B(X_t)\,\Delta W_t$$

$$\int_t^{t+\Delta t}\xi(t) = \Delta W_t \sim G(0,\Delta t) \quad \rightarrow \quad P(\Delta W_t \in [w, w+dw]) = \frac{1}{2\pi\Delta t}e^{-\frac{w^2}{2\Delta t}}dw$$

Stochastic processes

Norbert Wiener (1894-1964)

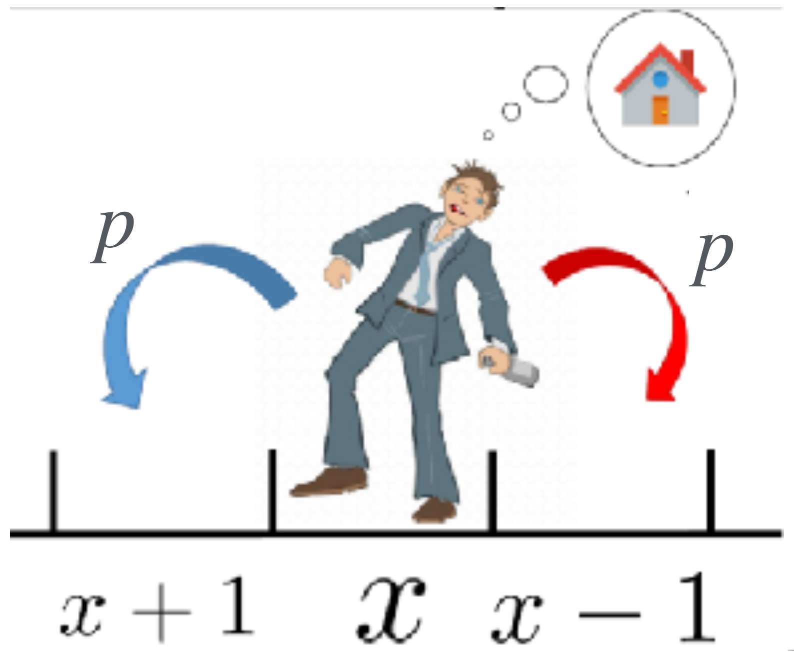
Paul Langevin (1872-1946)

Case of study:

Diffusion, random walks

Markov chains, Fokker-Planck eq.

Discrete space and time



Continuous space and time

$$\partial_t P_t(x) = \frac{D^2}{2} \partial_x^2 P_t(x)$$

Trajectories??

$$\frac{d}{dt} X_t = ??$$

Deterministic dynamics:

ODE

$$\frac{d}{dt} X_t = A(X_t)$$

$$\downarrow$$

$$X_{t+\Delta t} = X_t + A(X_t)\Delta t$$

Euler Method

Stochastic dynamics:

Euler-Maruyama Method

$$\frac{d}{dt} X_t = A(X_t) + B(X_t)\xi(t) \longrightarrow X_{t+\Delta t} = X_t + A(X_t)\Delta t + B(X_t)\,\Delta W_t$$

$$\int_t^{t+\Delta t} \xi(t) = \Delta W_t \sim G(0,\Delta t) \longrightarrow P(\Delta W_t \in [w, w+dw]) = \frac{1}{2\pi\Delta t} e^{-\frac{w^2}{2\Delta t}} dw$$

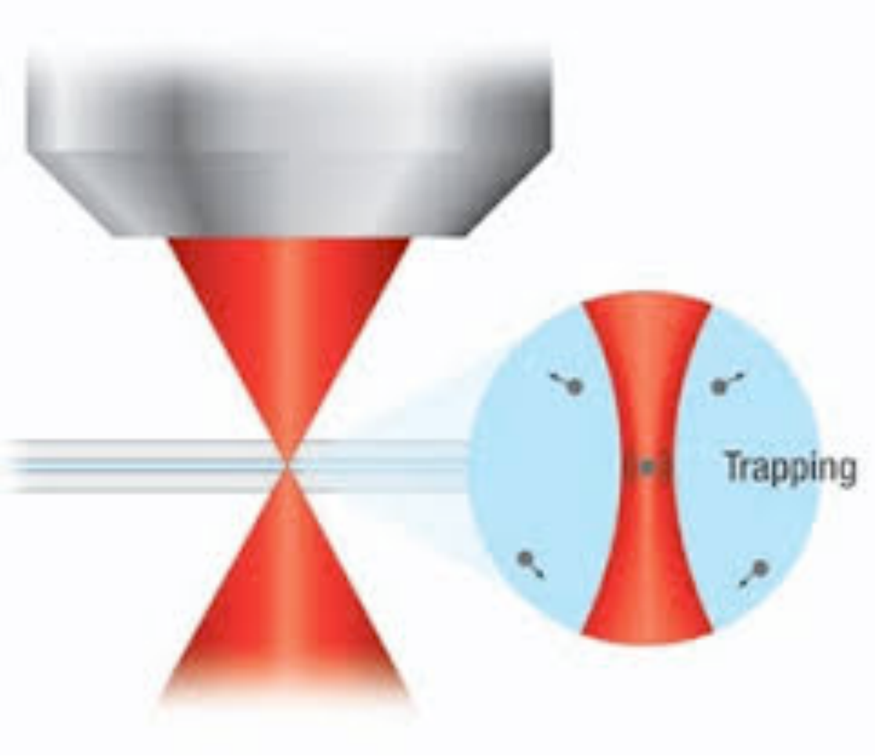
SDE

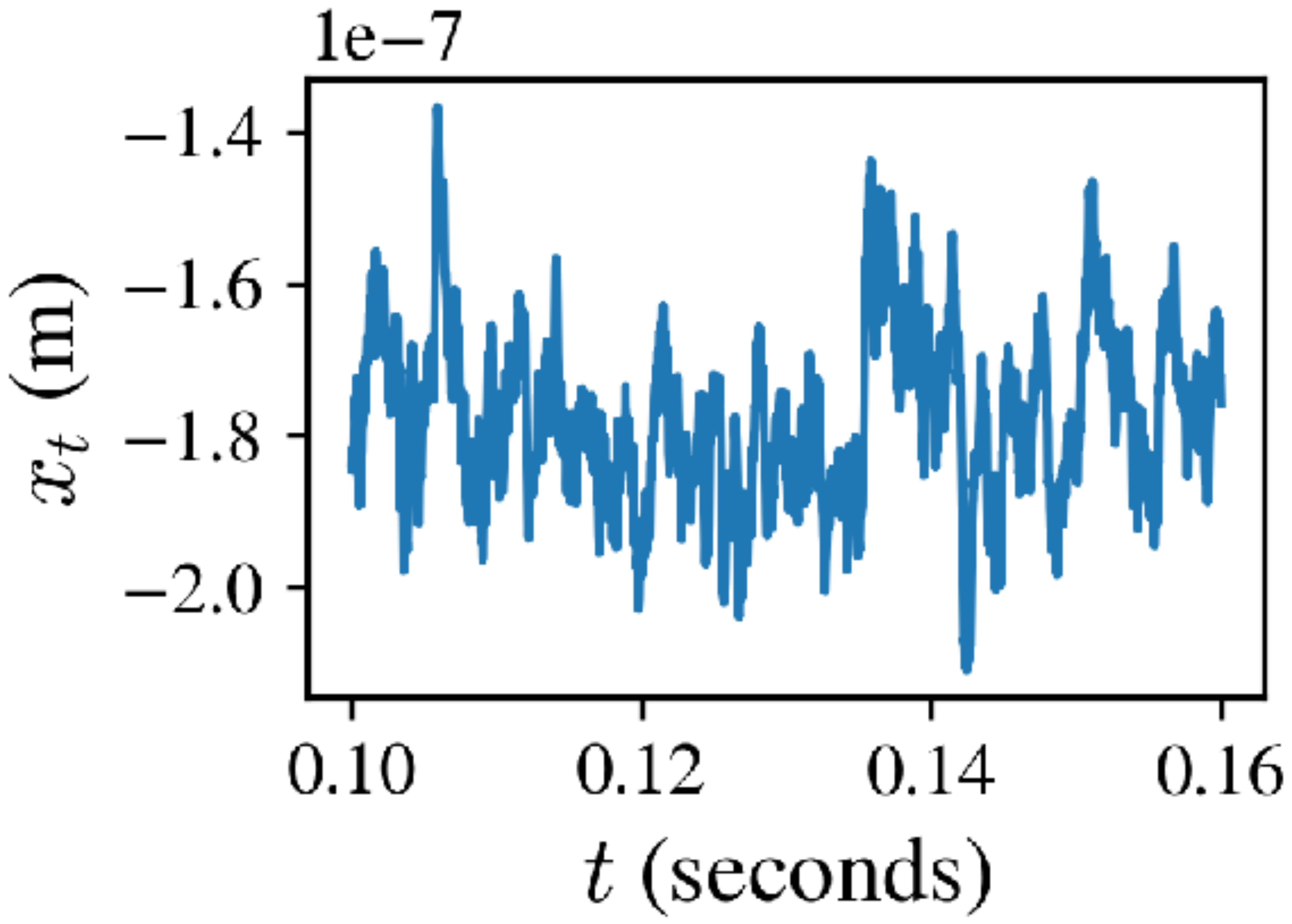
Fokker-Planck eq.

$$\frac{d}{dt} X_t = A(X_t) + B(X_t)\xi(t) \longleftrightarrow \partial_t \rho_t(x) = -\,\partial_x \left(A(x)\,\rho_t(x)\right) + \partial_x^2 \left(B^2(x)\rho_t(x)\right)$$

Optical tweezers

Ibáñez et.al. 2024





Miguel Ibáñez

Ornstein-Uhlenbeck process

$$\frac{d}{dt}X_t = -k(X_t - \mu) + D\xi(t) \longrightarrow \partial_t \rho_t(x) = k\partial_x \left((x-\mu)\,\rho_t(x)\right) + \frac{D^2}{2}\partial_x^2 \rho_t(x)$$

Stationary properties (with zero current)

$$\lim_{t\to\infty} \rho_t(x) = \rho(x) \longrightarrow 0 = \partial_x \left((x-\mu)\,\rho_t(x)\right) + \frac{D^2}{2k}\partial_x^2 \rho_t(x) \longrightarrow \rho(x) = \sqrt{\frac{2k}{2\pi D^2}}\exp\left(-\frac{2k(x-\mu)^2}{2D^2}\right)$$

In general (equilibrium stationary distributions)

$$\partial_t \rho_t(x) = -\partial_x \left(A(x)\rho_t(x)\right) + \partial_x^2 \left(B(x)\rho_t(x)\right) \longrightarrow \rho(x) = \frac{1}{B(x)}\exp\left(-\int^x \frac{A(s)}{B(s)}ds\right)$$

Probability

Modeling randomness

William Feller
(1906 – 1970)

Case of study:

Waiting time paradox

Exponential distribution
Poisson distribution

Exponential distribution

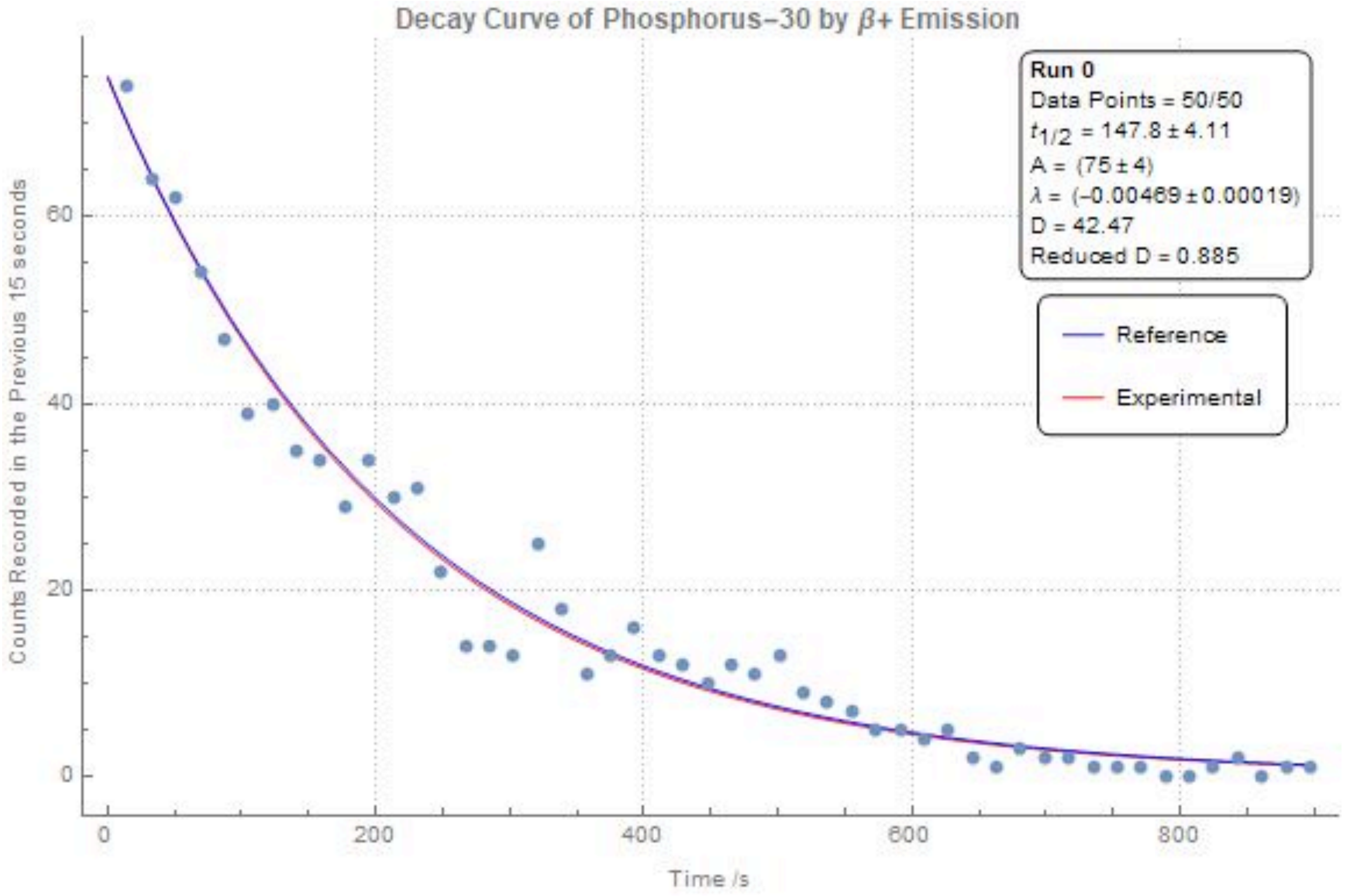
$$P(\Delta t) = w e^{-w\Delta t}$$

$$\boxed{E(\Delta t) = \frac{1}{w}}$$

Typical time scale

Memoryless property: $\dfrac{P(\Delta t)}{P(0)} = \dfrac{P(\Delta t + t)}{P(t)}$

Example: Radioactive decay



Probability
Modeling randomness

William Feller
(1906 – 1970)

Case of study:

Waiting time paradox

Exponential distribution

Poisson distribution

Exponential distribution

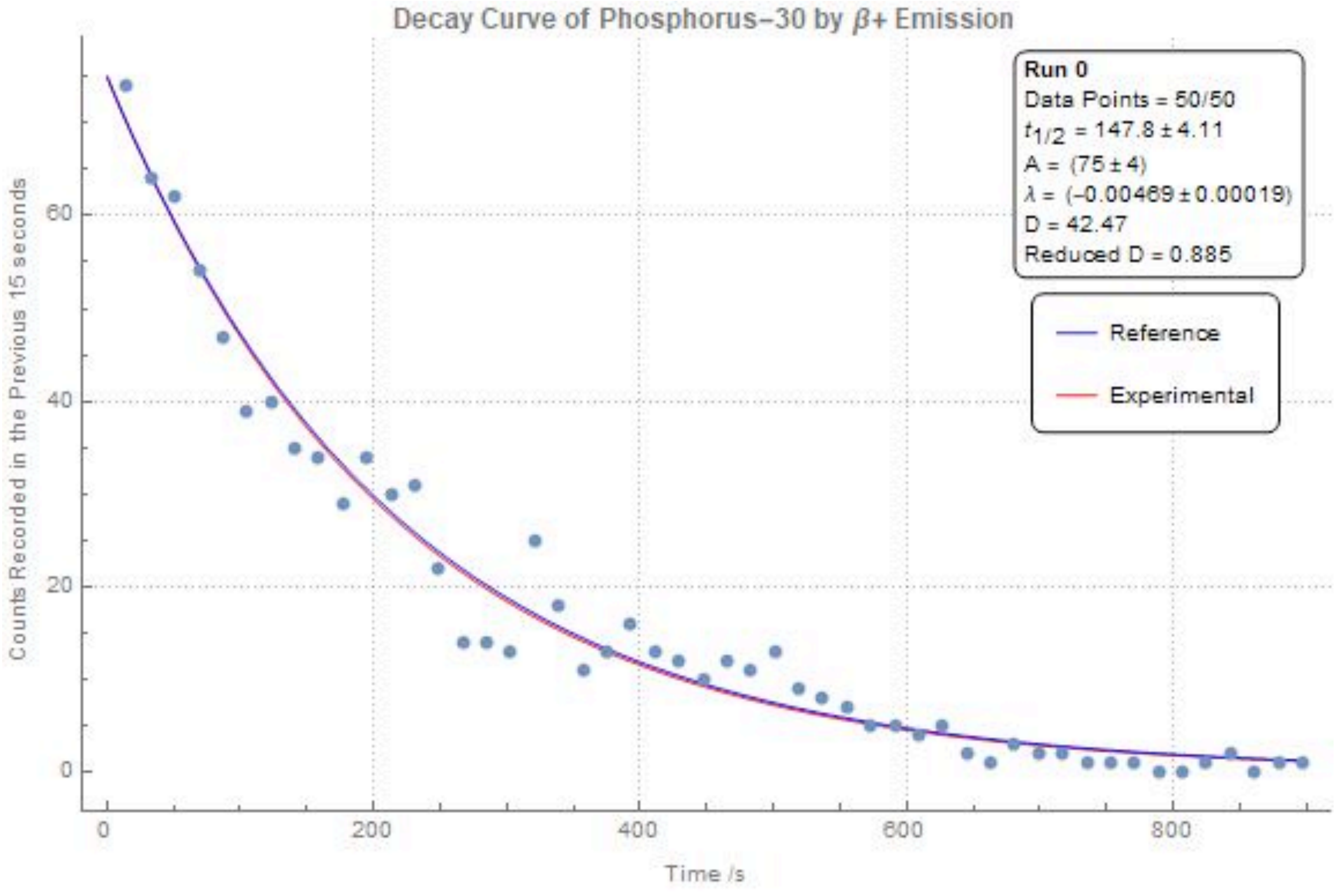
$$P(\Delta t) = w e^{-w\Delta t}$$

$$\boxed{E(\Delta t) = \frac{1}{w}}$$

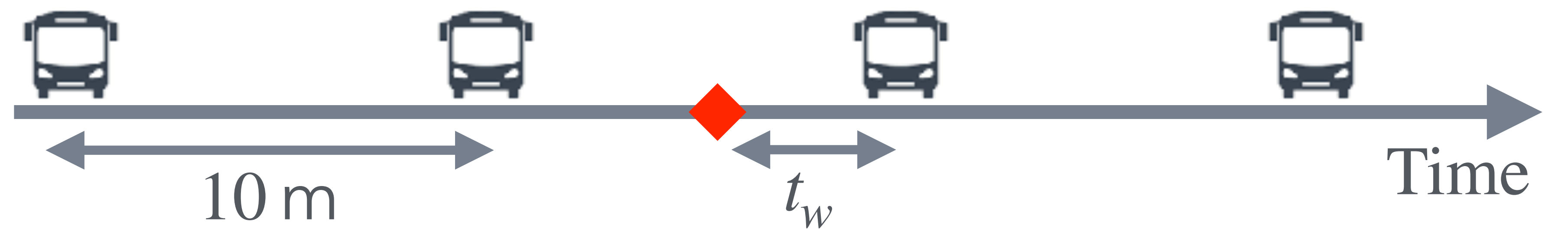
Typical time scale

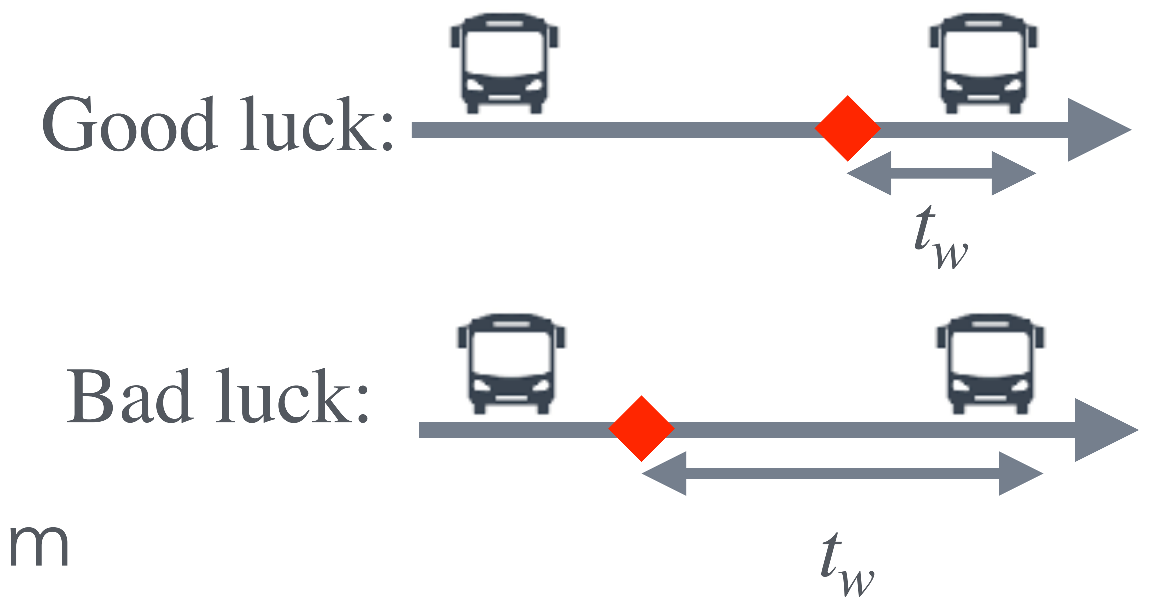
Memoryless property: $\dfrac{P(\Delta t)}{P(0)} = \dfrac{P(\Delta t + t)}{P(t)}$

Example: Radioactive decay



Waiting times with deterministic arrivals





Bad luck -good luck compensation: $E[t_w] = \dfrac{10\ \mathrm{m}}{2} = 5\ \mathrm{m}$

Probability
Modeling randomness

William Feller
(1906 – 1970)

Case of study:

Waiting time paradox
Exponential distribution
Poisson distribution

Exponential distribution

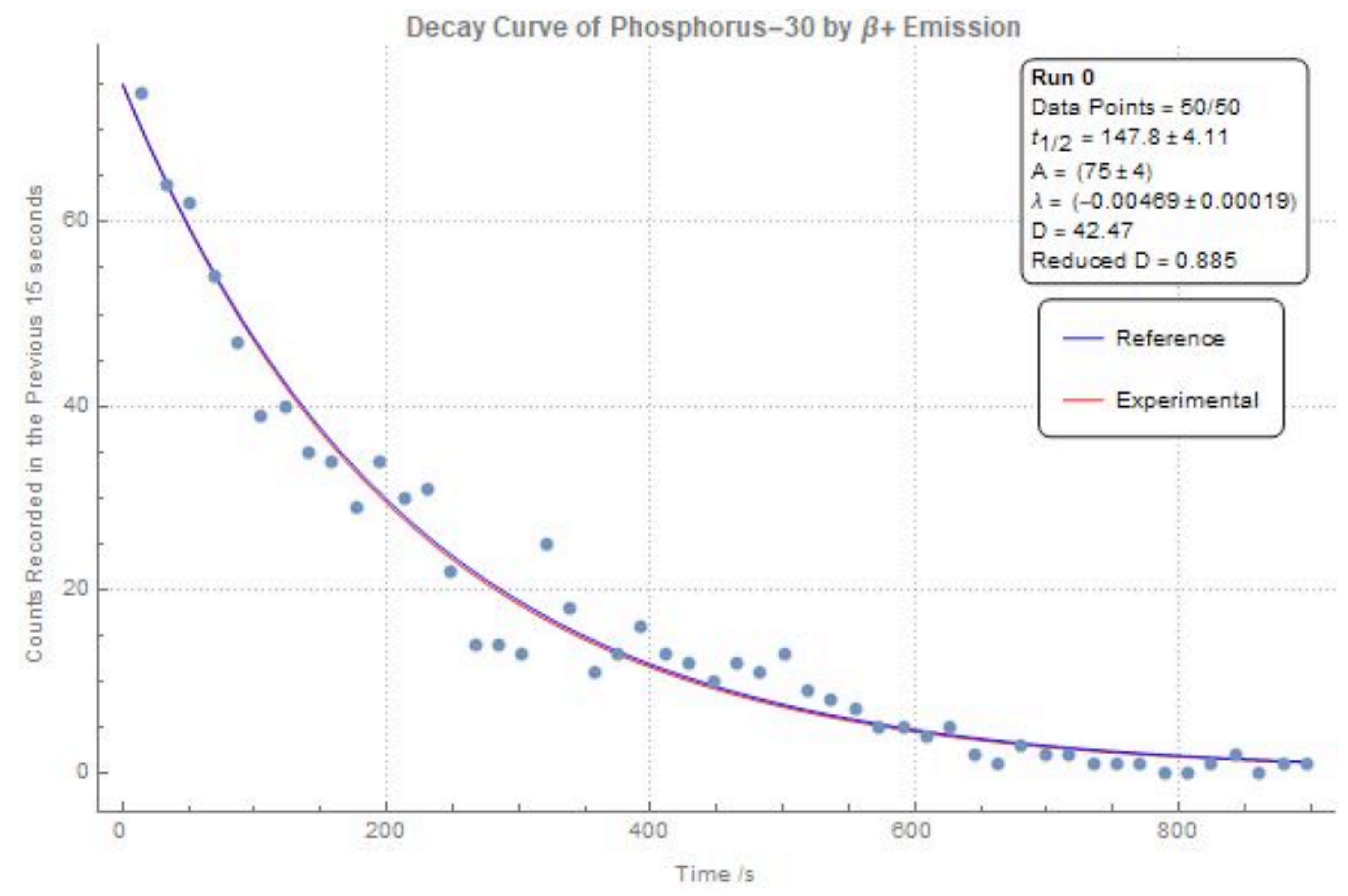
$$P(\Delta t) = w e^{-w\Delta t}$$

$$\boxed{E(\Delta t) = \frac{1}{w}}$$

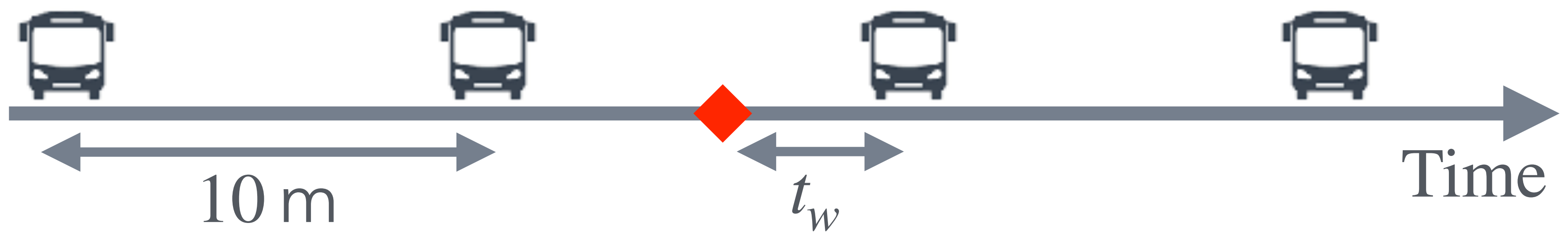
Typical time scale

Memoryless property: $\dfrac{P(\Delta t)}{P(0)} = \dfrac{P(\Delta t + t)}{P(t)}$

Example: Radioactive decay

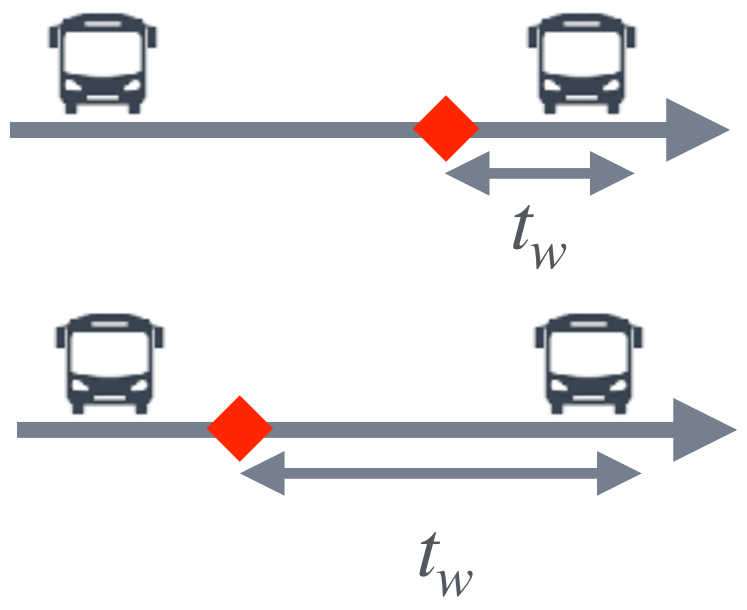


Waiting times with deterministic arrivals



Good luck:

Bad luck:



Bad luck -good luck compensation: $E[t_w] = \dfrac{10\text{ m}}{2} = 5\text{ m}$

Waiting times with random arrivals

$$P(\Delta t) = 10\, e^{-10\Delta t}\ \text{m}^{-1}$$

Bad luck -good luck compensation: $E[t_w] = \dfrac{E[\Delta t]}{2} = 5\text{ m}$

Memoryless property: Waiting time doesn't depend on arrival time $\boxed{\langle t_w \rangle = \langle t_a + \Delta t \,|\, t_a \rangle = 10\text{ m}}$

Probability

Modeling randomness

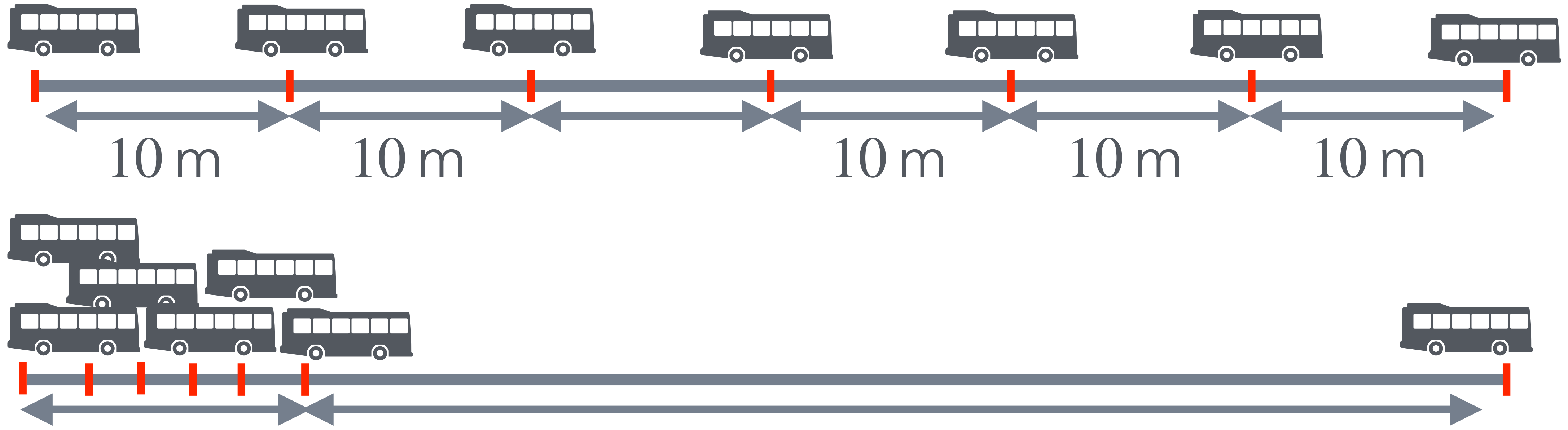
William Feller
(1906 – 1970)

Case of study:

Waiting time paradox

Exponential distribution
Poisson distribution

6 arrivals in one hour:



Probability

Modeling randomness

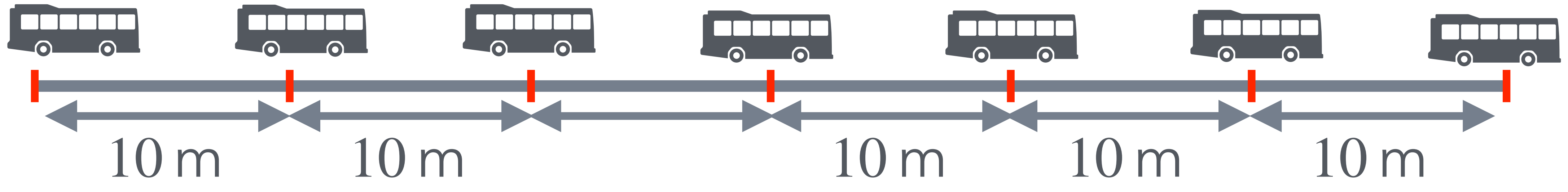
William Feller
(1906 – 1970)

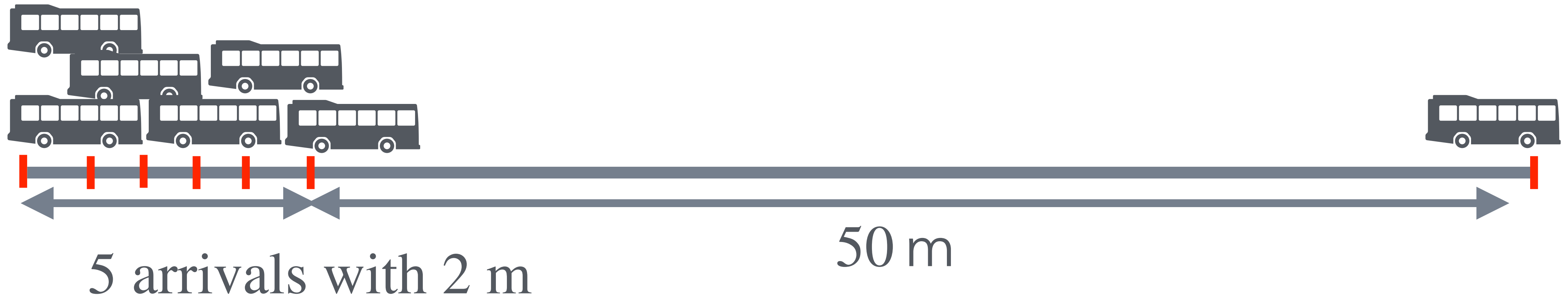
Case of study:

Waiting time paradox

Exponential distribution
Poisson distribution

6 arrivals in one hour:





$$\text{Average time between arrivals} = \frac{1}{6}\left(5 \cdot 2 + 50\right) = 10\,\mathrm{m}$$

Prob. Of hitting 2- and 50- time intervals

$$\text{Average observed T.B.A.} = 5\boxed{\frac{2}{60}}2 + \boxed{\frac{50}{60}}50 = 42\,\mathrm{m}$$

It is more likely to arrive in long intervals between arrivals !!!

Probability

Modeling randomness

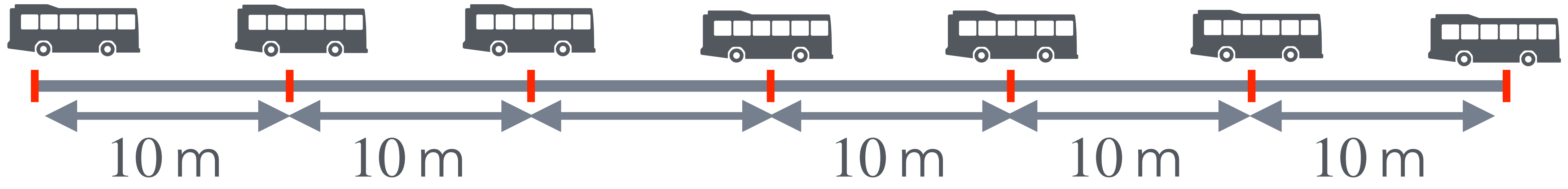
William Feller
(1906 – 1970)

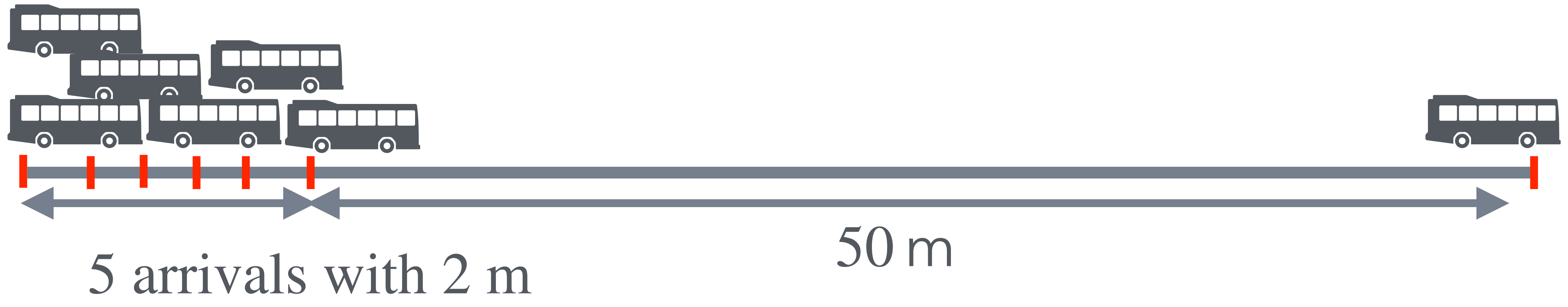
Case of study:

Waiting time paradox

Exponential distribution
Poisson distribution

6 arrivals in one hour:





How we gather data alter statistics (Length bias)

Average time between arrivals $= \frac{1}{6}\left(5 \cdot 2 + 50\right) = \boxed{10\text{ m}}$ → Event-based observer

Prob. Of hitting 2- and 50- time intervals

Average observed T.B.A.$= 5\boxed{\frac{2}{60}}2 + \boxed{\frac{50}{60}}50 = \boxed{42\text{ m}}$ → Random-time observer

$\neq$

It is more likely to arrive in long intervals between arrivals !!!

Stochastic processes

$$P(\Delta t \in [0,dt]) \propto dt \quad \rightarrow \quad w = \lim_{dt \to 0} \frac{P(\Delta t \in [t, t+dt])}{dt}$$

M possible transitions; $w_\nu(n) = w\,(n \to n+\nu), \quad \nu = 1,\ldots,M$

Jumping process: SIS model

Constant population: $N = n + n_s \rightarrow n_s = N - n$

Recovery: $I \rightarrow S$

$$w\,(n \to n-1) = \lambda\, n$$

Infection: $I + S \rightarrow I + I$

$$w\,(n \to n+1) = \beta \frac{N-n}{N} n$$

Stochastic processes

$$\Delta t$$

$$P(\Delta t \in [0,dt]) \propto dt \;\longrightarrow\; w = \lim_{dt\to 0} \frac{P(\Delta t \in [t, t+dt])}{dt}$$

Jumping process: SIS model

Constant population: $N = n + n_s \;\longrightarrow\; n_s = N - n$

Recovery: $I \longrightarrow S$

$$w\,(n \to n-1) = \lambda\, n$$

Infection: $I + S \longrightarrow I + I$

$$w\,(n \to n+1) = \beta \frac{N-n}{N} n$$

M possible transitions; $w_\nu(n) = w\,(n \to n+\nu), \quad \nu = 1,\ldots,M$

$$\lim_{dt\to 0} P(\nu - \text{reaction happens in } dt) = \lim_{dt\to 0} w_\nu(n)\, dt$$

$$\lim_{dt\to 0} P(\text{something happens in } dt) = \lim_{dt\to 0} \sum_\nu w_\nu(n)\, dt$$

$$\lim_{dt\to 0} P(\text{nothing happens in } dt) = \lim_{dt\to 0} 1 - W(n)\, dt$$

$$P(\text{nothing happens in } \Delta t) = e^{-W(n)\Delta t}$$

$$P(\text{nothing happens in } \Delta t, \nu \text{ happens in } [\Delta t, \Delta t + dt]) = \boxed{W(n)\, e^{-W(n)\Delta t} dt}\,\boxed{\frac{w_\nu(n)}{W(n)}}$$

Time between consecutive events are exponentially distributed

Prob. of next state is weighted with the transition rate

Gillespie Algorithm

$$N_t = n \;\longrightarrow\; \Delta t \sim E(W(n)) \;\longrightarrow\; P(N_{t+\Delta t} = n + \nu \,|\, N_t = n) = \frac{w_\nu(n)}{W(n)}$$

Stochastic processes

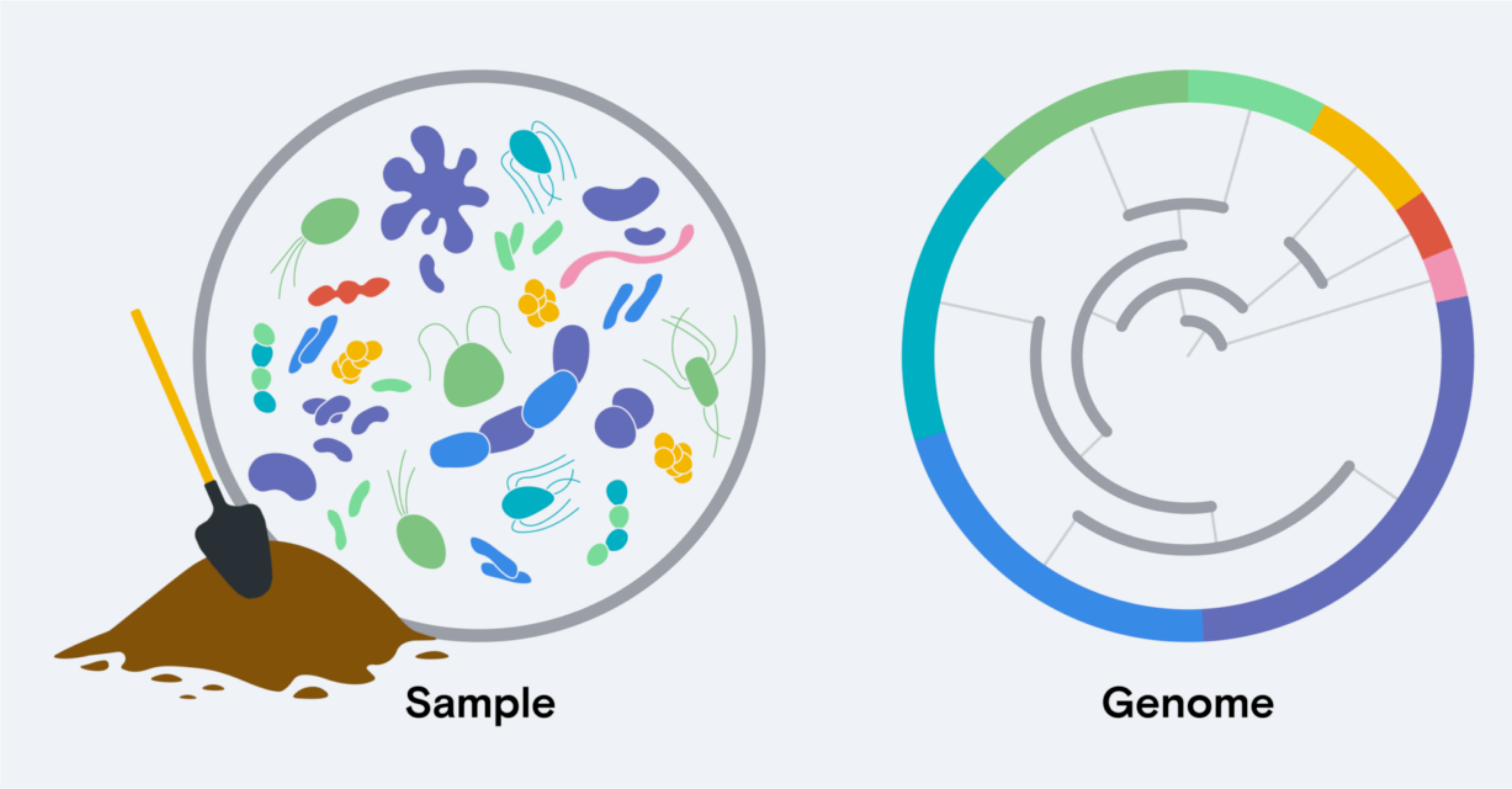
Van Kampen's expansion $w_\nu(n) = w\,(n \to n + \nu) = \Omega f_\nu(x), \quad x = \dfrac{n}{N}$

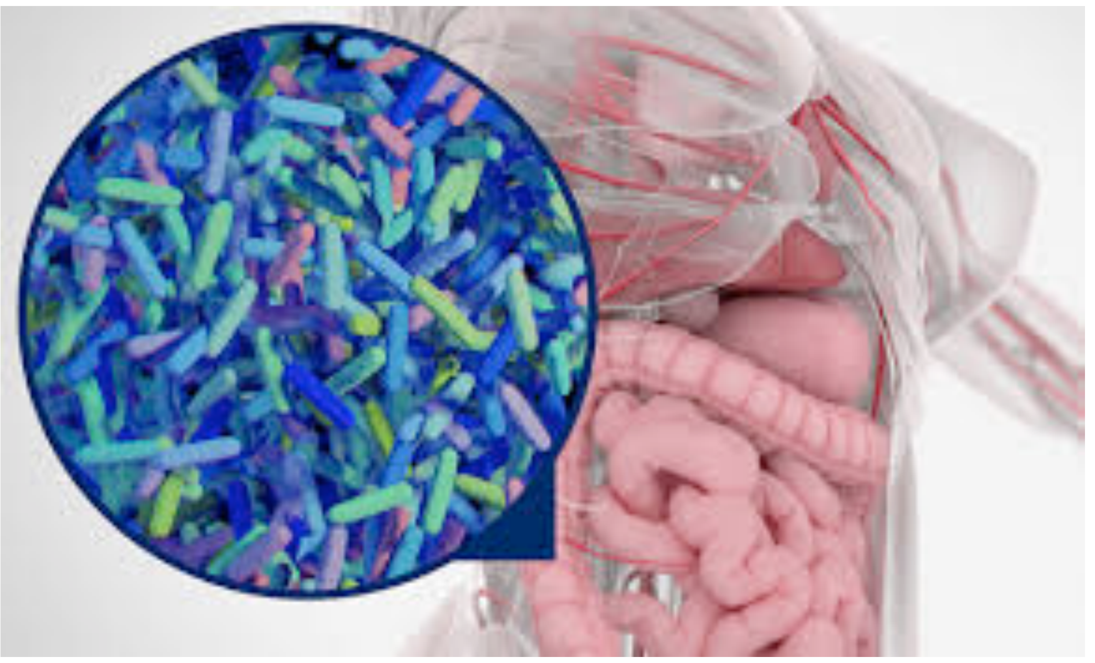
Stochastic processes

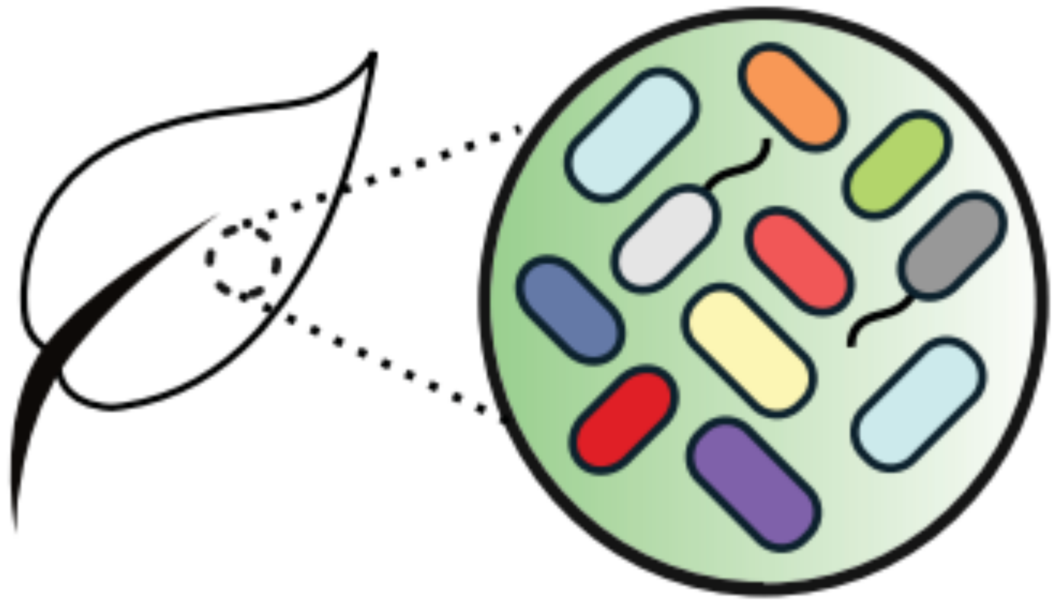
Mean field approximation: Deterministic limit

The metagenomic era

Sequencing of genetic material from all organisms in a particular environment







Goldford et.al. 2018

Human microbiome

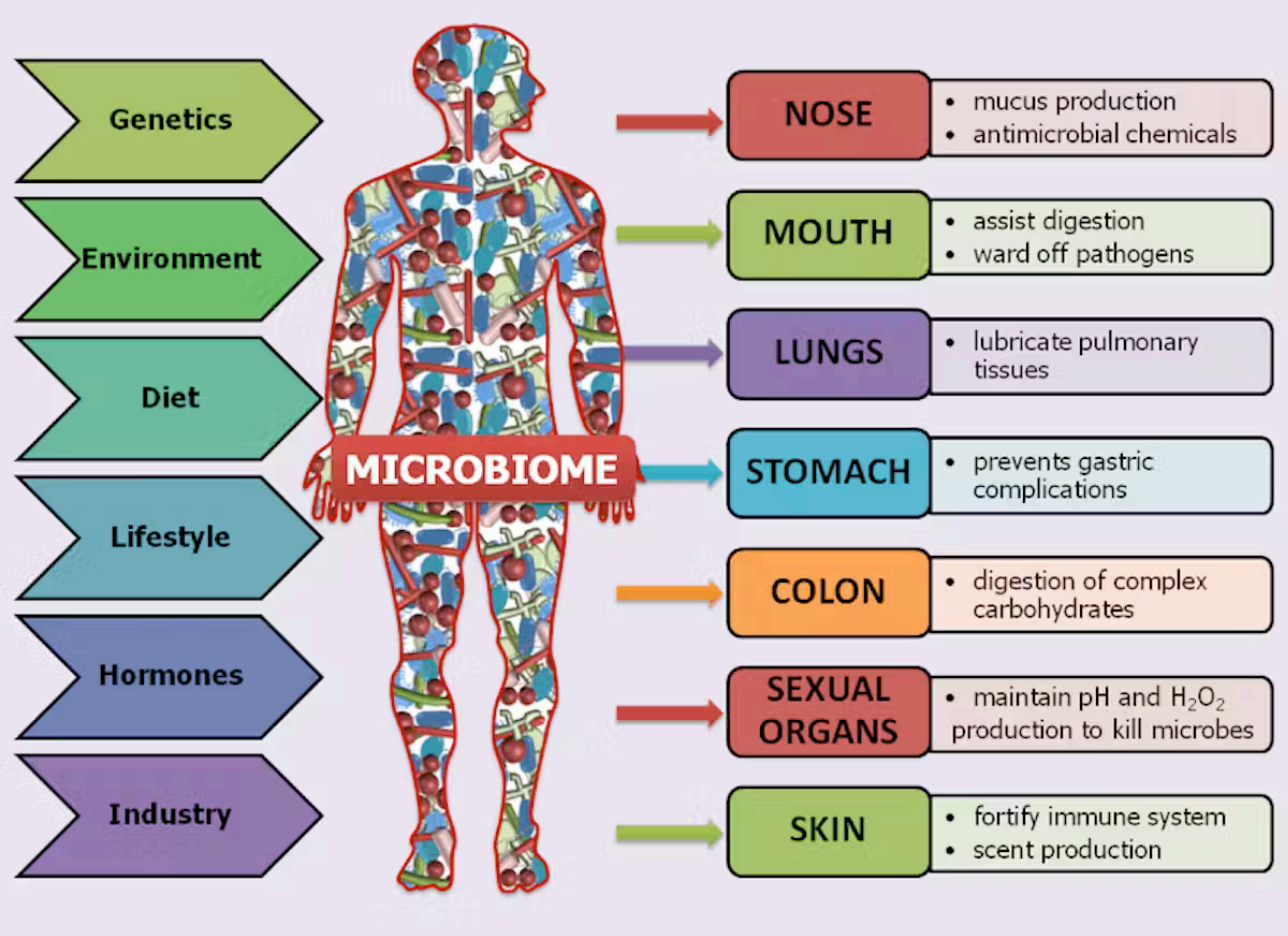
$$\frac{\text{Microbiome}}{\text{human cells}} \approx \frac{10}{1}$$

Largest 'organ'

Unique Microbial Fingerprint

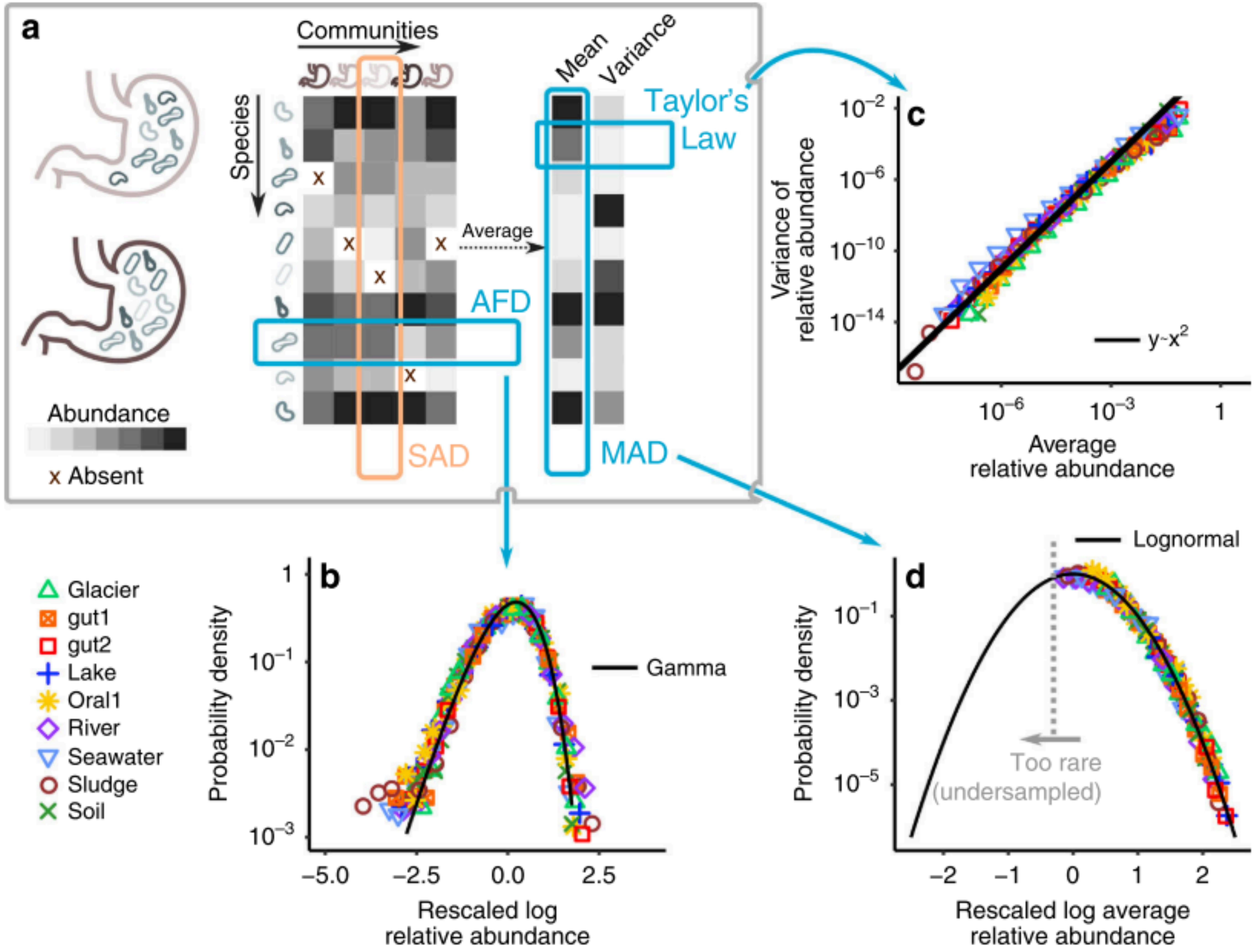
The 'Second Brain'

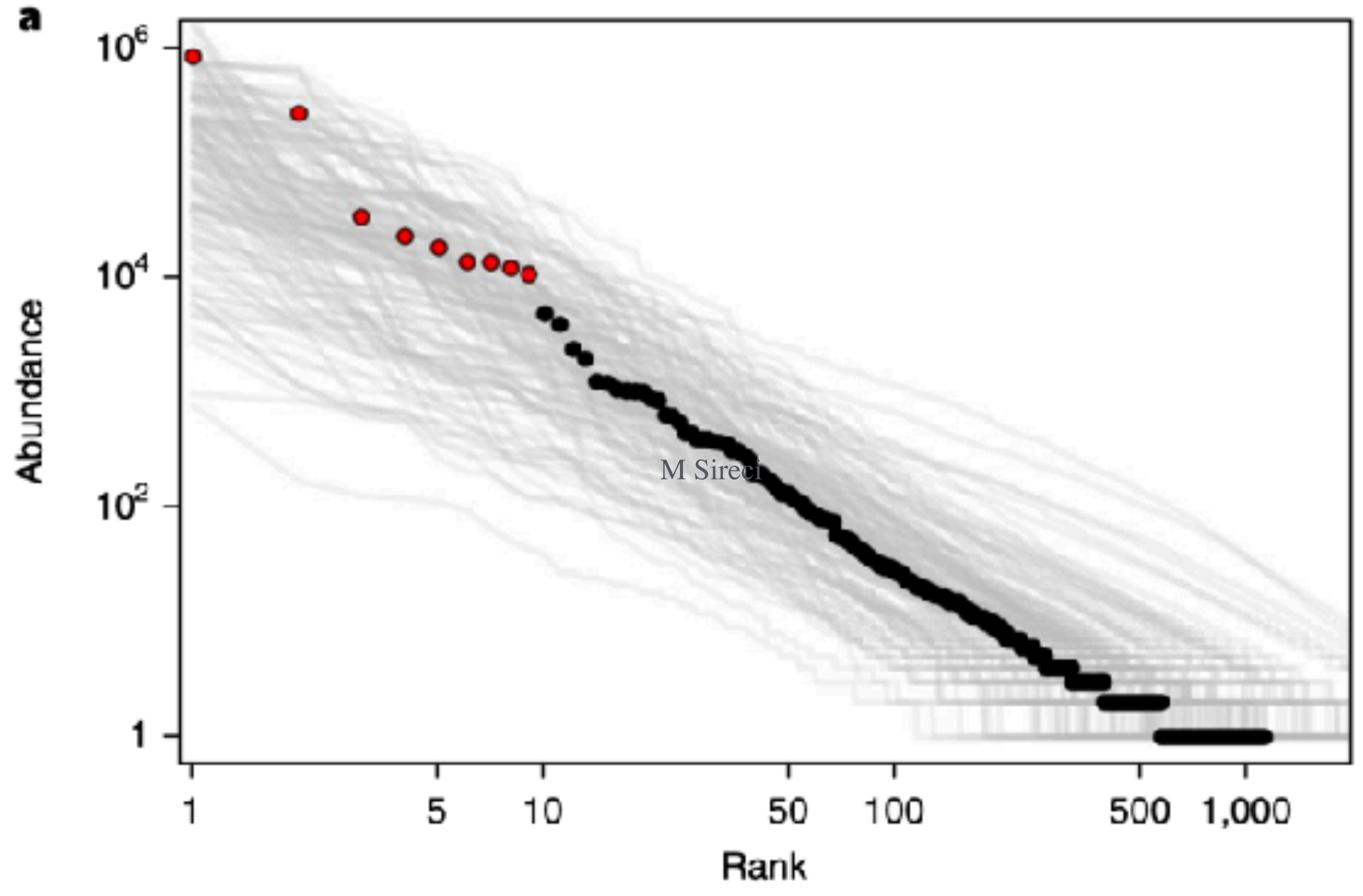
Immune System Powerhouse

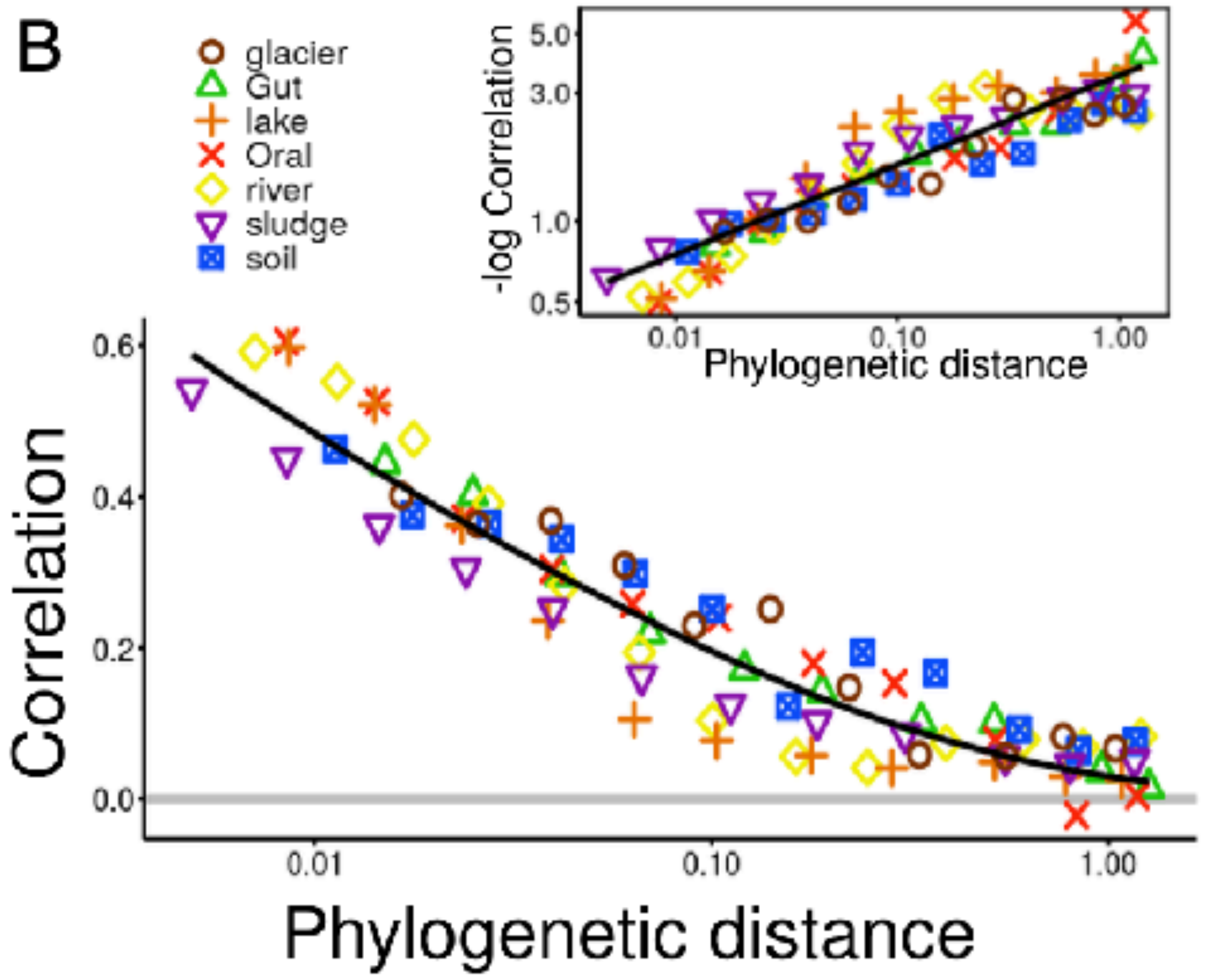


Appanna 2018

Universal macroecological patterns





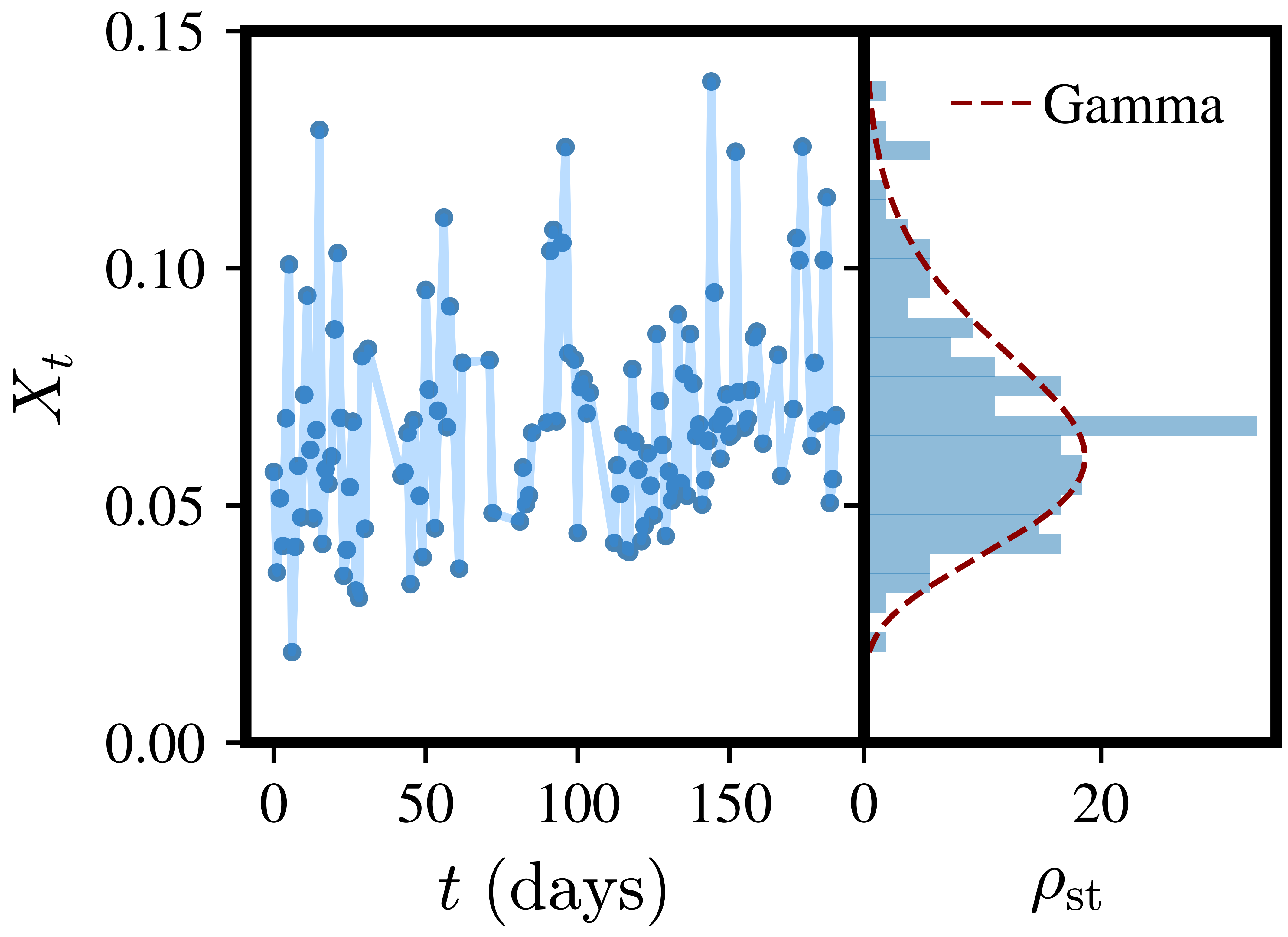


Grilli 2020

Sireci, Muñoz, Grilli 2023

Ser-Giacomi et al. 2018

Modeling microbes density evolution: decoupled theories



$$d X_t = A\left(X_t\right)\, dt + B\left(X_t\right)\, dW_t$$

Caporaso et al. 2011

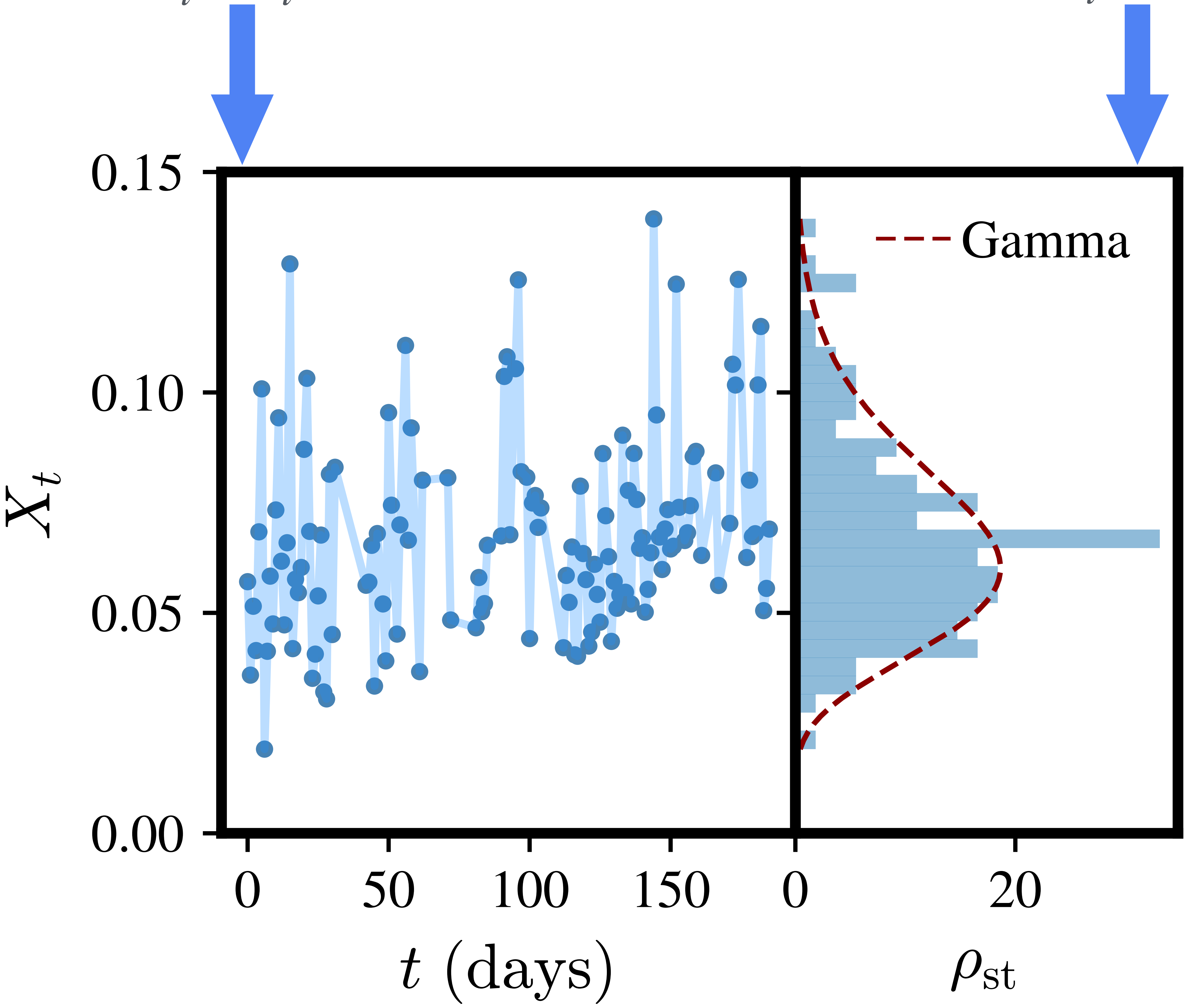
Modeling microbes density evolution: decoupled theories

Environmental fluctuations

$$dX_t = k\left(\mu - X_t\right)\, X_t\, dt + D\, X_t\, dW_t$$

Demographic fluctuations

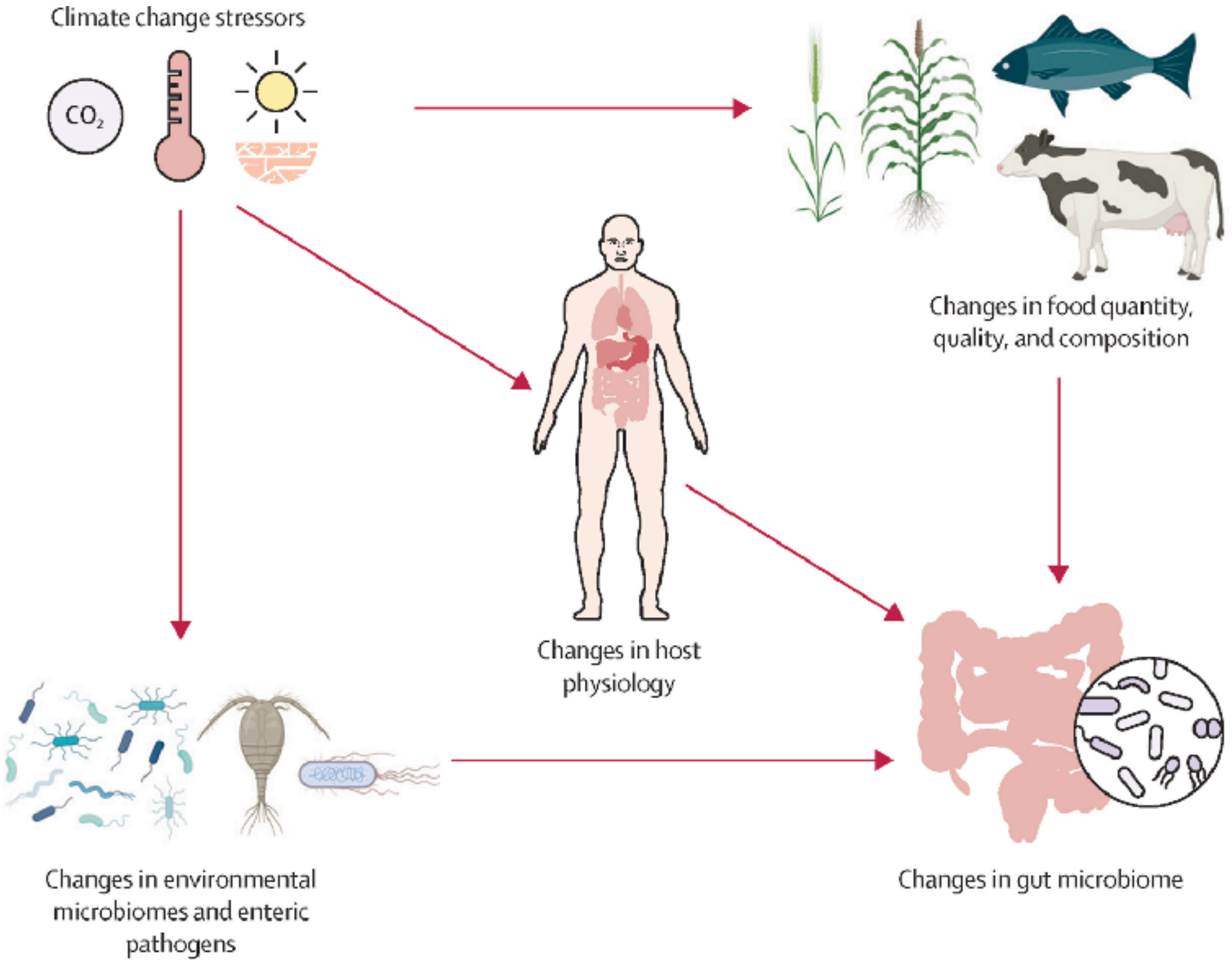
$$dX_t = k\left(\mu - X_t\right)\, dt + D\sqrt{X_t}\, dW_t$$



Modeling microbes density evolution: decoupled theories

Environmental fluctuations

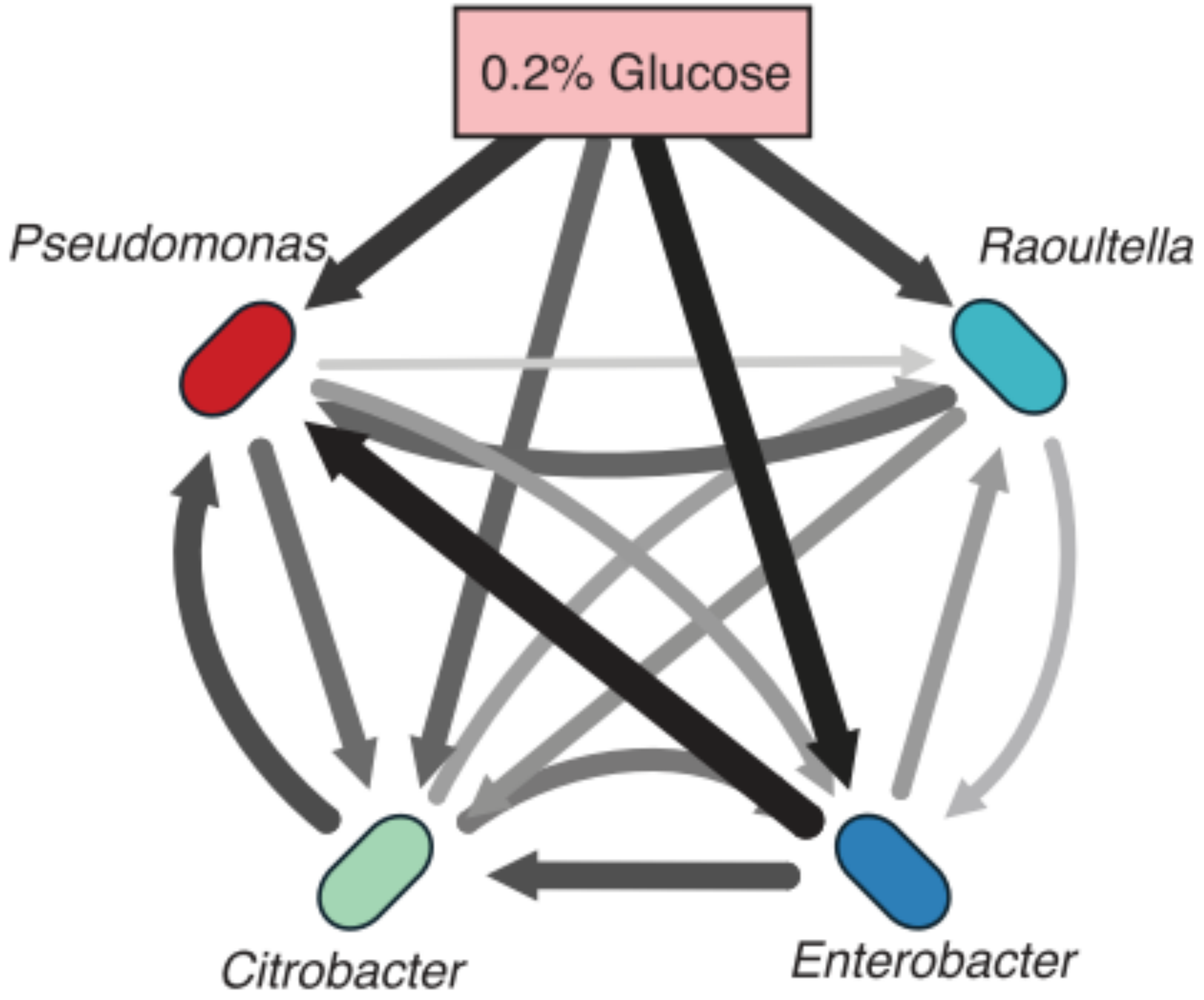
$$d X_t = k\left(\mu - X_t\right) X_t\, dt + D X_t\, dW_t$$



Grilli 2020

Demographic fluctuations

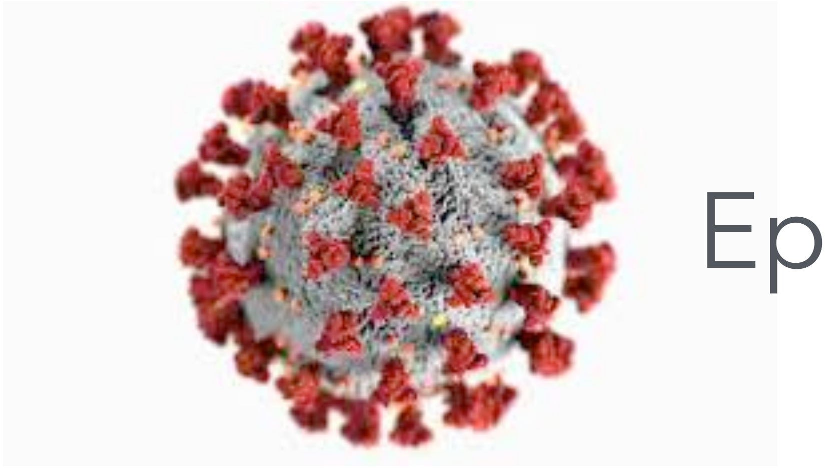
$$d X_t = k\left(\mu - X_t\right) dt + D\sqrt{X_t}\, dW_t$$



Goldford et.al. 2018

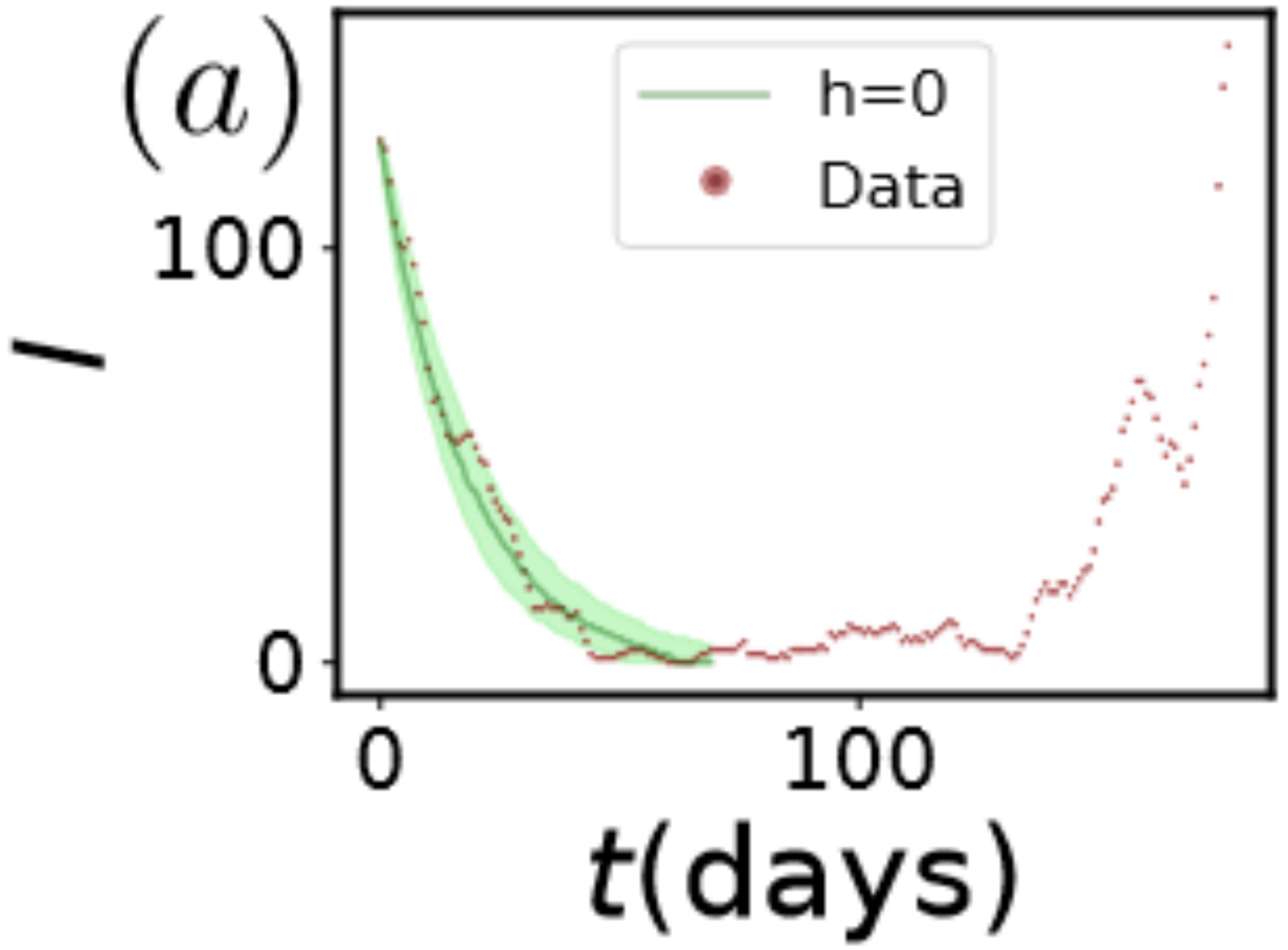
Azaele et.al. 2006 Ashish, O'Dwyer 2023

Multiple modeling choices in complex systems

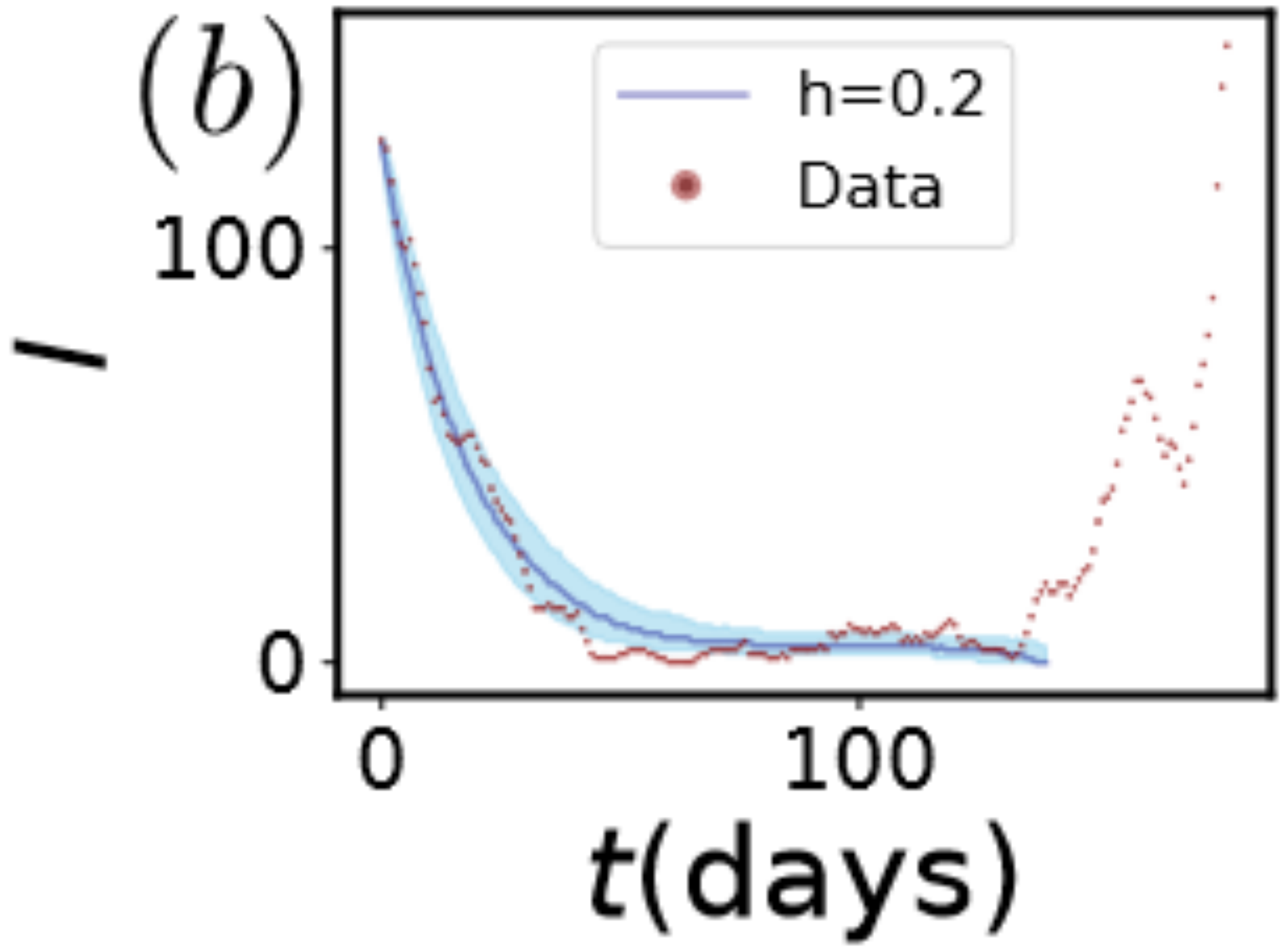


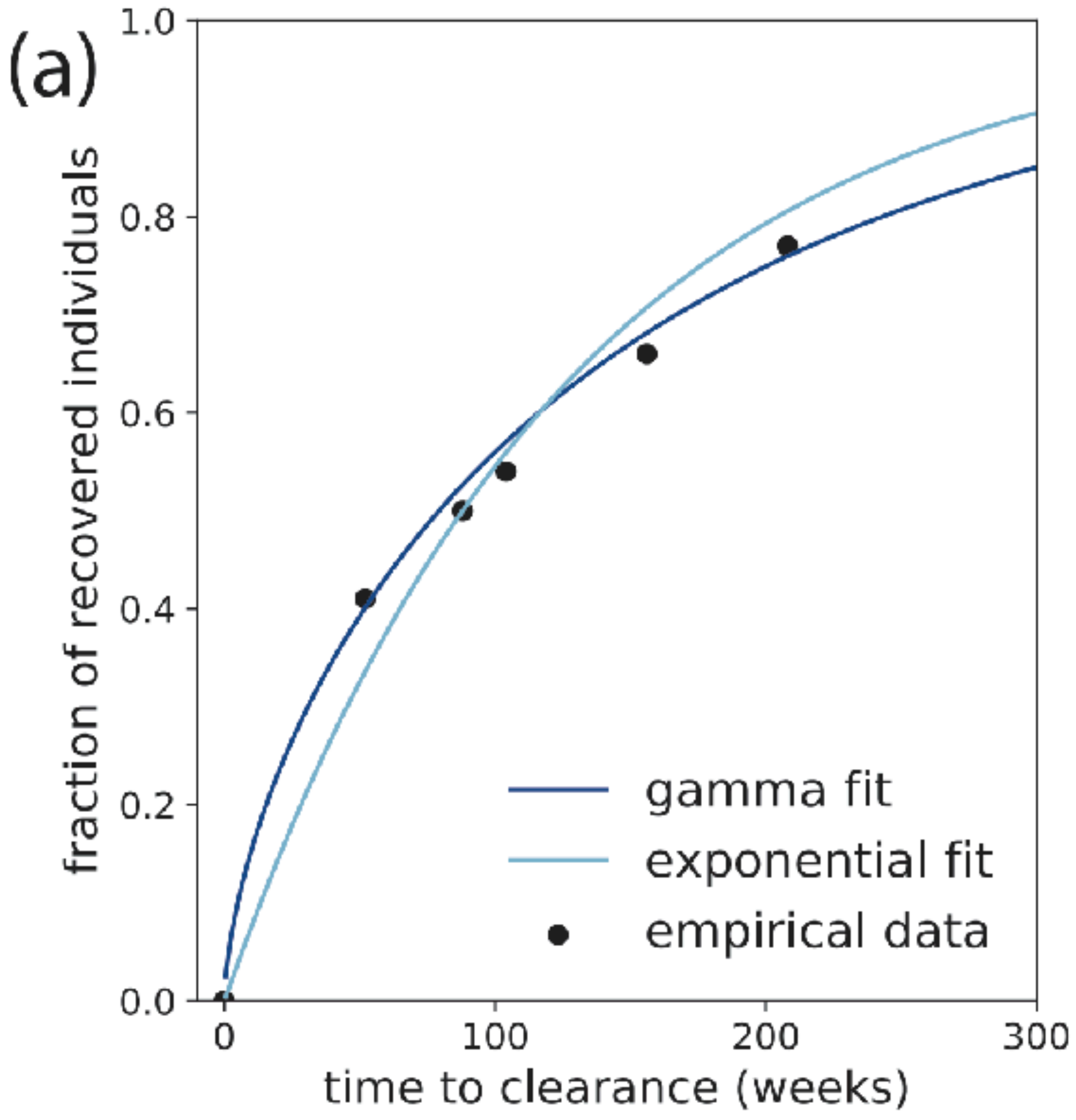
Epidemic modeling

Demographic



Additive





Aguilar et.al. 2023

Di Domenico et.al. 2024

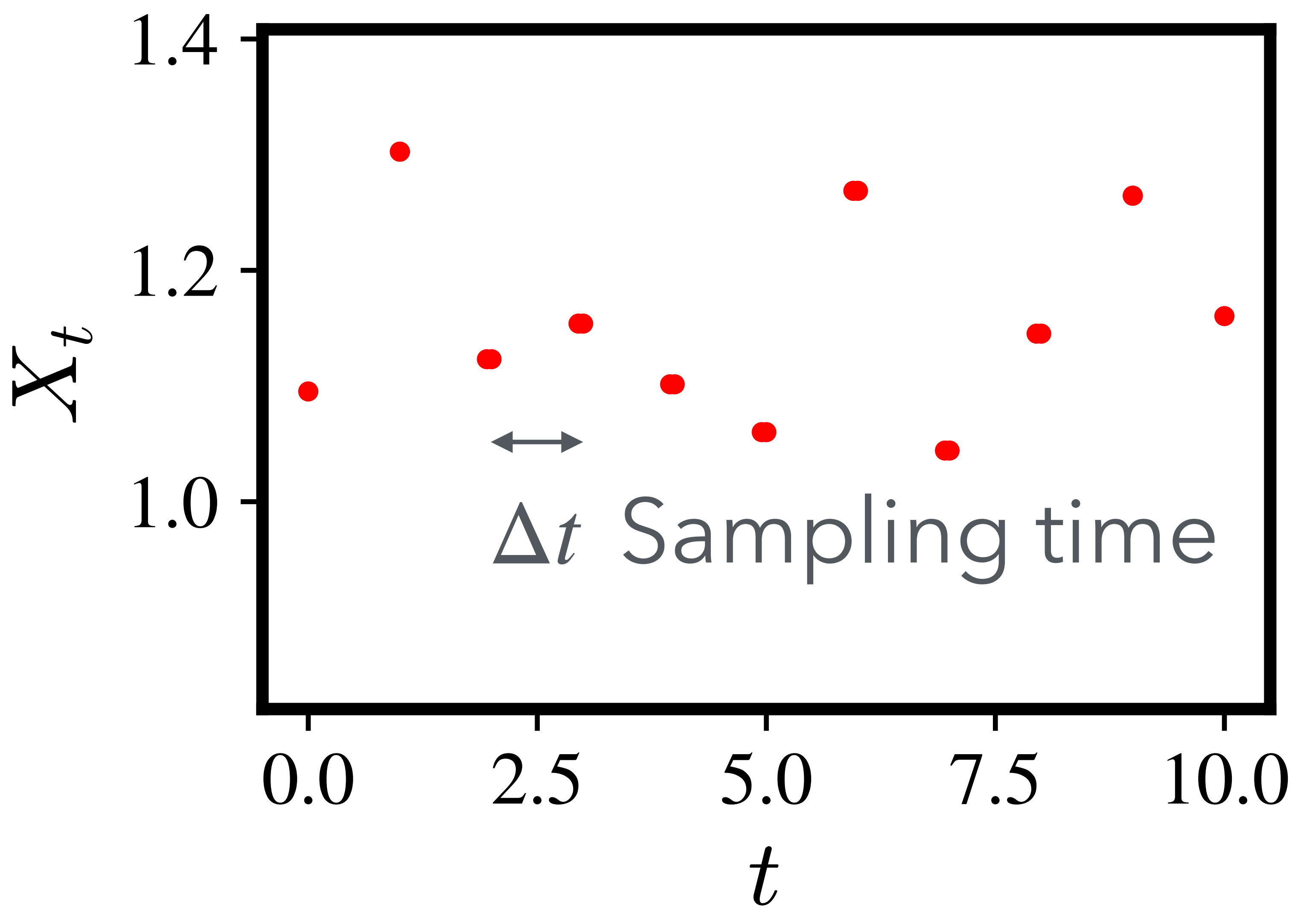
Question:

Models encoding different mechanisms provide good fits for the same data !!

Can we infer which is the best description?

Data

$\{x_i,\ t_i\}_{i=1\ldots,M}$

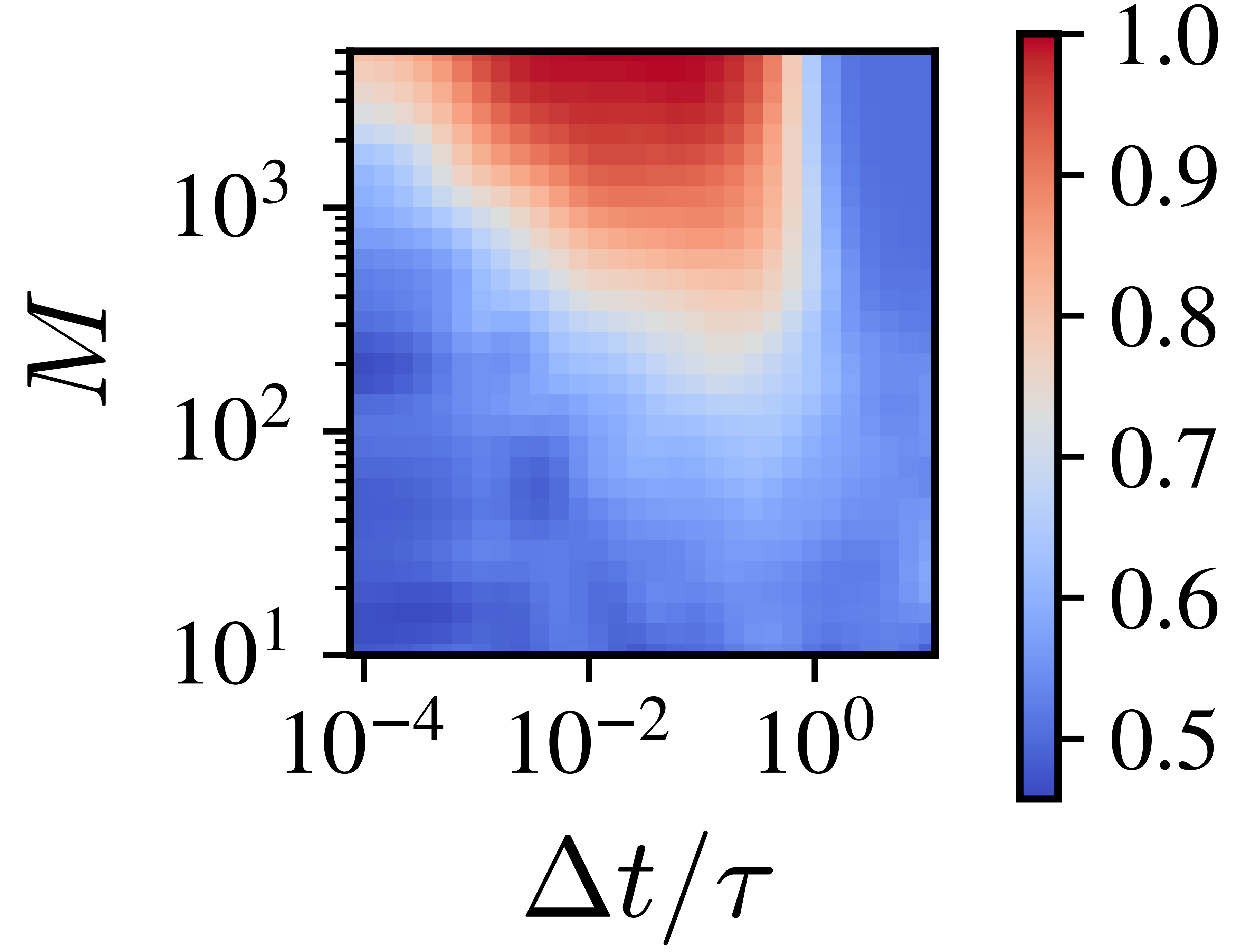


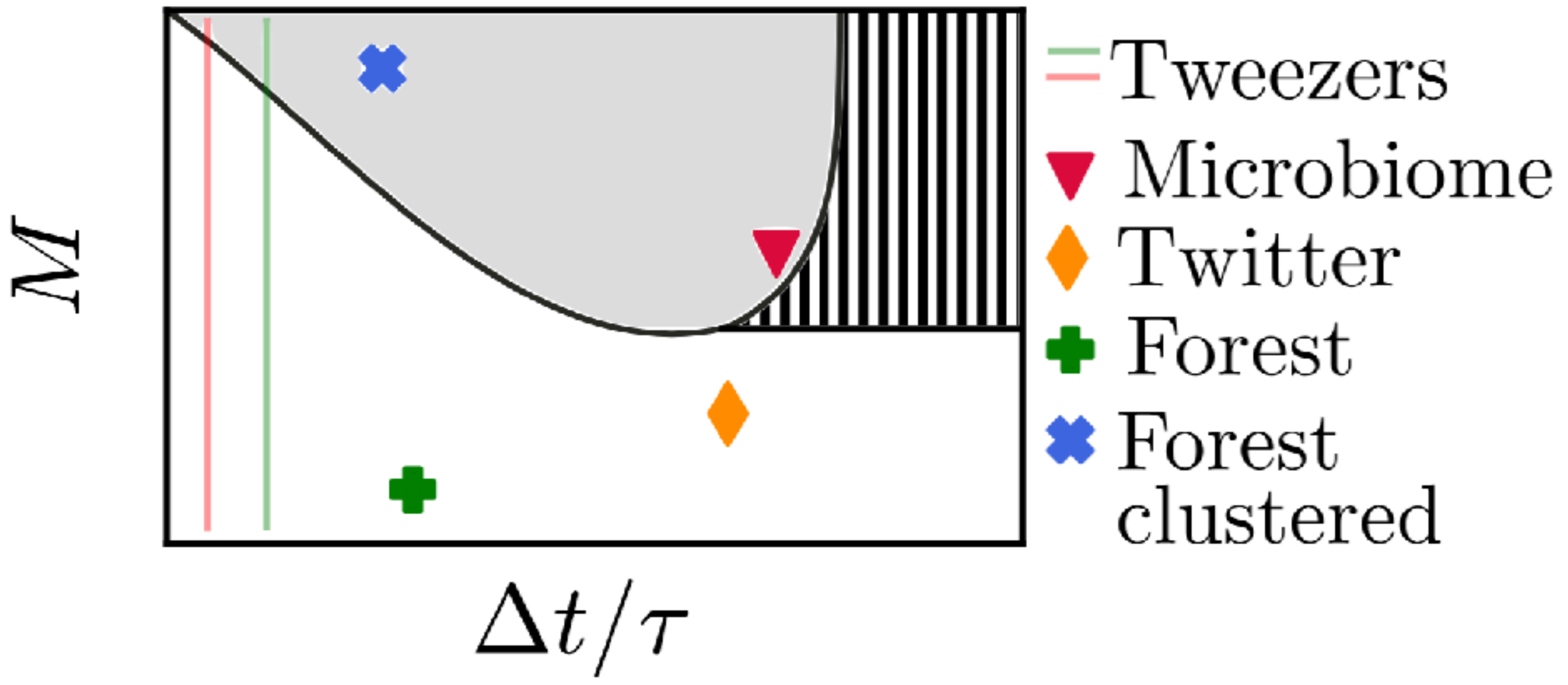
Phase transitions in model distinguishability

$$d X_t = k \left(\mu - X_t\right) X_t \, dt \; + D X_t \, dW_t$$ EN

vs.

$$d X_t = k \left(\mu - X_t\right) dt \; + D \sqrt{X_t} \, dW_t$$ DE





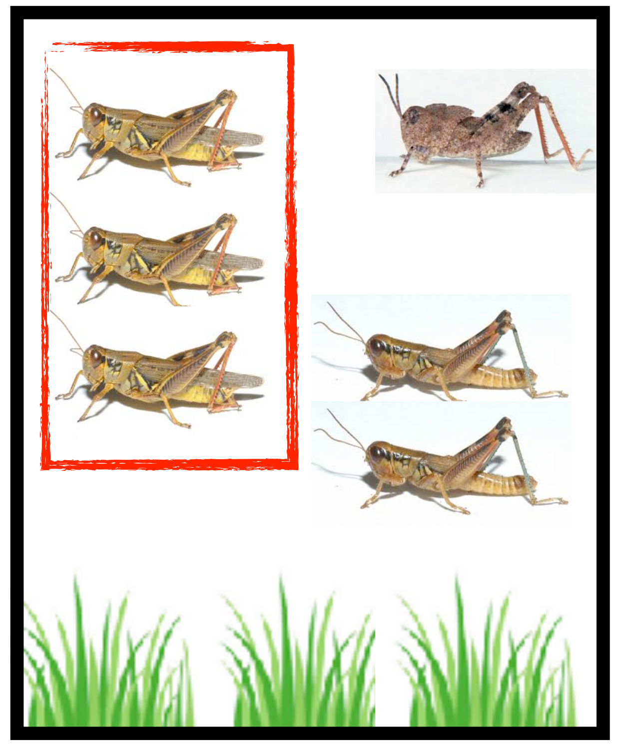
Measures of interactions: Perturbation experiments

M. femur-rubrum

S. collare

P. nebrascensis

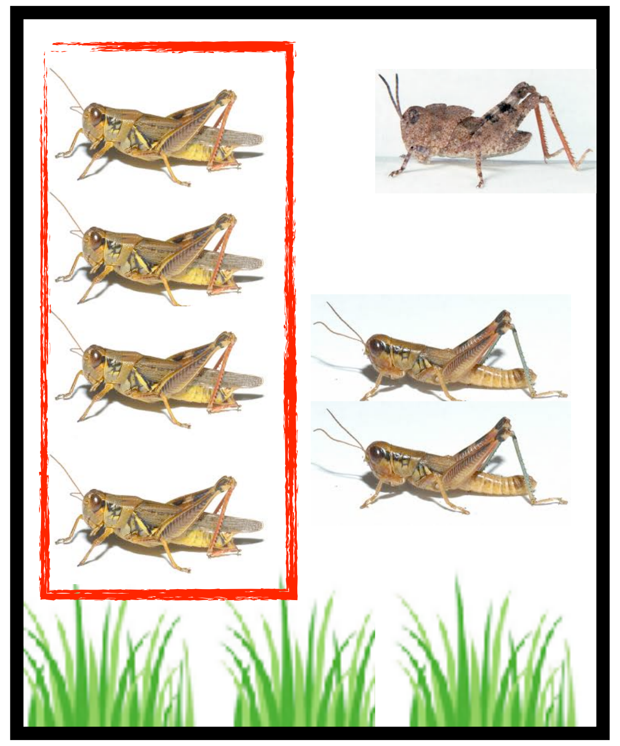
Initial Condition 1



Experiment duration

$g^{(E)}$ [Baseline]

Initial Condition 2



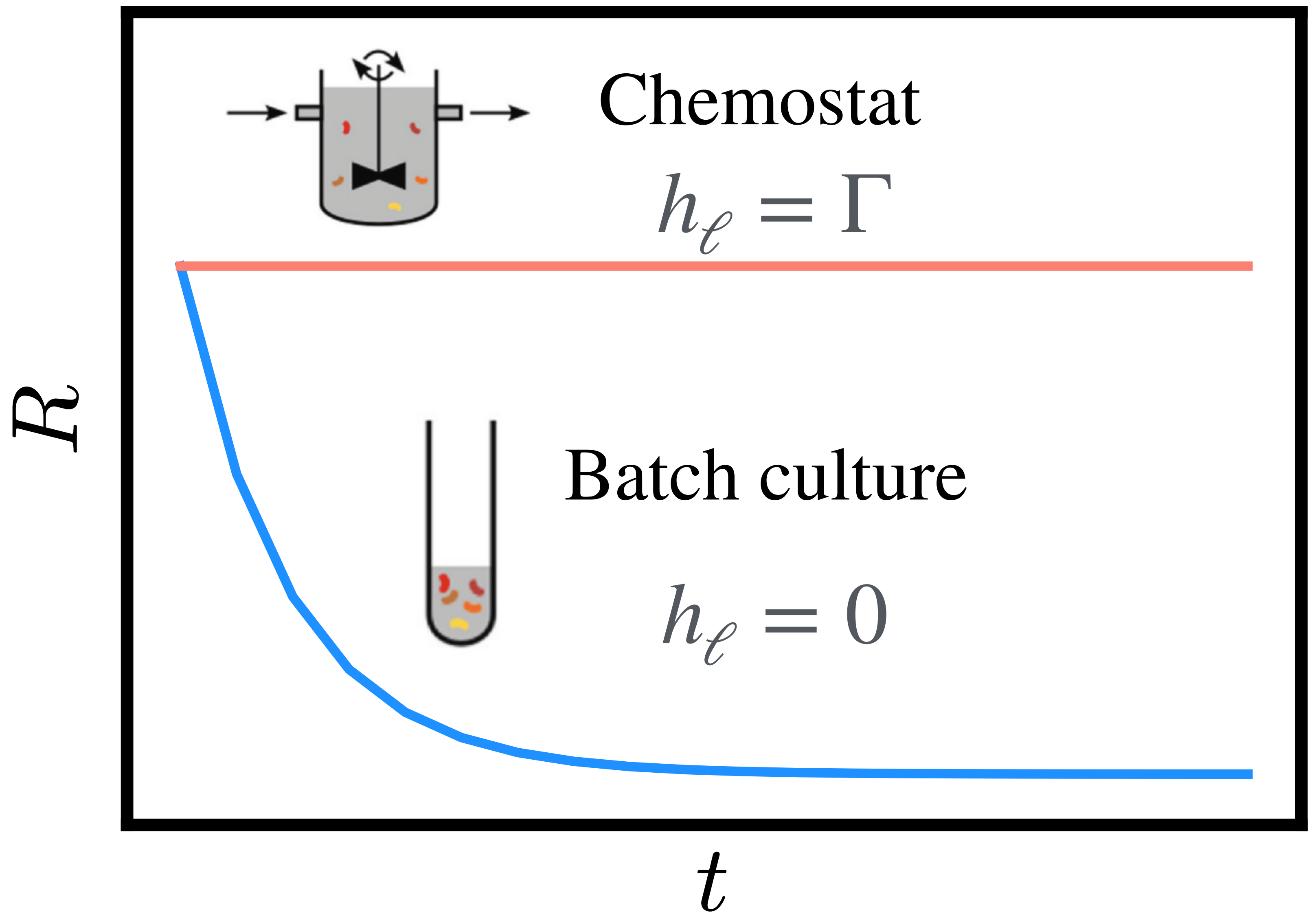
Experiment duration

$g^{(E)}$ [Perturbed]

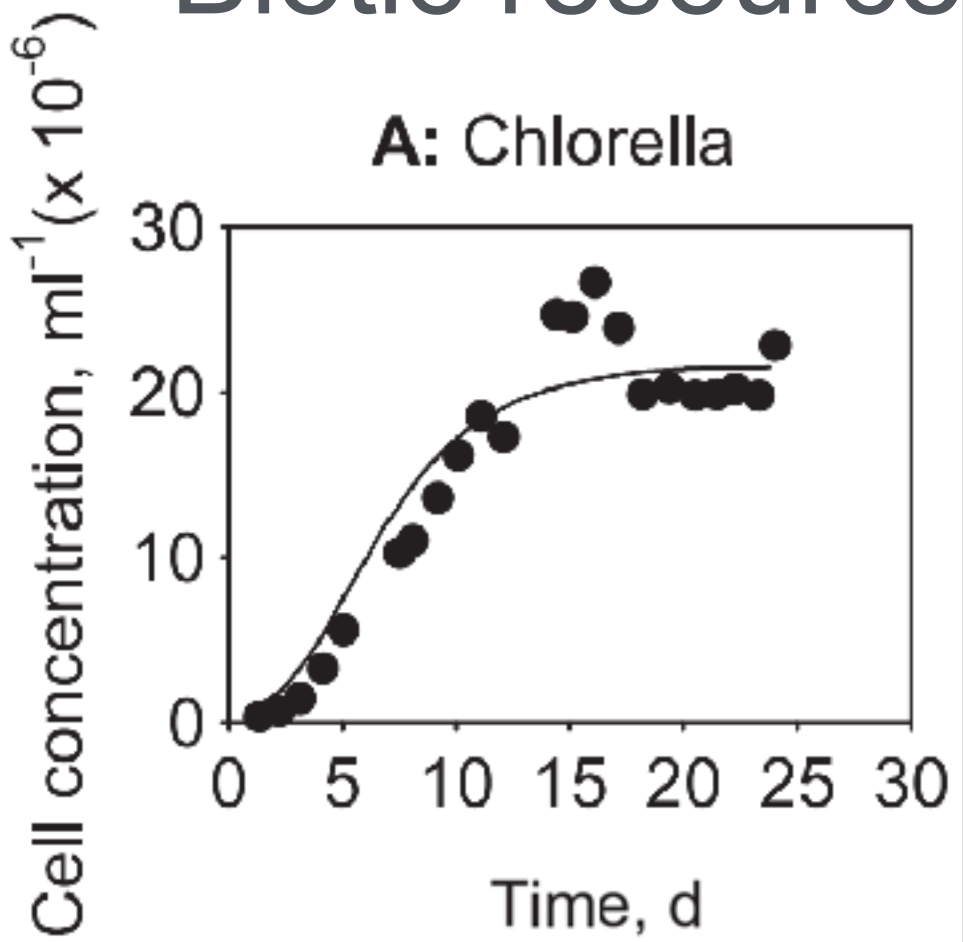
Ritchie and Tilman, 1993

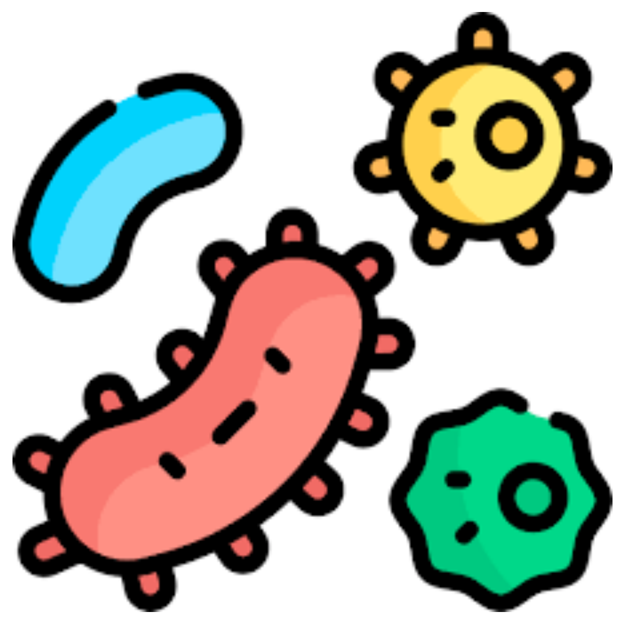
Interactions depend on the experimental protocol

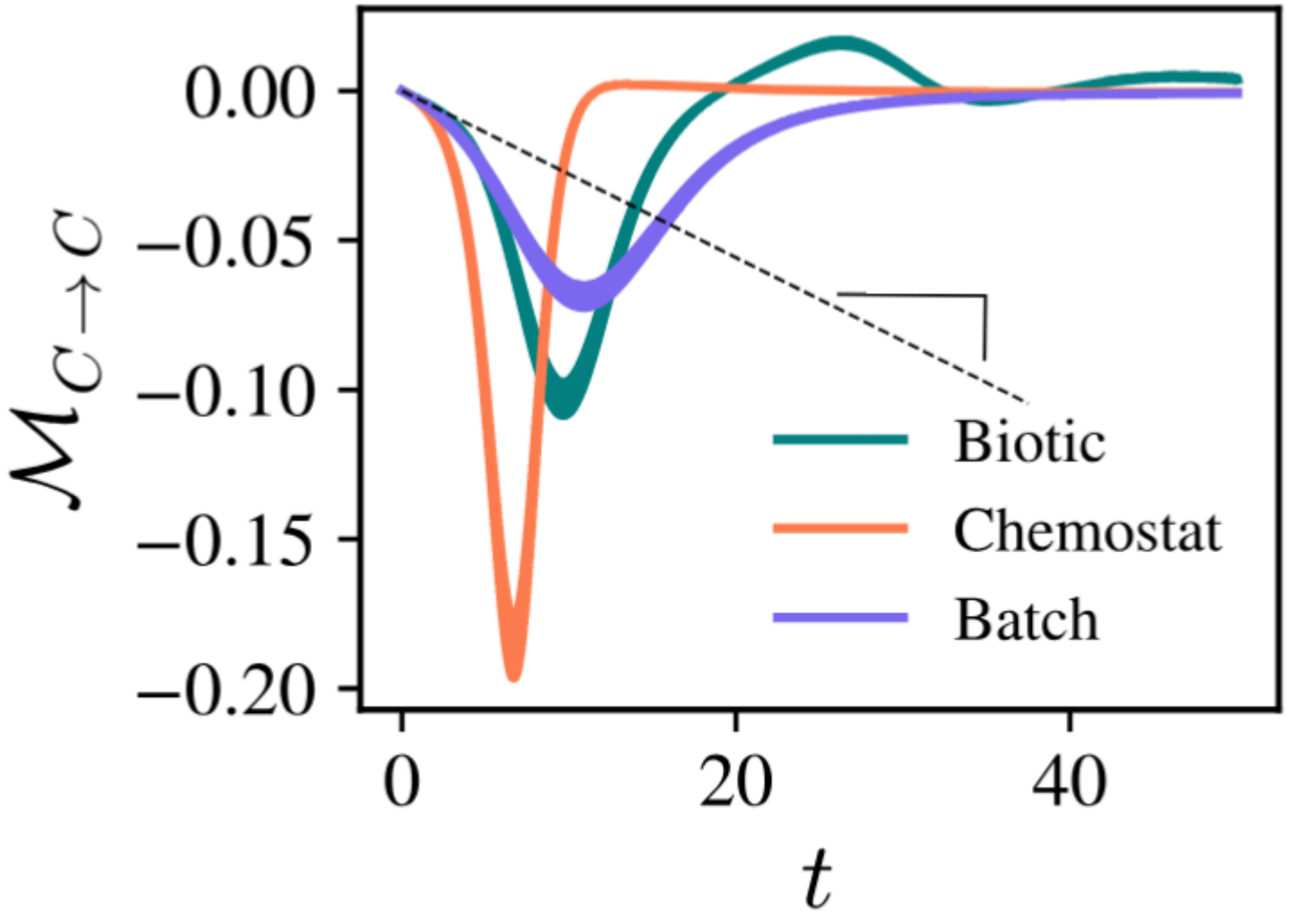
Abiotic resources



Biotic resources $h_\ell = r\,R_\ell\,(1 - \frac{R_\ell}{K})$







Kiørboe, 2008

Picot, Shibasaki, Meacock, Mitri 2023

EXCELENCIA MARÍA DE MAEZTU 2023-2027

IFISC

CONCEPTS AND TOOLS

PHOTONICS AND QUANTUM SYSTEMS

LIFE AND ENVIRONMENT

SOCIAL AND SOCIO-TECHNICAL SYSTEMS

Fellowships for the Master's Degree

1. The IFISC offers a limited number of mobility fellowships. One of them is sponsored by the Sicomoro Foundation. Each fellowship covers tuition fee and, additionally for applicants who are non-residents in Mallorca, travel costs up to 500€ and 7000€. Conditions for these grants can be found in the following link. You can apply to them using the form below. 2025/2026 Selected candidates can be found here
2. CSIC offers JAE-INTRO grants open to those who have scored a minimum of 8/10 in the bachelor's degree. For more information: https://jaeintro.csic.es/en/
3. The Spanish Ministry of Education offers general grants as well as collaboration grants. Master students can apply to both of them.
4. Additional grants, awards and general aid for master degree estudies at the webpage of the UIB Center of Postgraduate Studies
5. Fellowship offered by Fundación Carolina. Deadline March 12 (2025), 9:00 am Spanish time.
6. Santander Open Academy offers several grants for UIB postgraduate studies. Details can be found here .
7. Other options can be found at European Funding Guide

Bibliography

- Blitzstein, J. K.; Hwang, J. Introduction to Probability.
- Roussas, G. G.; Probability and statistics through the centuries.
- Florescu, I.; Probability. A (very) brief history.
- Spiegelhalter D.; The Art of Statistics
- Galton, F. Regression Towards Mediocrity in Hereditary Stature.
- Cole, T. J. Galton's Midparent Height Revisited.
- Newman, M. E. J. Power Laws, Pareto Distributions and Zipf's Law.
- Feynman, R. P. Statistical mechanics: A set of lectures.
- Van Kampen, N.G. Stochastic Processes in Physics and Chemistry.
- Risken, H. The Fokker-Planck Equation: Methods of Solution and Application.
- Wolfgang, P. and Baschnagel, J. Stochastic Processes. From physics to finance.
- Piantadosi, S. T. Zipf's Word Frequency Law in Natural Language: A Critical Review and Future Directions.
- Feller, W. An introduction to probability theory and its applications. Vol. {II}.
- Casella, G. And Berger R. L. Statistical inference.
- Cocco, S. , Monasson, R. and Zamponi F. From statistical physics to data-driven modeling.
- Goldstein R.E. Are theoretical results 'Results'?

Datasets and resources

- Kaggle Male & Female height and weight
- Galton height dataset
- Sally Clarck information