Learning from Sensor Data: Set I

Behnaam Aazhang
J.S. Abercombie Professor
Electrical and Computer Engineering
Rice University

Course Outline

- 1. Preliminaries
- 2. A probabilistic approach (books by Hajek and Mackay)
  - Statistical characteristics of data
  - Statistical analysis of the performance
- 3. Data (book by Hajek)
  - Continuous time
  - Discrete time

- 4. Frameworks for learning from data (MacKay)
  - Parametric models
  - Non-parametric—data driven

- 5. Estimating key statistical metrics from data (Bishop 2.4, 2.5)
  - Estimating probability mass function
  - Density estimation
    - Plugin estimators
      - Kernel density estimation (KDE)
      - K nearest neighbor (k-NN)

Set II

- 6. Data representation (Bishop 8)
  - Graphical modeling
    - Directed graphs
      - Bayesian network
    - Undirected graphs
      - Markov random fields
    - Factor graphs

1. Preliminaries

- Engineering is all about designing a system with constraints
  - or more often, “improving” the functionality of a physical system within some practical constraints
- The system could be anything from a bridge to the space station to the world wide web
- Examples of physical systems could be our environment, a biological system, or a factory
- The constraints could be the form factor, the cost, power, time, among others

- Engineers use fundamental tools like mathematics, physics, chemistry, and economics
- For years their starting point has been building a model
    - Model of the system
    - Model of the constraints
- The impact of their work has been limited by the accuracy of their model
- The model is often also used to evaluate the performance

- Despite possible limitations of models we have thousands of engineering marbles
  - Golden gate bridge
  - World wide web
  - Cellular LTE
  - Robots
  - …

- “Essentially all models are wrong but some are useful” G. Box (1987)
- A move from model based engineering to data based engineering
    - Can we engineer based on data?
    - A precursor is “inference” where we try to find the most appropriate explanation for data

- Over the last decade there has been a data deluge
    - Incredible connectivity
    - Cheap storage and computational machines
    - Availability of sensors
- There are many positives and negatives to the explosion of data
    - Let’s only focus on the positives

- Learning from data
  - A probabilist approach
    - Data could be noisy
    - Model could have inherent uncertainty
    - Insufficient size of data set
    - A probabilistic inference may be desirable
      - Example: 80% chance of rain

2. A probabilistic approach

- Input space = feature space = signal domain $\mathcal{X}$
- Output space = response space = signal range $\mathcal{Y}$
- Examples:
  - Classification
    $$\mathcal{X} = \Re^d \text{ and } \mathcal{Y} = \{0, 1\}$$
  - Estimation
    $$\mathcal{X} = \Re \text{ and } \mathcal{Y} = \Re \text{ where } Y = g(X) + Z$$

- In many systems and problems, input (data) denoted as *X* and output by *Y*
- Assume a joint distribution of (X, Y) as $F_{X,Y}$
  - Cumulative distribution function (CDF) and joint CDF

$$F_X(a) = P\{X \leq a\} \text{ and } F_{X,Y}(a,b) = P\{X \leq a \text{ and } Y \leq b\}$$

- Probability density function (PDF) and joint PDF if variables are continuous valued

$$F_X(a) = \int_{-\infty}^{a} f_X(x)dx$$
$$F_{X,Y}(a,b) = \int_{-\infty}^{a} \int_{-\infty}^{b} f_{X,Y}(x,y)dxdy$$

- For discrete data we define probability mass function (PMF)

$$F_X(a) = \sum_{x_i \leq a} p_X(x_i) \text{ where } p_X(x_i) = P(X = x_i)$$

- Joint probability mass function

$$F_{X,Y}(a,b) = \sum_{x_i \leq a} \sum_{y_j \leq b} p_{X,Y}(x_i, y_j) \text{ where } p_{X,Y}(x_i, y_j) = P(X = x_i, Y = y_j)$$

- Conditional distribution and conditional probability mass function

$$F_{Y|X}(b|x_i) = \sum_{y_j \leq b} p_{Y|X}(y_j|x_i)$$

- If *X* and *Y* are jointly discrete

$$p_{Y|X}(y_j|x_i) = \frac{p_{X,Y}(x_i, y_j)}{p_X(x_i)}$$

- Conditional distribution and conditional density

$$F_{Y|X}(y|x) \text{ and } F_{Y|X}(b|x) = \int_{-\infty}^{b} f_{Y|X}(y|x)dy$$

- If *X* and *Y* are jointly continuous then

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

- The expectation operator

$$E[g(X)] = \int_{\Re} g(x) dF_X = \int_{\Re} g(x) f_X dx$$

$$E[g(Y)|X] = \int_{\Re} g(y) dF_{Y|X} = \int_{\Re} g(y) f_{Y|X} dy$$

$$E[g(X,Y)] = \int_{\Re^2} g(x,y) dF_{X,Y} = \int_{\Re^2} g(x,y) f_{X,Y} dx dy$$

- Similarly if *X* is discrete

$$E[g(X)] = \int_{\Re} g(x) dF_X = \sum_i g(x_i) p_X(x_i)$$

- *X* and *Y* are independent if

$$F_{X,Y}(a,b) = F_X(a)F_Y(b)\ \forall a \text{ and } b$$
$$\text{or } f_{X,Y}(x,y) = f_x(x)f_Y(y)\ \forall x \text{ and } y$$

- Correlation between *X* and *Y*

$$R_{X,Y} = E[XY^*] \text{ and } C_{X,Y} = E[XY^*] - E[X]E[Y]^*$$

- Mutual information between *X* and *Y*

$$I(X;Y) = \int_{\Re^2} f_{X,Y} \log(\tfrac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)})dxdy$$

- *X* and *Y* are independent if

$$F_{X,Y}(a,b) = F_X(a)F_Y(b)\ \forall a \text{ and } b$$
$$\text{or } p_{X,Y}(x_i,y_j) = p_X(x_i)p_Y(y_j)\ \forall i \text{ and } j$$

- Correlation between *X* and *Y*

$$R_{X,Y} = E[XY^*] \text{ and } C_{X,Y} = E[XY^*] - E[X]E[Y]^*$$

- Mutual information between *X* and *Y*

$$I(X;Y) = \sum_{i,j} p_{X,Y}(x_i,y_j)\log\frac{p_{X,Y}(x_i,y_j)}{p_X(x_i)p_Y(y_j)}$$

- Correlation coefficients

$$-1 \leq \rho_{X,Y} = \frac{C_{X,Y}}{\sqrt{Var(X)Var(Y)}} \leq 1$$

- Mutual information

$$0 \leq I(X;Y)$$

- All these measure relationship among variables
  - Correlation, independence, and mutual information

- Example 2.1: If *X* and *Y* are independent
  - Then $$C_{X,Y} = 0 \text{ and } I(X;Y) = 0$$
  - If *X* is zero mean and has a symmetric density and *Y* is squared *X* then
    - Are *X* and *Y* independent?
    - Are they uncorrelated?
    - Is their mutual information zero?

- Mutual information seems to be a powerful metric of dependency
- The origin of mutual information dates back to late 1940s.
- It is based on the concept of entropy from thermodynamics and statistical mechanics from mid 1800s.

- We can define a triple probability space to describe uncertainty of our system

$$(\Omega, \mathcal{F}, P)$$

- The outcome of the experiment $w \in \Omega$

- The universal set of possible outcomes Ω

- A relevant event *A* as a collection of outcomes of interest $w \in A$

- The probability of an event *P(A)*

- A random variable $X : (\Omega, \mathcal{F}) \to (\Re, \mathbb{B}(\Re))$

- Information content of an event $-\log_2(P(A))$ where $A \in \mathcal{F}$

- Average information content of a discrete random variable

$$H(X) = -\sum_i p_X(x_i) \log p_X(x_i)$$

- It is the entropy

$$H(X) \geq 0$$

- Differential entropy of a continuous random variable

$$h(X) = -\int_x f_X(x) \log f_X(x) dx$$

- Differential entropy can be negative.

- It is best used comparing *h(X)* and *h(Y),* hence the concept of differential

- An alternative, formulation

$$I(X;Y) = H(Y) - H(Y|X) = H(X) - H(X|Y)$$

$$I(X;Y) = h(Y) - h(Y|X) = h(X) - h(X|Y)$$

- Yet another formulation based on a distance measure

- The “distance” between two probability measures (PDF or PMF)
  - Kullback-Leibler distance

$$D_{KL}(f_X||g_X) = \int_x f_X(x) \log \frac{f_X(x)}{g_X(x)} dx$$

$$D_{KL}(f||g) \geq 0$$

$$I(X;Y) = D_{KL}(f_{X,Y}||f_X f_Y)$$

- Recall inference is a critical outcome of many problems in data analysis

- In all inference problems, we have an objective, therefore, we have loss and risk

- Loss function

$$\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \Re$$

- Examples are

$$\text{if } \mathcal{Y} = \{0, 1\} \text{ then } \ell(y, \hat{y}) = 1 \text{ if } y \neq \hat{y}$$
$$\text{if } \mathcal{Y} = \Re \text{ then } \ell(y, \hat{y}) = (y - \hat{y})^2 \text{ or } E(Y - \hat{Y})^2$$

- Risk of inference

- Finding the output corresponding an input

$$g : \mathcal{X} \to \mathcal{Y}$$

- The performance of a given mapping

$$R(g) = E[\ell(Y, g(X))]$$

- The optimum mapping

$$R^* = \inf_g R(g) = \inf_g E[\ell(Y, g(X))]$$

- Example 2.2
- The connection between estimation and information theory
- Assume data $\mathbf{X} = (X_1, X_2, \ldots, X_n)^\top$
  - Data is assumed independent and identically distributed with probability mass function $p_{X_i}(x)$
- The objective:
  - Find a distribution for the data that maximizes the likelihood of the data

- Find the probability mass function that generated the data, that is,

$$p_{X_i} \text{ for observed } (x_1, x_2, \ldots, x_n)$$

- Can data provide a mechanism to find the underling distribution that generated the data?

- Find the model among the set of possible models that maximizes the likelihood of generating the data.

- The maximum likelihood estimate of the probability among a set is

$$\arg\max_{q \in \mathcal{Q}} q_{\mathbf{X}}(\mathbf{x}) = \arg\max_{q \in \mathcal{Q}} \log q_{\mathbf{X}}(\mathbf{x}) = \arg\min_{q \in \mathcal{Q}} -\log q_{\mathbf{X}}(\mathbf{x})$$

- *q* is a possible probability mass function that could have generated the data
  - q is the probability that x = 0
- An appropriate loss function could be the negative log loss

- The loss function

$$\ell(y, \hat{y}) = \ell(q, \mathbf{X}) = -\log q_{\mathbf{X}}$$

- The risk

$$\begin{aligned} R(q) =& E[\ell(y, \hat{y})] = E_p[\ell(q, \mathbf{X})] = -E_p[\log q_{\mathbf{X}})] \\ =& D_{KL}(p||q) + E_p[\ell(p, \mathbf{X})] \\ =& D_{KL}(p||q) + R(p) \end{aligned}$$

- The risk is minimized with *q* = *p*

- The minimum risk is

$$R^* = E_p[\ell(p, \mathbf{X})] = H(p)$$

- A specific case is binary independent identically distributed sequence of data

$$\mathbf{X} = (X_1, X_2, \ldots, X_n)^\top \text{ with } X_i \in \{0, 1\}$$

- Ground truth

$$p_{\mathbf{X}}(\mathbf{x}) = [p_{X_i}(x_i)]^n$$

- Find a distribution for the data that maximizes the likelihood of the data

$$\mathbf{x} = (0, 1, 0, 0, 0, 1)$$

- Since the data samples are independent

$$\arg\max_{q \in \mathcal{Q}} q_{\mathbf{X}}(\mathbf{x}) = \arg\max_{q \in \mathcal{Q}} \prod_{i=1}^{n} q_{X_i}(x_i)$$

- Since data are binary

$$\arg\max_{q \in \mathcal{Q}} q_{\mathbf{X}}(\mathbf{x}) = \arg\max_{q \in [0,1]} q^l (1-q)^{(n-l)}$$

- The maximum likelihood estimate of the probability *q* is derived

$$\frac{d(q^l (1-q)^{(n-l)})}{dq} = 0$$

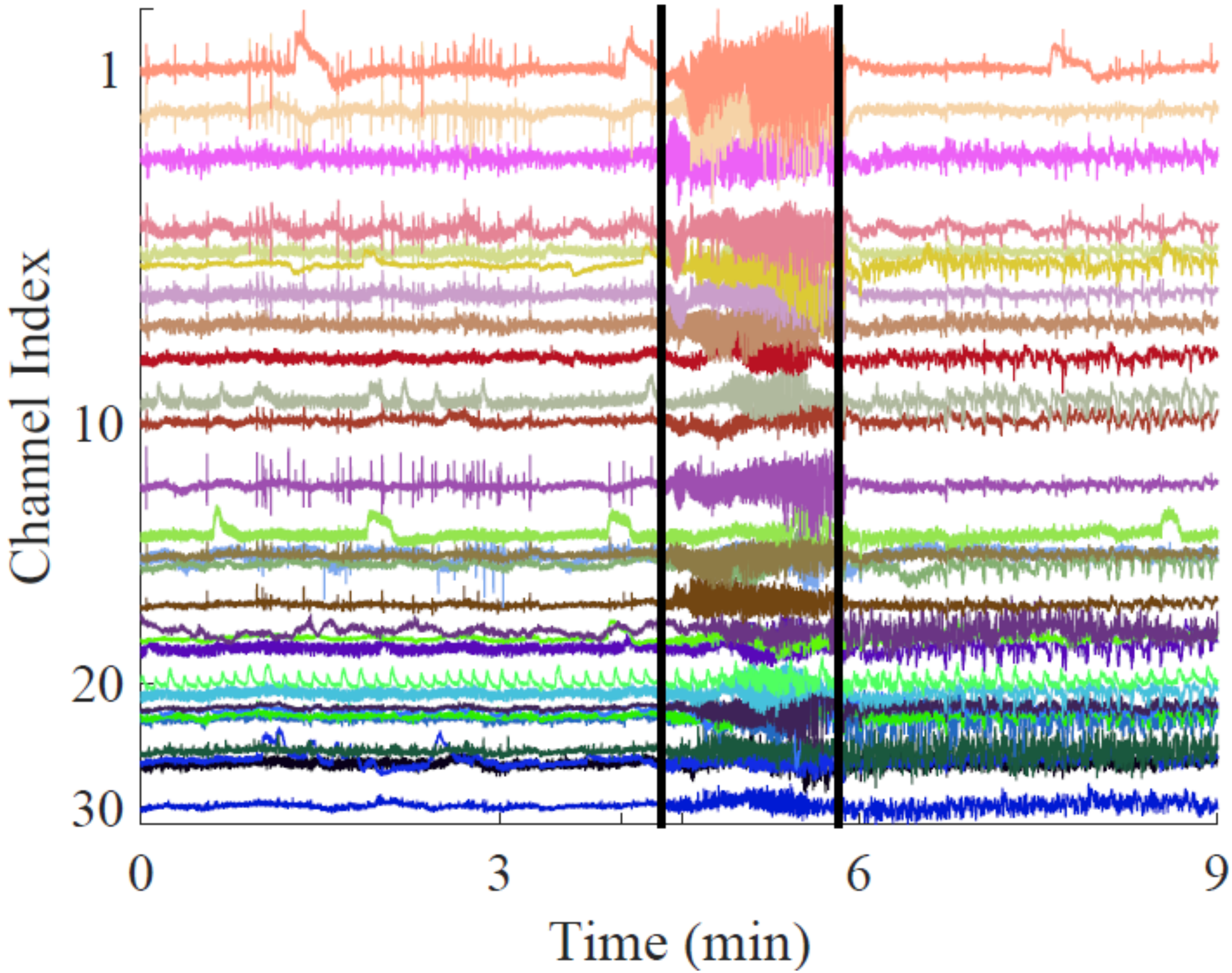
- The most likely probability is $q^* = \frac{l}{n}$

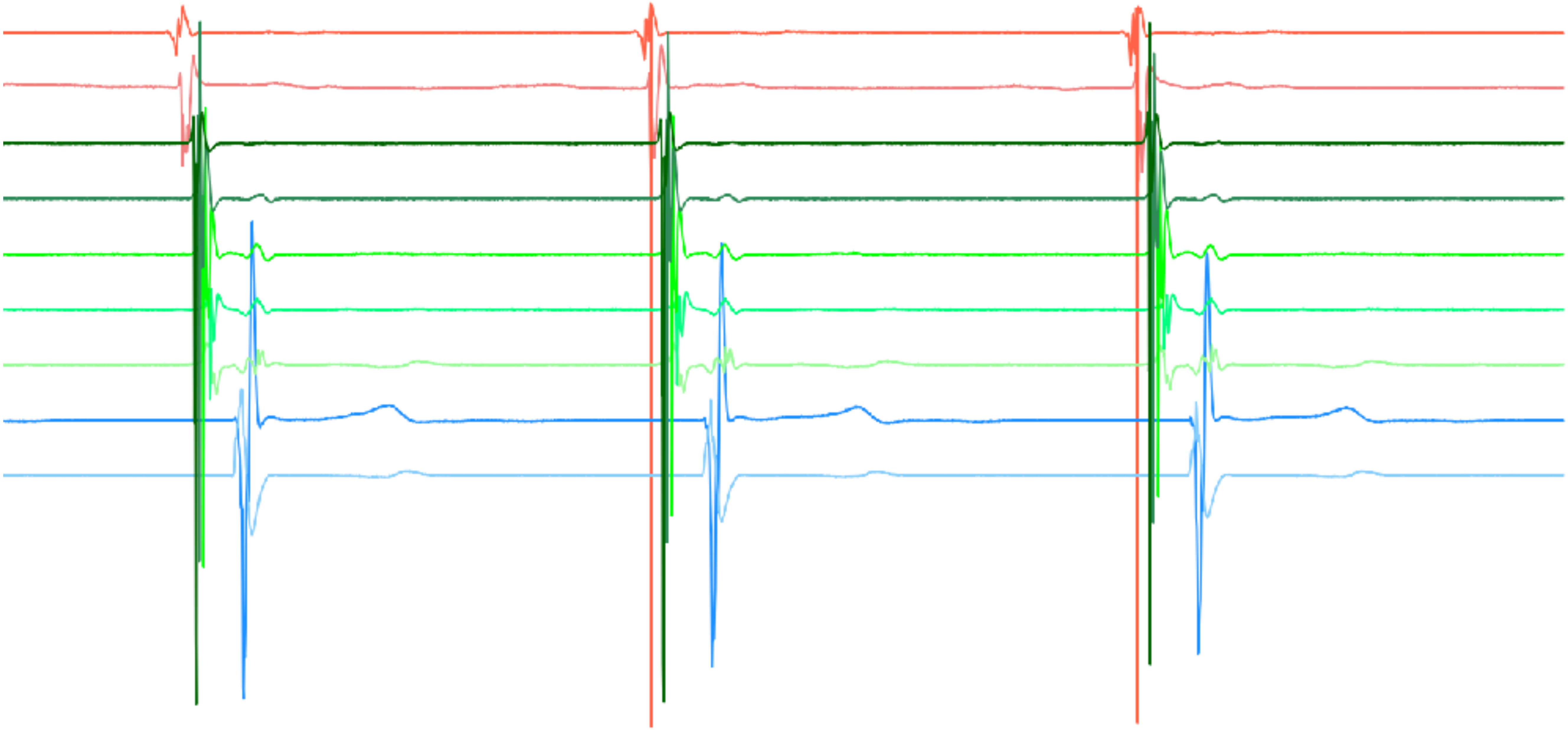
- In the specific case of $\mathbf{x} = (0, 1, 0, 0, 0, 1)$
- The ML estimate is $q^* = 2/3$
- Obviously the ground truth is not known.

3. Data

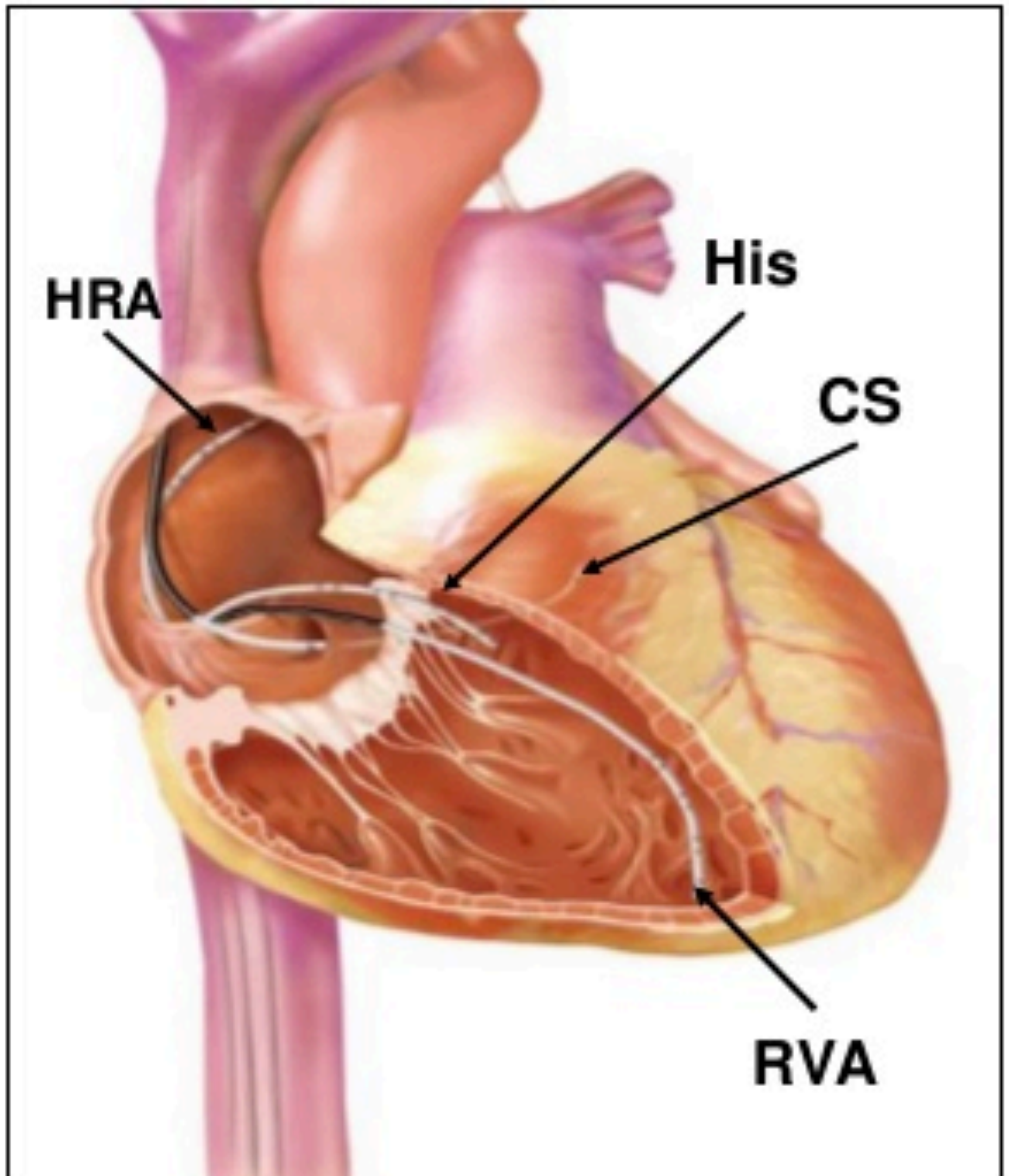
- Temporal observations $X_1, X_2, \ldots, X_n$
- Temporal relationships R_{X_i, X_j}
- Spatial observations $X^{(1)}, X^{(2)}, \ldots, X^{(d)}$
- Spatial relationships $R_{X^{(k)}, X^{(l)}}$

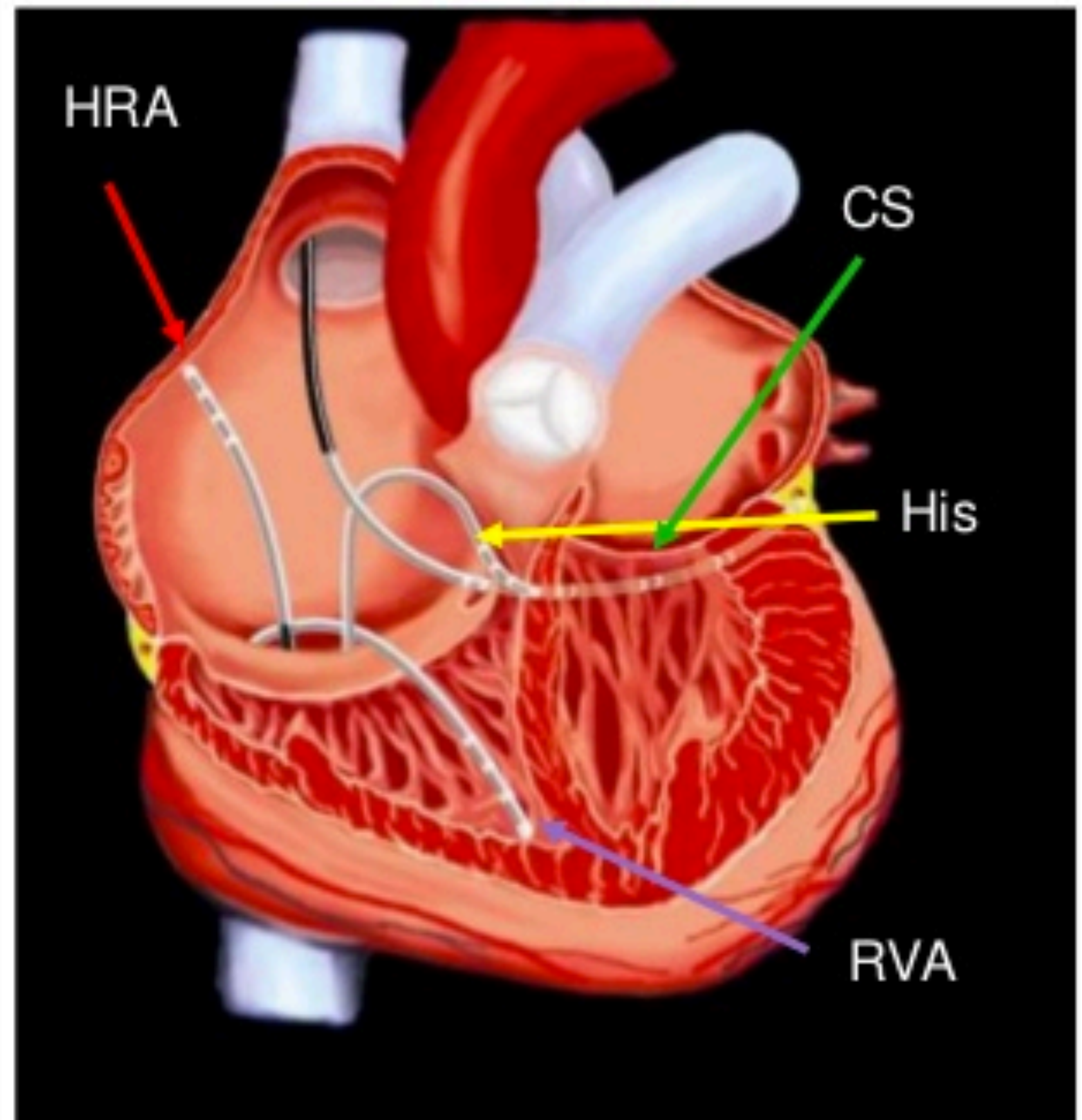
- 3 illustrative examples of data

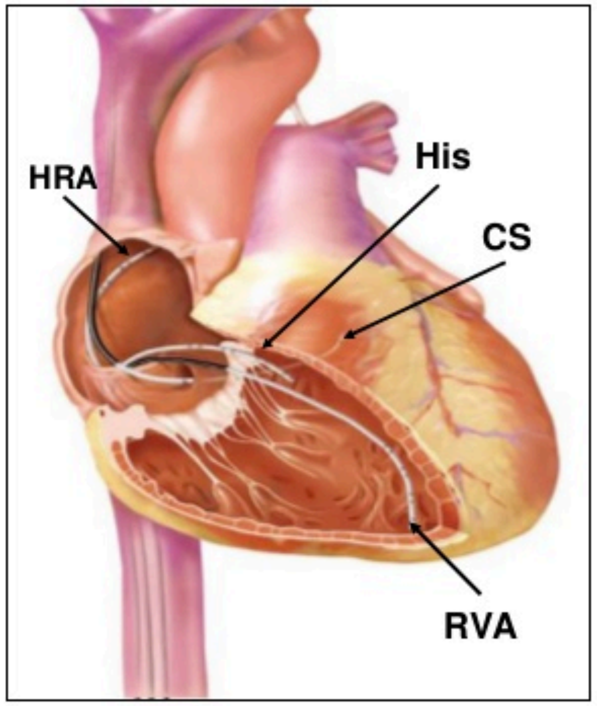

1
10
20
30
Channel Index
0
3
6
9
Time (min)

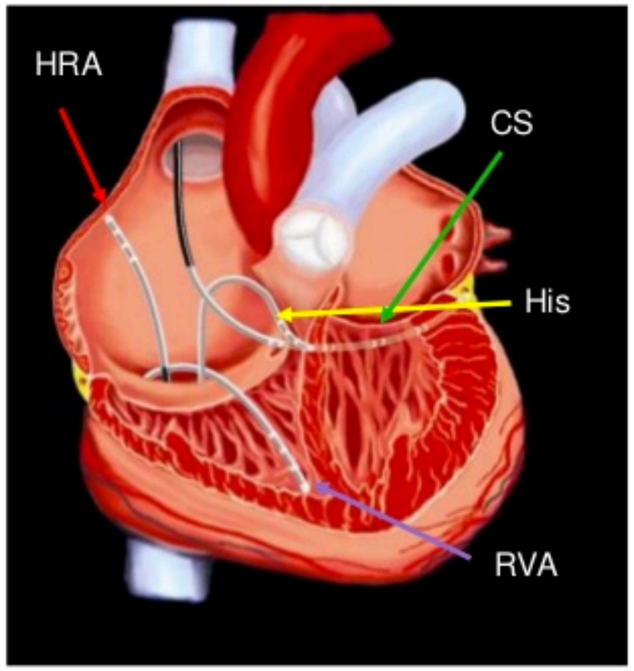


Intracardiac Electrogram Recordings – Catheter Placement





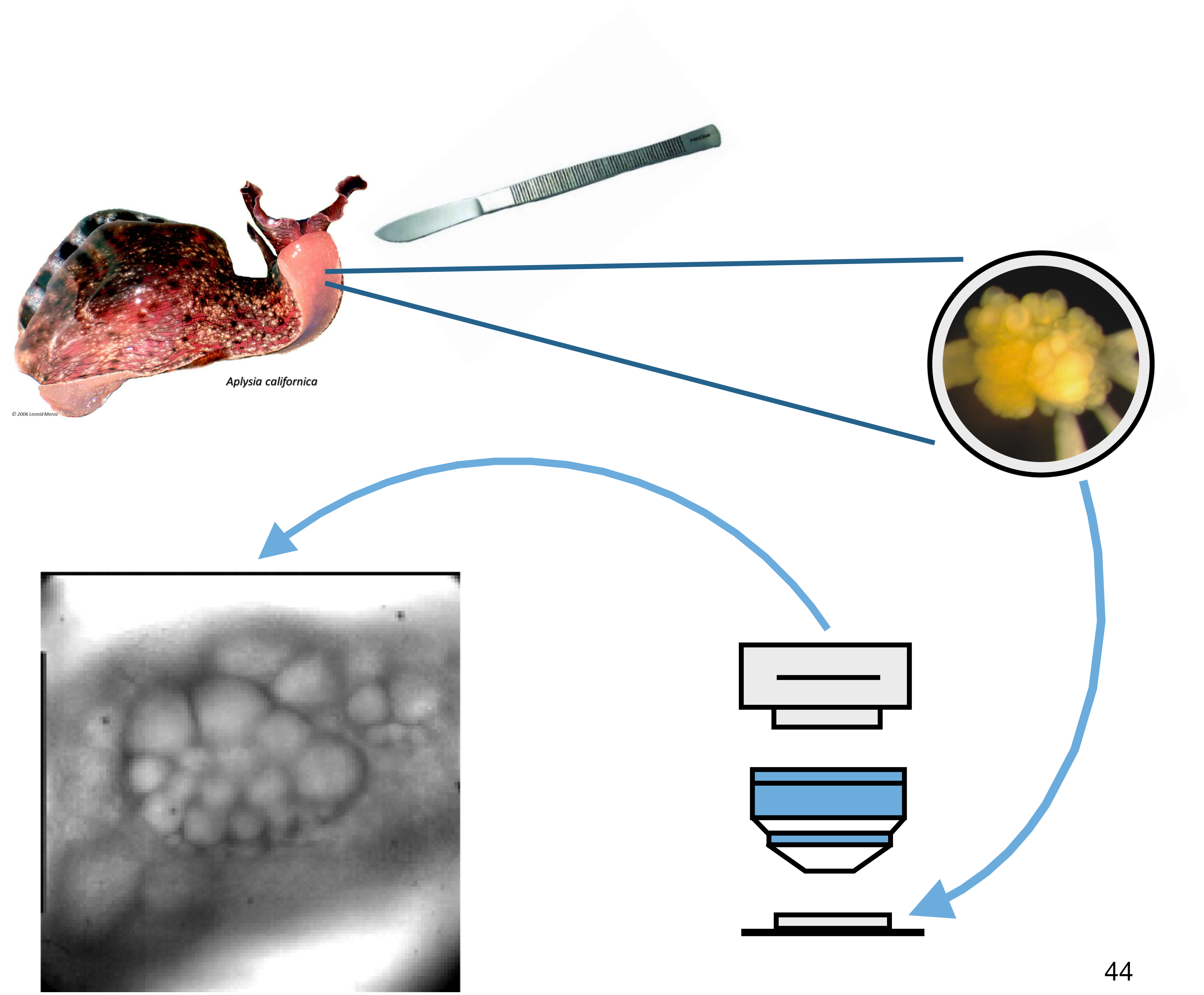


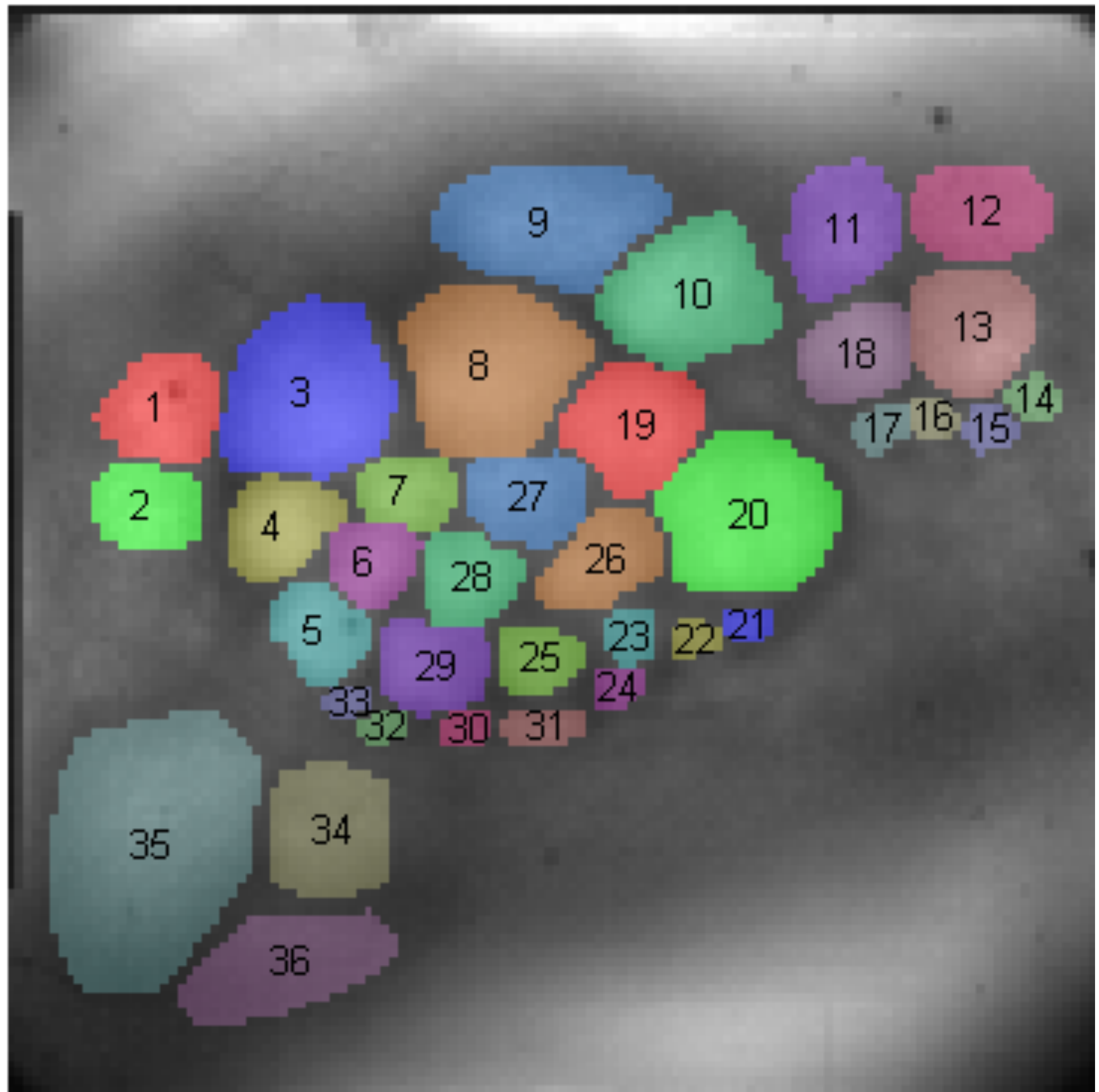


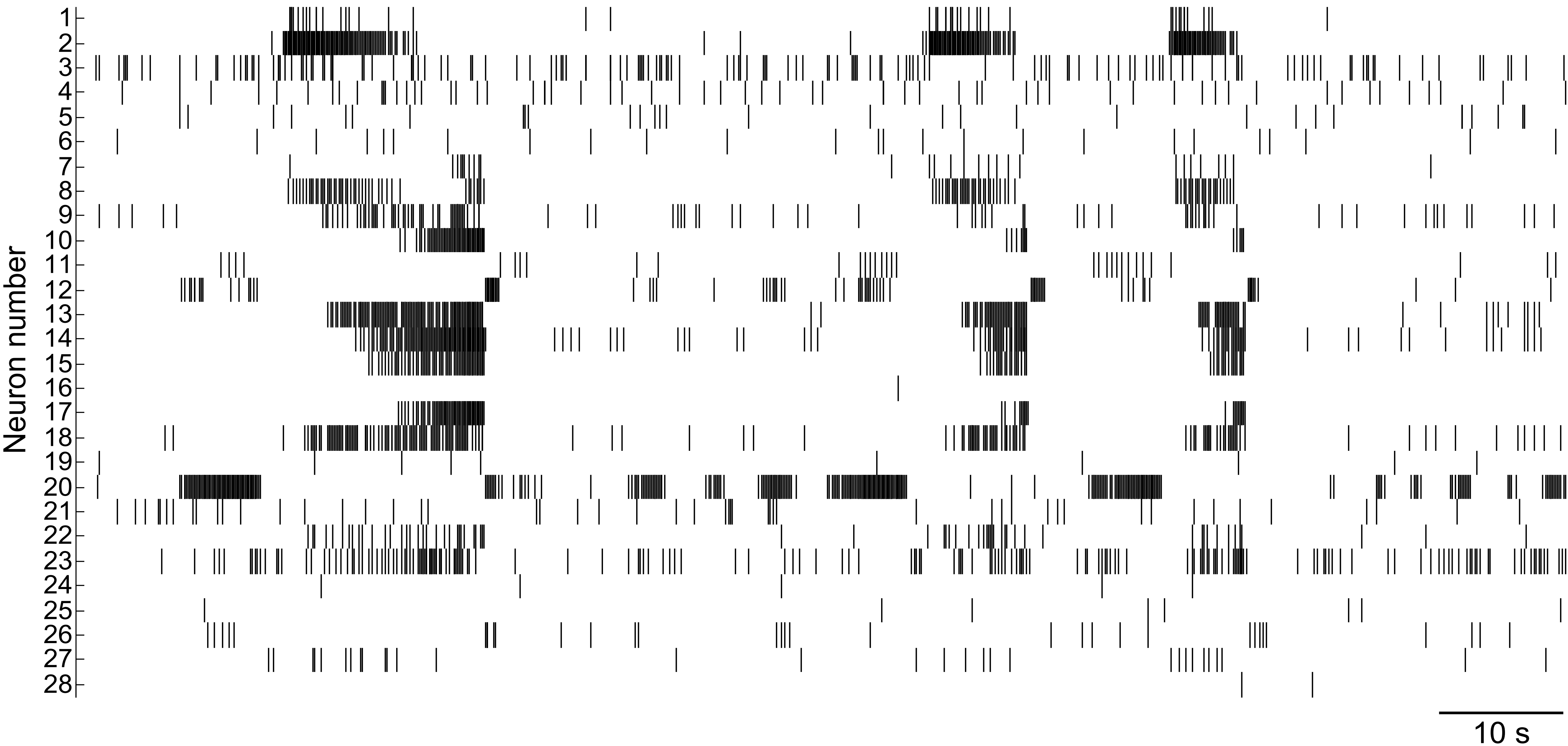
- High right atrial
- His* bundle
- Coronary sinus
- Right ventricle apex

* William His, Junior, a Swiss cardiologist, 1893

- A very different example,
    - Voltage sensitive dye




1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
Neuron number
10 s

- Often recorded data are continuous time signals

$$X_t^{(1)}(w), X_t^{(2)}(w), \dots, X_t^{(d)}(w) \ \forall t \text{ and } w \in \Omega$$

- where $w \text{ is an outcome of the random experiment and } \Omega$ $\text{is the set of all outcomes}$

- Discrete time data is often much more desirable
  - It can be stored
  - It is easy to analyze and process with digital filters

- Continuous time signals can be represented with discrete time data
  - With no loss of information

$$X_t(w)\ \forall t \rightarrow X_1(w), X_2(w), \ldots, X_n(w)$$

    - Sampling
    - Projection

- Sampling and reconstruction

$$X_t(w) = \sum_{n=-\infty}^{+\infty} X_{nT}(w) \frac{\sin(W[t - nT])}{W(t - nT)}$$

- Where *W* is the bandwidth of the power spectral density and $T = \dfrac{\pi}{W}$

- The power spectral density of the process is $S_X(f) = \mathcal{F}\{R_X(\tau)\}$

- The autocorrelation is $R_X(\tau) = E\{X_{t+\tau} X_t^*\}$

- The data signal is assumed to be wide sense stationary (wss)

- Example 3.1 : Assume that the process is ideally band limited, that is,

$$S_X(f) = \begin{cases} \frac{\mathcal{N}_0}{2} & \text{if } f \in [-W, W], \\ 0 & \text{otherwise} \end{cases}$$

- In this example,

$$R_X(\tau) = \frac{\mathcal{N}_0}{2T} \frac{\sin(W\tau)}{W\tau}$$

- Where $T = \dfrac{\pi}{W}$

- And $E[X_{nT} X^*_{mT}] = 0 \text{ if } m \neq n$

- If the data signal is wide sense stationary
  - That is,

$$R_X(\tau) = E\{X_{t+\tau} X_t^*\} \; \forall \tau \text{ not a function of } t$$

- The discrete samples carry all the information in the data signal

$$\ldots, X_{-T}, X_0, X_T, X_{2T}, \ldots, X_{nT}$$

- Since we have

$$X_t(w) = \sum_{n=-\infty}^{+\infty} X_{nT}(w) \frac{\sin(W[t - nT])}{W(t - nT)}$$

- The discrete samples carry all the information in the data signal

$$\ldots, X_{-T}, X_0, X_T, X_{2T}, \ldots, X_{nT}$$

- These samples will be uncorrelated (independent if the signal is Gaussian) if the spectrum is ideally band-limited.

- No need to carry the sampling period in the notation

$$\mathbf{X} = (X_1, X_2, \ldots, X_n)^\top$$

- In general, for band limited processes, the samples are correlated.
- The samples can be made uncorrelated using whitening linear filters.
- Define zero mean process $\mathbf{X} = (X_1, X_2, \ldots, X_n)^\top$
- The *n x n* covariance matrix $\Sigma_X = E[\mathbf{X}\mathbf{X}^\top]$
  - It is square
    - non-negative definite
    - Hermitian matrix

- The covariance matrix $\Sigma_X = E[\mathbf{X}\mathbf{X}^\top]$

- If the covariance matrix is positive definite

- Linear transformation $\mathbf{Y} = A\mathbf{X}$

- The matrix *A* could be an *m x n matrix* and ***Y*** will then be *m x1*

- Then, the *m x m* covariance matrix of ***Y*** is $\Sigma_Y = A\Sigma_X A^\top$

- If $\Sigma_X = CC^\top \text{ then } \mathbf{Y} = C^{-1}\mathbf{X} \text{ has } \Sigma_Y = I$

- Example 3.2

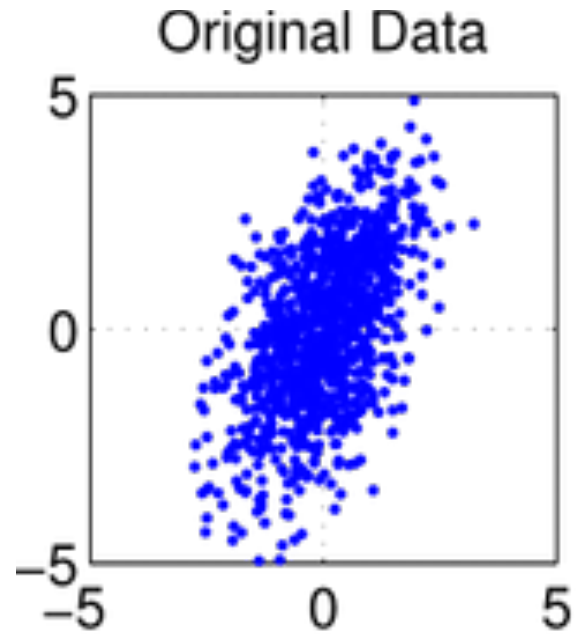
$$\Sigma_Y = A\Sigma_X A^\top$$

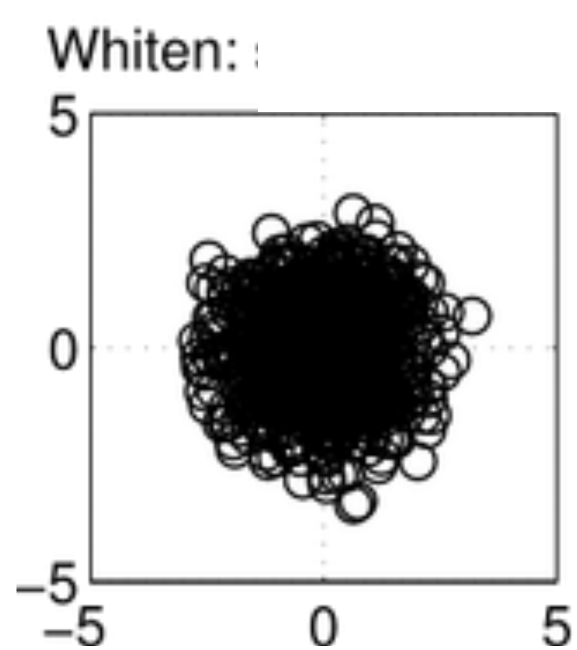
$$\Sigma_X = CC^\top \text{ then } \mathbf{Y} = C^{-1}\mathbf{X} \text{ has } \Sigma_Y = I$$

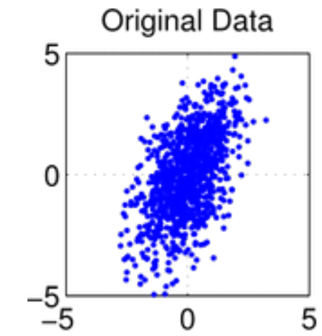
$$\begin{pmatrix} 4 & 12 & -16 \\ 12 & 37 & -43 \\ -16 & -43 & 98 \end{pmatrix} = \begin{pmatrix} 2 & 0 & 0 \\ 6 & 1 & 0 \\ -8 & 5 & 3 \end{pmatrix} \begin{pmatrix} 2 & 6 & -8 \\ 0 & 1 & 5 \\ 0 & 0 & 3 \end{pmatrix}$$

$$\mathbf{Y} = \begin{bmatrix} 1/2 & 0 & 0 \\ -3 & 1 & 0 \\ 19/3 & -5/3 & 1/3 \end{bmatrix} \mathbf{X}$$

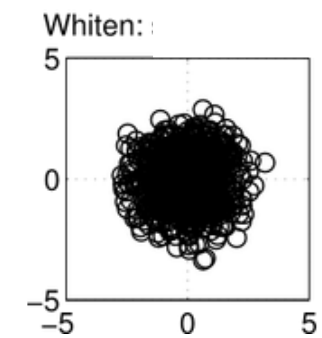
- Example 3.3







- Example 3.3
  - One interpretation
    - Different elements of the original data are correlated
      - if one element is 1.2 it is very likely that the other element is close to 1.
      - When data is whitened, then in the processed data, if one element is 1.2 the other one is still widely distributed
      - Still no information is lost



- Sampling “would not work” when the random signal is not wide sense stationary
  - Even if wss, the samples could be, often are, correlated
- Karhunen-Loeve expansion of a more general random signal

$$R_X(t, s) = E[X_t X_s^*]$$

- The autocorrelation

$$\int_{-\infty}^{+\infty} R_X(t, s)\alpha_n(s)ds = \lambda_n \alpha_n(t)\ \forall t$$

- Eigenfunctions of the autocorrelation function

- Example 3.4
  - A concept analogous to eigenvectors of a matrix $A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$

$$A\mathbf{x} = \lambda \mathbf{x}$$

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \mathbf{x}_1 = \lambda_1 \mathbf{x}_1 \text{ and } \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \mathbf{x}_2 = \lambda_2 \mathbf{x}_2$$

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \lambda_1 = 3 \text{ and } \mathbf{x}_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \lambda_2 = 1$$

  - The eigenvectors are orthogonal since *A* is a symmetric matrix

$$< \mathbf{x}_1, \mathbf{x}_2 >= 0$$

- Analogous to eigenvectors, eigenfunctions are also orthogonal

$$\int_{-\infty}^{+\infty} \alpha_n(t)\alpha_m^*(t)dt = \lambda_n\delta_{n,m}$$

$$\delta_{n,m} = \begin{cases} 1 & \text{if } n = m, \\ 0 & \text{otherwise} \end{cases}$$

- It is intuitive to expect that the projection of the data signal on these eigenfunctions would be orthogonal and uncorrelated if the random process was zero mean.

- Then, we can write

$$X_t(w) = \sum_{n=0}^{+\infty} \alpha_n(t) Z_n(w)$$

- Where $E[Z_n Z_m^*] = \delta_{n,m}$

$$Z_n(w) = \lambda_n^{-1} \int_{-\infty}^{+\infty} X_t(w) \alpha_n^*(t) dt$$

$$\alpha_n(t) = E[X_t Z_n^*] \; \forall t$$

$$\int_{-\infty}^{+\infty} \alpha_n(t) \alpha_m^*(t) dt = \lambda_n \delta_{n,m}$$

- Where

$$\delta_{n,m} = \begin{cases} 1 & \text{if } n = m, \\ 0 & \text{otherwise} \end{cases}$$

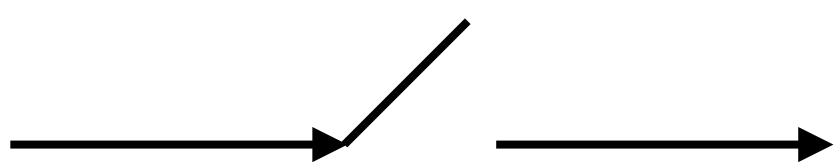
- The information is represented in $Z_0(w), Z_1(w), \ldots, Z_n(w), \ldots$
- The structure is represented in $\alpha_0(t), \alpha_1(t), \ldots, \alpha_n(t), \ldots$
- All because we have

$$X_t(w) = \sum_{n=0}^{+\infty} \alpha_n(t) Z_n(w)$$

- Similar to sampling

$$X_t(w) = \sum_{n=-\infty}^{+\infty} X_{nT}(w) \frac{\sin(W[t-nT])}{W(t-nT)}$$

- where samples carry all the information



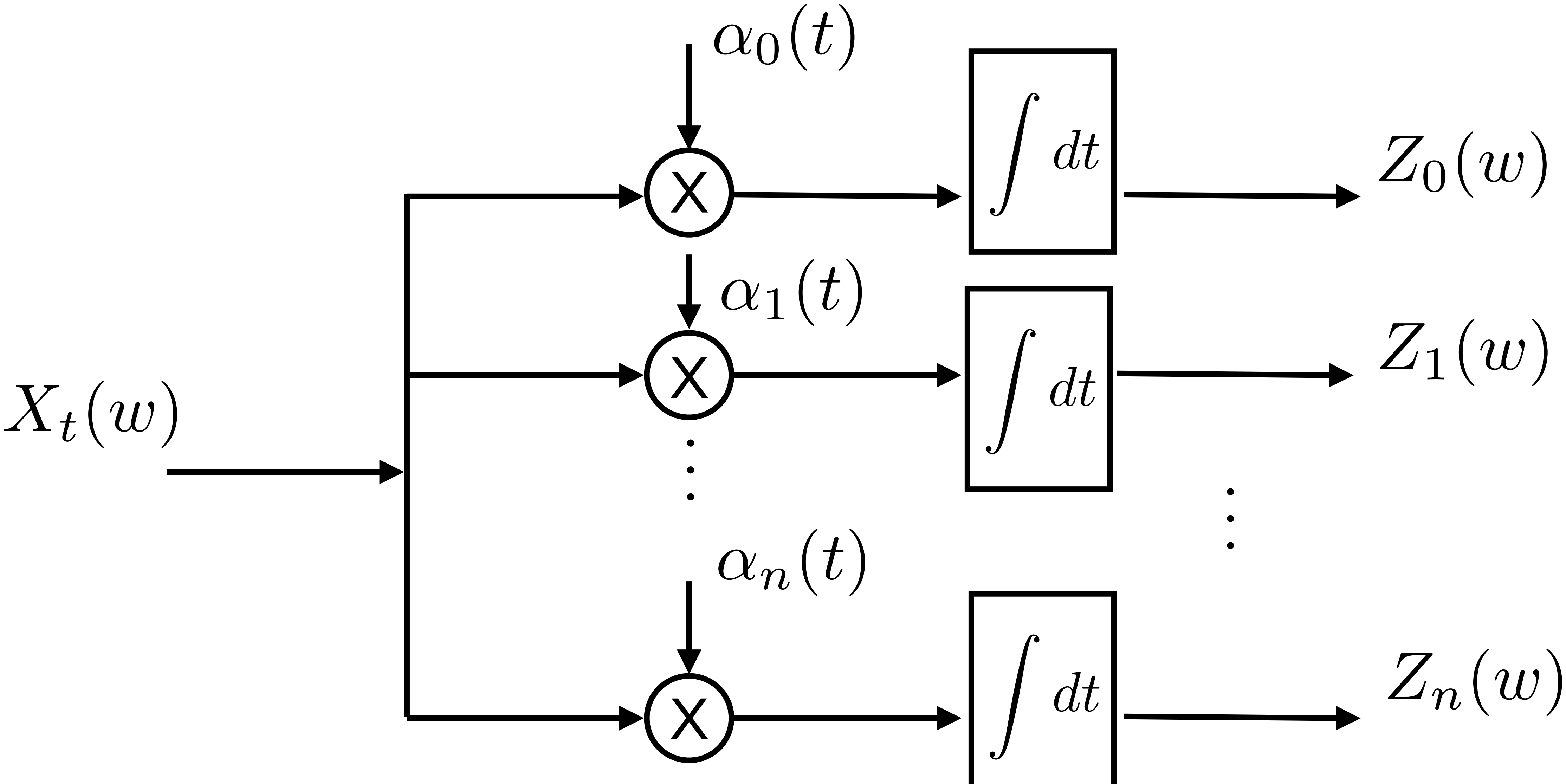
$X_t(w)$

sampling rate 1/T

$$\ldots, X_{-T}, X_0, X_T, X_{2T}, \ldots, X_{nT}, \ldots$$

- Projections on eigenfunctions carry all the information

$$Z_0, Z_1, Z_2, \ldots, Z_n, \ldots$$



- Assume that the data is discrete time stochastic process
  - Parametric models with a few parameters
    - Gaussian, linear, Poisson, …
  - Data driven—“model free”
- Discrete valued time series
- Continuous valued time series

- Assume the data is

$$X_1^n = (X_1, X_2, \ldots, X_n) \text{ where } X_i \in \Re$$

- Then the mutual information between two time series X_1^n and Y_1^n
  - The dependency of one set of data with another

- Example 3.5
  - Two small sets of data and their dependency

$$I(X_1, X_2; Y_1, Y_2) = I(X_1, X_2; Y_1) + I(X_1, X_2; Y_2|Y_1)$$

  - Where

$$I(X_1, X_2; Y_1) = I(X_1; Y_1) + I(X_2; Y_1|X_1)$$

  - Recall that

$$I(X_1; Y_1) = h(X_1) - h(X_1|Y_1) = h(Y_1) - h(Y_1|X_1)$$

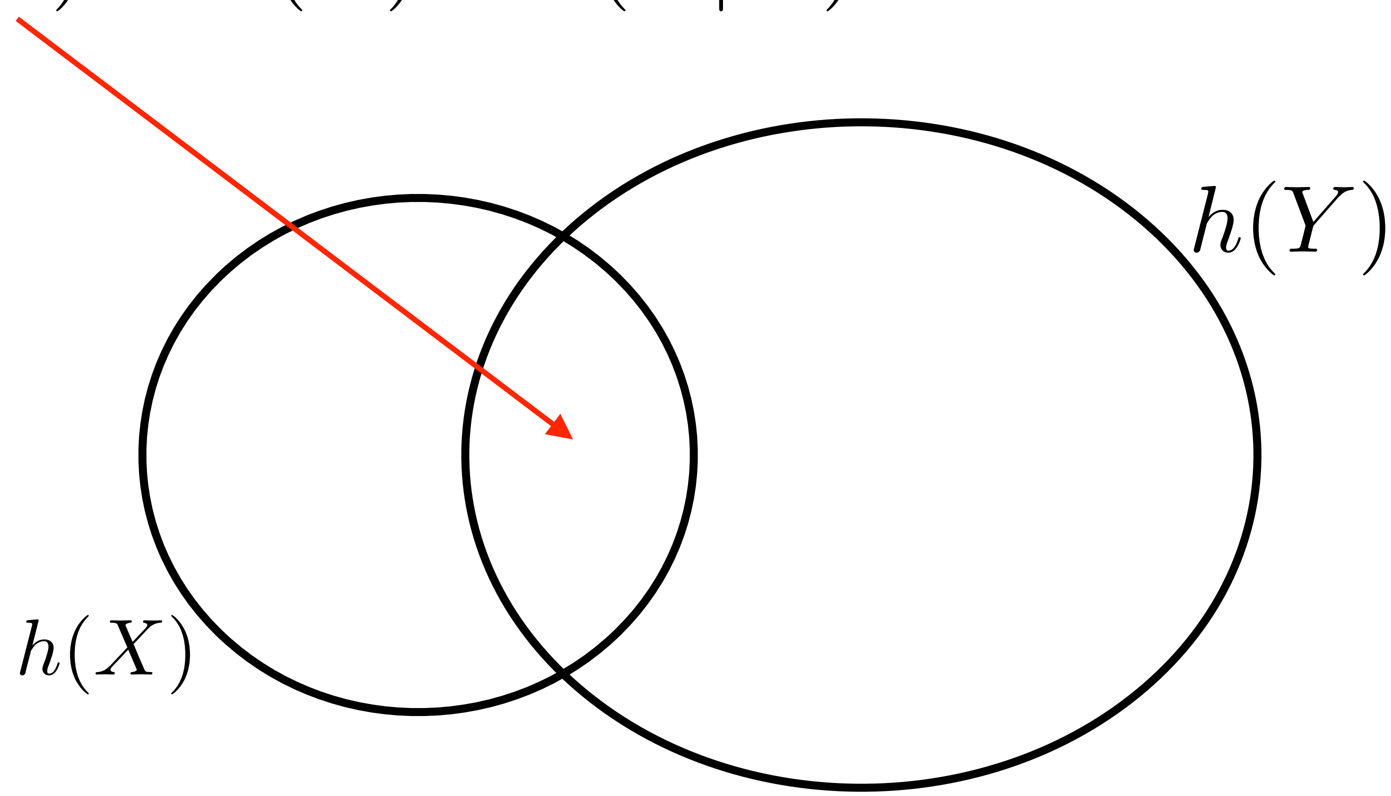
- Example 3.6
    - Lets start with dependencies between two single random variables X and Y.

- Example 3.6

- Assume *X* and *Z* are each a Gaussian random variable and independent

- The model $Y = X + Z$, that is, *Y* is a noisy but direct observation of *X*

$$I(X;Y) = h(Y) - h(Y|X)$$

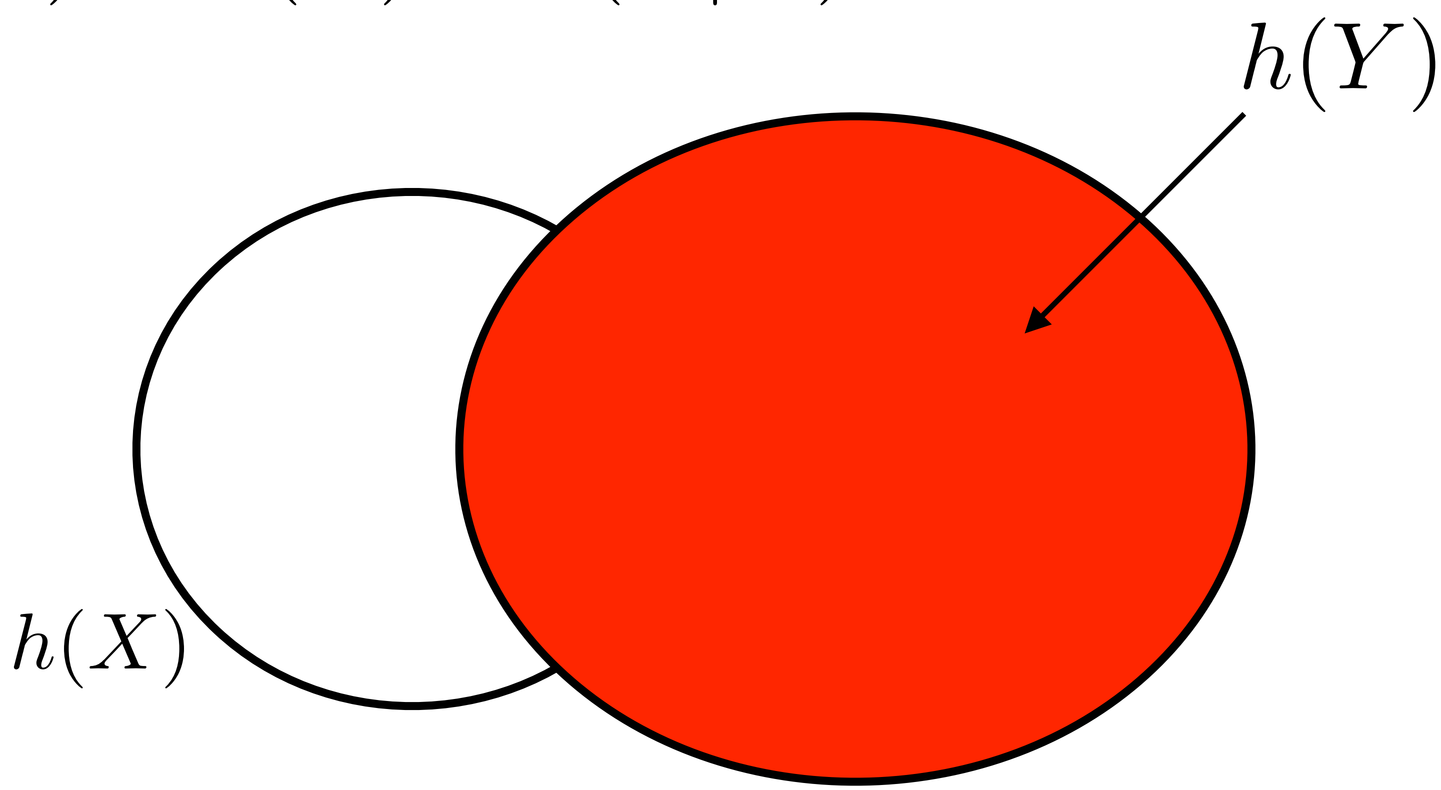


- Example 3.6

- Assume *X* and *Z* are each a Gaussian random variable and independent

- The model *Y* = *X* + *Z* , that is, *Y* is a noisy but direct observation of *X*

$$I(X;Y) = h(Y) - h(Y|X)$$

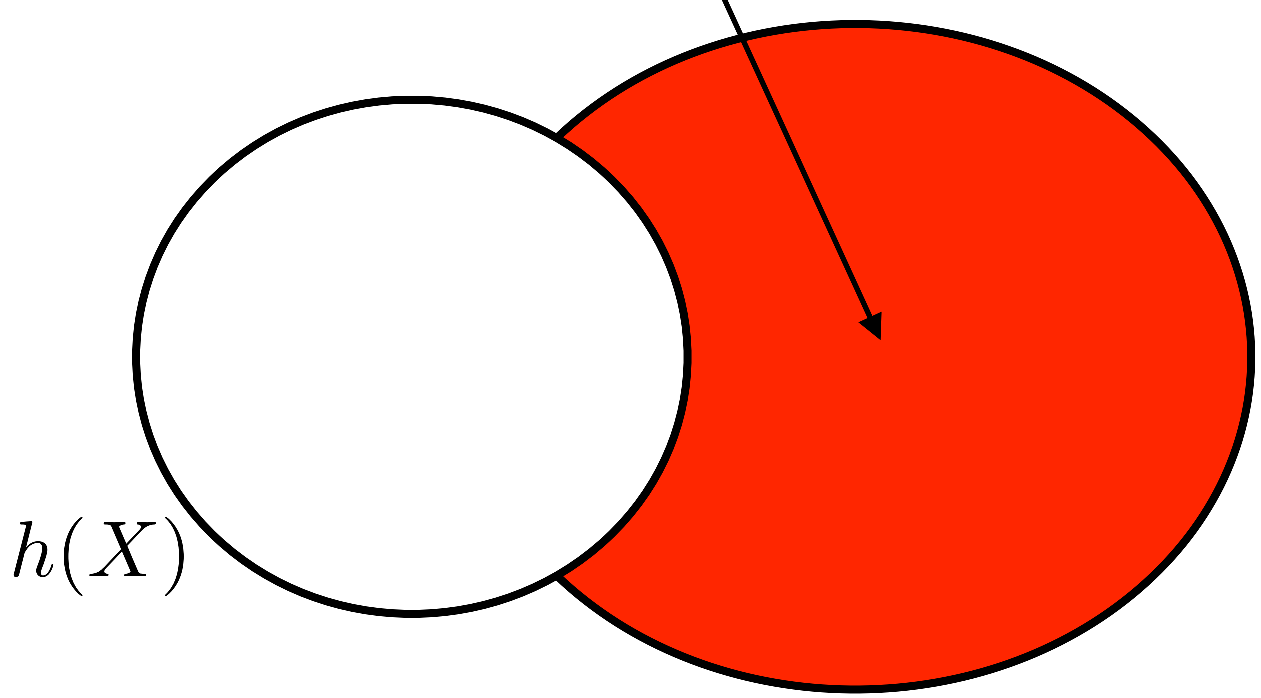


- Example 3.6

- Assume *X* and *Z* are each a Gaussian random variable and independent

- The model $Y = X + Z$, that is, *Y* is a noisy but direct observation of *X*

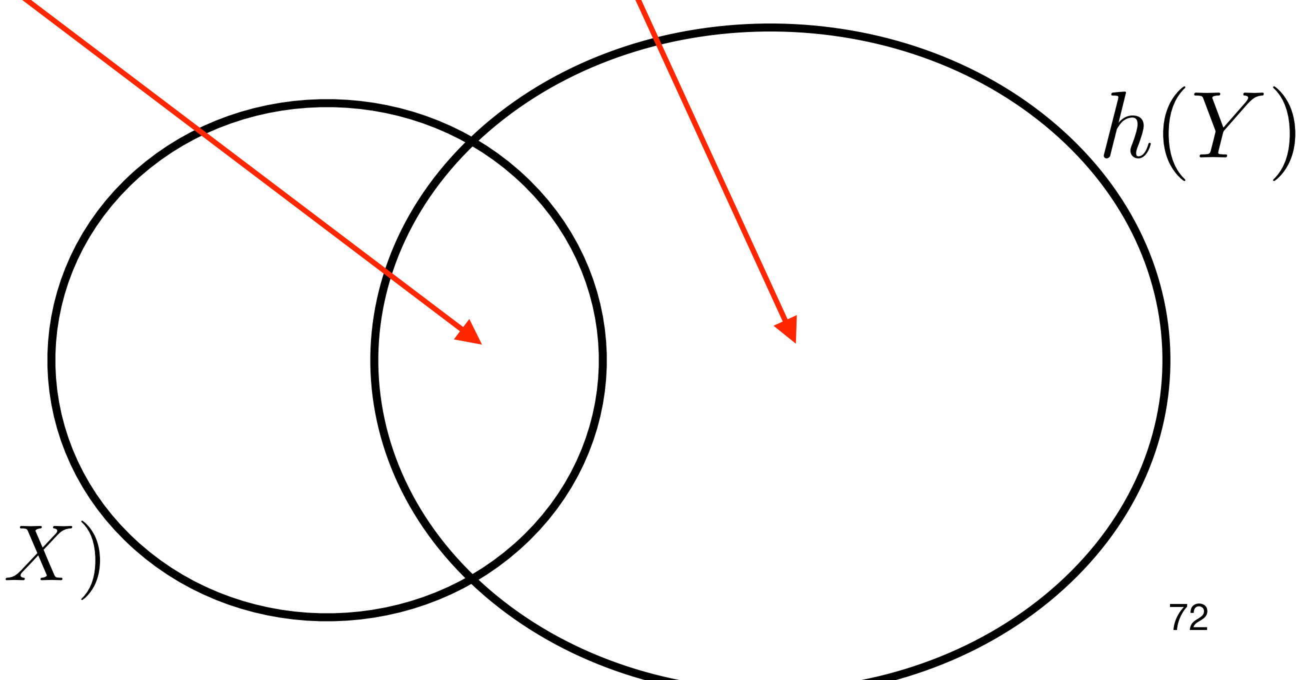
$$I(X;Y) = h(Y) - h(Y|X)$$



- Note that both X and Z are Gaussian and that $h(Y|X) = h(Z)$

$$h(X) = -\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma_X^2}} e^{-x^2/2\sigma_X^2} \log\left(\frac{1}{\sqrt{2\pi\sigma_X^2}} e^{-x^2/2\sigma_X^2}\right) dx$$

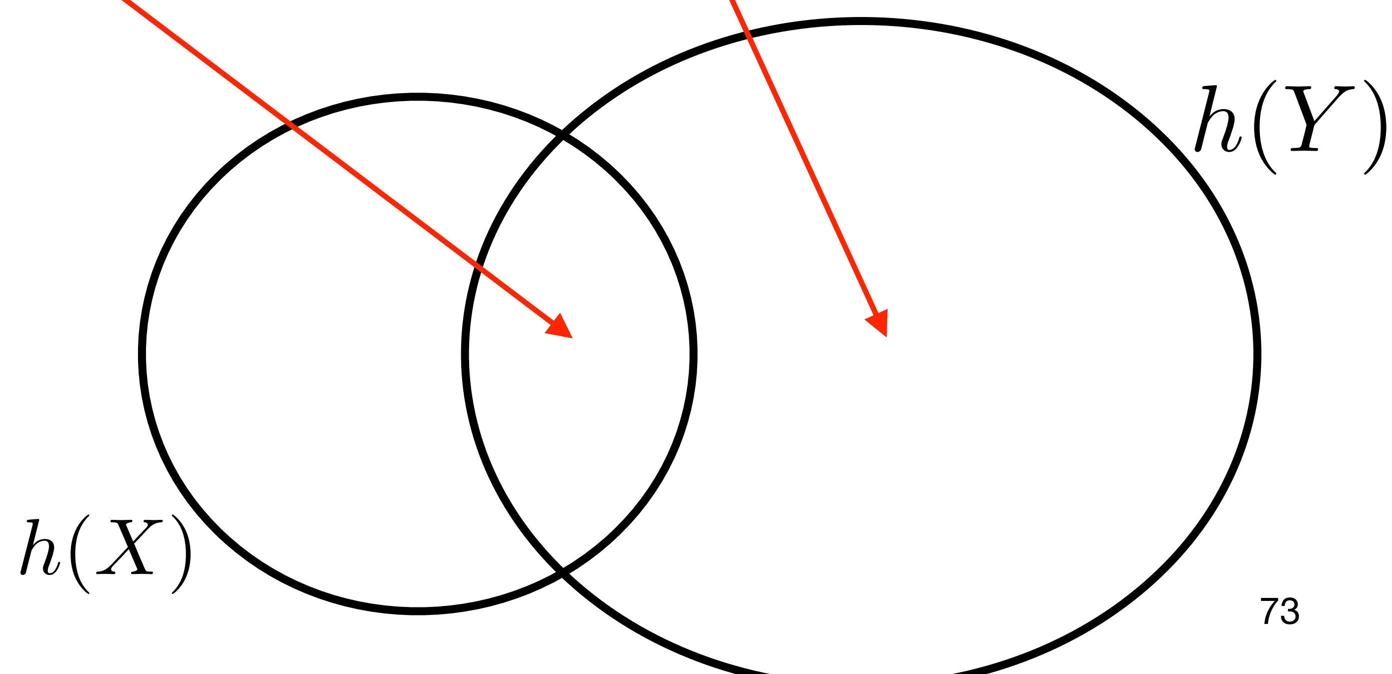
$$I(X;Y) = h(Y) - h(Y|X)$$



- Note that both X and Z are Gaussian and that $h(Y|X) = h(Z)$

$$h(X) = -E_X[\log \frac{1}{\sqrt{2\pi\sigma_X^2}} + \log e^{-X^2/2\sigma_X^2}]$$

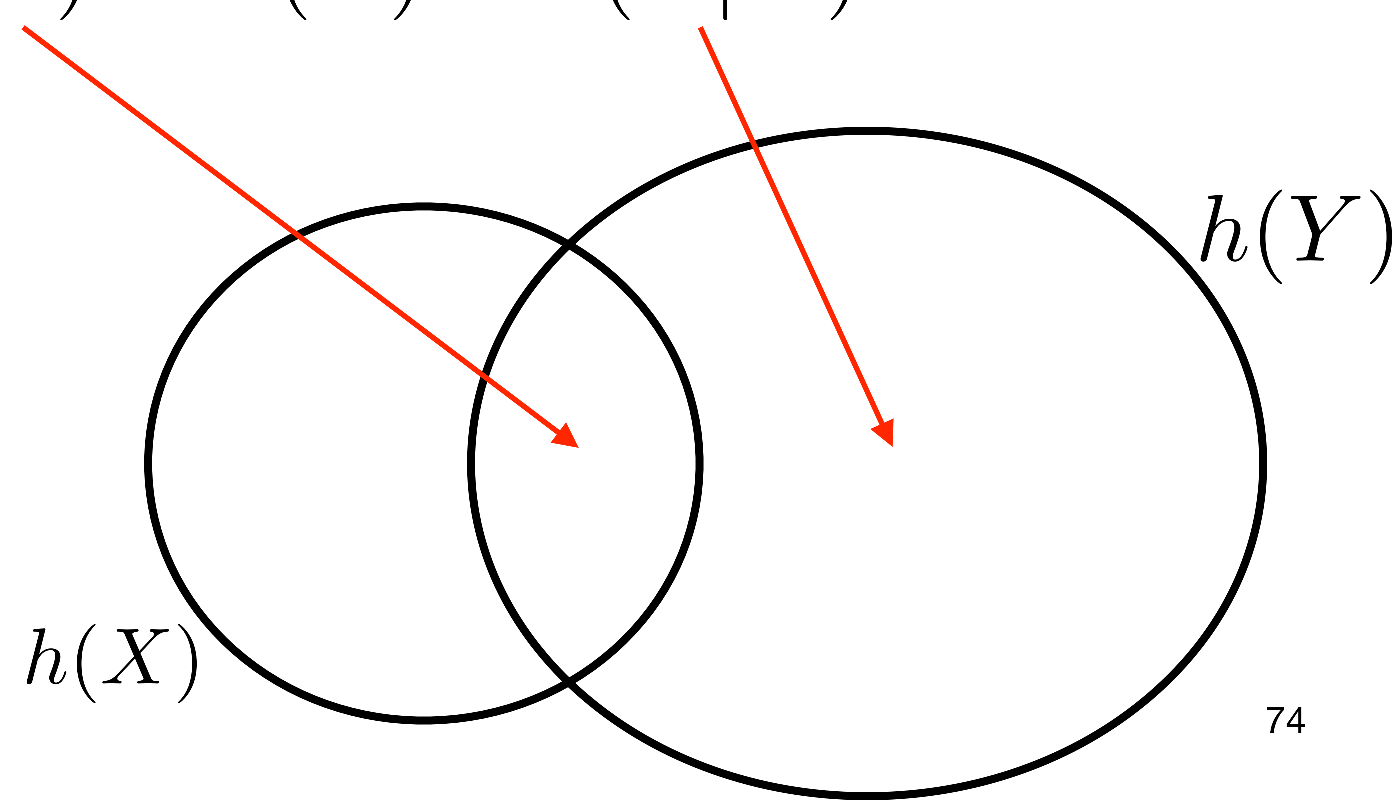
$$I(X;Y) = h(Y) - h(Y|X)$$



- Note that both X and Z are Gaussian and that $h(Y|X) = h(Z)$

$$h(X) = \frac{1}{2}\log 2\pi\sigma_X^2 + \frac{E[X^2]}{2\sigma^2}\log e = \frac{1}{2}[\log 2\pi\sigma_X^2 + \log e]$$

$$I(X;Y) = h(Y) - h(Y|X)$$

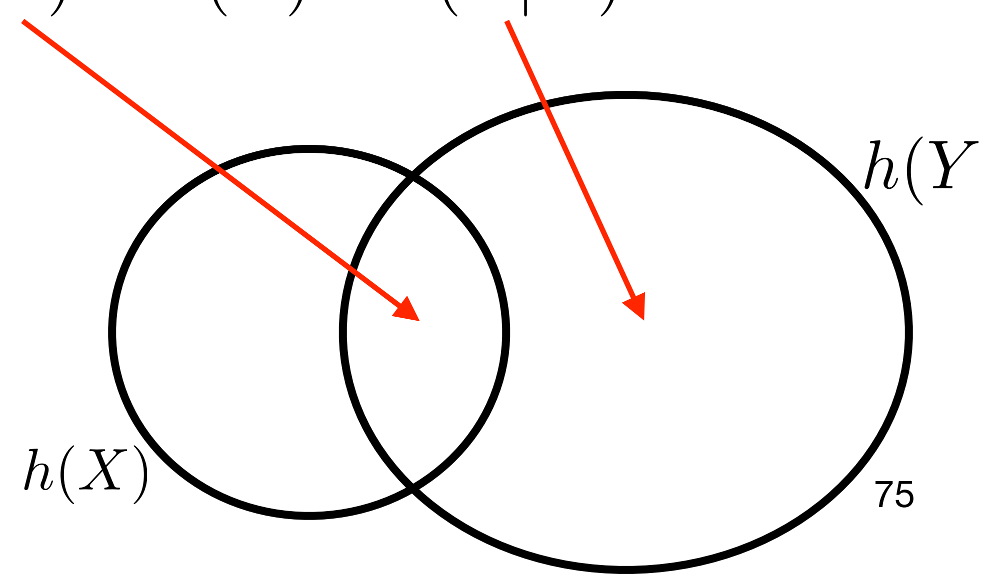


- Note that both X and Z are Gaussian and that $h(Y|X) = h(Z)$

$$h(Y) = \frac{1}{2}\log 2\pi e[\sigma_X^2 + \sigma_Z^2] \qquad h(Y|X) = \frac{1}{2}\log 2\pi e\sigma_Z^2$$

$$I(X;Y) = \frac{1}{2}\log\left(1 + \frac{\sigma_X^2}{\sigma_Z^2}\right)$$

$$I(X;Y) = h(Y) - h(Y|X)$$



- For comparison, let's examine the correlation between *X* and *Y*

$$R_{X,Y}?$$

- For comparison, let's examine the correlation between *X* and *Y*.

- Recall

$$R_{X,Y} = E[XY^*] \text{ and } C_{X,Y} = E[XY^*] - E[X]E[Y]^*$$

- In this example,

$$R_{X,Y} = E[XY^*] = E[X(X+Z)^*] = E[X(X+Z)] = \sigma_X^2$$

- Compared to,

$$I(X;Y) = \frac{1}{2}\log\left(1 + \frac{\sigma_X^2}{\sigma_Z^2}\right)$$

- Back to time series

$$X_1^n = (X_1, X_2, \dots, X_n) \text{ where } X_i \in \Re$$

- Then the mutual information between two time series $X_1^n \text{ and } Y_1^n$

$$I(X_1^n; Y_1^n) = \sum_{i=1}^{n} I(X_1^n; Y_i | Y_1^{i-1})$$

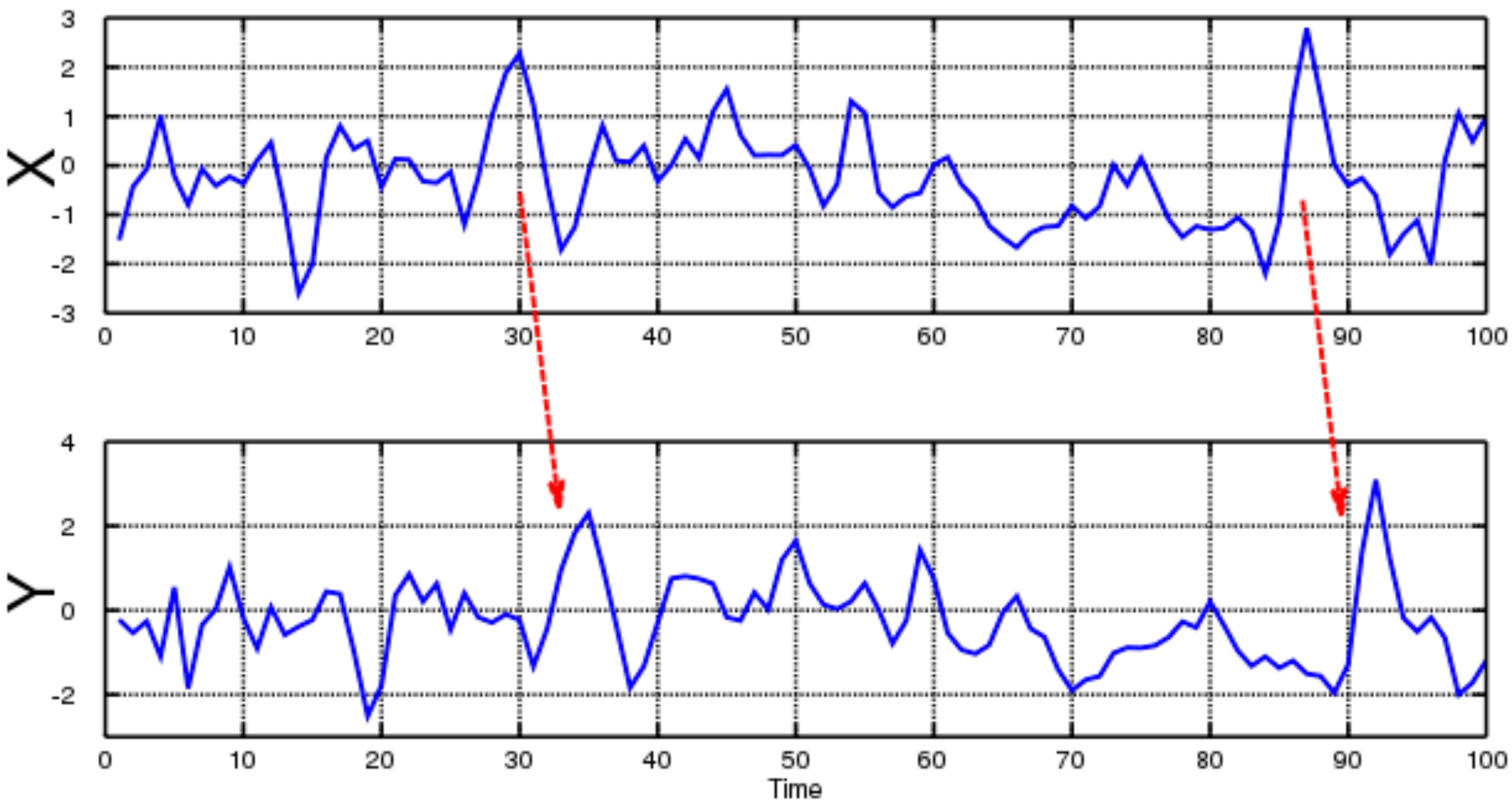
$$= I(X_1^n; Y_1) + I(X_1^n; Y_2 | Y_1) + I(X_1^n; Y_3 | Y_1^2) + \dots$$

- Also

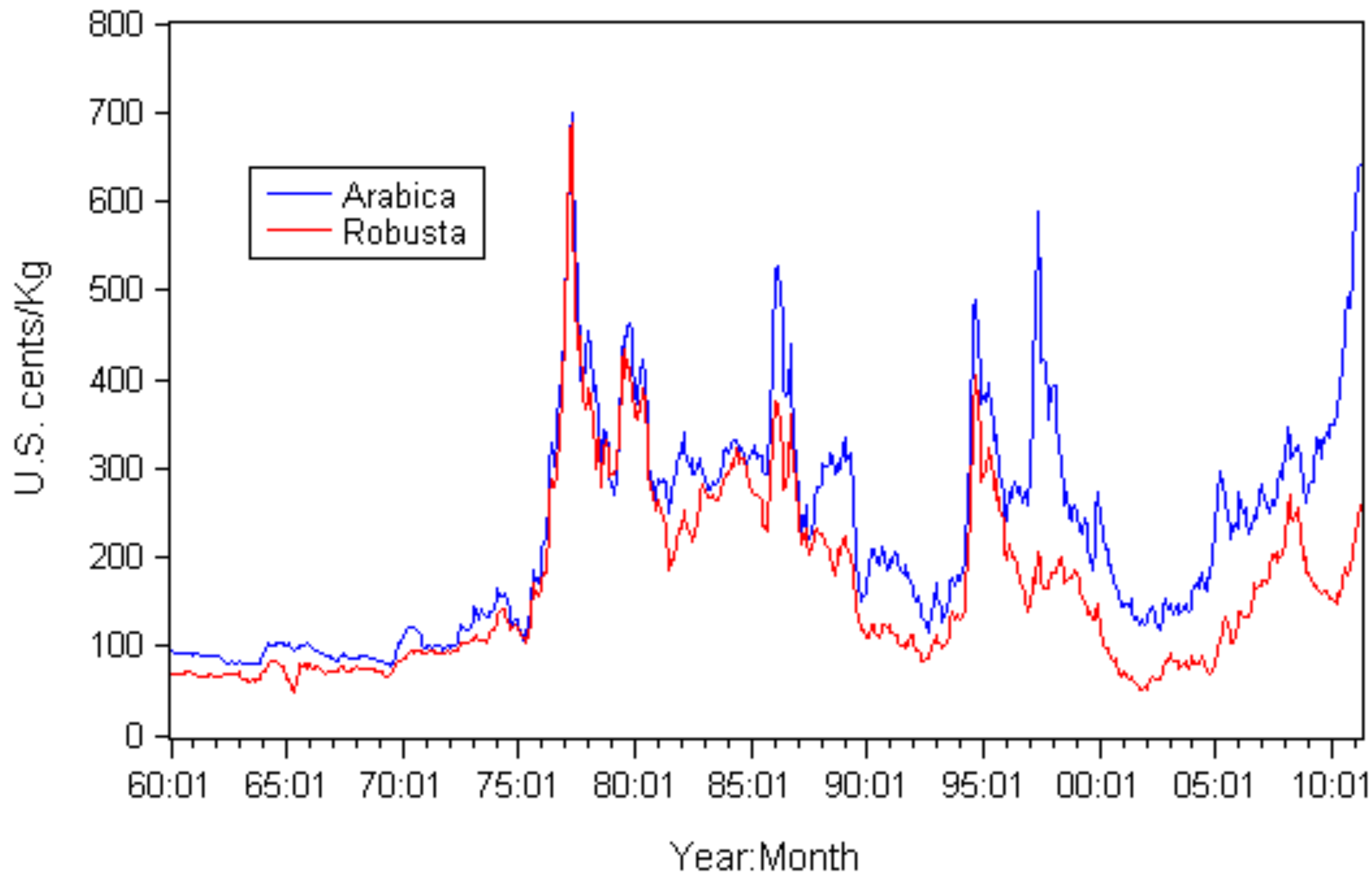
$$I(X_1^n; Y_1^n) = h(Y_1^n) - h(Y_1^n | X_1^n)$$

- Mutual information of two time series measure general “dependence” of the two time series as a whole
  - No temporal information, no influence, nor causality
- It is often critical to measure causality.
  - One data forecasting or influencing another
    - Stock market
    - Transportation
    - Economics

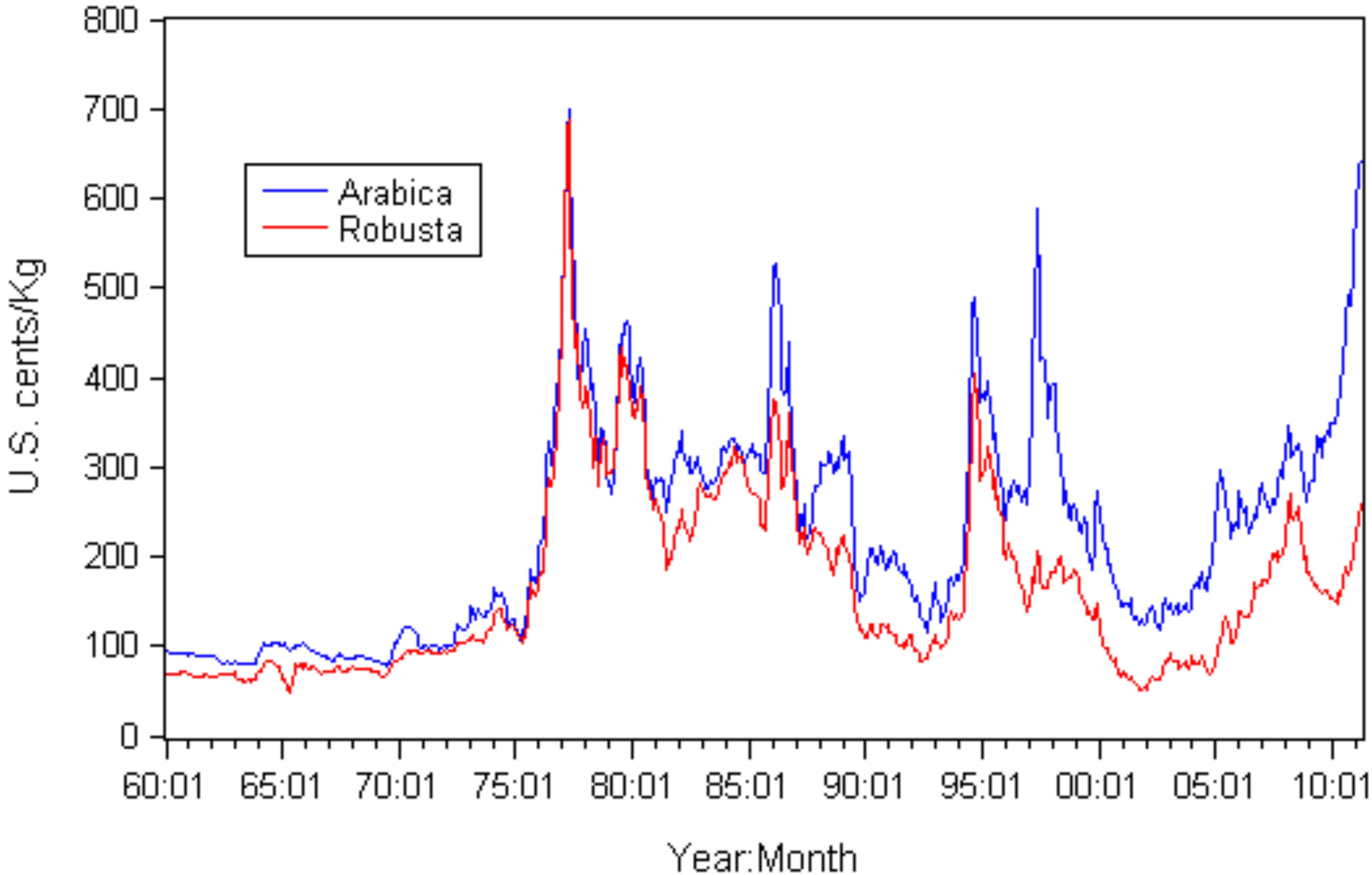
- In this example it is easy to guess that *X* causes *Y*



- In this example it is not easy



Price of Arabica Granger causes price of Robusta



- Grainger causality
  - If signal *X* causes signal *Y* then passed values of *X* should contain information that helps predict *Y* above and beyond the information contained in past values of *Y* alone
  - Granger is defined based on a linear model assumption where *Z* is noise

$$Y_{k+1} = a_0 Y_k + a_1 Y_{k-1} + \ldots + b_0 X_k + b_1 X_{k-1} + \ldots + Z_k$$

$$X_{k+1} = c_0 X_k + c_1 X_{k-1} + \ldots + d_0 Y_k + d_1 Y_{k-1} + \ldots + Z'_k$$

- Example 3.7

- If the relationship were based on a linear autoregressive model

$$X_{k+1} = 0.3X_k + Z'_k$$

$$Y_{k+1} = 0.1Y_k + 0.2X_k + Z_k$$

- Does *X* cause *Y* or does *Y* cause *X*?
  - Past and current values of *X* can help better predict the future values of *Y*

$$Y_{k+1} = a_0 Y_k + a_1 Y_{k-1} + \ldots + b_0 X_k + b_1 X_{k-1} + \ldots + Z_k$$

$$X_{k+1} = c_0 X_k + c_1 X_{k-1} + \ldots + d_0 Y_k + d_1 Y_{k-1} + \ldots + Z'_k$$

- Testing hypotheses
  - If the coefficients, *b*’s, are zero then *X* does not Granger cause *Y*
  - If the coefficients, *d*’s, are zero then *Y* does not Granger cause *X*
- Granger causality quantifies the impact of coefficients *b*’s and *d*’s.

$$Y_{k+1} = a_0 Y_k + a_1 Y_{k-1} + \ldots + b_0 X_k + b_1 X_{k-1} + \ldots + Z_k$$

$$X_{k+1} = c_0 X_k + c_1 X_{k-1} + \ldots + d_0 Y_k + d_1 Y_{k-1} + \ldots + Z'_k$$

- Test the hypothesis that setting *b*'s to zero increases the residual variance of estimating

$$C_G(X \to Y) = \log \frac{\sigma^2_{\hat{Y}}(\mathbf{0})}{\sigma^2_{\hat{Y}}(\mathbf{b})}$$

$$C_G(Y \to X) = \log \frac{\sigma^2_{\hat{X}}(\mathbf{0})}{\sigma^2_{\hat{X}}(\mathbf{d})}$$

- Shortcomings of Granger casualty
  - The data is assumed to be linearly dependent in time.
    - Autoregressive
  - The two data sets are assumed to be linearly dependent
  - The data sets are assumed to be Gaussian
  - Stationarity is assumed
    - The impact of using Granger on non-stationary data is not known

- Recall that mutual information does not capture temporal information

$$
\begin{aligned}
I(X_1^n; Y_1^n) &= \sum_{i=1}^{n} I(X_1^n; Y_i | Y_1^{i-1}) \\
&= I(X_1^n; Y_1) + I(X_1^n; Y_2 | Y_1) + I(X_1^n; Y_3 | Y_1^2) + \ldots
\end{aligned}
$$

- A careful adjustment

$$
\begin{aligned}
I(X_1^n \to Y_1^n) &= \sum_{i=1}^{n} I(X_1^i; Y_i | Y_1^{i-1}) \\
&= I(X_1; Y_1) + I(X_1^2; Y_2 | Y_1) + I(X_1^3; Y_3 | Y_1^2) + \ldots
\end{aligned}
$$

- Directed information is a measure of causality in relation between X and Y
  - It is a universal quantity measuring
    - influence
    - predictability
    - information flow

- Example 3.8

$$Y_n = X_n + Z_n$$

- with i.i.d.

$$X_n \sim \mathrm{Gaussian}(0, \sigma_X^2)$$

$$Z_n \sim \mathrm{Gaussian}(0, \sigma_Z^2)$$

independent

$$Y_n = X_n + Z_n$$

- with i.i.d.

$$X_n \sim \text{Gaussian}(0, \sigma_X^2)$$

$$Z_n \sim \text{Gaussian}(0, \sigma_Z^2)$$

$$X_n \sim \text{Gaussian}(0, \sigma_X^2)$$

$$Z_n \sim \text{Gaussian}(0, \sigma_Z^2)$$

$$Y_n = X_n + Z_n$$

$$\begin{aligned}
I(X_1^n \to Y_1^n) &= \sum_{i=1}^{n} I(X_1^i; Y_i | Y_1^{i-1}) \\
&= I(X_1; Y_1) + I(X_1^2; Y_2 | Y_1) + I(X_1^3; Y_3 | Y_1^2) + \ldots \\
&= I(X_1; Y_1) + I(X_1; Y_2 | Y_1) + I(X_2; Y_2 | Y_1, X_1) + \ldots \\
&= \frac{1}{2}\log(1 + \frac{\sigma_X^2}{\sigma_Z^2}) + 0 + \frac{1}{2}\log(1 + \frac{\sigma_X^2}{\sigma_Z^2}) + \ldots \\
&= \frac{n}{2}\log(1 + \frac{\sigma_X^2}{\sigma_Z^2})
\end{aligned}$$

$$X_n \sim \text{Gaussian}(0, \sigma_X^2)$$

$$Z_n \sim \text{Gaussian}(0, \sigma_Z^2)$$

$$Y_n = X_n + Z_n$$

$$\begin{aligned}
I(X_1^n \to Y_1^n) &= \sum_{i=1}^{n} I(X_1^i; Y_i | Y_1^{i-1}) \\
&= I(X_1; Y_1) + I(X_1^2; Y_2 | Y_1) + I(X_1^3; Y_3 | Y_1^2) + \dots \\
&= I(X_1; Y_1) + I(X_1; Y_2 | Y_1) + I(X_2; Y_2 | Y_1, X_1) + \dots \\
&= \frac{1}{2} \log(1 + \frac{\sigma_X^2}{\sigma_Z^2}) + 0 + \frac{1}{2} \log(1 + \frac{\sigma_X^2}{\sigma_Z^2}) + \dots \\
&= \frac{n}{2} \log(1 + \frac{\sigma_X^2}{\sigma_Z^2})
\end{aligned}$$

$$X_n \sim \mathrm{Gaussian}(0, \sigma_X^2)$$

$$Z_n \sim \mathrm{Gaussian}(0, \sigma_Z^2)$$

- The normalized, per time, mutual information and directed information
- $$Y_n = X_n + Z_n$$

$$I(X \to Y) = I(Y \to X) = I(X;Y) = \frac{1}{2}\log(1 + \frac{\sigma_X^2}{\sigma_Z^2})$$

- Example 3.9

$$Y_n = X_{n-1} + Z_n$$

- With i.i.d.

$$X_n \sim \mathrm{Gaussian}(0, \sigma_X^2)$$

$$Z_n \sim \mathrm{Gaussian}(0, \sigma_Z^2)$$

independent

$$Y_n = X_{n-1} + Z_n$$

- With i.i.d.

$$X_n \sim \text{Gaussian}(0, \sigma_X^2)$$

$$Z_n \sim \text{Gaussian}(0, \sigma_Z^2)$$

$$X_n \sim \text{Gaussian}(0, \sigma_X^2)$$

independent

$$Z_n \sim \text{Gaussian}(0, \sigma_Z^2)$$

$$Y_n = X_{n-1} + Z_n$$

$$\begin{aligned}
I(X_1^n \to Y_1^n) &= \sum_{i=1}^{n} I(X_1^i; Y_i | Y_1^{i-1}) \\
&= I(X_1; Y_1) + I(X_1^2; Y_2 | Y_1) + I(X_1^3; Y_3 | Y_1^2) + \ldots \\
&= I(X_1; Y_1) + I(X_1; Y_2 | Y_1) + I(X_2; Y_2 | Y_1, X_1) + \ldots \\
&= 0 + \frac{1}{2}\log(1 + \frac{\sigma_X^2}{\sigma_Z^2}) + 0 + \frac{1}{2}\log(1 + \frac{\sigma_X^2}{\sigma_Z^2}) + \ldots \\
&= \frac{n}{2}\log(1 + \frac{\sigma_X^2}{\sigma_Z^2})
\end{aligned}$$

$$X_n \sim \mathrm{Gaussian}(0, \sigma_X^2)$$

independent

$$Z_n \sim \mathrm{Gaussian}(0, \sigma_Z^2)$$

$$Y_n = X_{n-1} + Z_n$$

$$\begin{aligned}
\dot{I}(X_1^n \to Y_1^n) &= \sum_{i=1}^{n} I(X_1^i; Y_i | Y_1^{i-1}) \\
&= I(X_1; Y_1) + I(X_1^2; Y_2 | Y_1) + I(X_1^3; Y_3 | Y_1^2) + \ldots \\
&= I(X_1; Y_1) + I(X_1; Y_2 | Y_1) + I(X_2; Y_2 | Y_1, X_1) + \ldots \\
&= 0 + \frac{1}{2}\log(1 + \frac{\sigma_X^2}{\sigma_Z^2}) + 0 + \frac{1}{2}\log(1 + \frac{\sigma_X^2}{\sigma_Z^2}) + \ldots \\
&= \frac{n}{2}\log(1 + \frac{\sigma_X^2}{\sigma_Z^2})
\end{aligned}$$

$$X_n \sim \text{Gaussian}(0, \sigma_X^2)$$

$$Z_n \sim \text{Gaussian}(0, \sigma_Z^2)$$

$$Y_n = X_{n-1} + Z_n$$

-

$$\begin{aligned}
I(Y_1^n \to X_1^n) &= \sum_{i=1}^{n} I(Y_1^i; X_i | X_1^{i-1}) \\
&= I(Y_1; X_1) + I(Y_1^2; X_2 | X_1) + I(Y_1^3; X_3 | X_1^2) + \ldots \\
&= I(Y_1; X_1) + I(Y_1; X_2 | X_1) + I(Y_2; X_2 | X_1, Y_1) + \ldots \\
&= 0 + 0 + \ldots
\end{aligned}$$

$$X_n \sim \text{Gaussian}(0, \sigma_X^2)$$

$$Z_n \sim \text{Gaussian}(0, \sigma_Z^2)$$

$$Y_n = X_{n-1} + Z_n$$

•

$$\begin{aligned} I(Y_1^n \to X_1^n) &= \sum_{i=1}^{n} I(Y_1^i; X_i | X_1^{i-1}) \\ &= I(Y_1; X_1) + I(Y_1^2; X_2 | X_1) + I(Y_1^3; X_3 | X_1^2) + \dots \\ &= I(Y_1; X_1) + I(Y_1; X_2 | X_1) + I(Y_2; X_2 | X_1, Y_1) + \dots \\ &= 0 + 0 + \dots \end{aligned}$$

$$X_n \sim \mathrm{Gaussian}(0, \sigma_X^2)$$

$$Z_n \sim \mathrm{Gaussian}(0, \sigma_Z^2)$$

- Recall

$$Y_n = X_{n-1} + Z_n$$

- then

$$I(X \to Y) = \frac{1}{2}\log(1 + \frac{\sigma_X^2}{\sigma_Z^2})$$
$$I(Y \to X) = 0$$

- In these two examples Granger causality and directed information result in similar measures
  - Since time series are
    - Linearly related
    - Gaussian
- It is not clear if Granger causality is the right metric in the coffee price example since the linearity model may or may not be valid.

- A nonlinear model

$$Y_k = \beta_1 X_k^2 + \beta_2 X_{k-1}^2 + Z_k$$

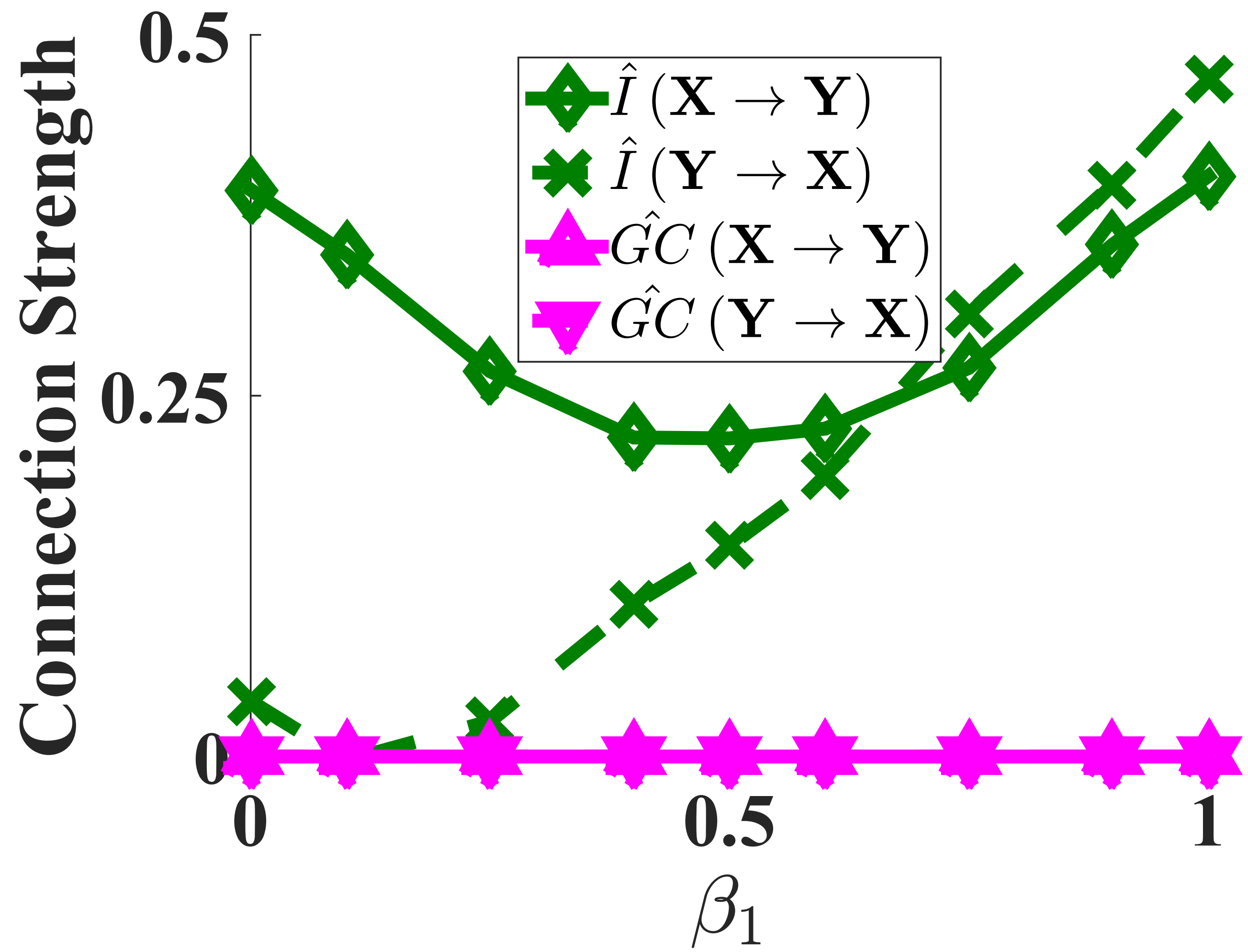
- where *Z* is Gaussian noise
- Can *X* help predict *Y*?
- Can *Y* help predict *X*?
- How about in these cases?

$$Y_k = X_k^2 + Z_k \text{ or } Y_k = X_{k-1}^2 + Z_k$$

- A nonlinear model

$$Y_k = \beta_1 X_k^2 + \beta_2 X_{k-1}^2 + Z_k$$

- where Z is Gaussian noise



(b) $\beta_2 = 1 - \beta_1$

- Directed information is a measure of causality in relation between *X* and *Y*
  - It is a universal quantity measuring
    - Influence
    - Predictability
    - Information flow
    - Another important metric of relation between time series

- Coherence
  - Another concept measuring relationship between two data sets
  - Consider two zero mean random vectors *X* and *Y*
  - The cross correlation is defined as

$$R_{X,Y}(m, m') = E[X_m Y^*_{m'}]$$

  - If the series are jointly wide sense stationary

$$R_{X,Y}(m, m') = R_{X,Y}(m - m')$$

- The cross power spectral density is defined as

$$S_{X,Y}(f) = \mathcal{F}\{R_{X,Y}(k)\} = \sum_{k=-\infty}^{\infty} R_{X,Y}(k) e^{j2\pi k f}$$

- Recall autocorrelation of a time series is

$$R_X(m, m') = E[X_m X_{m'}^*]$$

- If the times series is wide sense stationary then

$$R_X(m, m') = R_X(m - m')$$

- The power spectral density is

$$S_X(f) = \mathcal{F}\{R_X(k)\} = \sum_{k=-\infty}^{\infty} R_X(k) e^{j2\pi k f}$$

- The coherence at a given frequency between two time series is defined as

$$C_{X,Y}(f) = \frac{|S_{X,Y}(f)|^2}{S_X(f)S_Y(f)}$$

- The coherence estimates the extend that *Y* can be predicted by *X* using optimum linear estimator

- It can be shown that $0 \leq C_{X,Y}(f) \leq 1$

- If *Y* is a noiseless linear function of time series *X,* i.e., *Y* = *h* * *X,* what is the coherence between *X* and *Y*?

- If *Y* is a linear estimator of *X,* then *Y* = *h* * *X* with no noise then

$$S_{X,Y}(f) = H(f)S_X(f) \text{ and } S_Y(f) = |H(f)|^2 S_X(f)$$

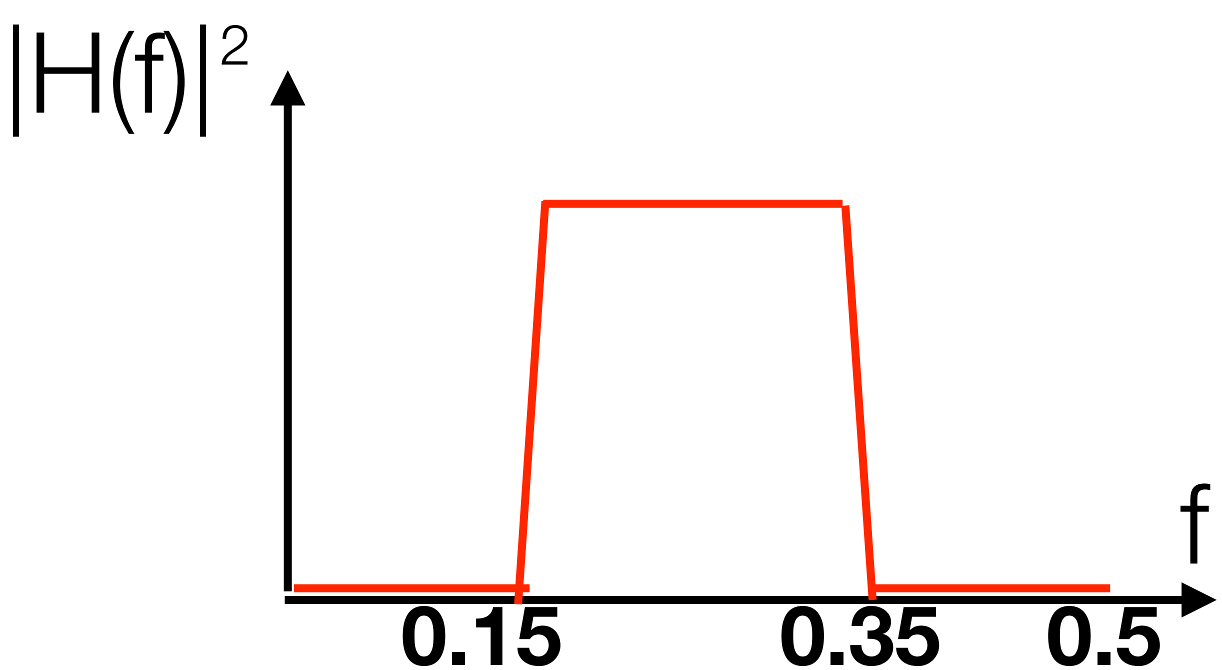
- And the coherence is 1.

- Any nonlinearity or noise in the system will reduce the coherence.

- Reduction in information or estimation accuracy due to nonlinearity or noise at a given frequency

$$1 - C_{X,Y}(f)$$

- Example 3.10
  - A linear system where $Y = h * X + Z$ where Z is noise
  - The filter is a 33 tap bandpass filter between [0.15, 0.35] normalized frequencies
  - How effectively can X at frequency 2.5 be estimated from Y?



- Example 3.11
    - Two nonlinearly related signals, assume *f = 4 Hz*

$$X_i = A\cos(2\pi f i + \theta) \ \forall i = 1, 2, \ldots, n$$
$$Y_i = X_i^2 + Z_i$$

    - Are *X* and *Y* coherent at frequency *4 Hz*?

- Mutual information quantifies relationship between data sets
  - Ignores relative timing and causality
  - Ignores frequency content of the data

$$
\begin{aligned}
I(X_1^n; Y_1^n) &= \sum_{i=1}^{n} I(X_1^n; Y_i | Y_1^{i-1}) \\
&= I(X_1^n; Y_1) + I(X_1^n; Y_2 | Y_1) + I(X_1^n; Y_3 | Y_1^2) + \ldots
\end{aligned}
$$

- In many scenarios the frequency content of the data is a critical element in the analysis or inference
  - Data from music
  - Auditory neurological data
  - Neurological data in different frequency bands have different significances
    - Alpha, theta, beta, gamma, and high gamma bands

- Mutual information in frequency

$$MI_{X,Y}(f_i, f_j) = I(d\tilde{X}_{f_i}; d\tilde{Y}_{f_j})$$

- That is, mutual information between Fourier transforms of the two time series

$$X_i = \int_0^1 e^{j2\pi i f} d\tilde{X}_f$$

$$Y_i = \int_0^1 e^{j2\pi i f} d\tilde{Y}_f$$

- Here *i = 1, 2, … , n*

$$X_i = \int_0^1 e^{j2\pi i f} d\tilde{X}_f$$

$$X_1^n = (X_1, X_2, \ldots, X_n) \text{ is the recoded data and}$$

$$\tilde{X}_f \text{ for } f \in [0,1] \text{ is spectral representation of data}$$

- Note that mutual information can be computed for any data set with time as the index or frequency or space.

- It has been shown that when *X* and *Y* have a linear relationship then

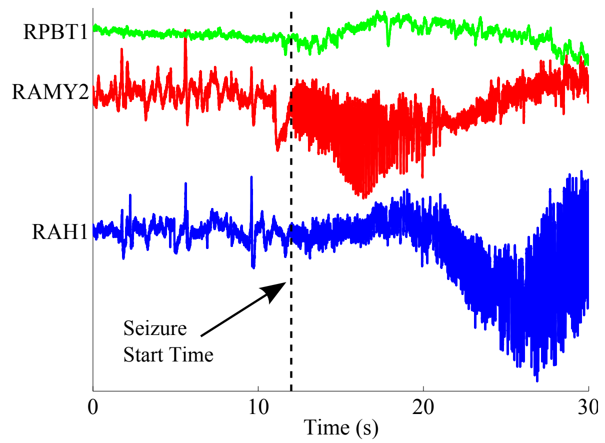
$$MI_{X,Y}(f,f) = I(d\tilde{X}_f; d\tilde{Y}_f) = -\log[1 - C_{X,Y}(f)]$$

- Note that coherence was defined for linear systems as

$$C_{X,Y}(f) = \frac{|S_{X,Y}(f)|^2}{S_X(f)S_Y(f)}$$

- Since it is related to mutual information in frequency it can be generalized to any data sets

$$MI_{X,Y}(f,f) = I(d\tilde{X}_f; d\tilde{Y}_f) = -\log[1 - C_{X,Y}(f)]$$

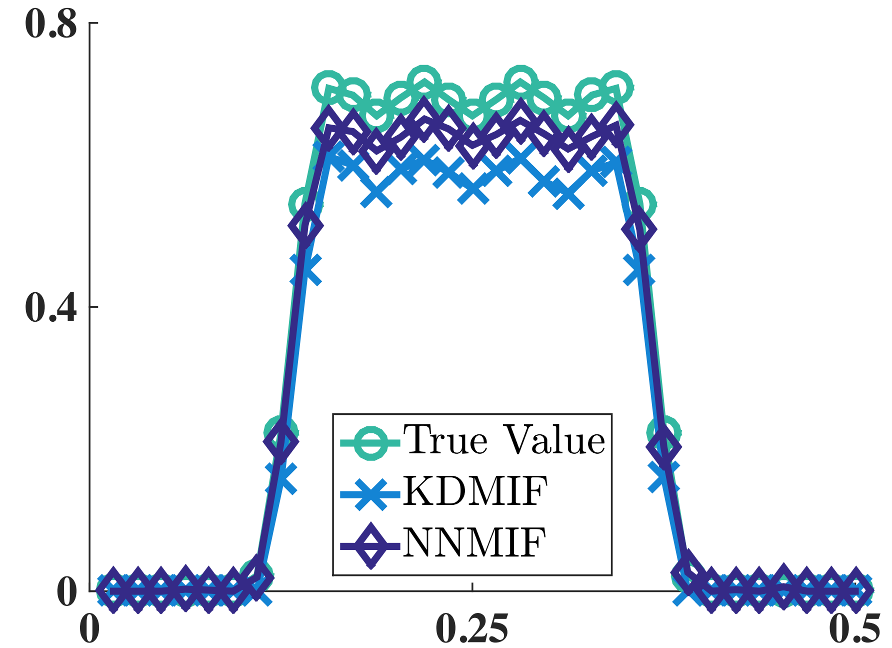


$\longrightarrow$ **DFT** $\longrightarrow \tilde{X}_f, \tilde{Y}_f \rightarrow \hat{f}_{\tilde{X},\tilde{Y}} \rightarrow \hat{H}(\tilde{X},\tilde{Y}) \rightarrow \hat{I}(d\tilde{X}; d\tilde{Y})$

- Note that for range of frequencies, similar to time periods, the mutual information in frequency is defined as

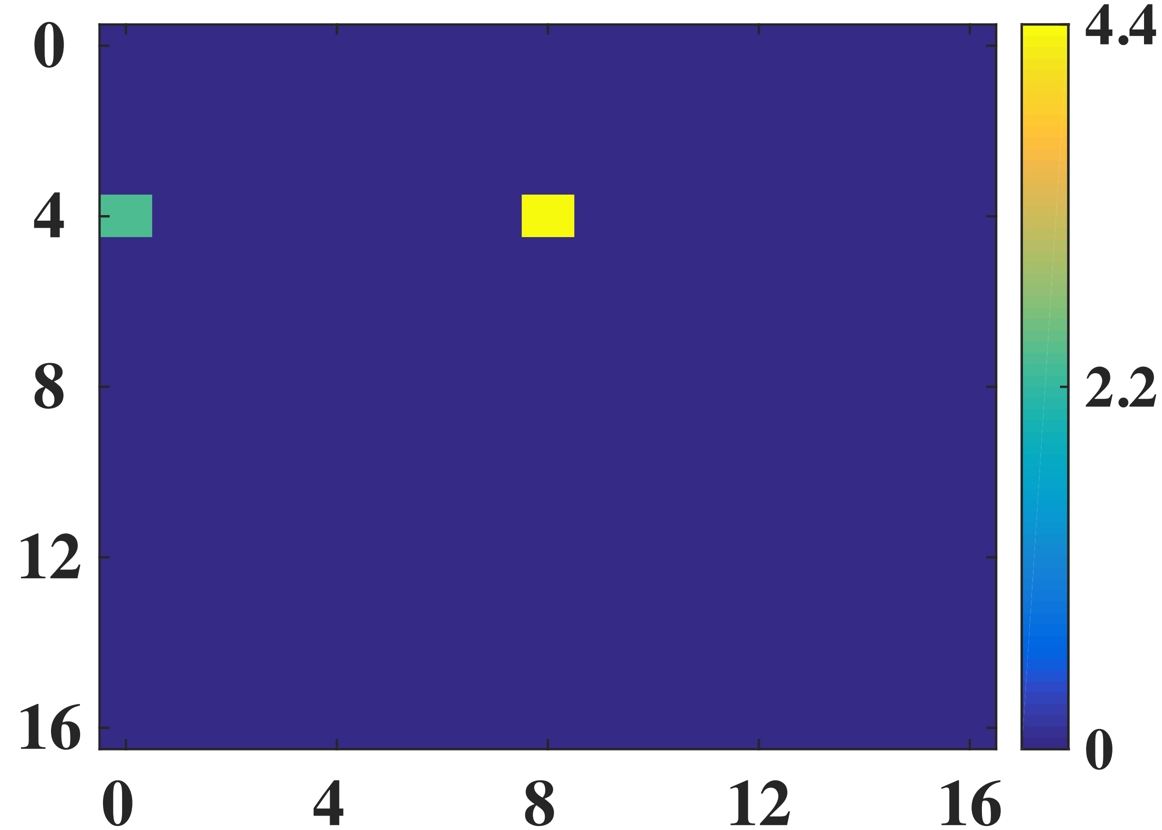
$$MI_{X,Y}(f, f') = I(d\tilde{X}_{f_1}^{f_n}; d\tilde{Y}_{f_1'}^{f_n'})$$

- Example 3.12
  - A linear system where $Y = h * X + Z$ where Z is noise
  - The filter is a 33 tap bandpass filter between [0.15, 0.35] normalized frequencies
  - The mutual information between X and Y



- Example 3.13
  - Two nonlinearly related signals, assume *f = 4 Hz*

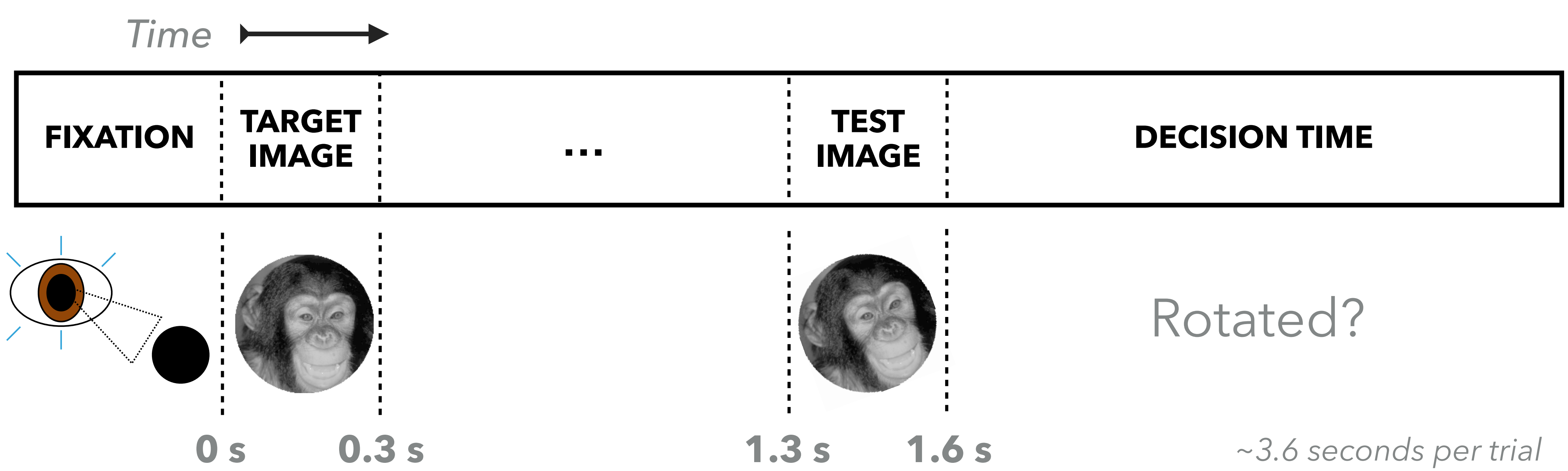
$$X_i = A\cos(2\pi f i + \theta)\ \forall i = 1, 2, \ldots, n$$
$$Y_i = X_i^2 + Z_i$$



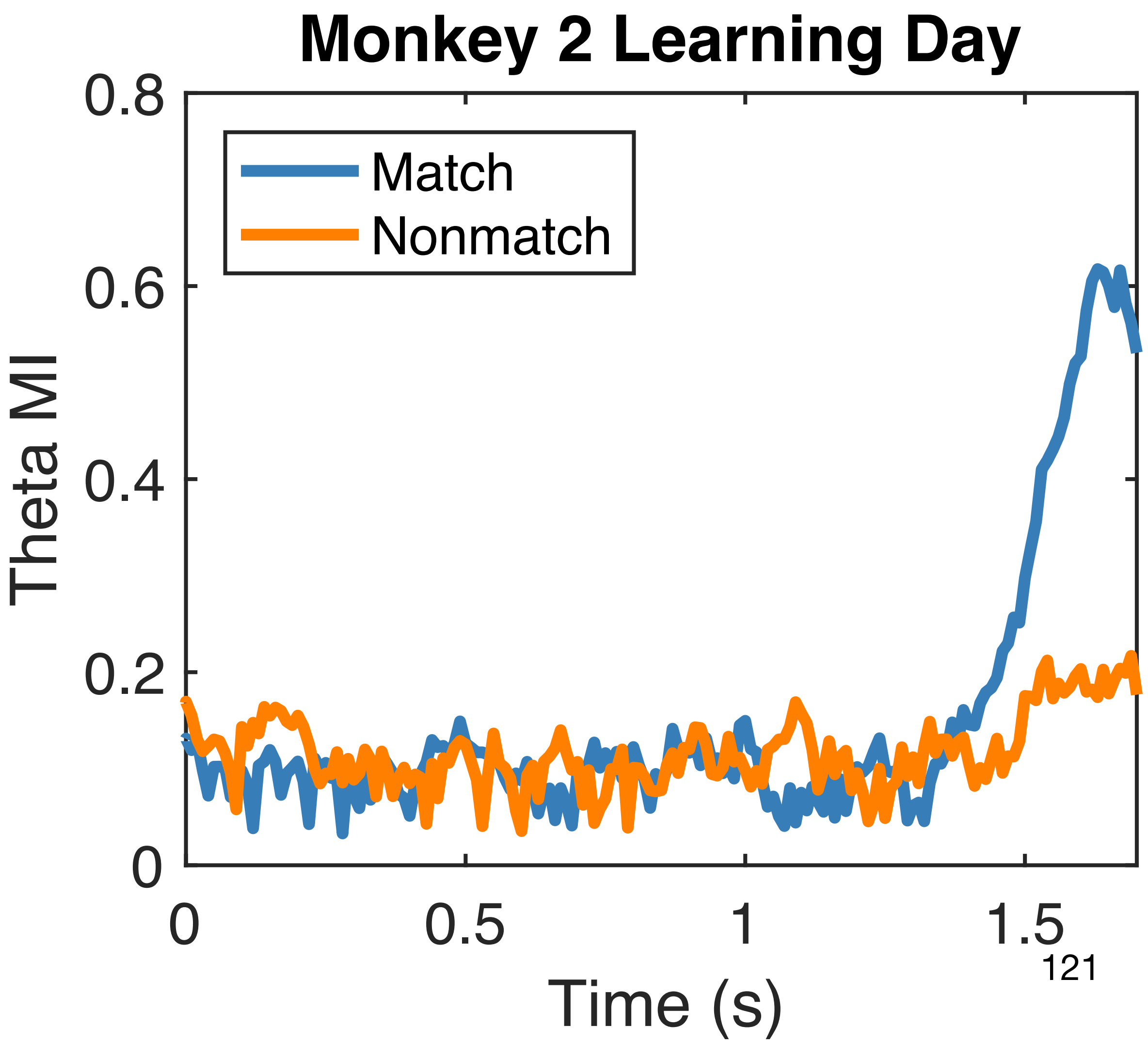
- Example 3.14

- An experiment with no known ground truth

- A visual task, one trial, one monkey, non-matched (rotated image)



- Local field potential recordings from visual cortex about 500 trials
- Increase in Coherency between recorded time series
  - Theta band (3-8 Hz)
  - Matched trials
  - As the 2nd scene is processed

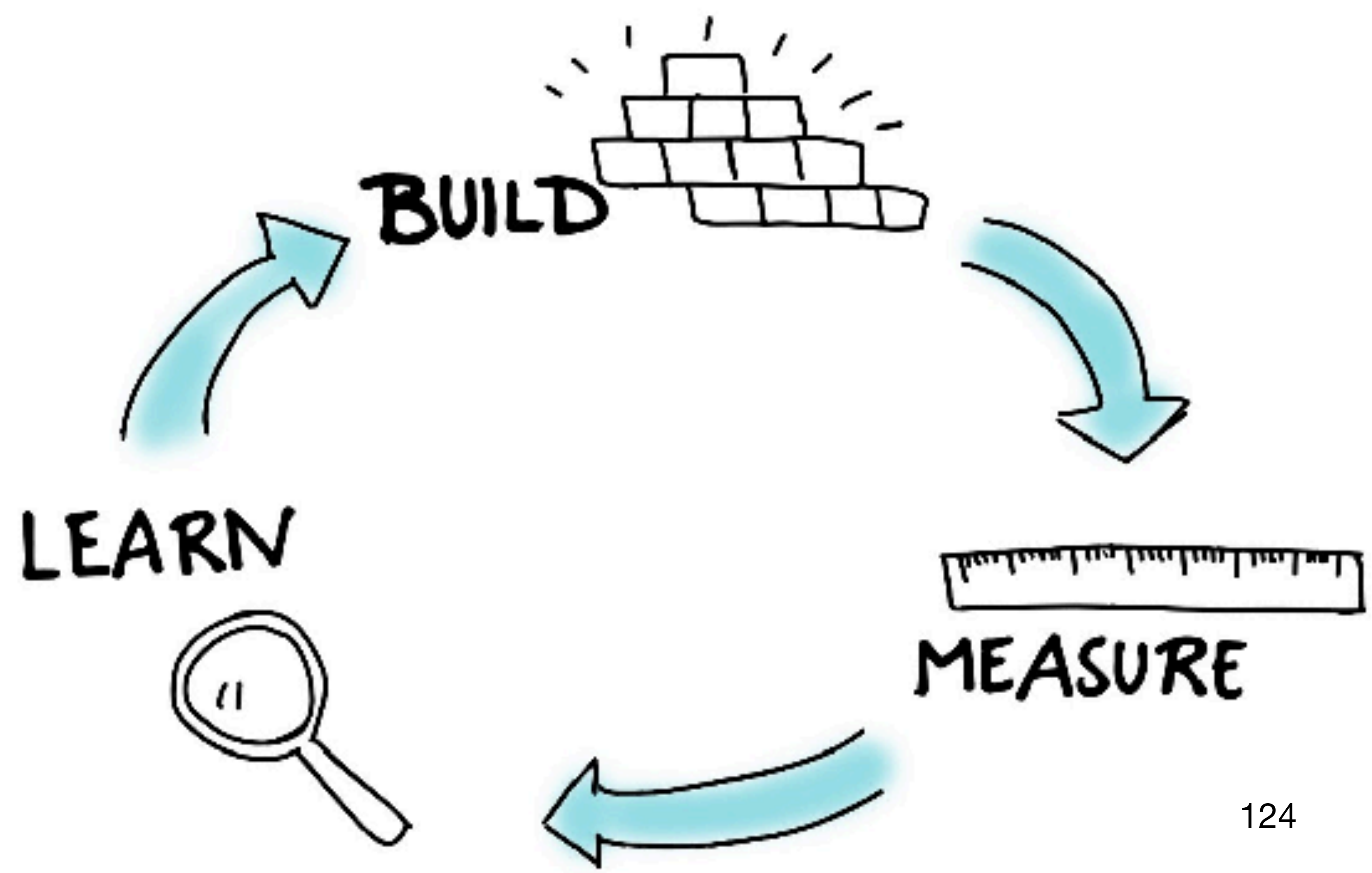


4. Frameworks for Learning from Data

- Parametric models
  - Accuracy of the model
  - Complexity of the model
    - Linear
    - Gaussian
    - Poisson

- Non-parametric, data driven, model free, universal, …
  - Issues
    - The size of the data
    - Relevance of the data
    - Overfitting
  - Merits
    - Not limited by the model

- Generate sufficient amount of data
  - to explore relevant features of the physical system
  - to use the features to manipulate the system



5. Estimating Key Statistical Metrics from Data

- A critical step for
  - Model based
  - Data driven
    - Estimating correlation, dependencies, coherence, and other measures among recordings, i.e., time series

- Entropy of discrete valued random variables

$$H(X) = -\sum_i p_X(x_i) \log p_X(x_i)$$

- Estimating the entropy
  - Plugin estimator

$$\hat{H}_n(X) = -\sum_{a=1}^{A} \hat{p}_a \log \hat{p}_a \text{ where } \hat{p}_a = \frac{\#\text{ occurrences of symbol } a}{n}$$

$$x_i \in \{1, 2, \ldots, A\}$$

- The random variables are assumed independent and identically distributed (i.i.d)

- It can be shown that

$$E\{[\hat{H}_n(X) - H(X)]^2\} = O(\frac{1}{n})$$

- Example 5.1

- The binary random variables.

- The random variables are assumed independent and identically distributed (i.i.d)

- $$\hat{H}_n(X) = -\hat{p}_0 \log \hat{p}_0 - \hat{p}_1 \log \hat{p}_1$$

- The binary random variables. $\hat{H}_n(X) = -\hat{p}_0 \log - \hat{p}_1 \log \hat{p}_1$

$$\hat{p}_0 = \frac{\text{\# of occurrences of symbol 0}}{n}$$

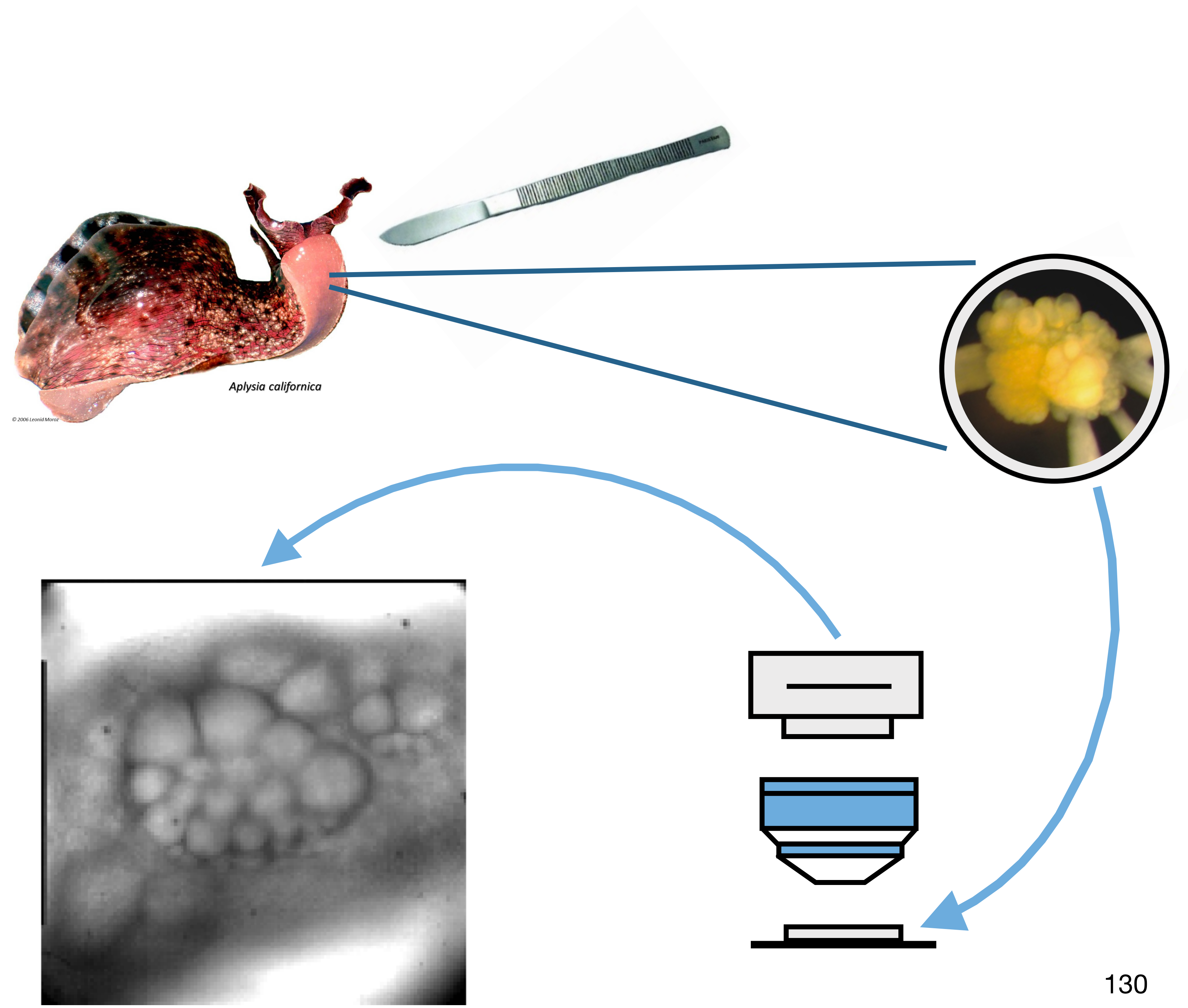
$$\hat{p}_1 = \frac{\text{\# of occurrences of symbol 1}}{n}$$

- Example with

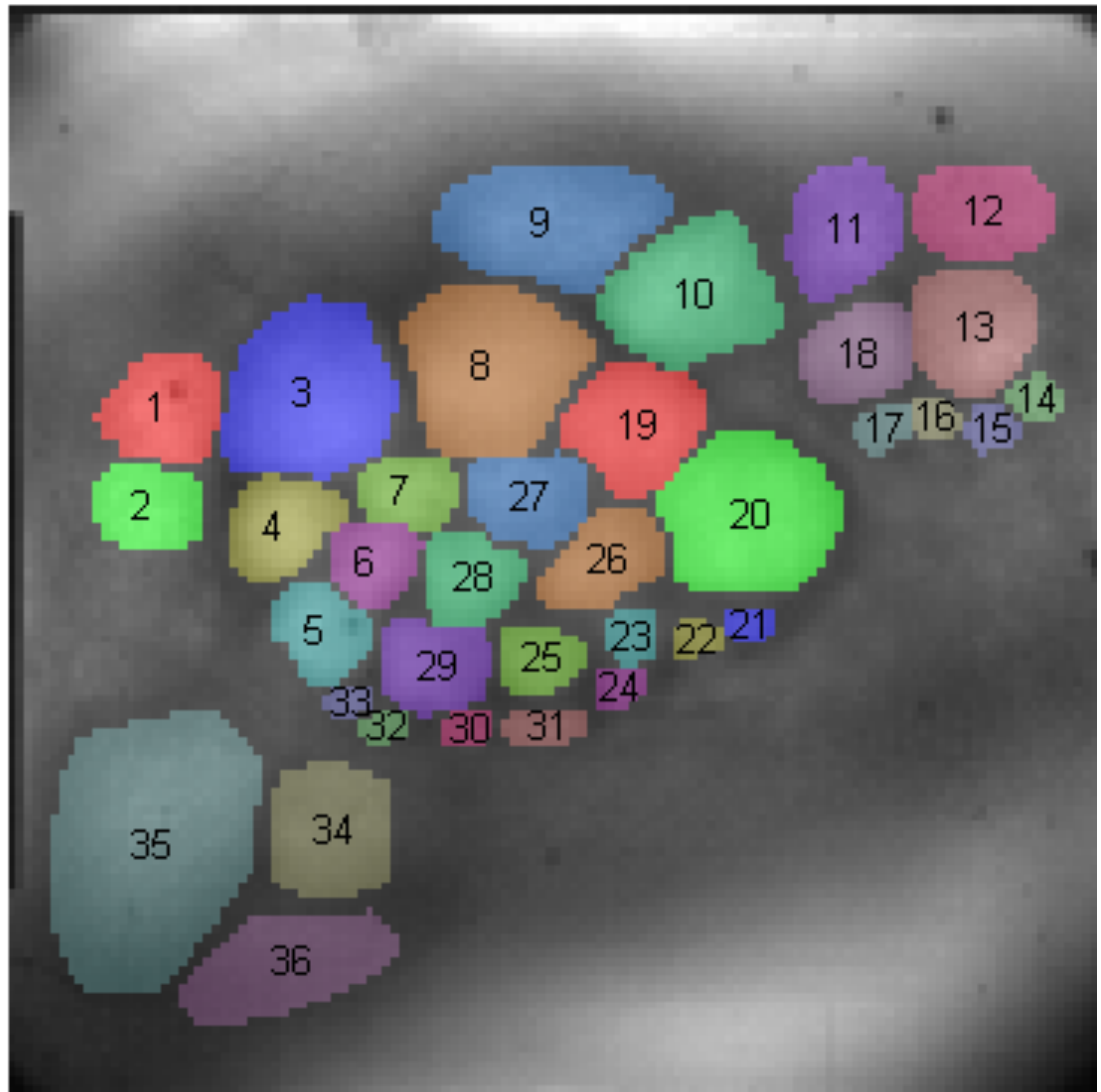
$$\mathbf{x} = (0, 1, 0, 0, 0, 1)$$

$$\hat{H}_n(X) = \frac{2}{3} log \frac{3}{2} + \frac{1}{3} log 3$$

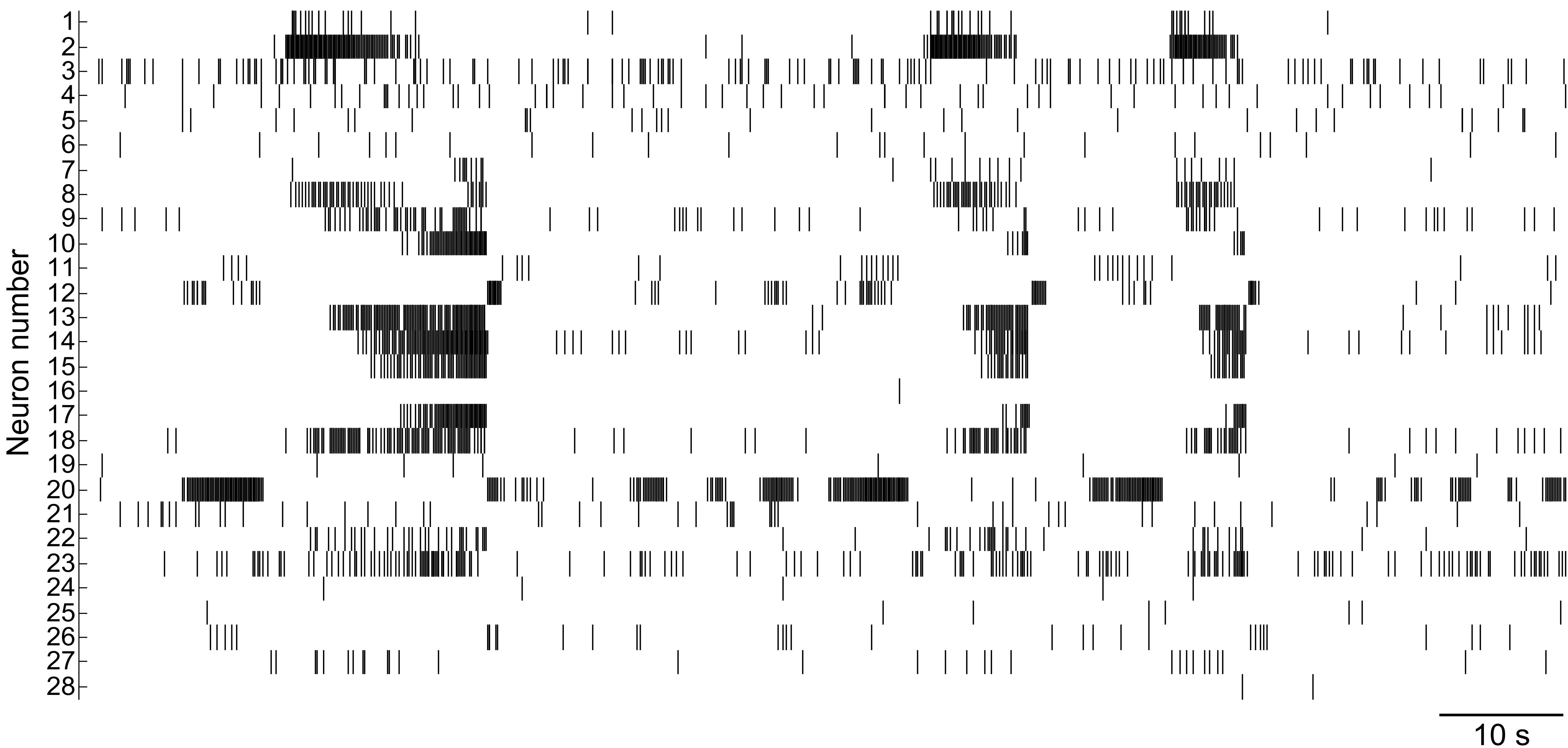
• Example 5.2



- Example 5.2

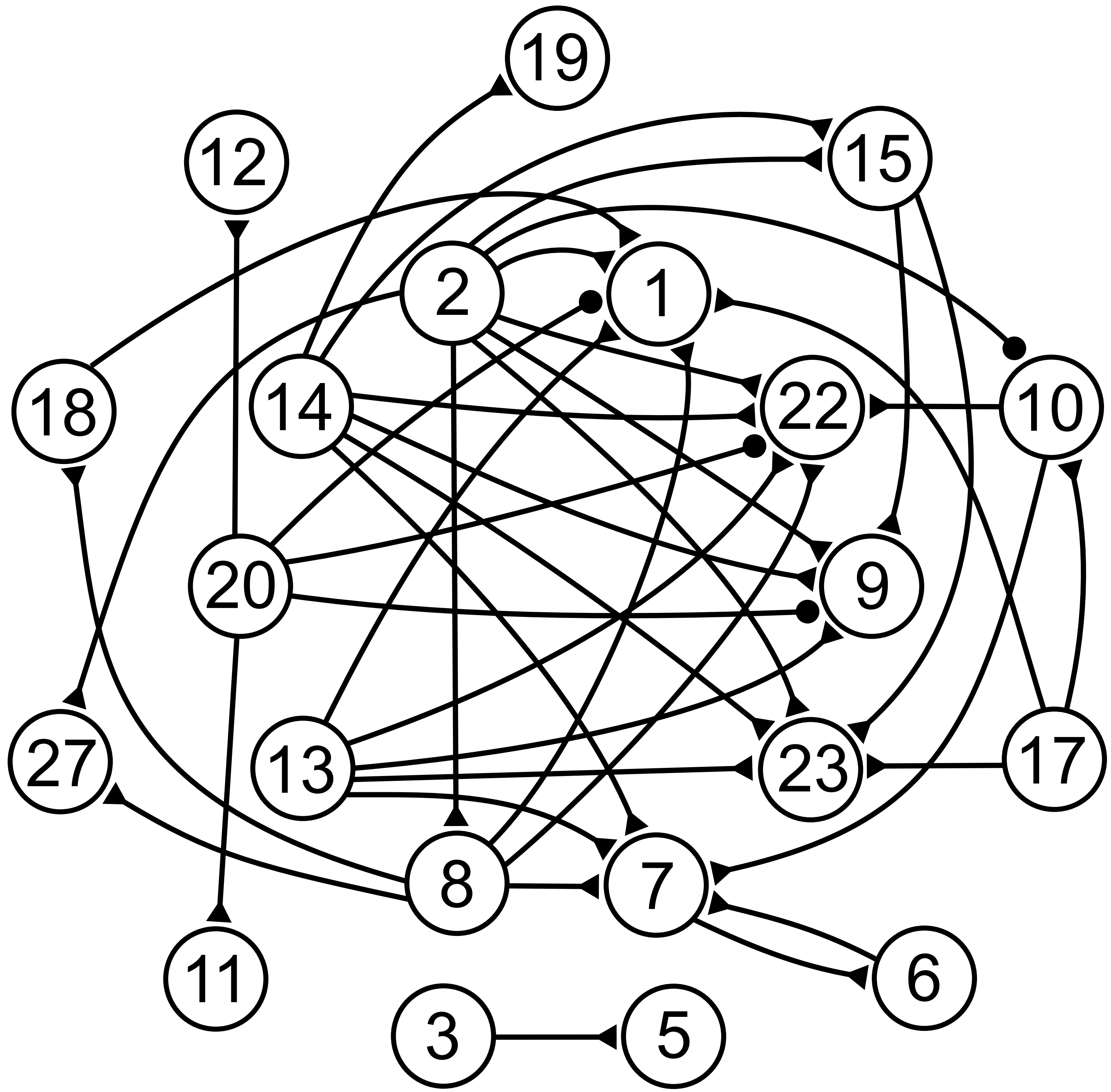

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36

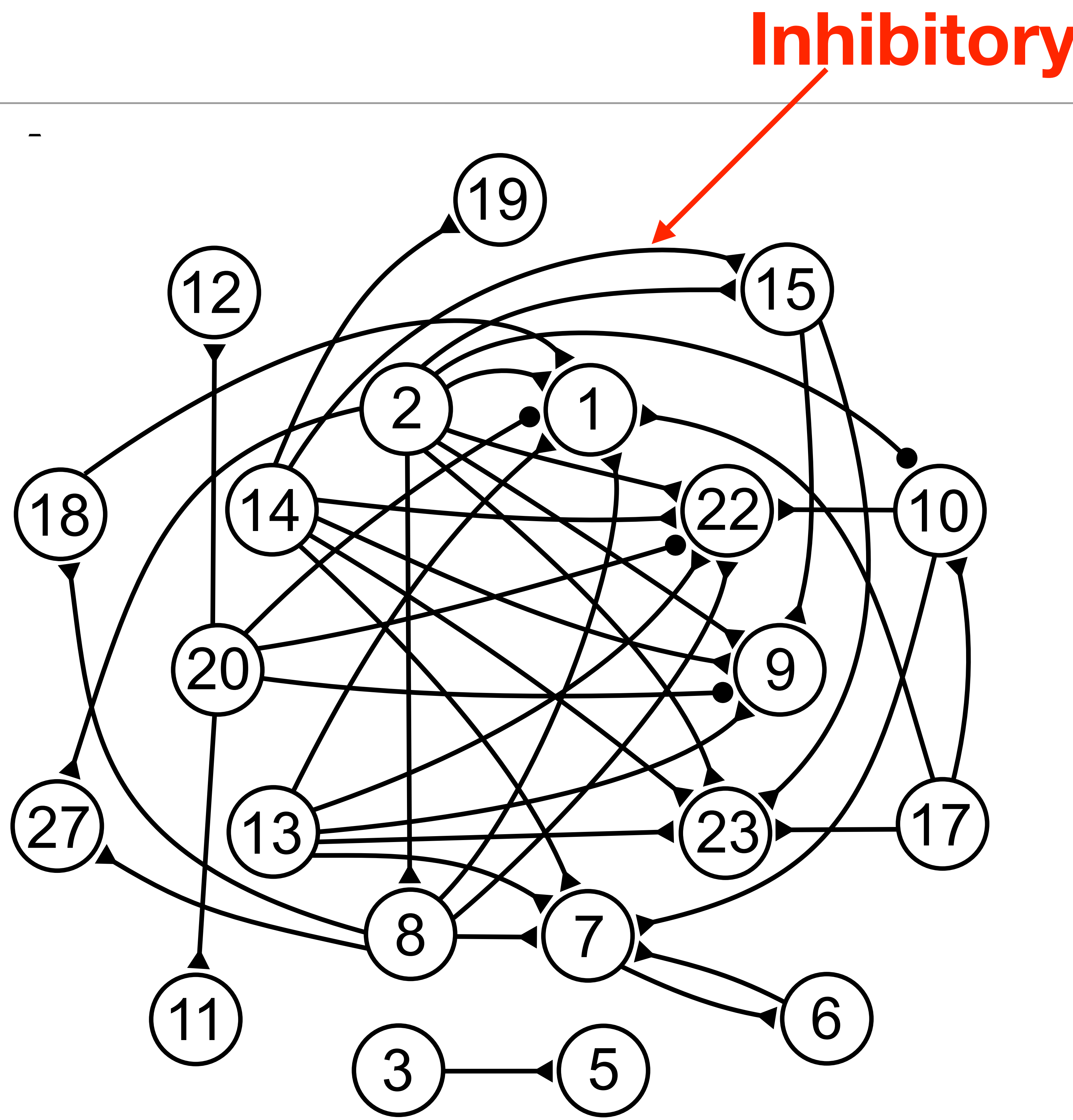
- Example 5.2

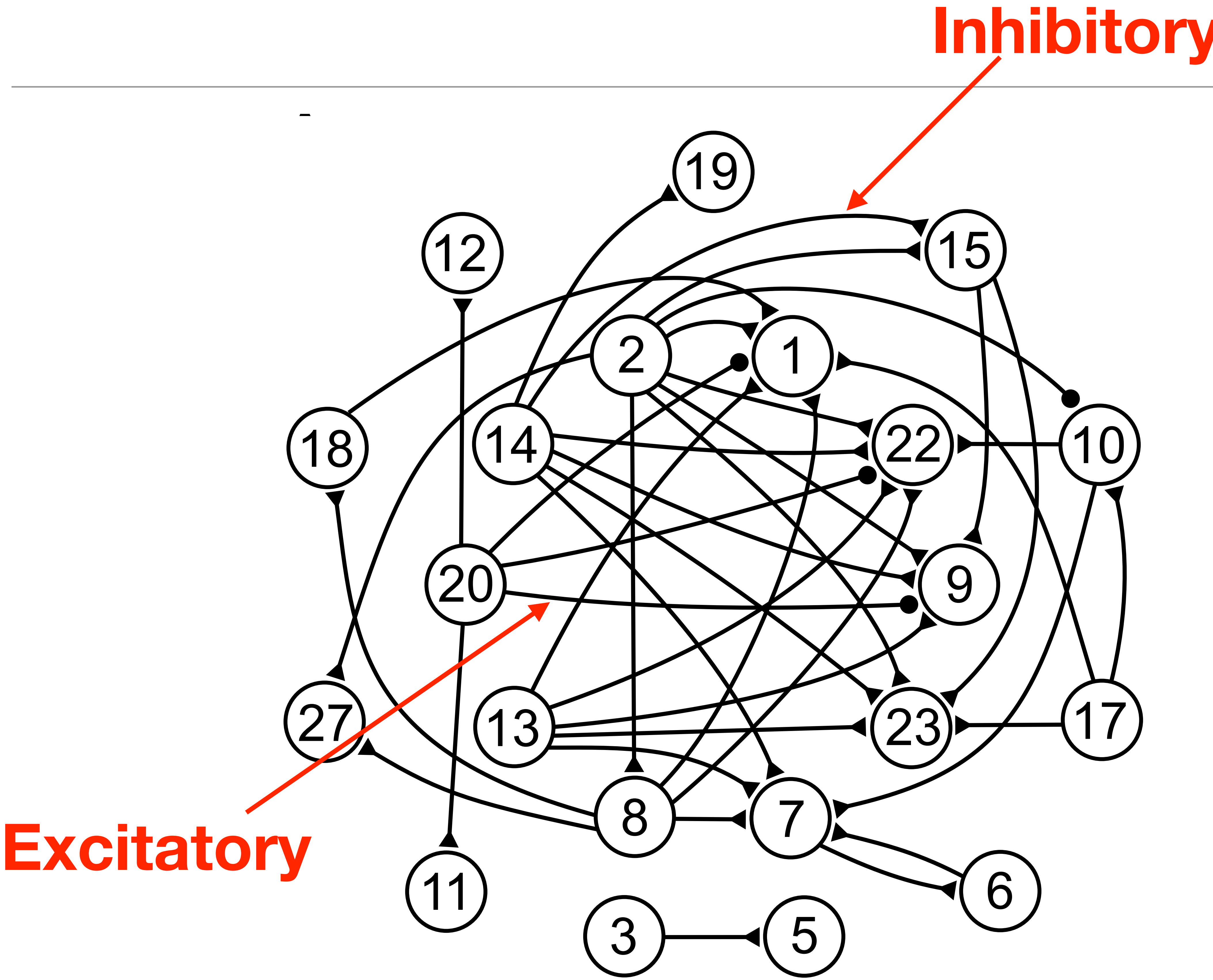


Neuron number
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
10 s

- What are the statistical properties of firing of each neuron?
- Are the spikes in different neurons related?
- Is one neuron’s spike excites another neuron to spike?
- Is one neuron’s spike inhibits another neuron from firing?
- What is the anatomical connectivity graph of these neurons?
- What is the functional connectivity graph of these neurons?


19
12
15
2
1
18
14
22
10
20
9
27
13
23
17
8
7
11
6
3
5


Inhibitory
19
12
15
2
1
18
14
22
10
20
9
27
13
23
17
8
7
11
6
3
5


Inhibitory
Excitatory
19
12
15
2
1
18
14
22
10
20
9
27
13
23
17
8
7
11
6
3
5

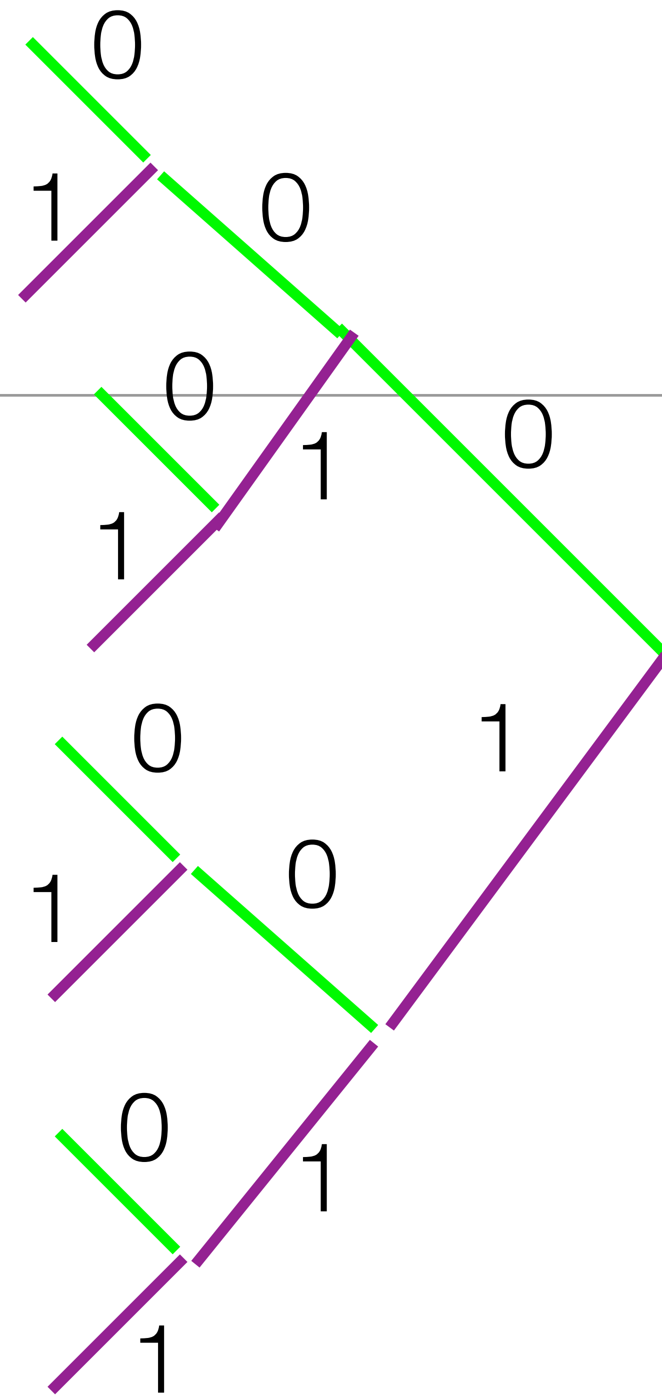
- Neurons do not independently fire and their spike probabilities are not identically distributed
  - The stimulus and the functionality is coded in the spike pattern of a population of neurons

- In many physical systems, the data symbols in time are not independent or identically distributed.

$$p_{X_i}(x) \neq p_{X_j}(x) \text{ or } p_{X_i|s}(x) \neq p_{X_i}(x)$$

- Here *s* is the context, that is the past observed values
- Krichevsky–Trofimov (KT) estimator is a powerful technique to estimate probability of sequences.
    - For discrete valued data
    - Data driven with no assumptions on independence and identically distributed symbols

- Example 5.3
- Assume binary data

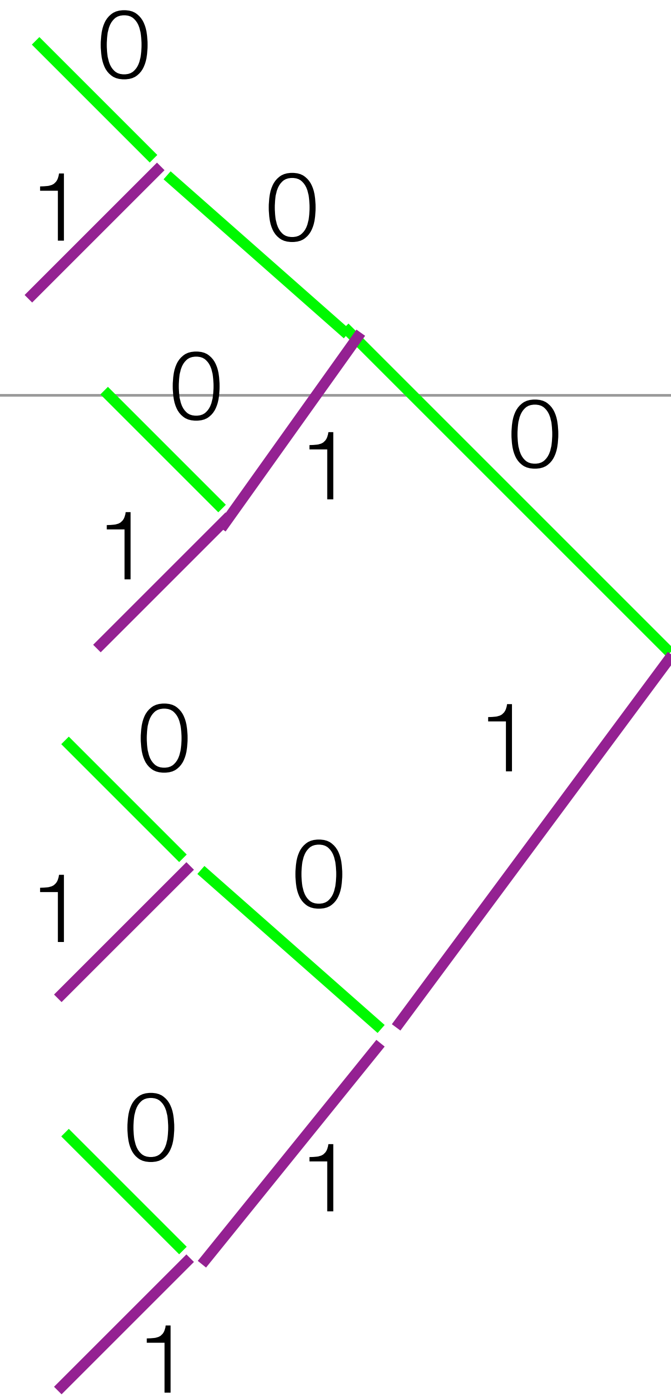


probability of this symbol?

01001010?..

past values: the context

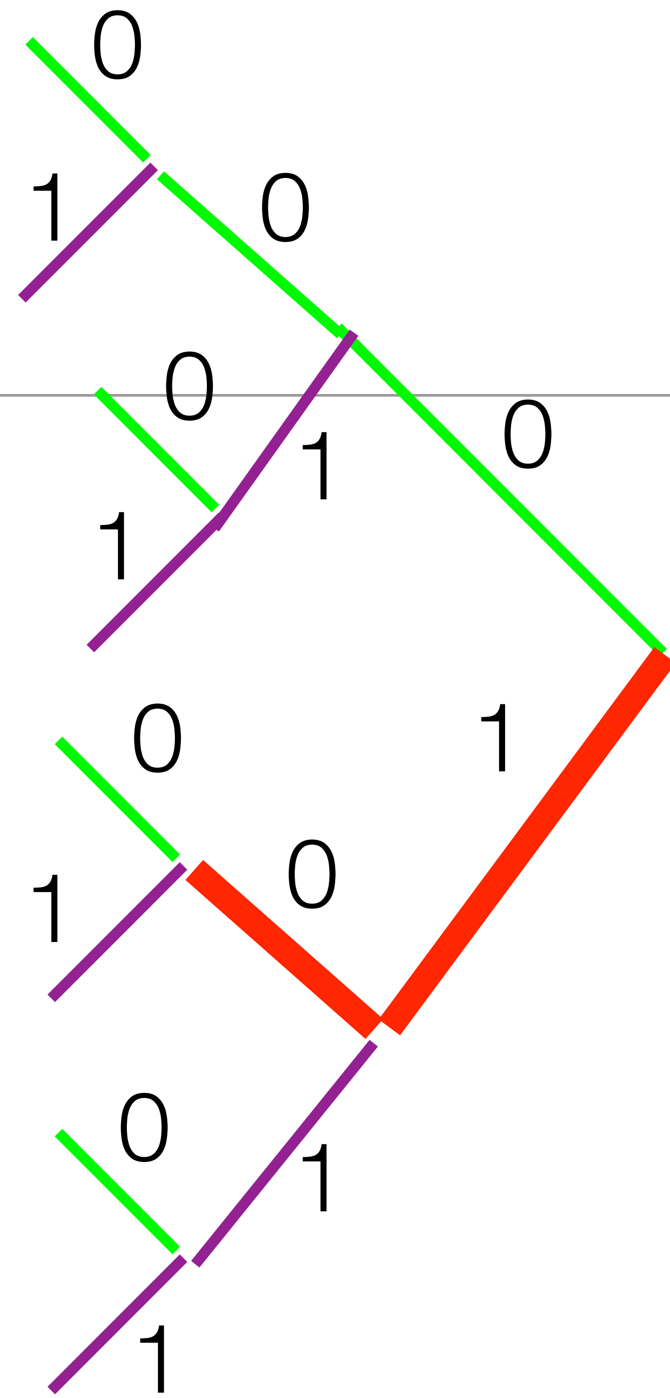
- KT on a tree



probability of this symbol?

010001010…

past values: the context



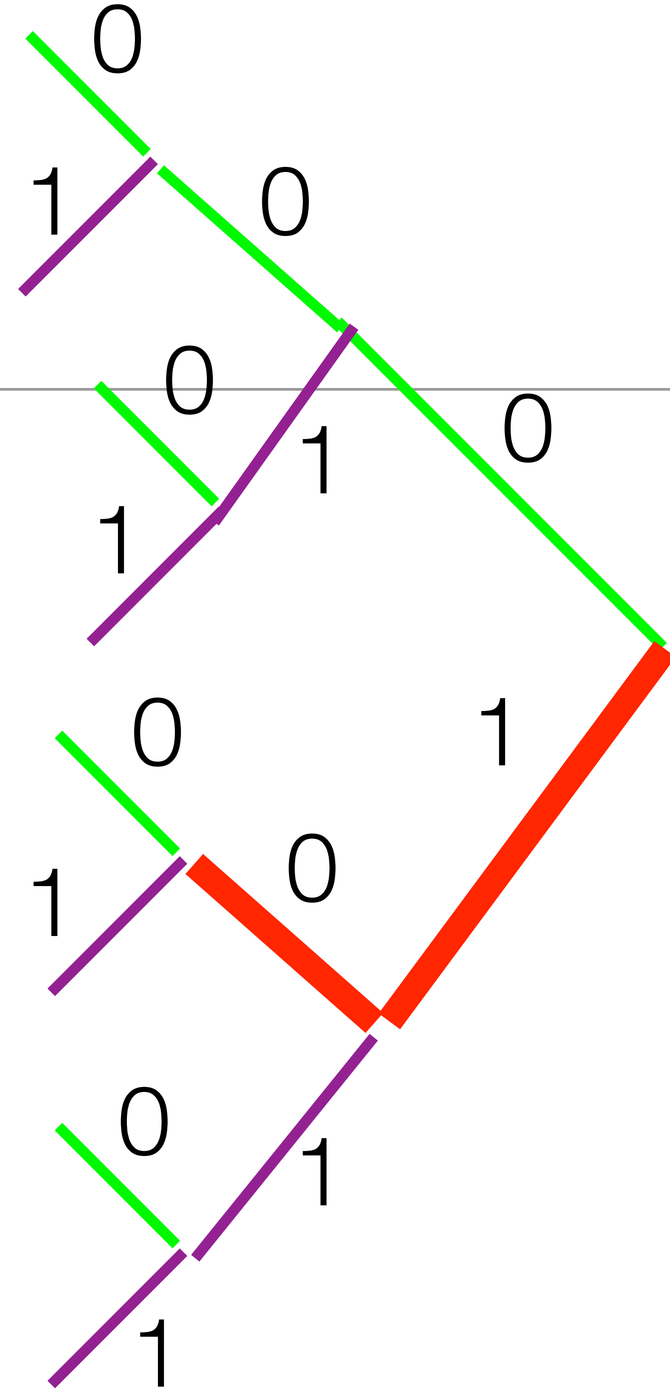
probability of this symbol?

010001010…

past values: the context

a parameter to fudge to have probabilities

$$p(X_3 = 0 | X_1 = 0, X_2 = 1) = \frac{0 + 1/2}{0 + 1} = 1/2$$



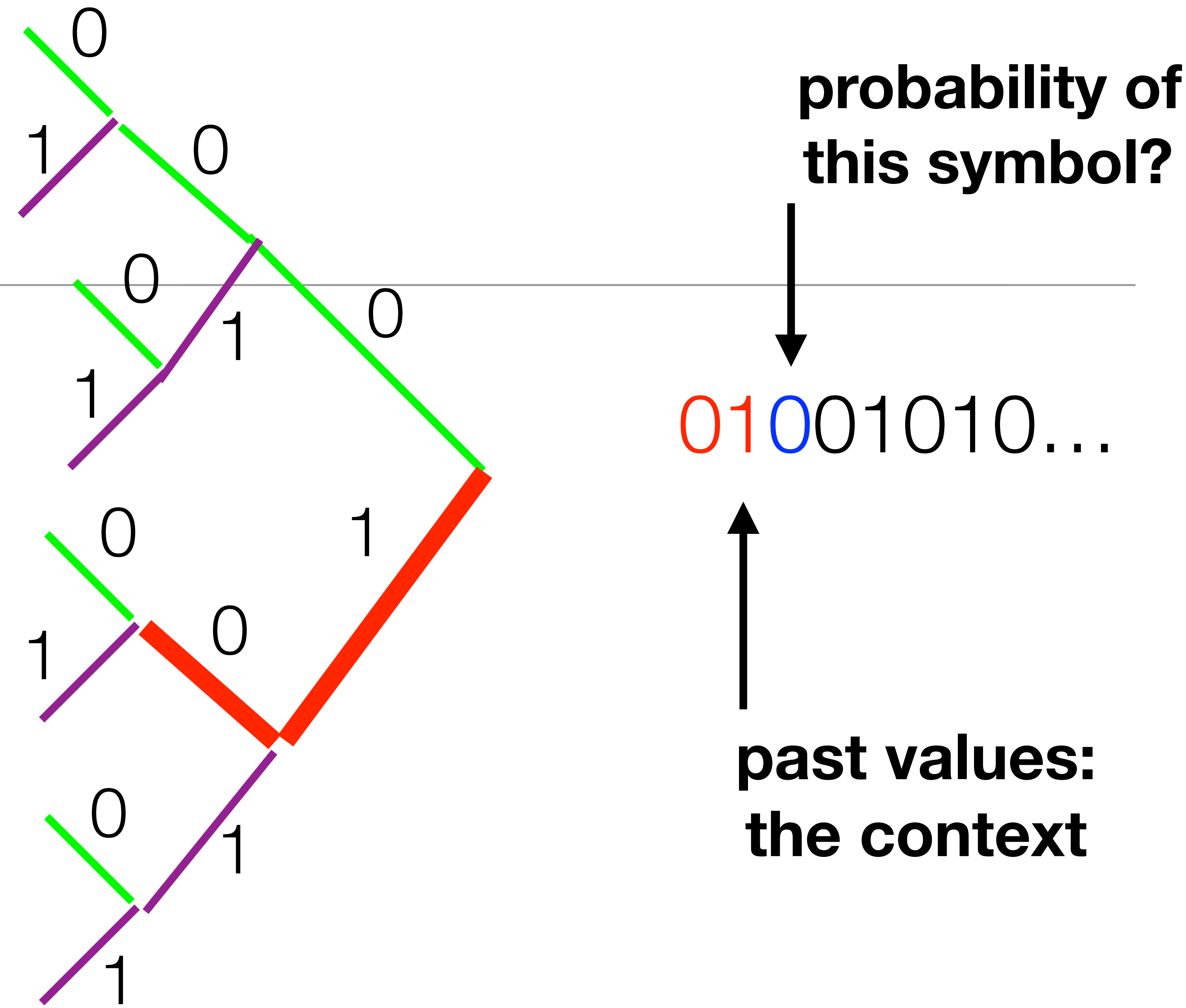
probability of this symbol?

010 01010…

past values: the context

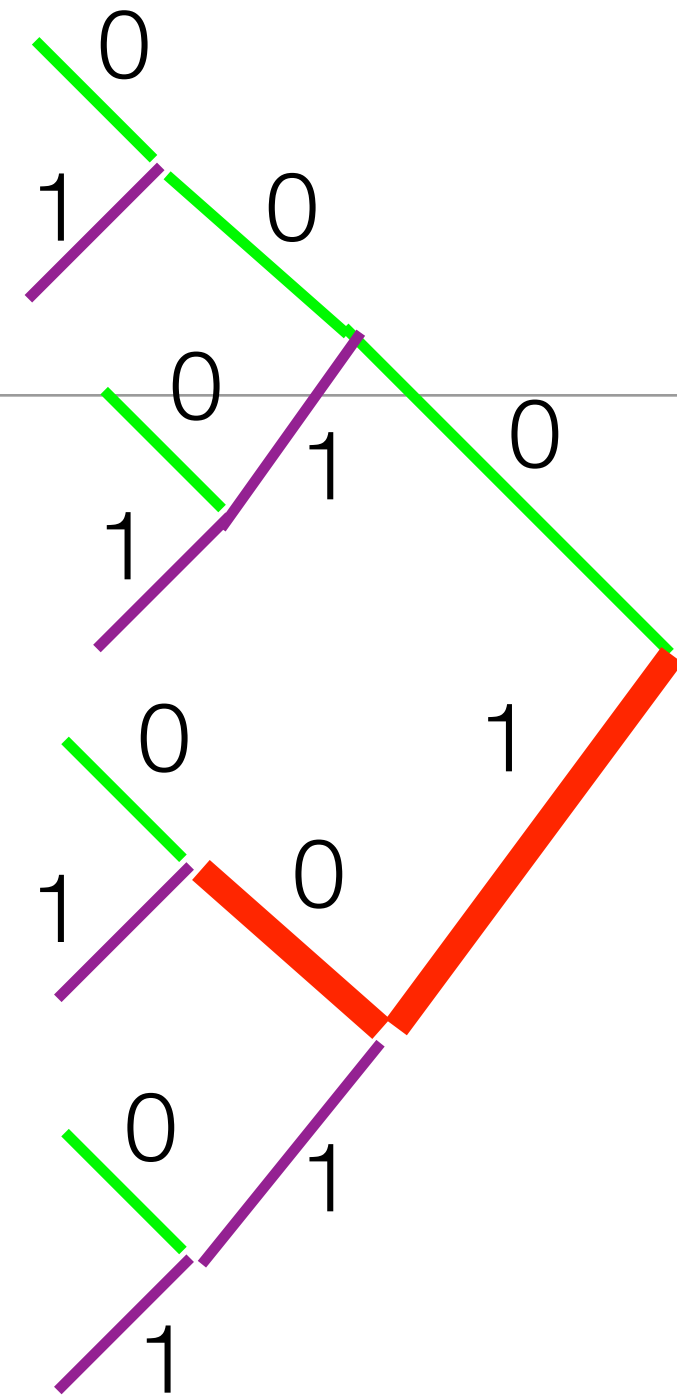
that 1/2 fudge parameter times the size of the alphabet

$$p(X_3 = 0 | X_1 = 0, X_2 = 1) = \frac{0 + 1/2}{0 + 1} = 1/2$$



how many times we have seen 0 given this context?

$$p(X_3 = 0 | X_1 = 0, X_2 = 1) = \frac{0 + 1/2}{0 + 1} = 1/2$$



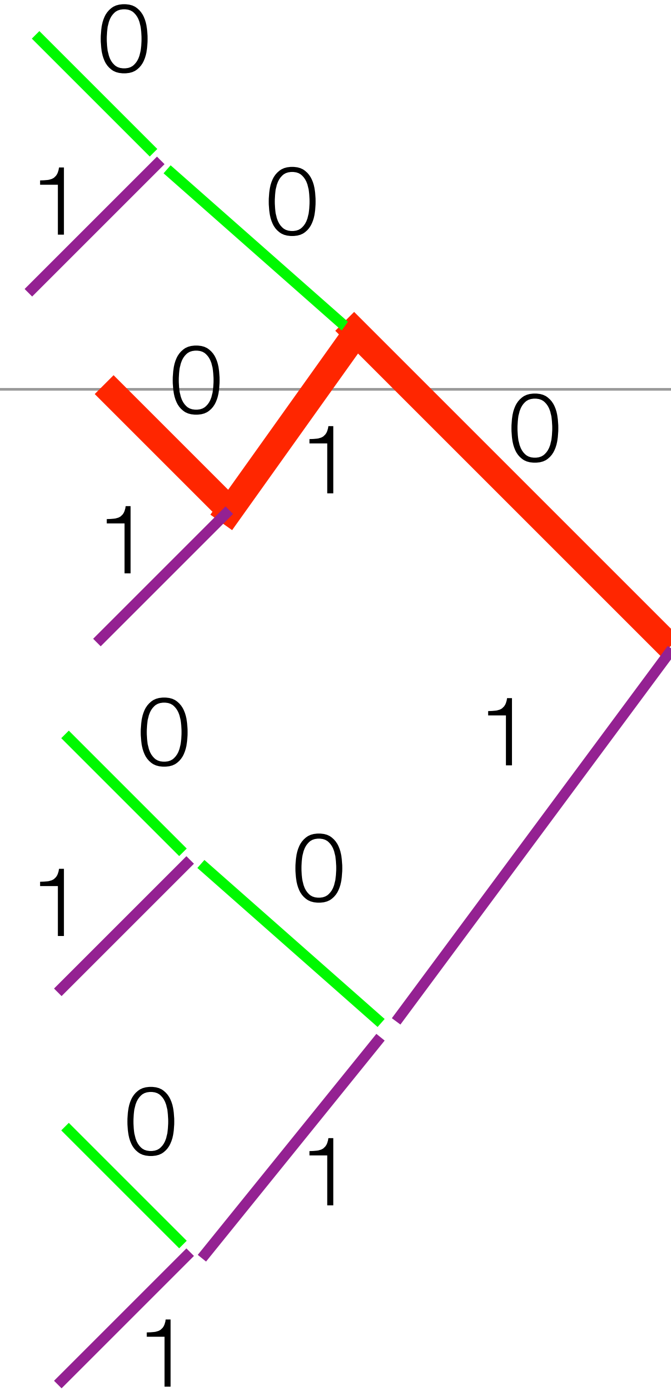
probability of this symbol?

010001010…

past values: the context

how many times we have seen this context?

$$p(X_3 = 0 | X_1 = 0, X_2 = 1) = \frac{0 + 1/2}{0 + 1} = 1/2$$



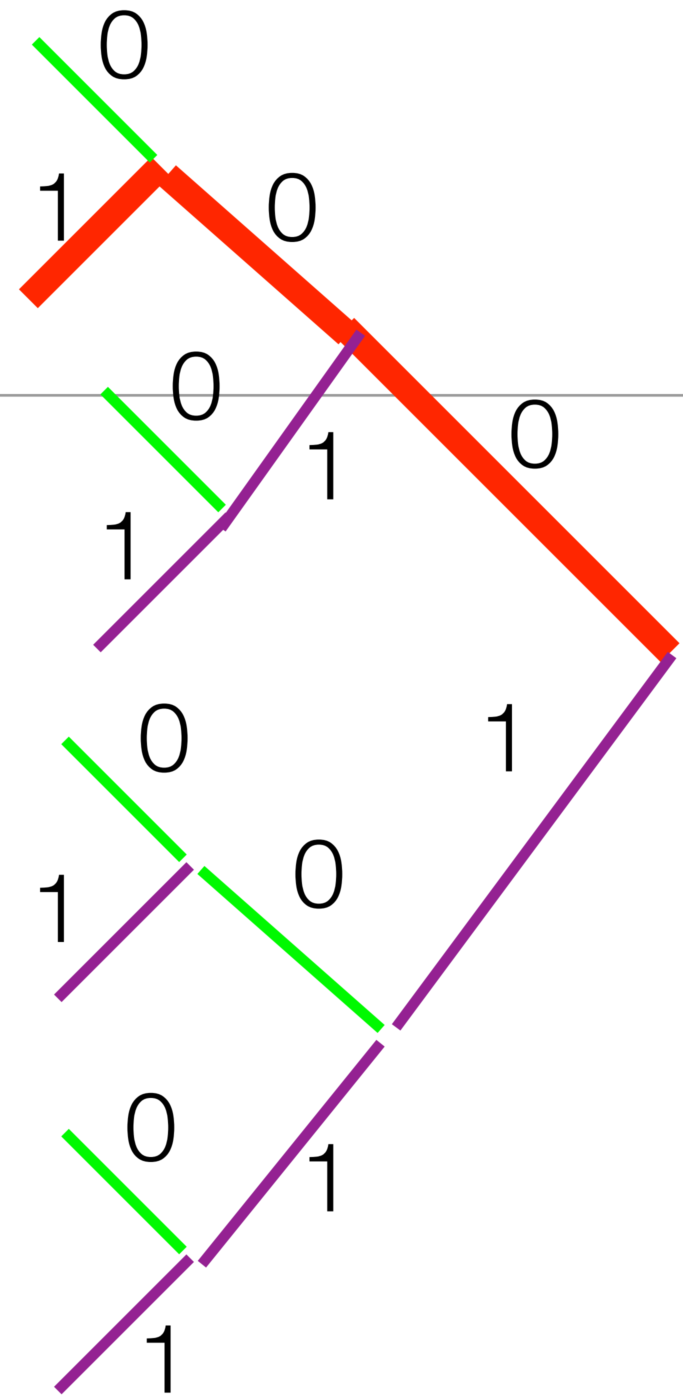
probability of this symbol?

010010101010...

past values: the context

how many times we have seen 0 given this context?

$$p(X_4 = 0 | X_1 = 0, X_2 = 1, X_3 = 0) = \frac{0 + 1/2}{0 + 1} = 1/2$$



probability of this symbol?

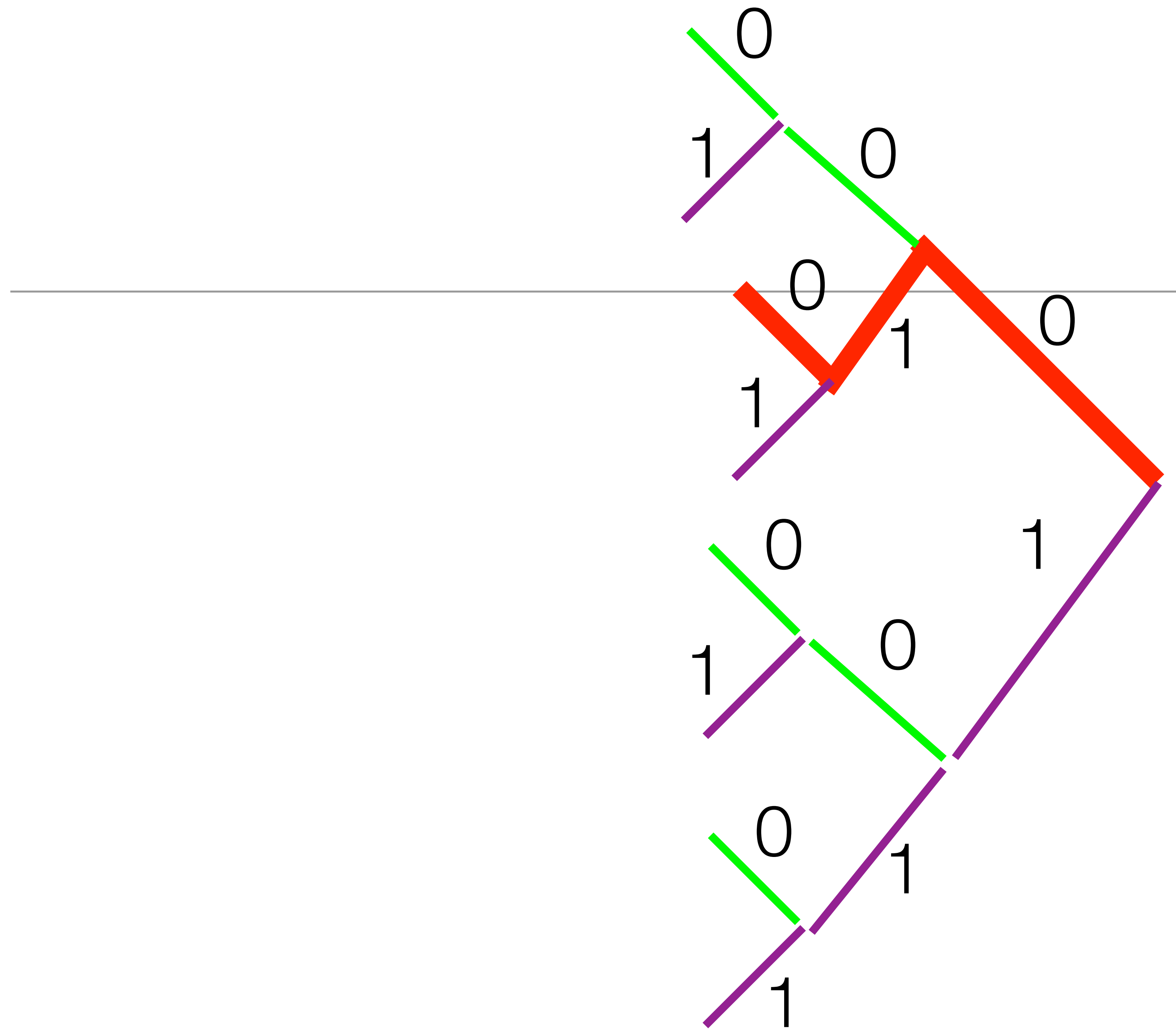
01001010…

past values: the context

how many times we have seen 0 given this context?

$$p(X_5 = 0 | X_2 = 1, X_3 = 0, X_4 = 0) = \frac{0 + 1/2}{0 + 1} = 1/2$$

- After a few steps, a familiar context appears



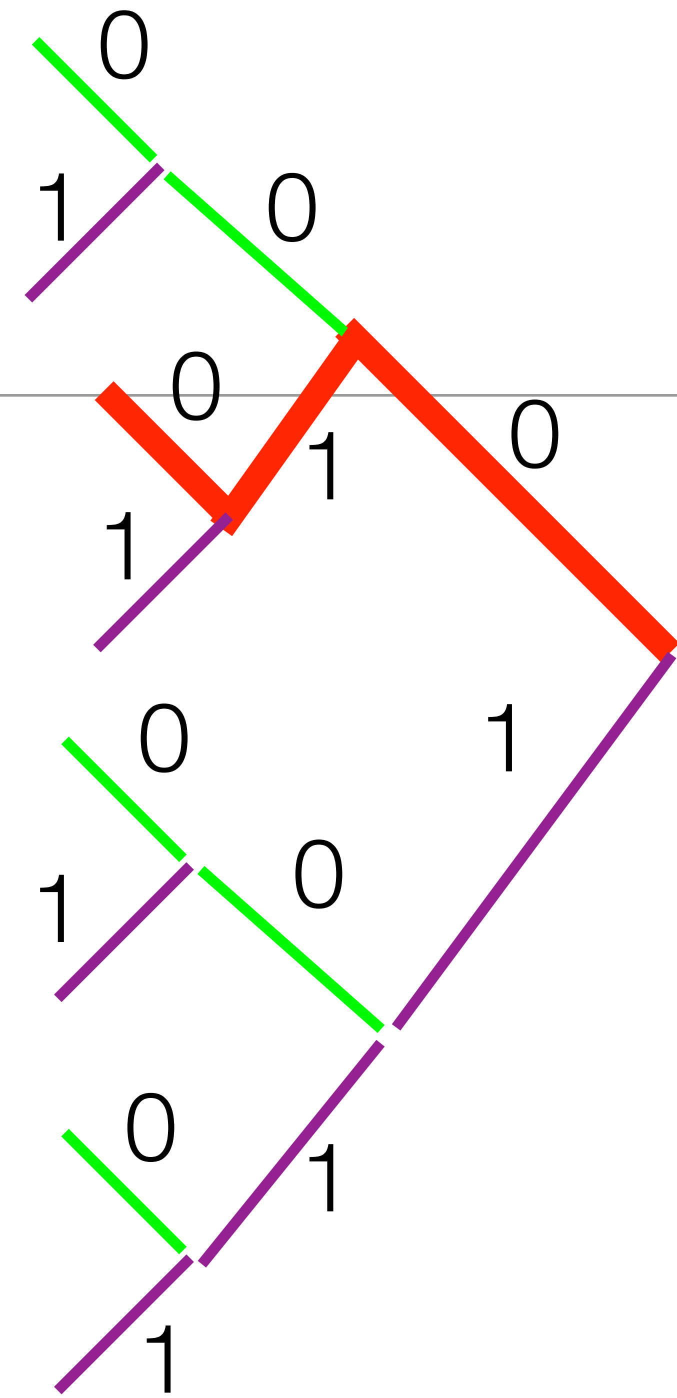
probability of this symbol?

01001010...

past values: the context

how many times we have seen 0 given this context?

$$p(X_9 = 0 | X_6 = 0, X_7 = 1, X_8 = 0) = \frac{1 + 1/2}{2 + 1} = 1/2$$



probability of this symbol?

01001010...

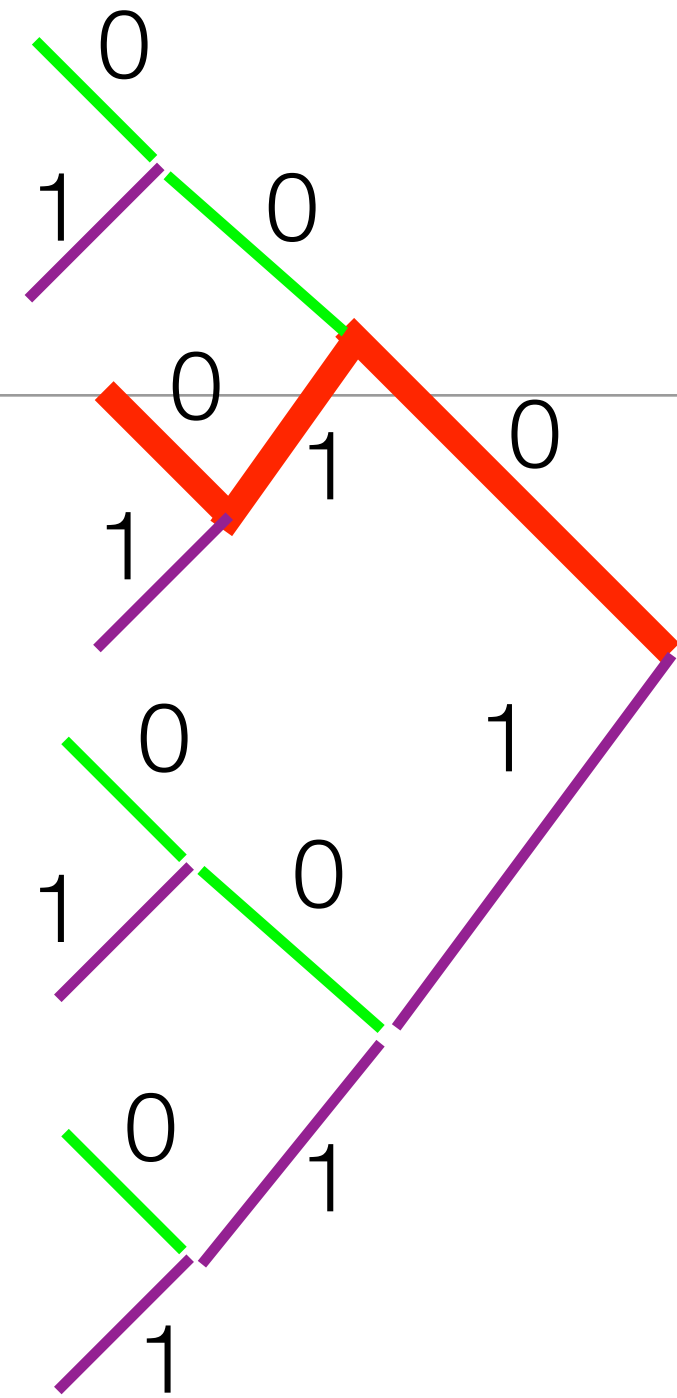
past values: the context

how many times we have seen this context?

$$p(X_9 = 0 | X_6 = 0, X_7 = 1, X_8 = 0) = \frac{1 + 1/2}{2 + 1} = 1/2$$

- If data was assumed to be i.i.d.
    - Best estimate of probability of zero = 5/8
- Without i.i.d assumption and with our context
    - Best estimate of probability of zero = 1/2
    - If the context was a little different—in one value
        - Best estimate of probability of zero = 1/4

- Example 5.4



probability of this symbol?

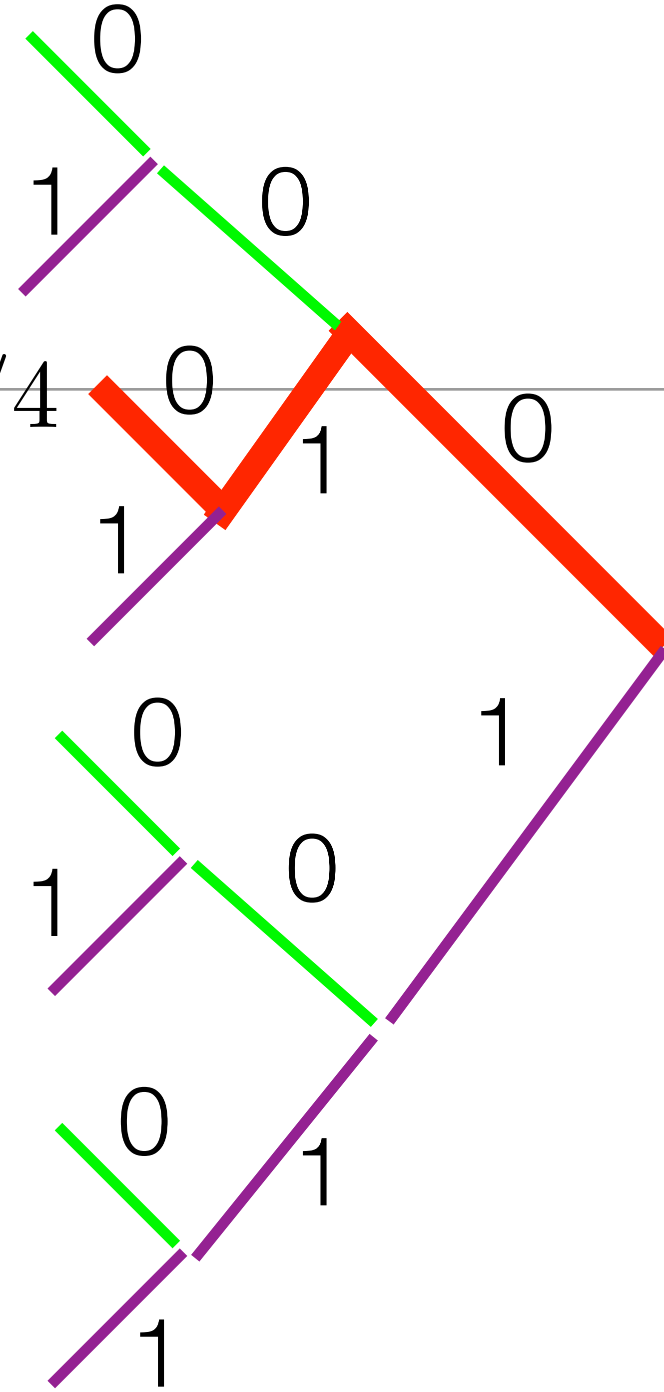
01011010...

past values: the context

how many times we have seen this context?

$$p(X_9 = 0 | X_6 = 0, X_7 = 1, X_8 = 0) = \frac{0 + 1/2}{1 + 1} = 1/4$$

$\hat{p}(X = 0|010) = 1/4$



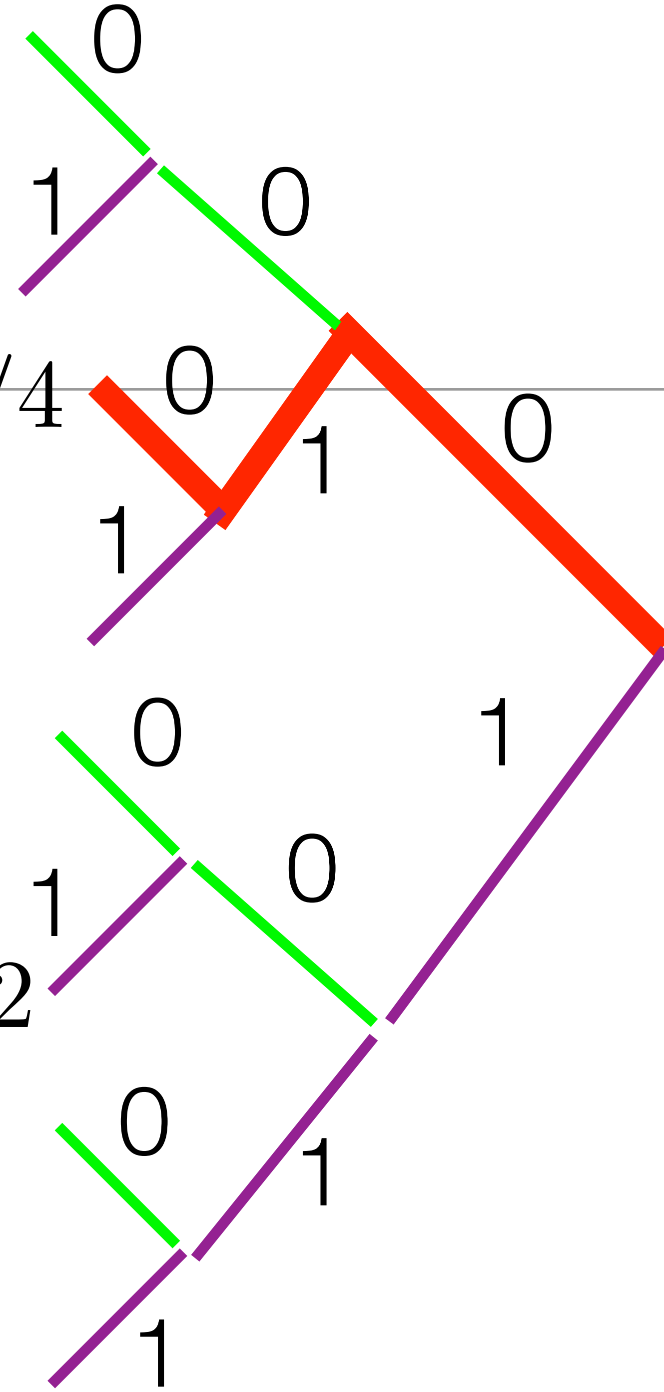
probability of this symbol?

01011010…

past values: the context

$$\hat{p}(X=0|010) = 1/4$$

$$\hat{p}(X=0|101) = 1/2$$

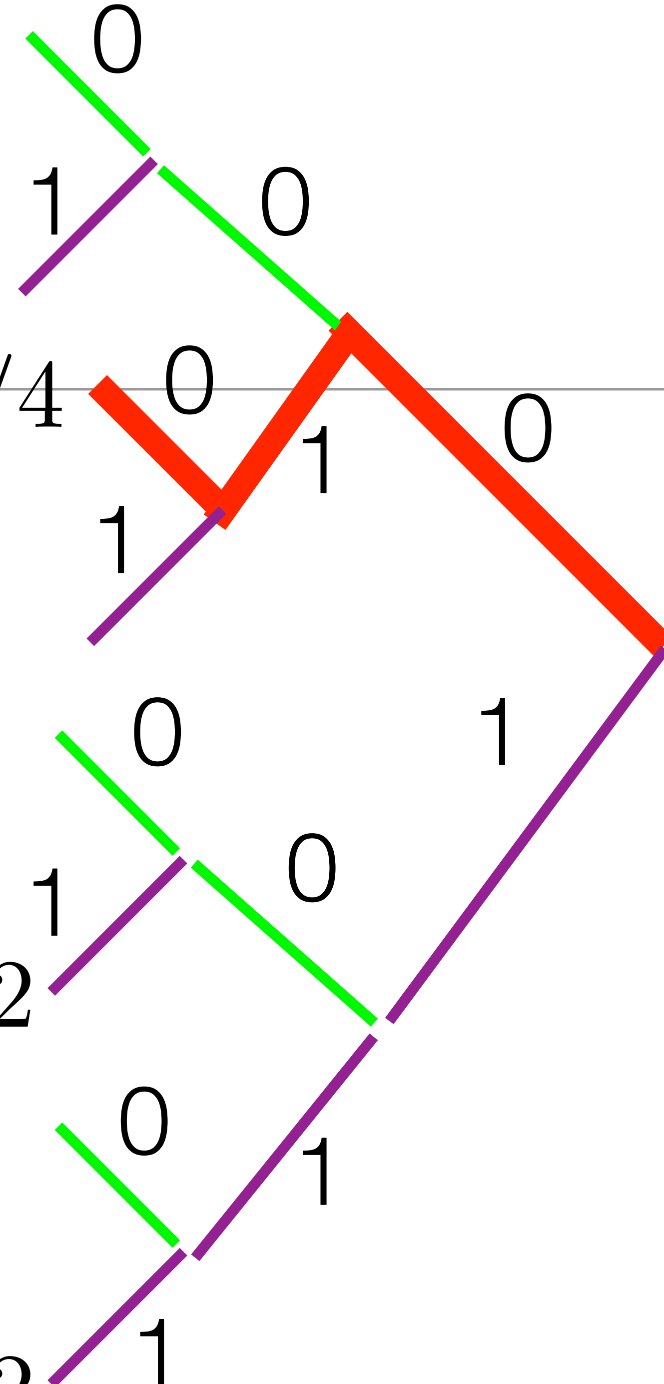


probability of this symbol?

01011010...

past values: the context

$$\hat{p}(X = 0|010) = 1/4$$



$$\hat{p}(X = 0|101) = 1/2$$

$$\hat{p}(X = 0|111) = 1/2$$

probability of this symbol?

01011010...

past values: the context

- A universal method to compute the joint probability

$$\hat{p}_{\mathbf{X}} = \hat{p}_{X_n|X_1^{(n-1)}} \hat{p}_{X_1^{(n-1)}} = \hat{p}_{X_n|X_1^{(n-1)}} \hat{p}_{X_{(n-1)}|X_1^{(n-2)}} \hat{p}_{X_1^{(n-2)}}$$
$$= \hat{p}_{X_n|X_1^{(n-1)}} \hat{p}_{X_{(n-1)}|X_1^{(n-2)}} \dots \hat{p}_{X_2|X_1} \hat{p}_{X_1}$$

- Where $X_1^n = (X_1, X_2, \dots, X_n)$

- The density estimator
  - The KT algorithm
  - The tree structure
    - Converges to the true density
- Plugin estimator

$$\hat{H}(\mathbf{X}) = - \sum_{i \in \{1, \ldots, n\}} \hat{p}_{\mathbf{X}} \log \hat{p}_{\mathbf{X}}$$

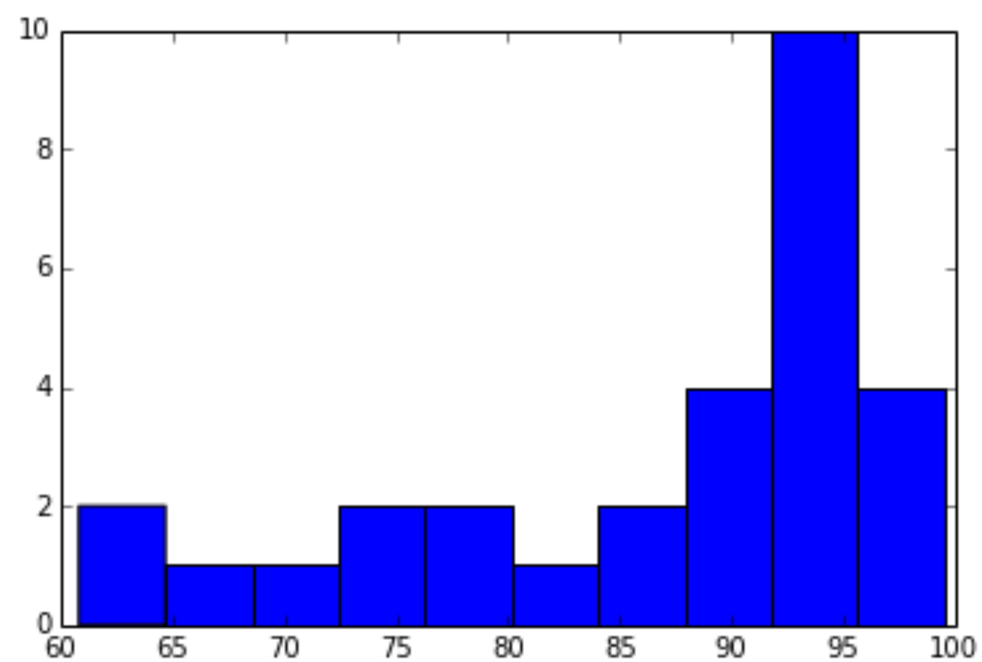
- Entropy of continuous valued random variables

$$h(X) = -\int_x f_X(x) \log f_X(x) dx$$

- Estimating the entropy
  - Plugin estimator
  - How does Histogram estimate perform?

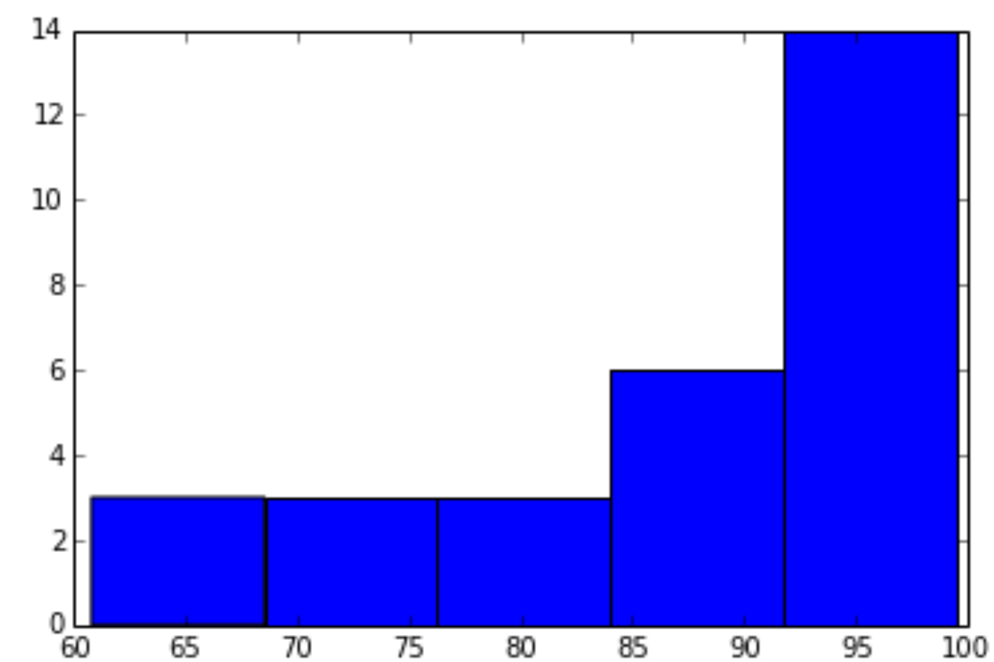
- Example 5.5

- Data: 93.5,93,60.8,94.5,82,87.5,91.5,99.5,86,93.5,92.5,78,76,69,94.5,89.5,92.8,78,65.5,98,98.5,92.3,95.5,76,91,95,61.4,96,90



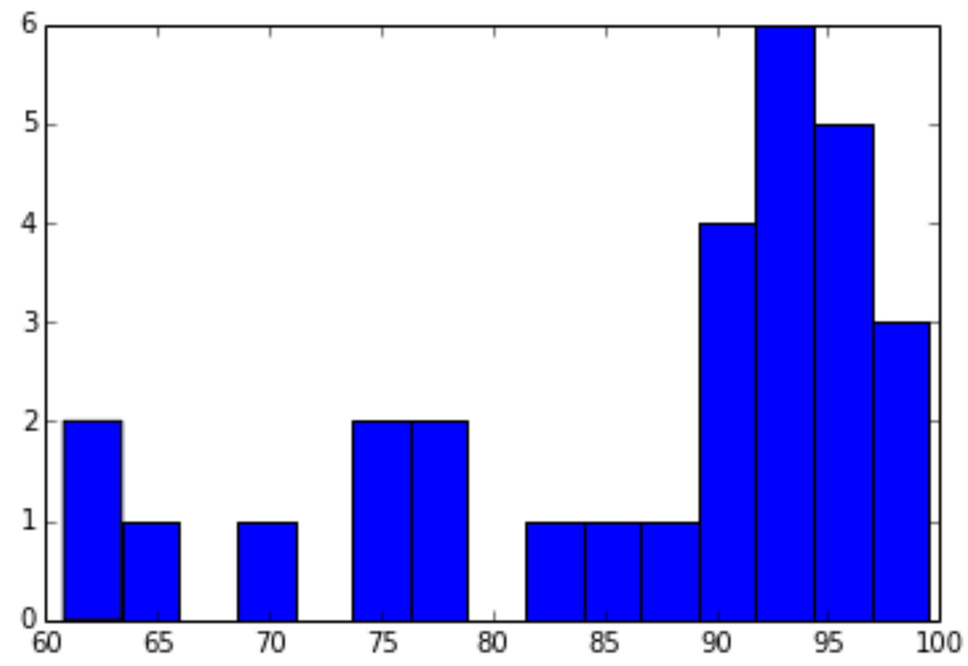
- Histogram of data

- Data:
93.5,93,60.8,94.5,82,87.5,91.5,99.5,86,93.5,92.5,78,76,69,94.5,89.5,92.8,78,65.5,98,98.5,92.3,95.5,76,91,95,61.4,96,90



- Histogram of data

- Data: 93.5,93,60.8,94.5,82,87.5,91.5,99.5,86,93.5,92.5,78,76,69,94.5,89.5,92.8,78,65.5,98,98.5,92.3,95.5,76,91,95,61.4,96,90



- Entropy of continuous valued random variables

$$h(X) = -\int_x f_X(x) \log f_X(x) dx$$

- Estimating the entropy
  - Plugin estimator
  - Histogram estimator performs poorly for high dimensional data
    - Extreme dependence on bin size, even in one dimensional data

$$\mathbf{X}_1^n = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n) \text{ where } \mathbf{X}_i \in \Re^d$$

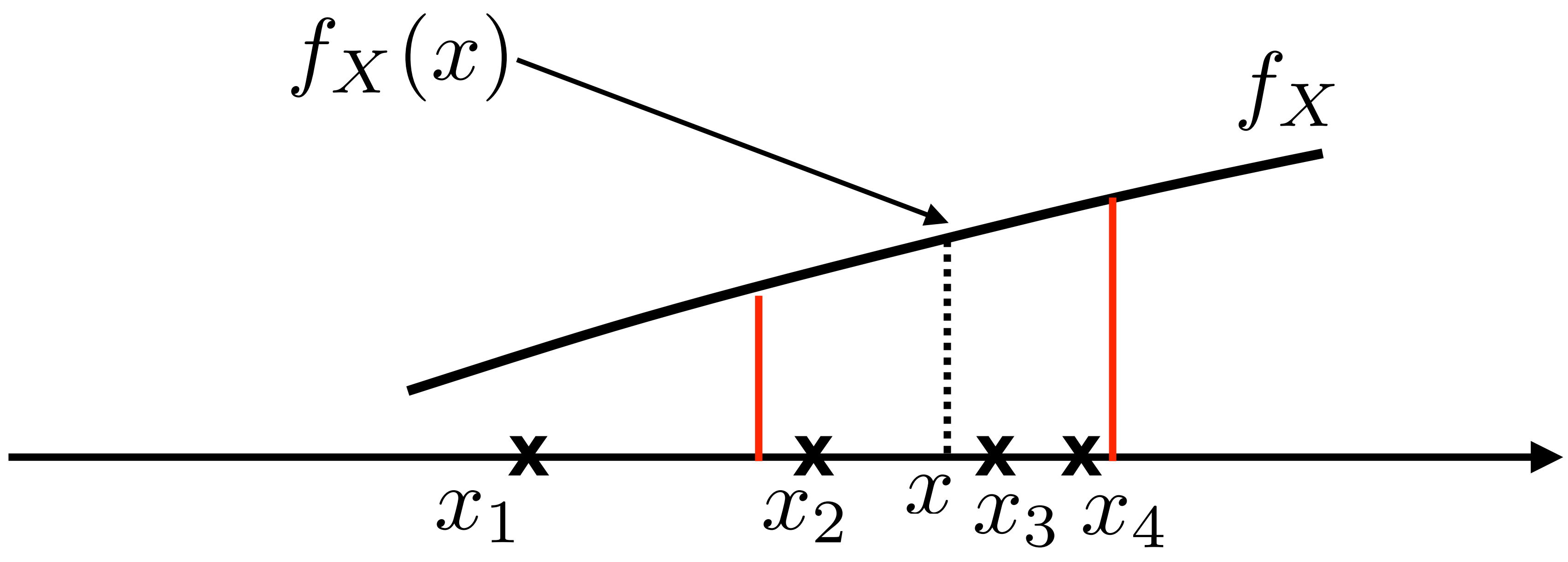
- Kernel density estimation (Parzen's window)
  - Based on *n* samples of *d* dimensional data

$$\mathbf{X}_1^n = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n) \text{ where } \mathbf{X}_i \in \Re^d$$

- The concept:
  - Consider the probability of a mass in a region

$$P = \int_A f_{\mathbf{X}}(\mathbf{x})d\mathbf{x}$$

    - That is, the probability of a point being inside of area *A*



- The concept:
  - Consider the probability "mass" in a region

$$P = \int_A f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}$$

    - That is, the probability of **x** being inside of area *A*
  - The total number of data points is *n*
  - The probability of *k* points being inside region *A* is P^k

$$P = \int_A f_{\mathbf{X}}(\mathbf{x})d\mathbf{x}$$

- The total number of data points is *n*
- Probability of *k* out of *n* be inside region *A* is

$$\Pr(n, k) = \binom{n}{k} P^k (1 - P)^{n-k}$$

$$P = \int_A f_{\mathbf{X}}(\mathbf{x})d\mathbf{x}$$

- For large *n,* the (average) number of points inside the region

$$k \approx nP$$

- If the region is assumed to be small then the density will be approximately constant

$$P \approx f_{\mathbf{X}} V_A$$

$$\text{where } V_A \text{ is the volume of } A$$

$$P = \int_A f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}$$

- If the region is assumed small then the density will be approximately constant

$$P \approx f_{\mathbf{X}} V_A$$

- The probability density function over a small region, however, with enough points inside is

$$f_{\mathbf{X}} \approx \frac{k}{n V_A}$$

- If we fix the volume and determine *k* from the data
  - We will have KDE (Parzen's window)
- If we fix *k* and determine the volume
  - We will have K-nearest neighbor (k-NN)

- Recall the density was approximated as

$$f_{\mathbf{x}} \approx \frac{k}{nV_A}$$

- Then, in KDE the volume is fixed. Example of fixed volume hypercube

$$K(\mathbf{x}) = \begin{cases} 1 & \text{if } |x^{(m)}| \leq 1/2, m = 1, 2, \ldots, d \\ 0 & \text{otherwise} \end{cases}$$

- If the data falls inside the cube it counts as one.

- If the region was a hypercube with side *h* then

$$K(\frac{\mathbf{X} - \mathbf{X_i}}{h}) \text{ will be 1}$$

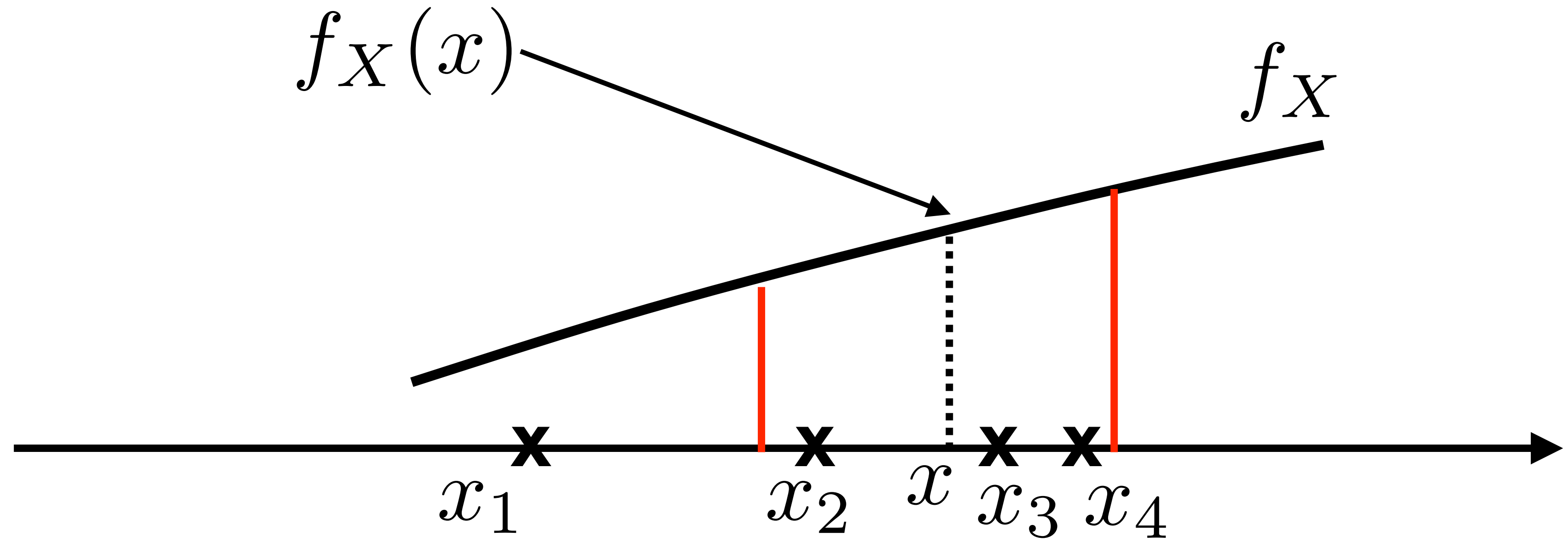
- Since the point $\mathbf{x}_i$ is inside the hypercube

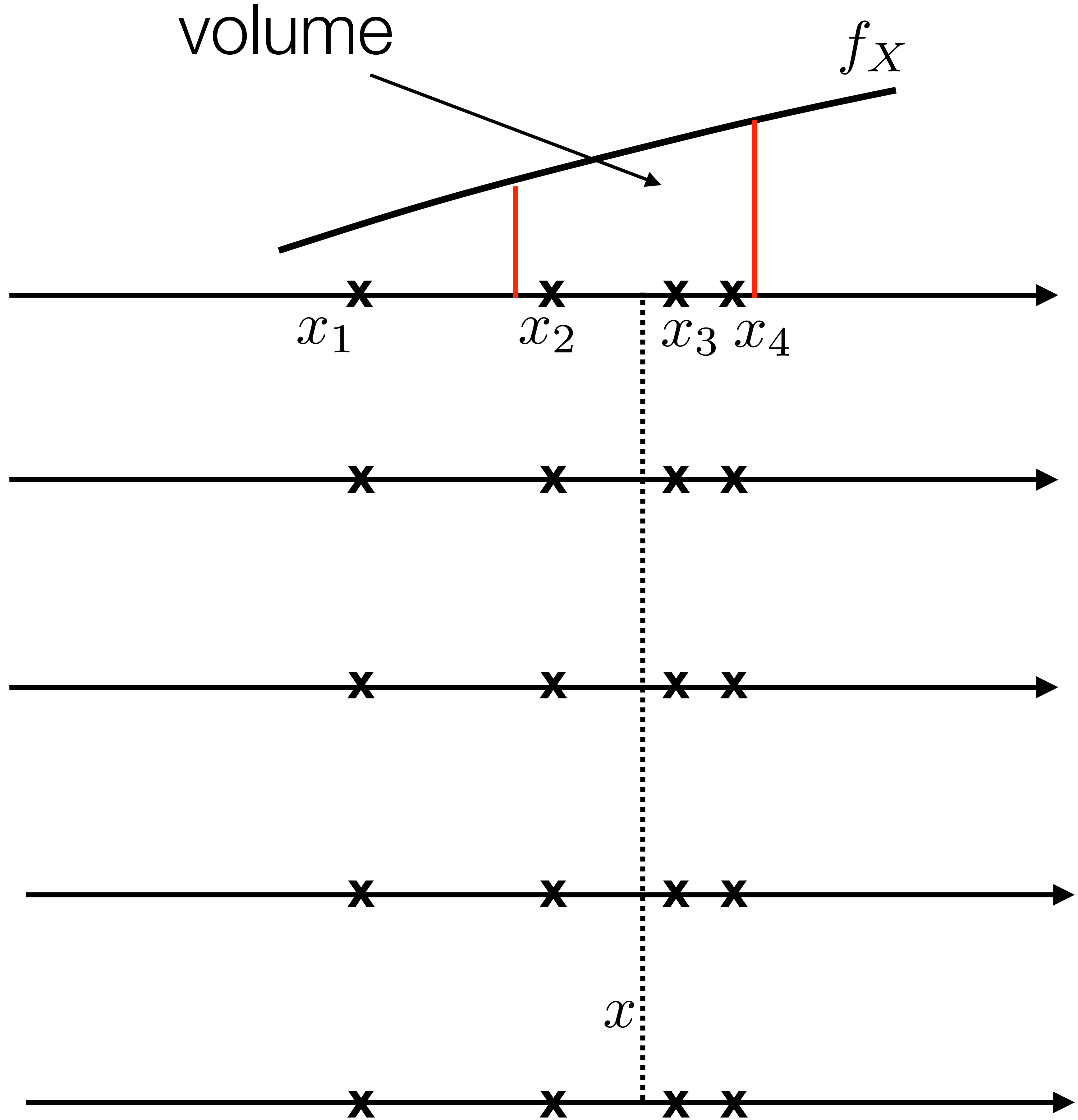
- Then, the total number of data points inside the kernel is

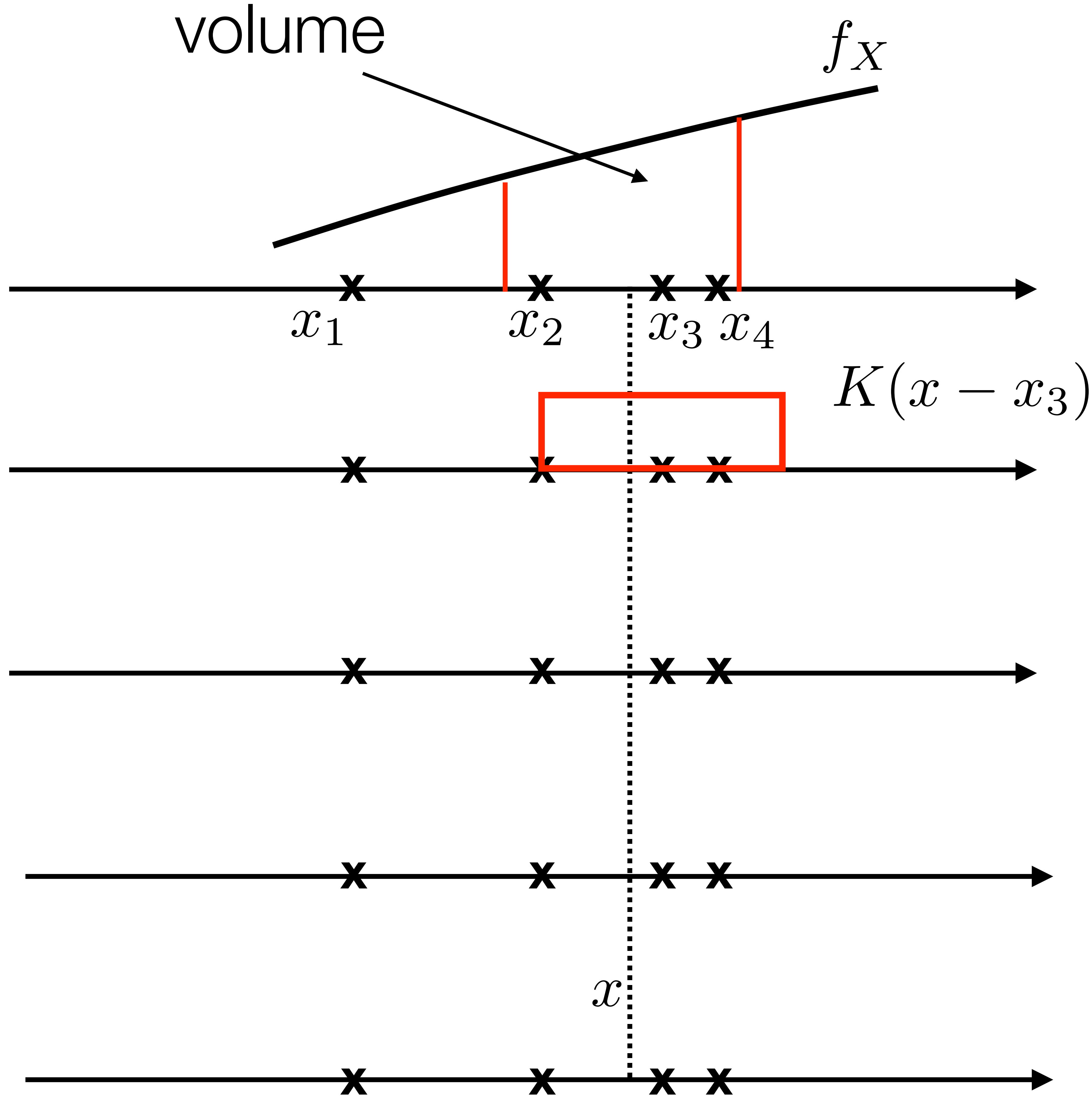
$$k = \sum_{i=1}^{n} K(\frac{\mathbf{X} - \mathbf{X_i}}{h})$$

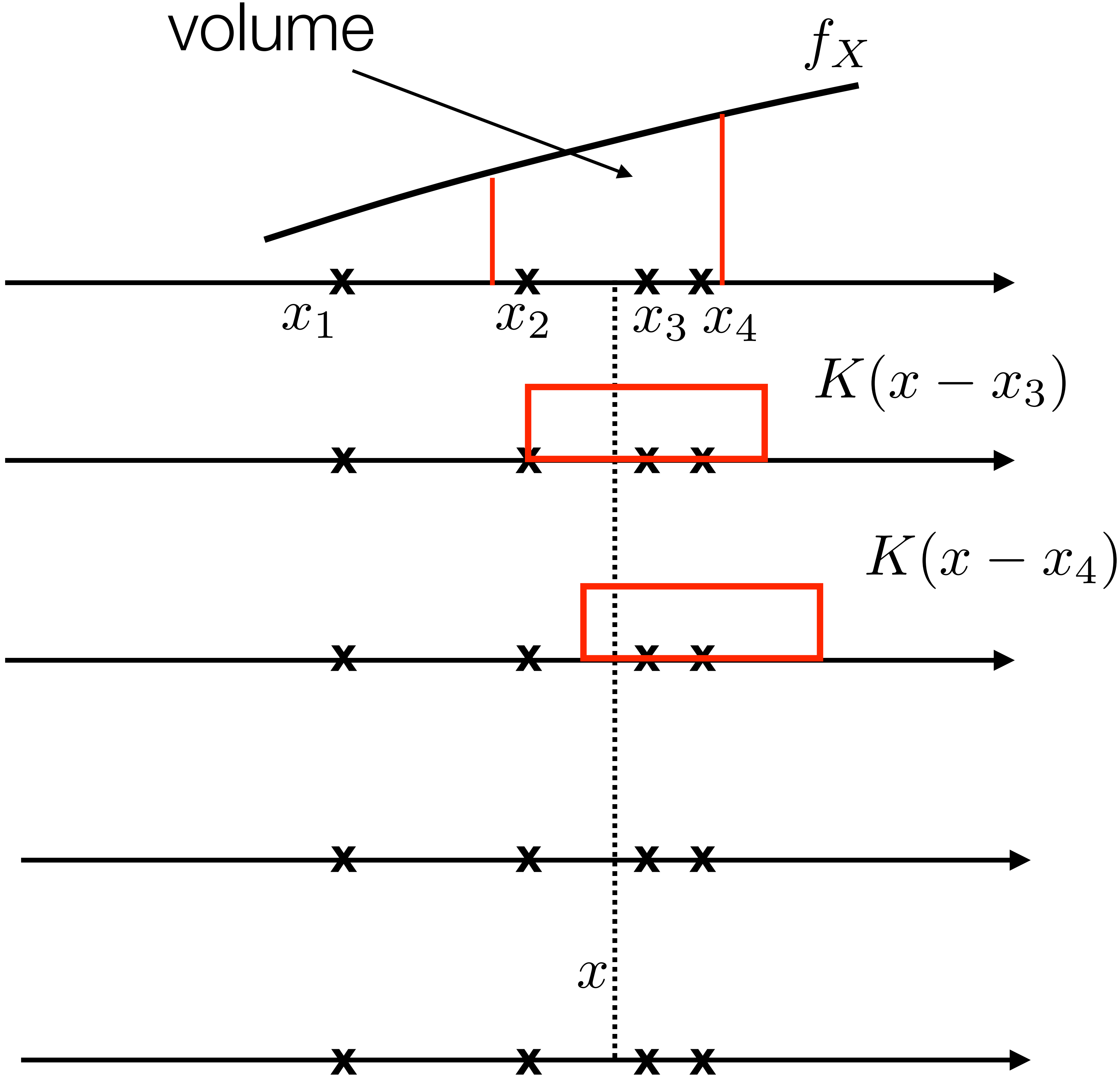
- Example 5.6

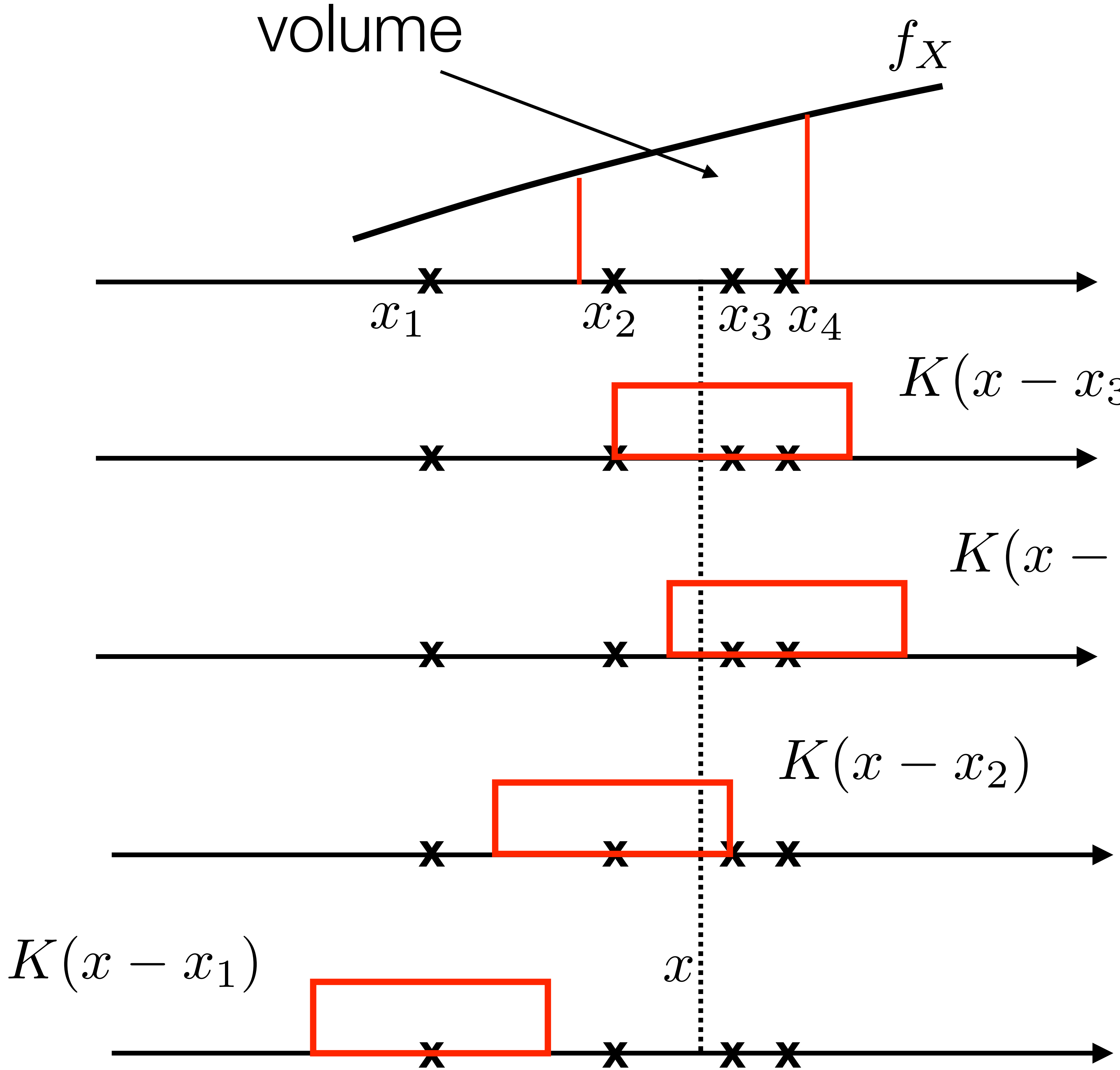
- An illustrative example


$f_X(x)$
f_X
x_1
x_2
x
x_3
x_4


volume
f_X
x_1
x_2
x_3
x_4
x


volume
f_X
x_1
x_2
x_3
x_4
$K(x - x_3)$
x


volume
f_X
x_1
x_2
x_3
x_4
$K(x - x_3)$
$K(x - x_4)$
x


volume
f_X
x_1
x_2
x_3
x_4
$K(x - x_3)$
$K(x - x_4)$
$K(x - x_2)$
$K(x - x_1)$
x

$$f_{\mathbf{x}} \approx \frac{k}{nV_A}$$

- The KDE

$$\hat{f}_h(\mathbf{x}) = \frac{1}{nh^d} \sum_{i=1}^{n} K(\frac{\mathbf{x} - \mathbf{x}_i}{h})$$

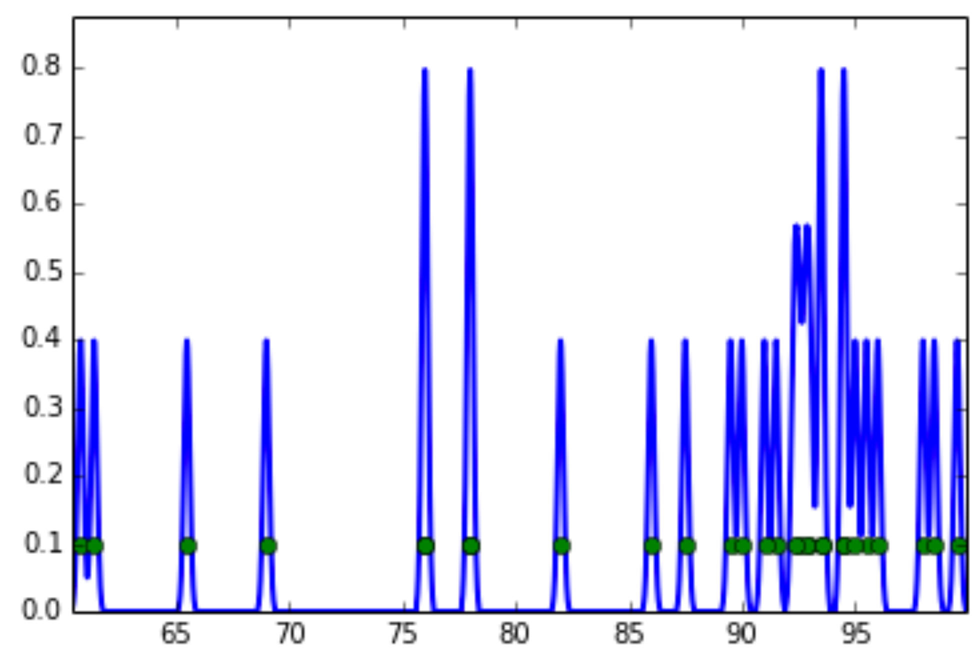
- Using hypercube has similar rough boundaries as histogram approach does

- A candidate kernel is Gaussian

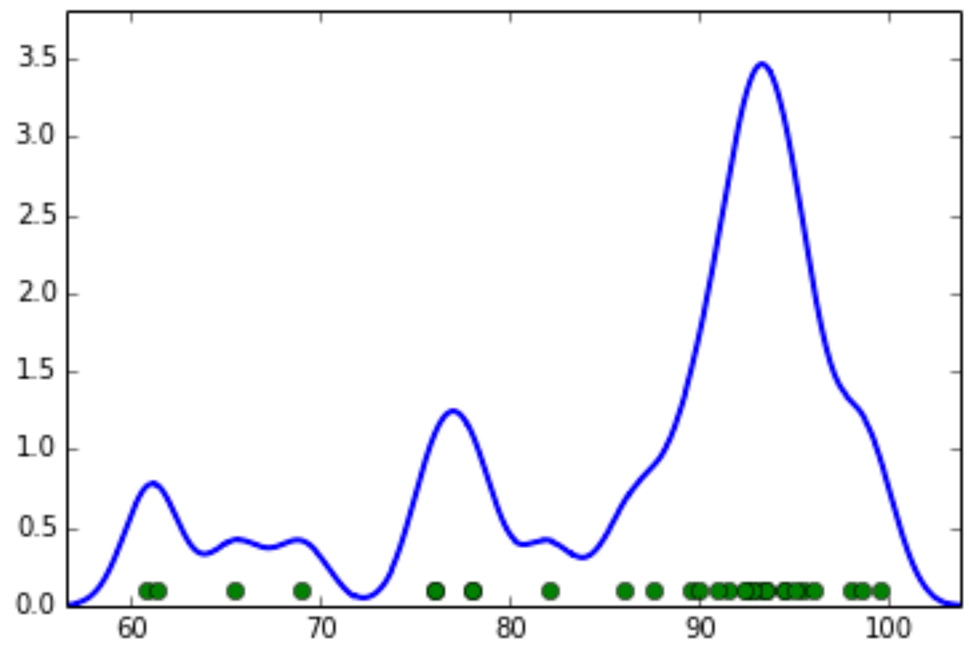
$$K(x) \propto e^{-x^2}$$

- Example 5.7

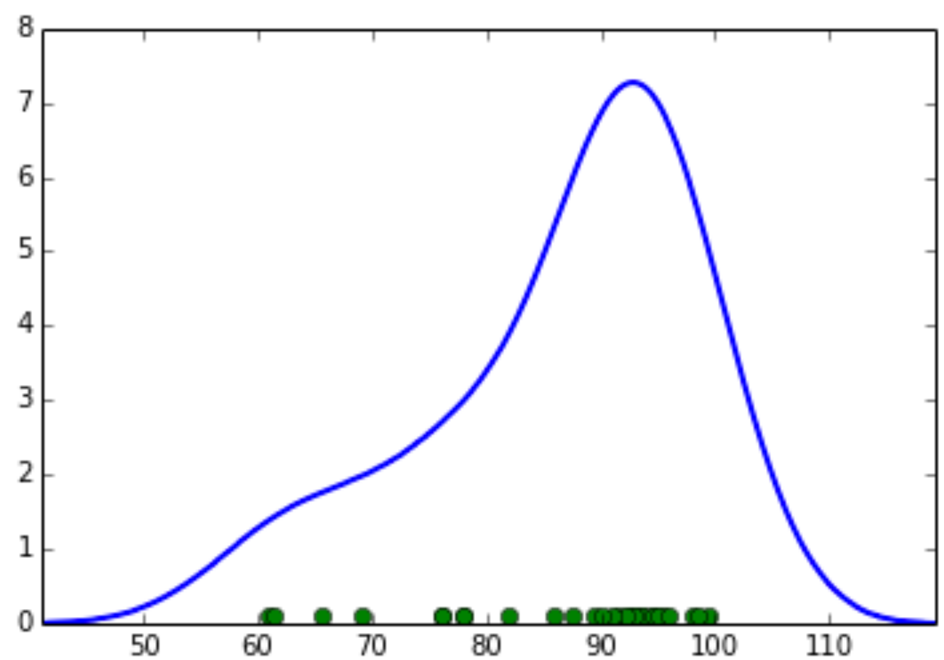
- KDE example with small *h*



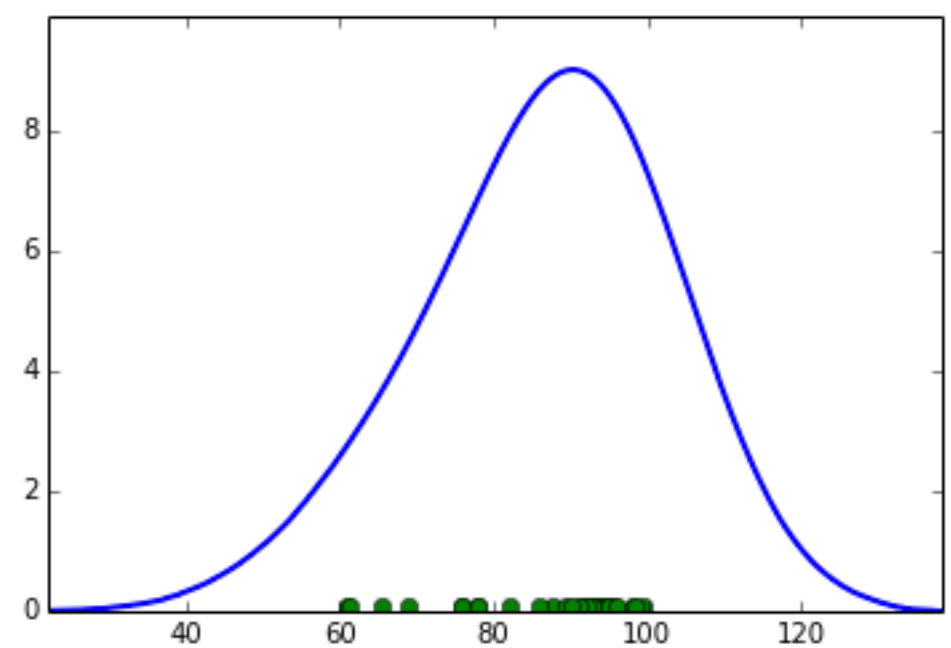
- Moderately small *h*



- Mid range value of *h*



- Large *h*



- The parameter h controls smoothness of resulting estimate
  - Choice of h is critical
  - There are still issues with KDE
    - Large dimensions
      - We can not guarantee to have enough points in each area A

- K nearest neighbor is a powerful alternative to KDE
  - With k-NN, we fix the number of points in a region
  - The k-NN estimate is

$$f_{\mathbf{X}} \approx \frac{k}{nV_A}$$

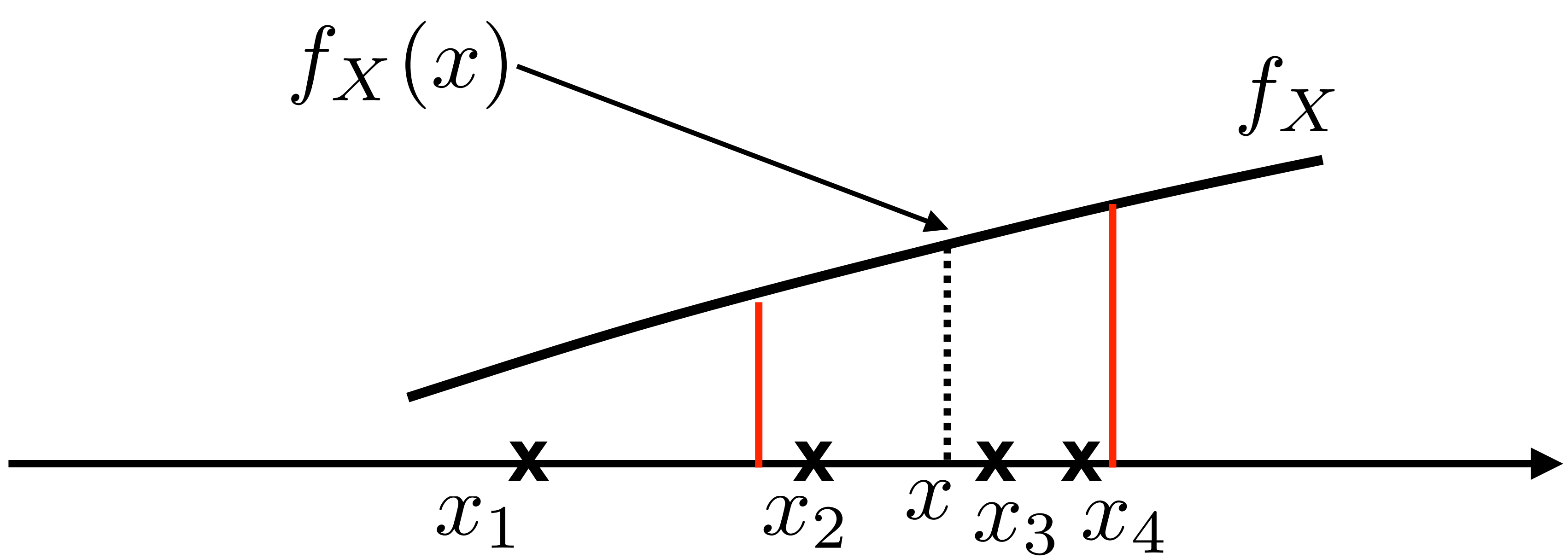
$$\hat{f}_{\mathbf{X}}(\mathbf{x}) = \frac{k}{nV} \text{ with } V \text{ as the volume with k points}$$

- For a point ***x*** to calculate density of the random vector at ***x***, that is, $f_{\mathbf{X}}(\mathbf{x})$
- The distance

$$D_i = ||\mathbf{x} - \mathbf{x}_i||_2 = \sqrt{\sum_{m=1}^{d} (x^{(m)} - x_i^{(m)})^2}$$

- Choose k nearest neighbors among all points

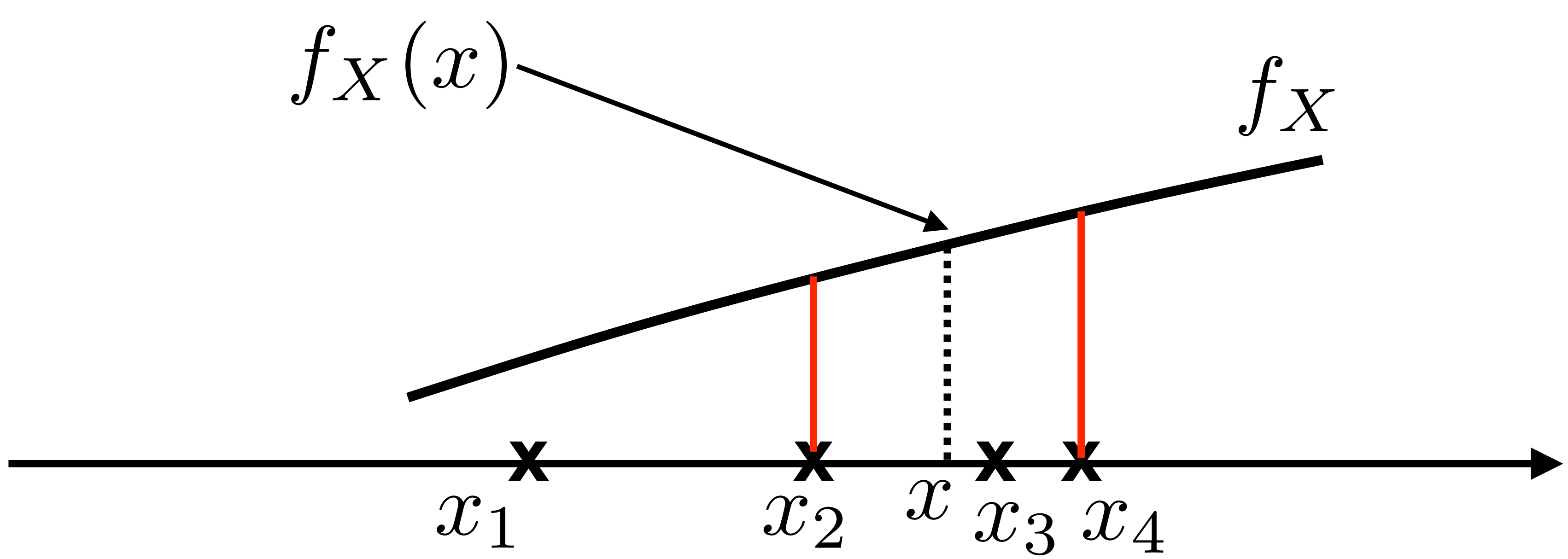
$$0 \leq D_3 \leq D_4 \leq D_2 \leq D_1$$



- For a point ***x*** to calculate density of the random vector at ***x***, that is, $f_{\mathbf{X}}(\mathbf{x})$

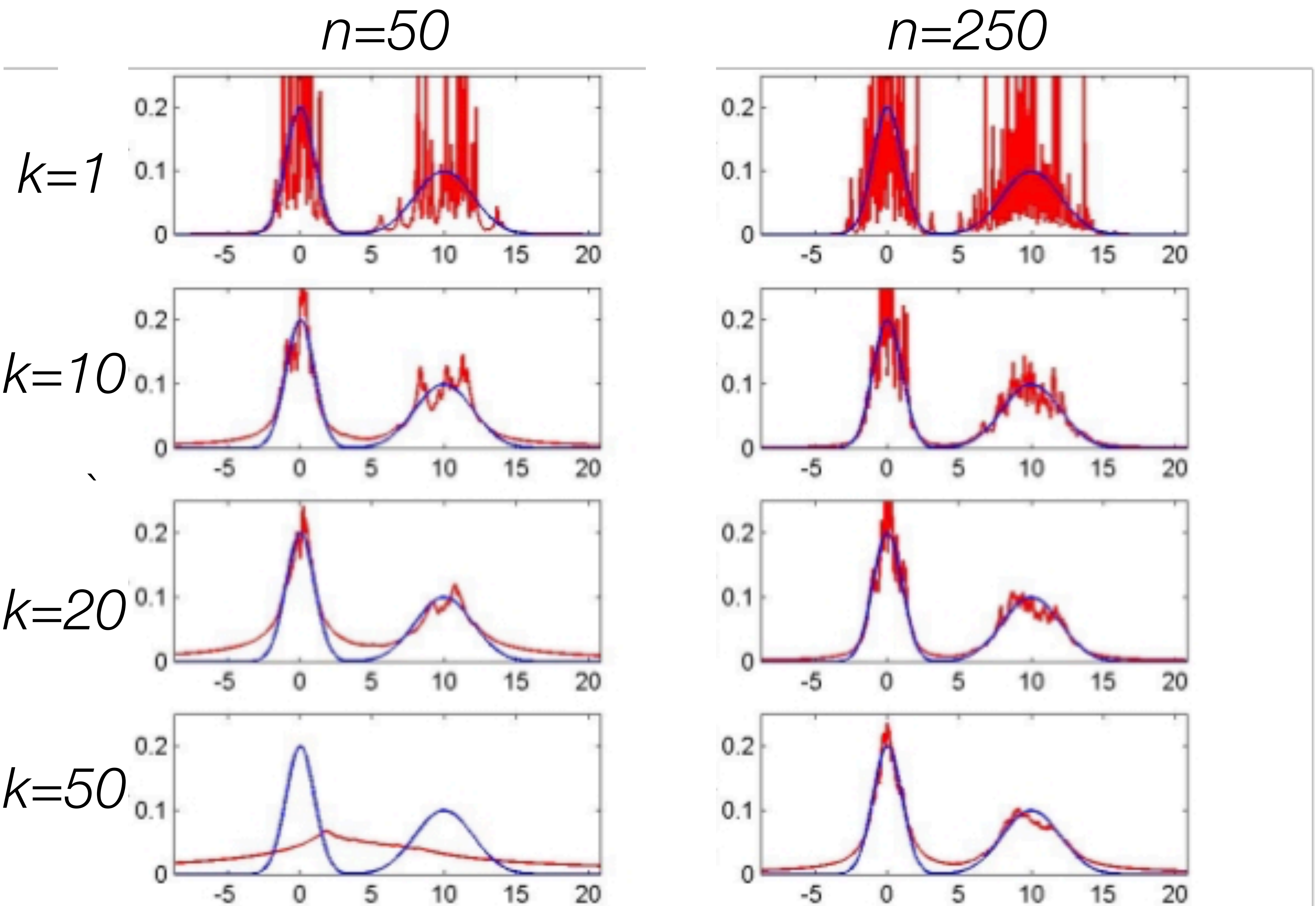
- Choose 3 nearest neighbors among all points then calculate the volume

$$\hat{f}_{\mathbf{X}}(\mathbf{x}) = \frac{k}{nV} \text{ with } V \text{ as the volume with k points}$$



- Blue density is the ground truth

- Red is the k-NN estimated density



n=50
n=250
k=1
k=10
k=20
k=50
0.2
0.1
0
-5
0
5
10
15
20

- These are examples of two density estimators as plugins for estimating
  - Entropy
$$\hat{h}(X) = -\int_x \hat{f}_X(x) \log \hat{f}_X(x) dx$$
  - Mutual information
$$\hat{I}(X;Y) = \hat{h}(X) - \hat{h}(X|Y)$$
  - Directed information
$$\hat{I}(X_1^n \to Y_1^n) = \hat{h}(Y_1^n) - \hat{h}(Y_1^n || X_1^n)$$
  - Coherence and mutual information in frequency
$$MI_{X,Y}(f,f) = I(d\tilde{X}_f; d\tilde{Y}_f) = -\log[1 - C_{X,Y}(f)]$$

Summary for Set I

- A probabilistic approach to dealing with recorded signals and data
- Avoid unnecessary assumption of a model
- Data driven techniques to estimate features in data
  - correlation, dependence, causality, coherence