Machine Learning for Signal Processing
Independent Component Analysis

Instructor: Bhiksha Raj

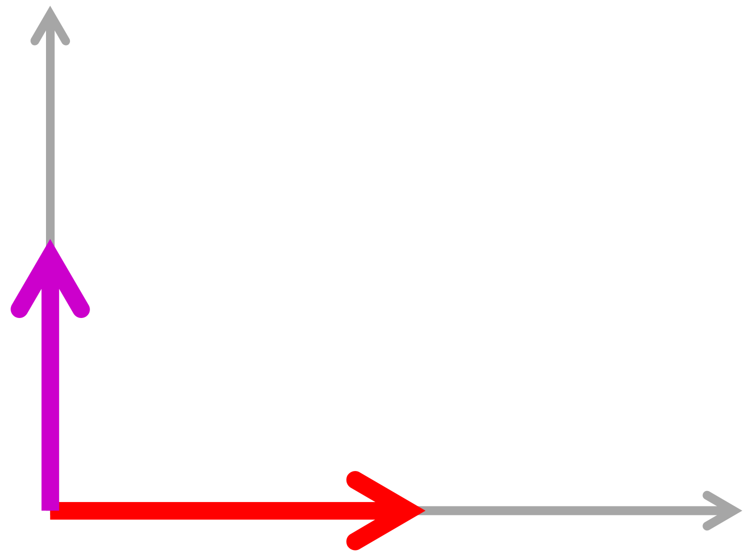
Revisiting the Covariance Matrix

- **Assuming centered data**

- $C = \Sigma_X XX^T$
- $= X_1X_1^T + X_2X_2^T + ...$

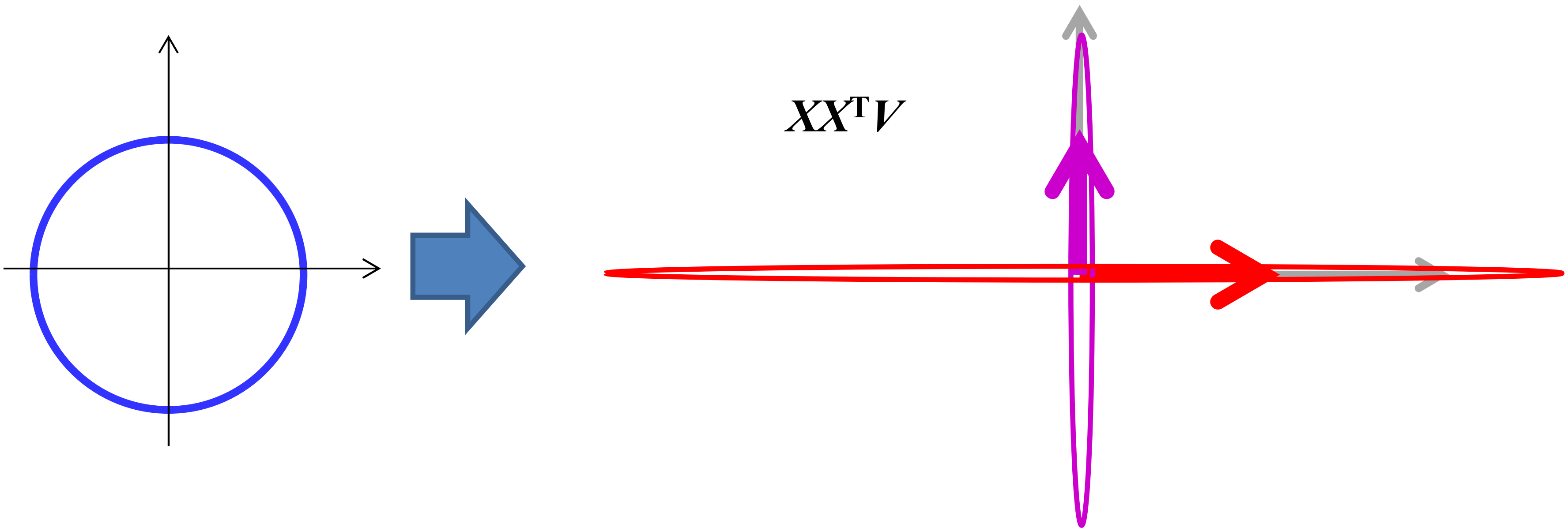
- Let us view C as a transform..

Covariance matrix as a transform



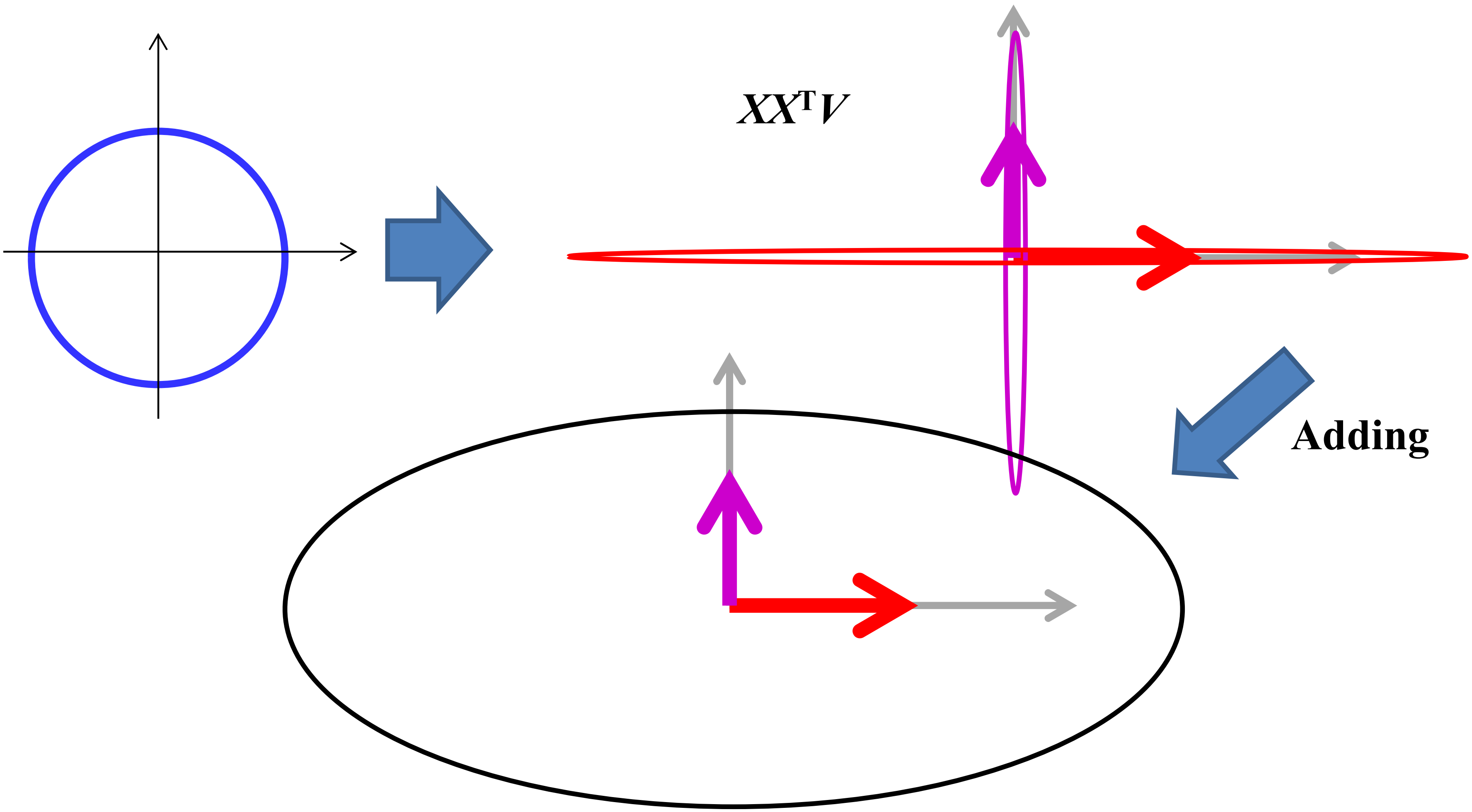
- $(X_1X_1^T + X_2X_2^T + \ldots)\ V = X_1X_1^TV + X_2X_2^TV + \ldots$
- Consider a 2-vector example
  - In two dimensions for illustration

Covariance Matrix as a transform



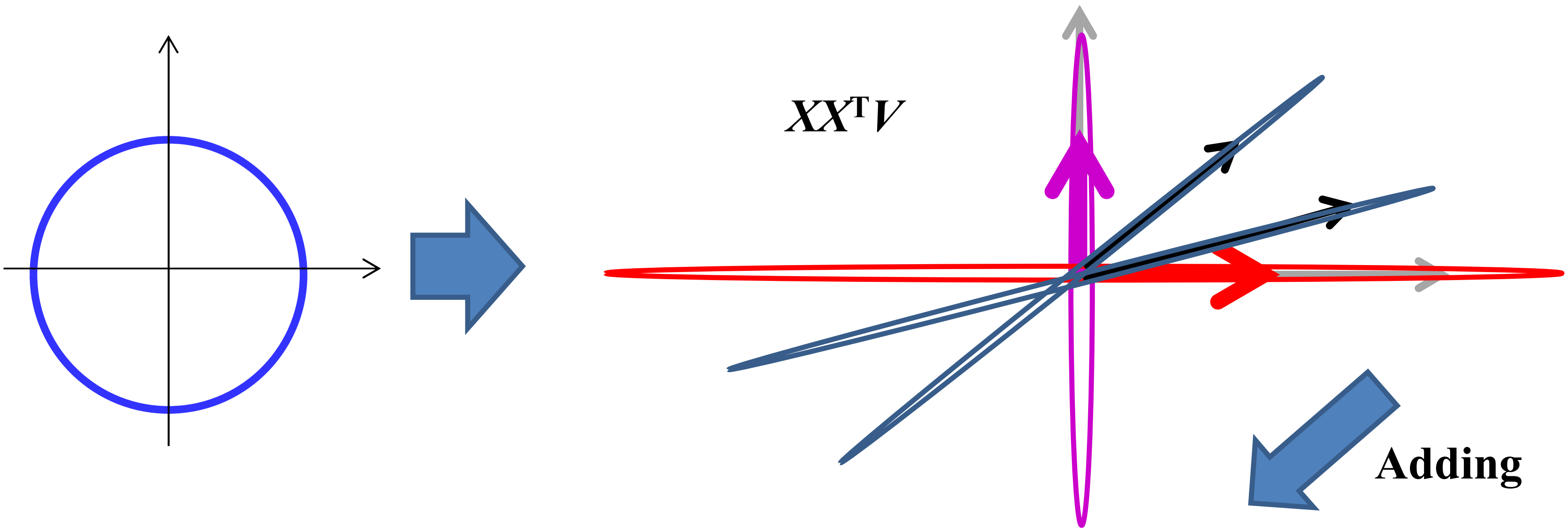
- Data comprises only 2 vectors..
- *Major axis of component ellipses proportional to the squared length of the corresponding vector*

Covariance Matrix as a transform



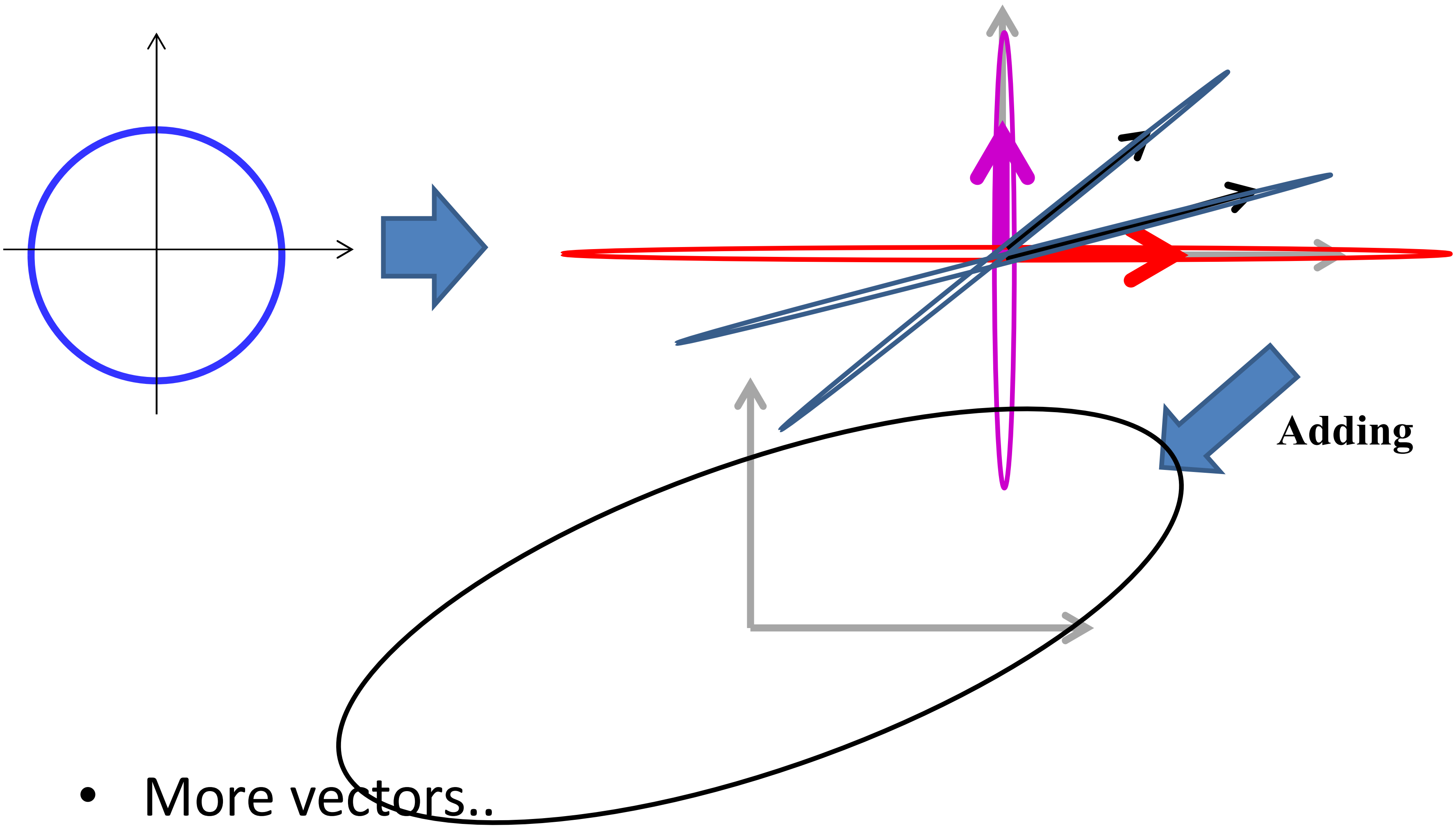
- Data comprises only 2 vectors..
- *Major axis of component ellipses proportional to the squared length of the corresponding vector*

Covariance Matrix as a transform



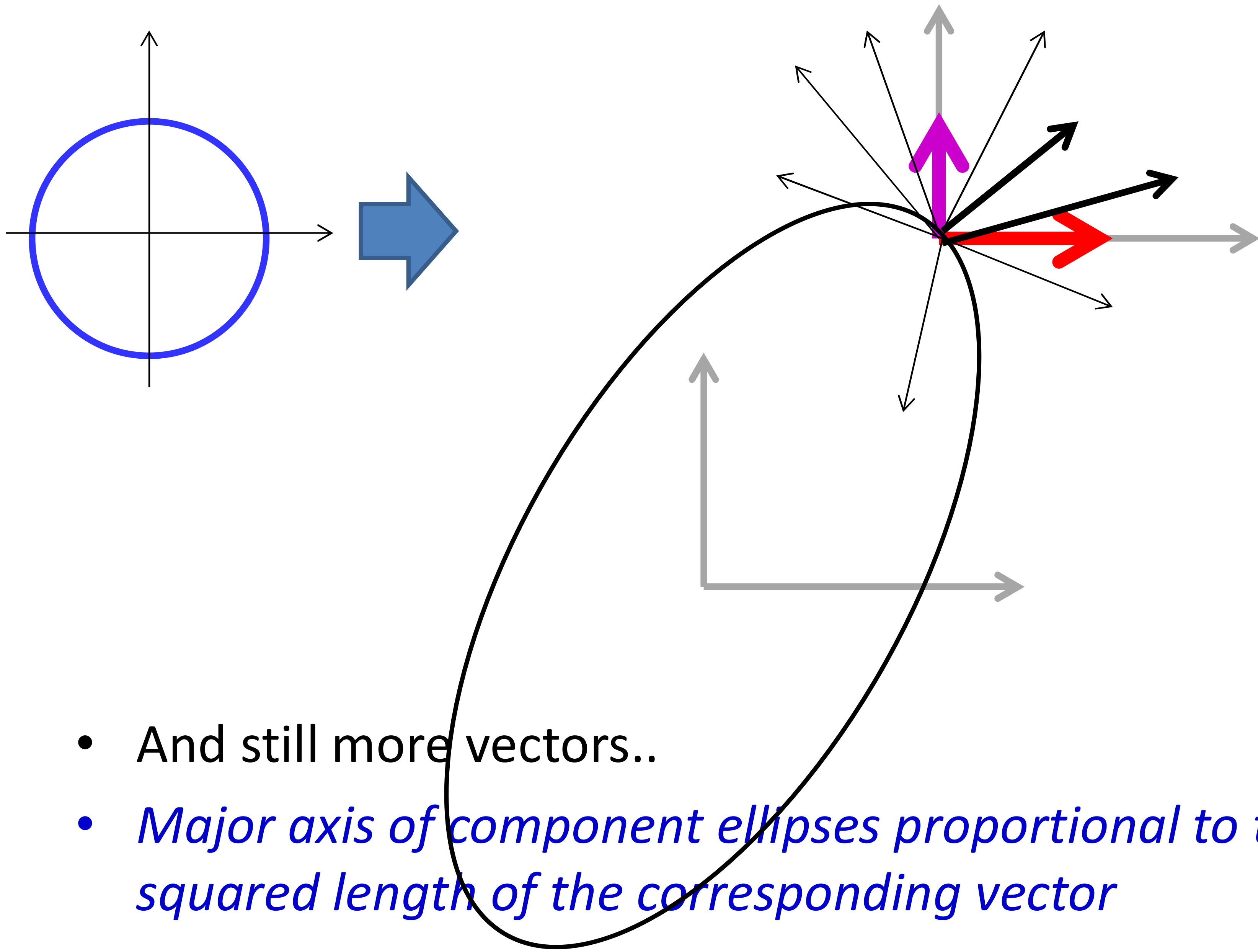
- More vectors..
- *Major axis of component ellipses proportional to the squared length of the corresponding vector*

Covariance Matrix as a transform



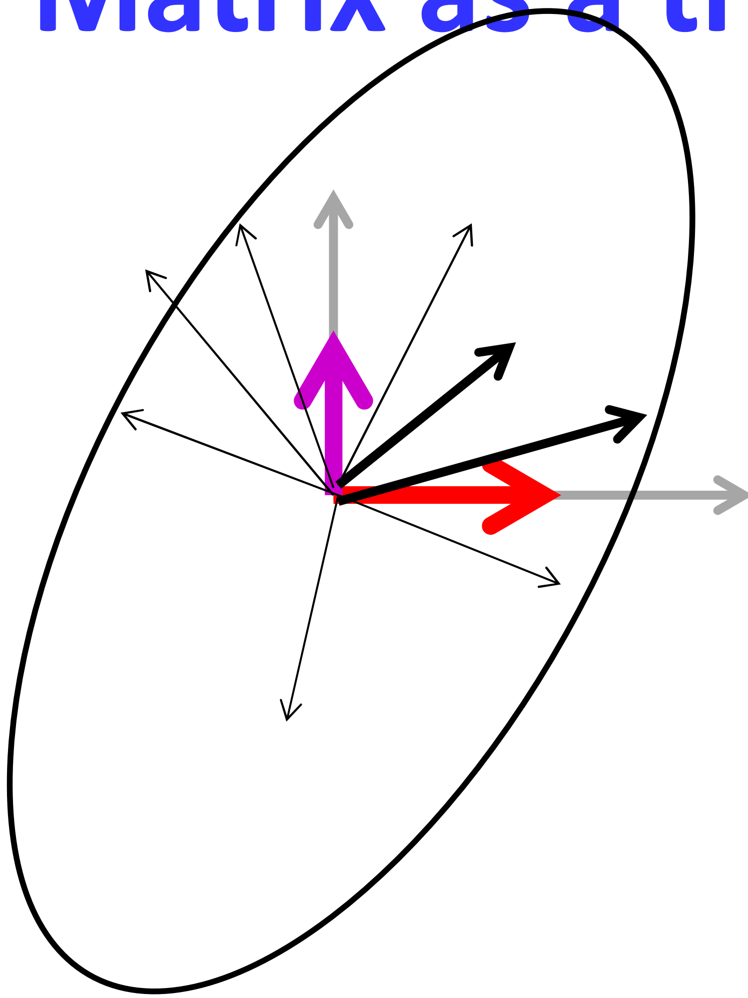
- More vectors..
- *Major axis of component ellipses proportional to the squared length of the corresponding vector*

Covariance Matrix as a transform



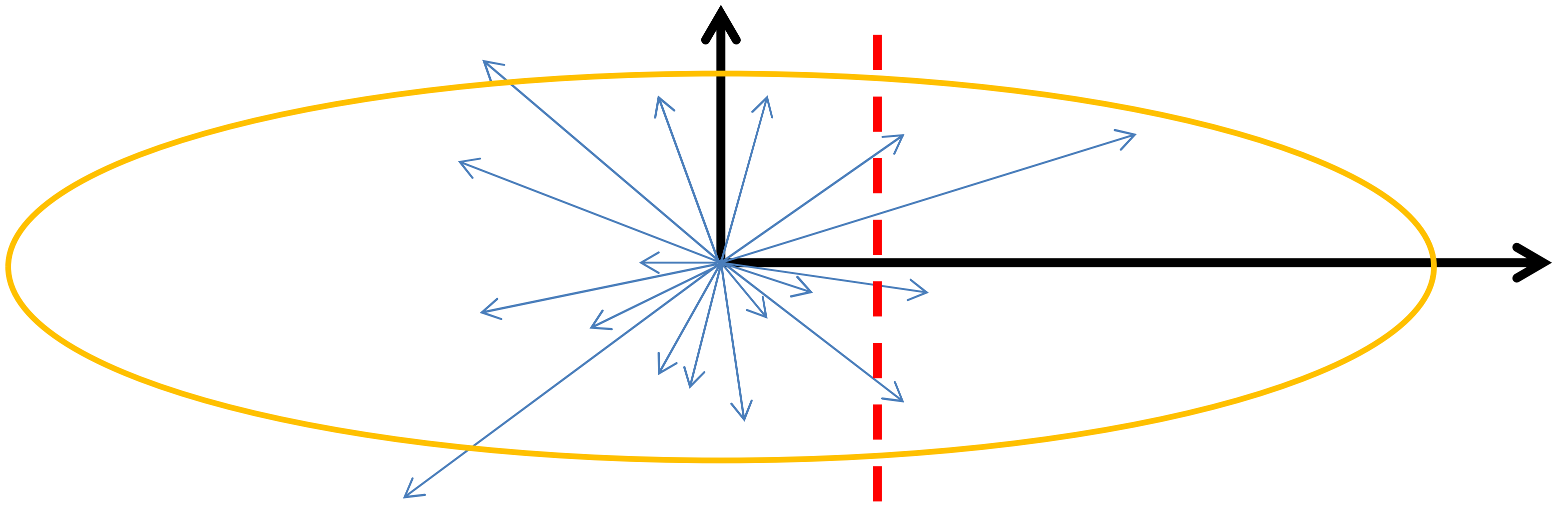
- And still more vectors..
- *Major axis of component ellipses proportional to the squared length of the corresponding vector*

Covariance Matrix as a transform



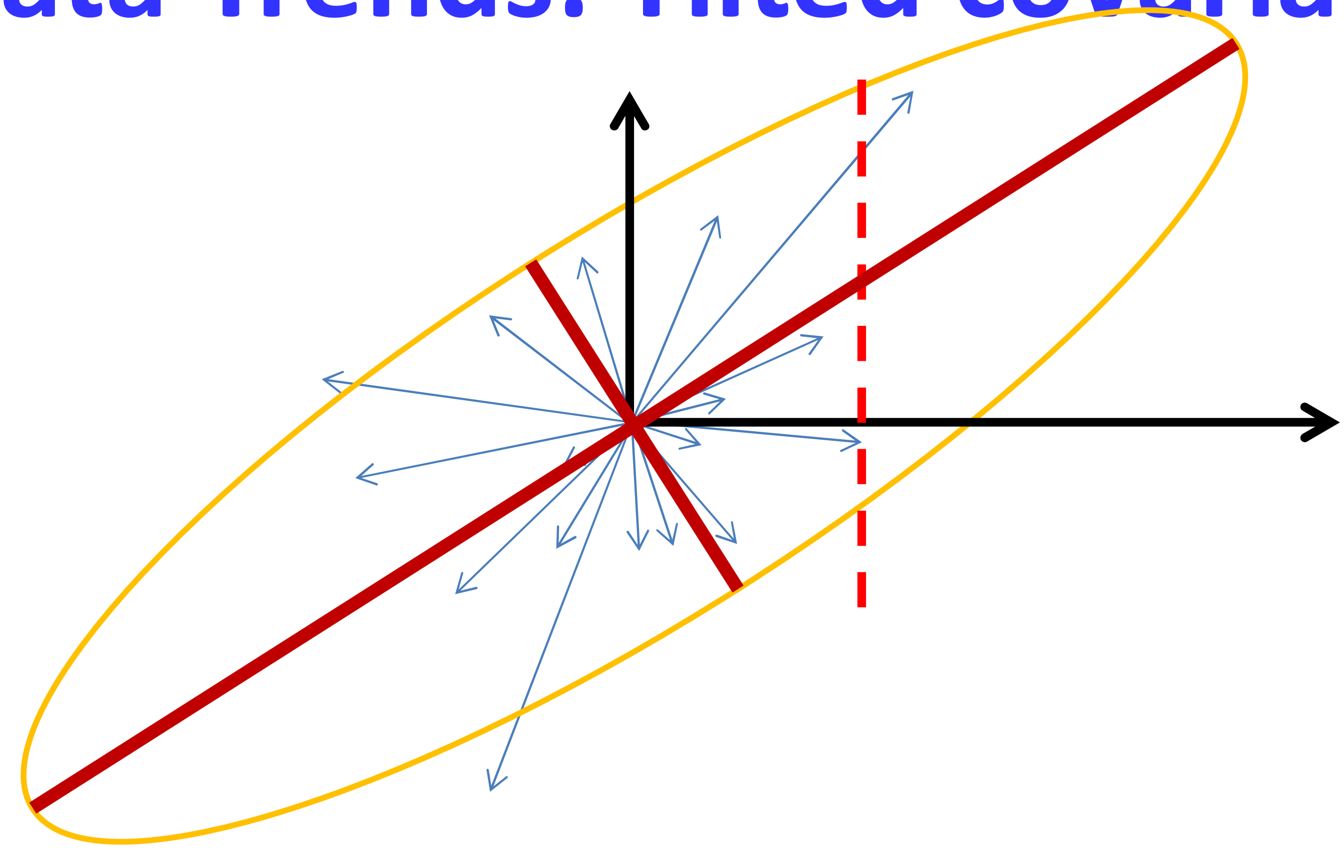
- The covariance matrix captures the directions of maximum variance
- What does it tell us about trends?

Data Trends: Axis aligned covariance



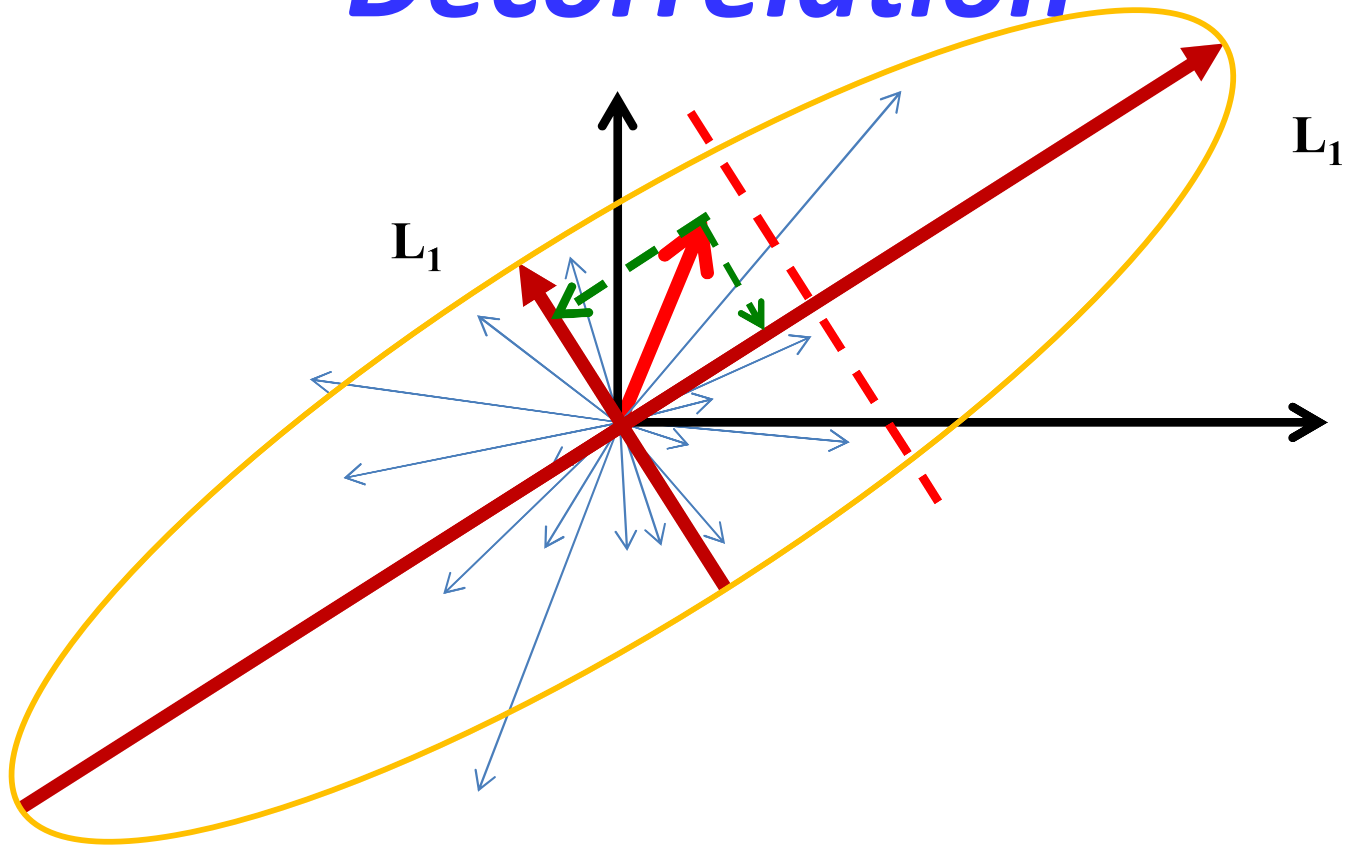
- Axis aligned covariance
- At any X value, the average Y value of vectors is 0
  - X cannot predict Y
- At any Y, the average X of vectors is 0
  - Y cannot predict X
- The X and Y components are ***uncorrelated***

Data Trends: Tilted covariance



- Tilted covariance
- The average Y value of vectors at any X varies with X
  - X predicts Y
- Average X varies with Y
- The X and Y components are ***correlated***

Decorrelation



- Shifting to using the major axes as the coordinate system
  - L_1 does not predict L_2 and vice versa
  - In this coordinate system the data are uncorrelated
- We have ***decorrelated*** the data by rotating the axes

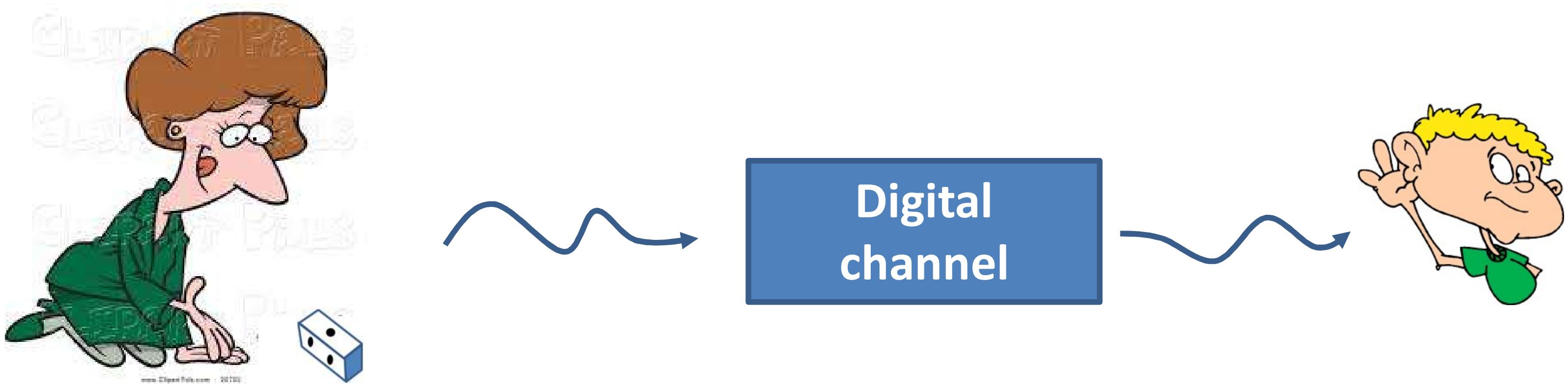
A note on bits..

- You flip a coin. You must inform your friend in the next room about whether the outcome was heads or tails

- How many bits will you have to send?

A note on bits..

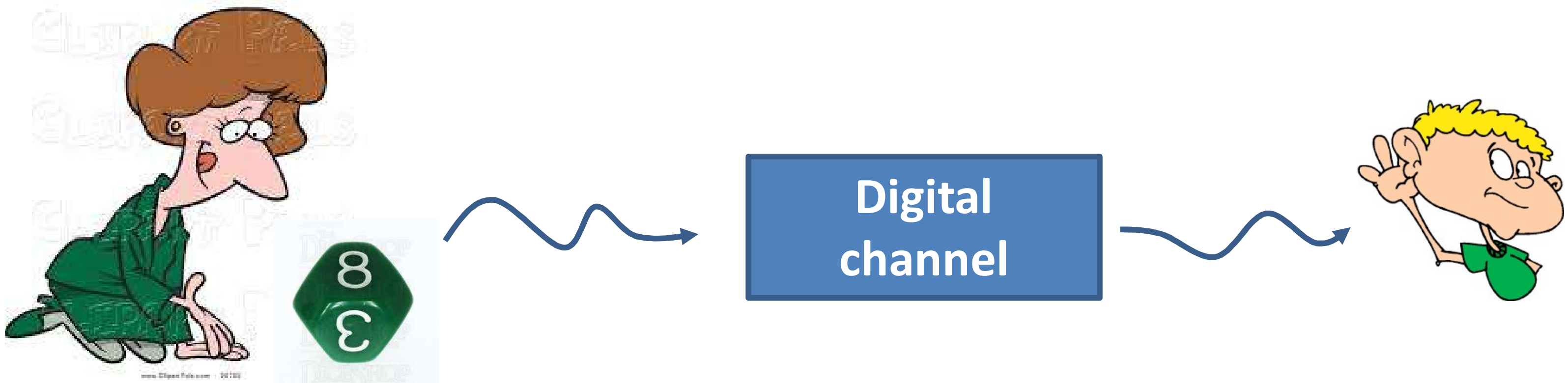
- You roll a four-side dice. You must inform your friend in the next room about the outcome



- How many bits will you have to send?

A note on bits..

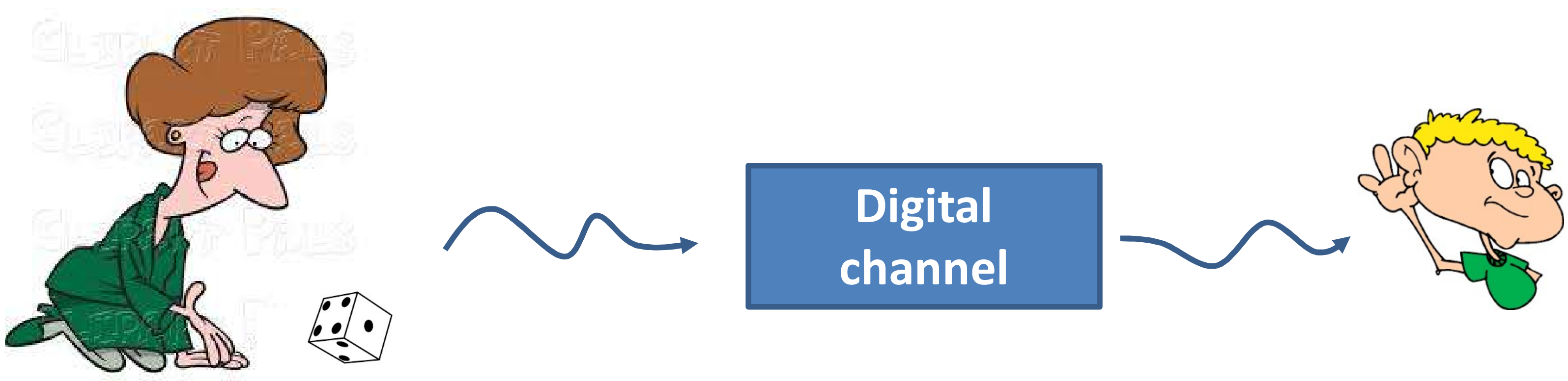
- You roll an *eight-sided octahedral* dice. You must inform your friend in the next room about the outcome



- How many bits will you have to send?

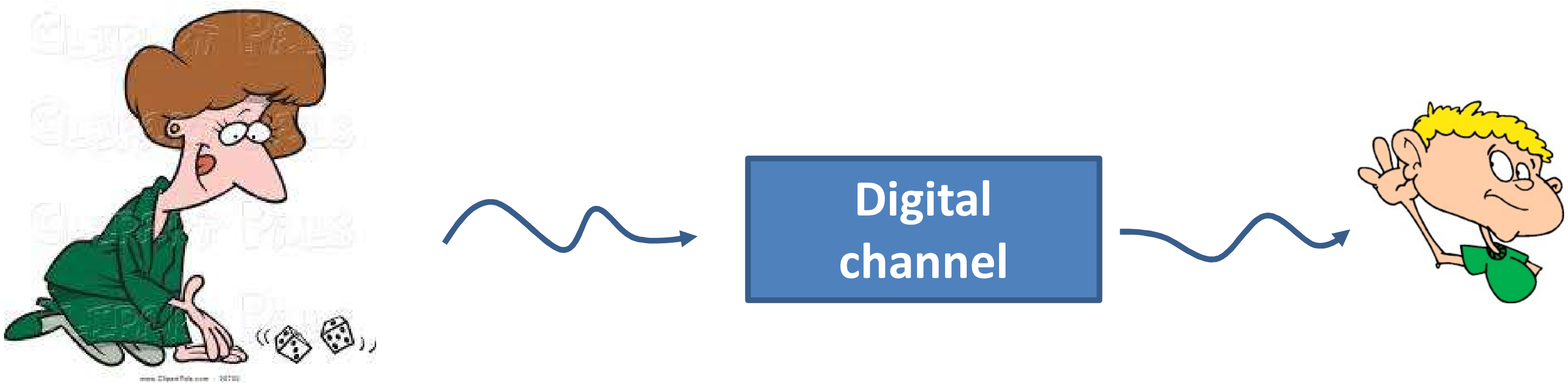
A note on bits..

- You roll a *six-sided* dice. You must inform your friend in the next room about the outcome



- How many bits will you have to send?

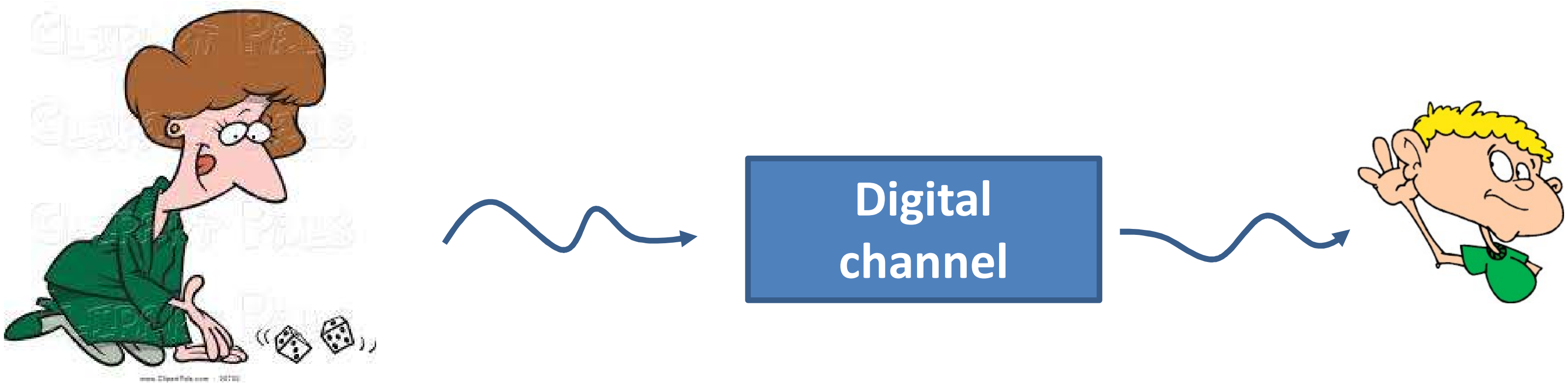
Batching up 6-sided dice rolls



- Instead of sending individual rolls, you roll the dice *twice*
  - And send the *pair* to your friend
- How many bits do you send *per roll?*

Roll 1	Roll 2
1	1
1	2
1	3
..	..
2	1
2	2
..	..
6	6

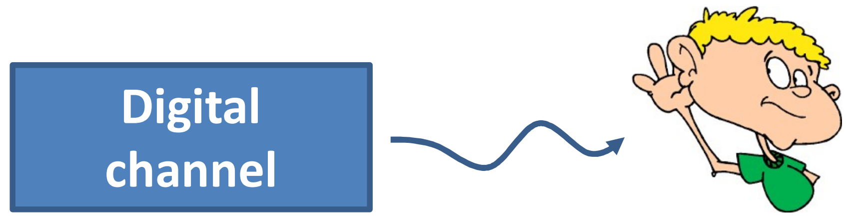
Batching up 6-sided dice rolls



- Instead of sending individual rolls, you roll the dice *twice*
  - And send the *pair* to your friend
- How many bits do you send *per roll?*
- 36 combinations: 6 bits per pair of numbers
  - Still 3 bits per roll

Roll 1	Roll 2
1	1
1	2
1	3
..	..
2	1
2	2
..	..
6	6

Batching up 6-sided dice rolls

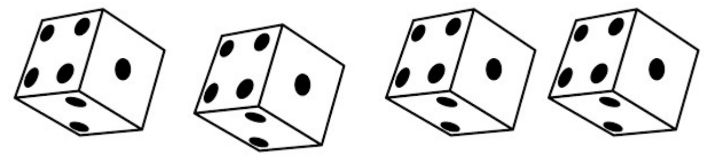


- Instead of sending individual rolls, you roll the dice ***three times***
  - And send the *triple* to your friend
- How many bits do you send *per roll?*
- 216 combinations: 8 bits per triple
  - Still 2.666 bits per roll
  - *Now we're talking!*

Roll 1	Roll 2	Roll 3
1	1	1
1	1	2
..	..	..
1	6	3
..		..
2	1	1
2	1	2
..		..
6	6	6

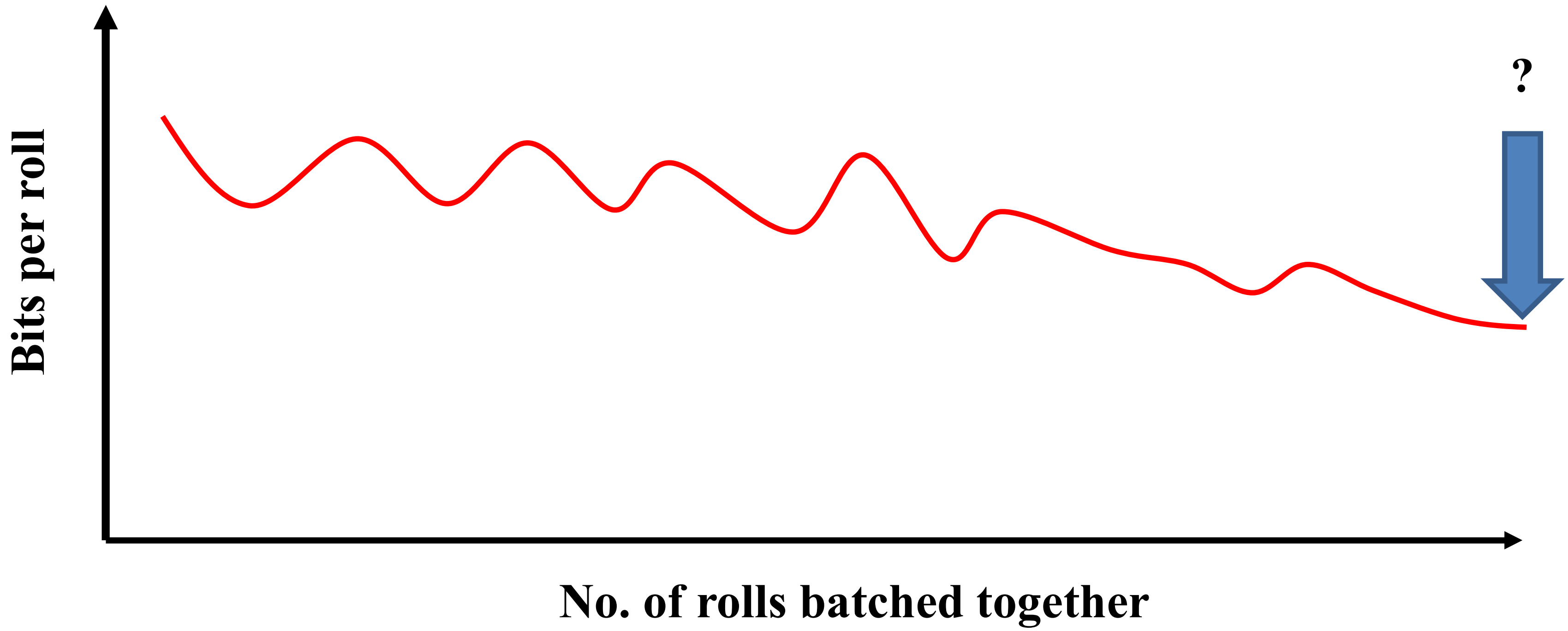
Batching up 6-sided dice rolls

- Batching ***four rolls***
  - 1296 combinations
  - 11 bits per outcome (4 rolls)
  - 2.75 bit per roll



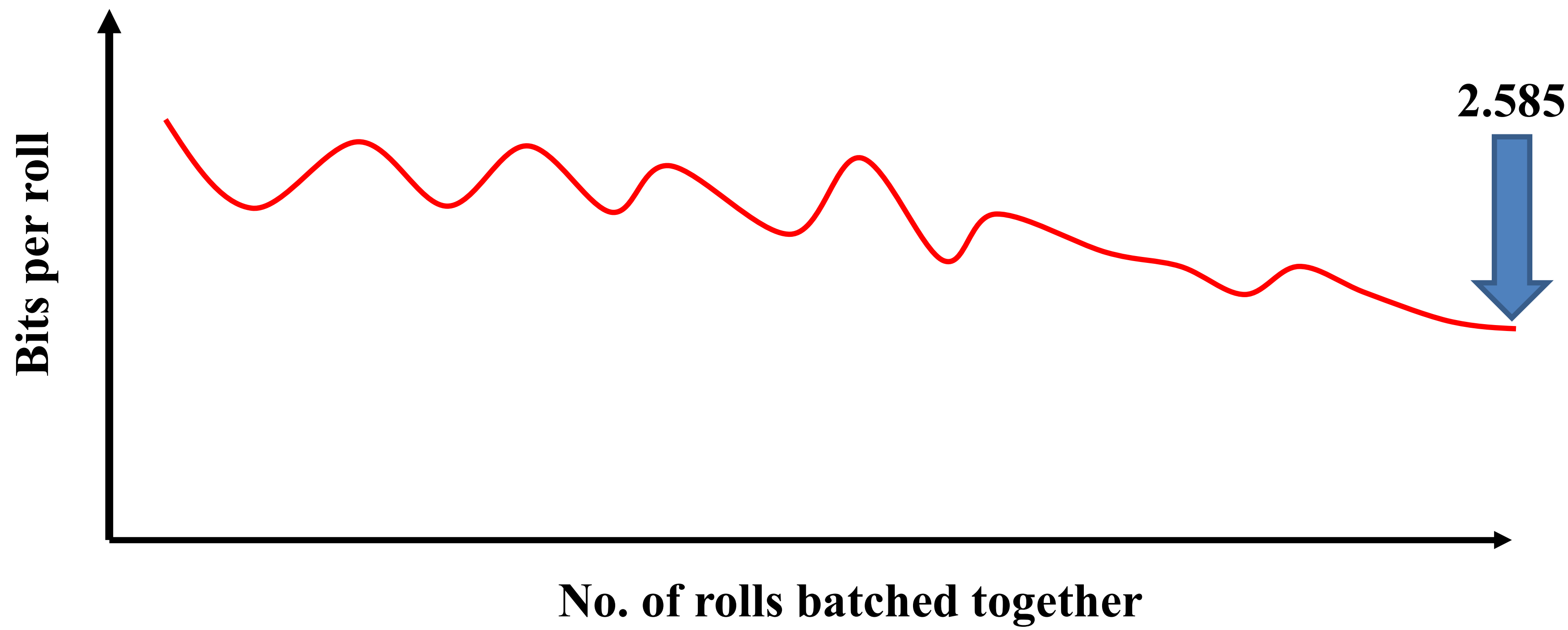
- Batching ***five rolls***
  - 7776 combinations
  - 13 bits per outcome (5 rolls)
  - 2.6 bits per roll

Batching up 6-sided dice rolls



- Where will it end?

Batching up 6-sided dice rolls



- Where will it end?
- $\lim_{k \to \infty} \frac{\lceil k \log 2(6) \rceil}{k} = \log 2(6)$ bits per roll in the limit
  - This is the absolute minimum – no simple batching will give you less than these many bits per outcome with this scheme

Can we do better?

- A four-sided die needs 2 bits per roll
- But then you find not all sides are equally likely

- P(1) = 0.5, P(2) = 0.25, P(3) 0.125, P(4) = 0.125
- *Can you do better than 2 bits per outcome*

Can we do better?

- You have

P(1) = 0.5, P(2) = 0.25, P(3) 0.125, P(4) = 0.125

- You use:

1	0
2	1 0
3	1 1 0
4	1 1 1

  - Note receiver is *never in any doubt as to what they received*
- What is the average number of bits per outcome

Can we do better?

- You have

P(1) = 0.5, P(2) = 0.25, P(3) 0.125, P(4) = 0.125

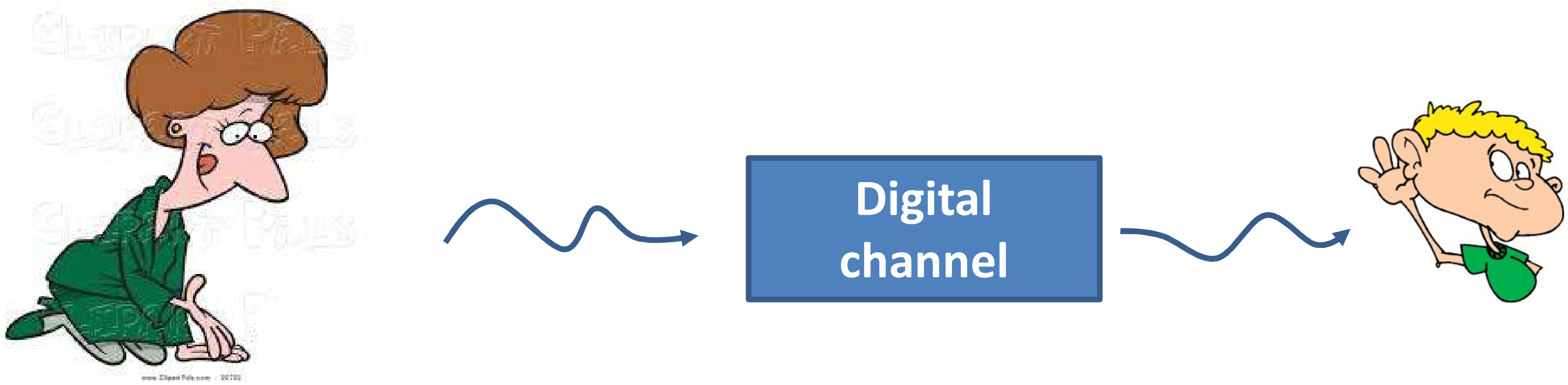
- You use:

1	0
2	1 0
3	1 1 0
4	1 1 1

  – Note receiver is *never in any doubt as to what they received*

- An outcome with probability p is equivalent to obtaining one of $1/p$ equally likely choices
  – Requires $\log 2\left(\frac{1}{p}\right)$ bits on average

Entropy

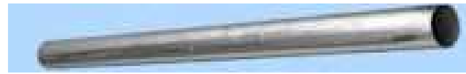


- The average number of bits per symbol required to communicate a random variable over a digitial channel *using an optimal code* is

$$H(p) = \sum_i p_i \log \frac{1}{p_i} = -\sum_i p_i \log p_i$$

- You can’t do better
  - Any other code will require more bits
- This is the *entropy of the random variable*

A brief review of basic info. theory



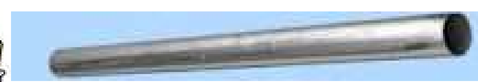
T(all), M(ed), S(hort)…

$$H(X)=\sum_{X}P(X)[-\log P(X)]$$

- Entropy: The *minimum average* number of bits to transmit to convey a symbol

X

T, M, S…



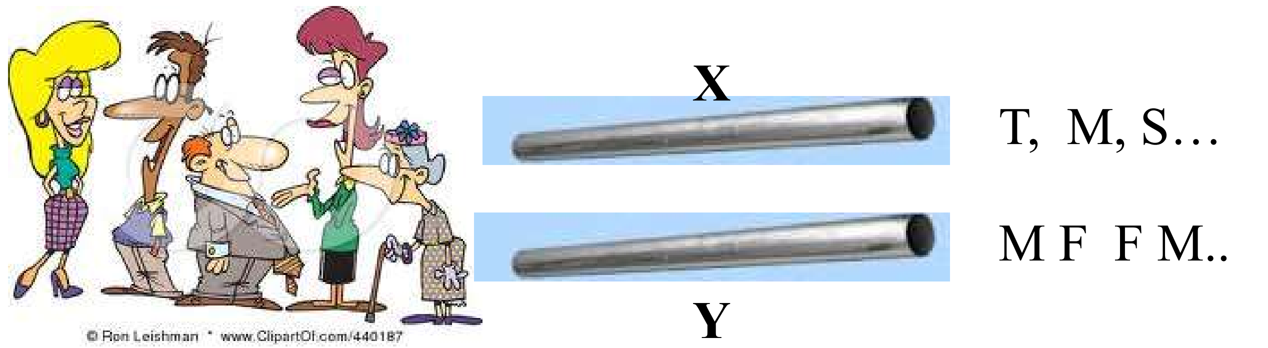
M F F M..

Y

$$H(X,Y)=\sum_{X,Y}P(X,Y)[-\log P(X,Y)]$$

- Joint entropy: The *minimum average* number of bits to convey sets (pairs here) of symbols

A brief review of basic info. theory



$$H(X \mid Y) = \sum_{Y} P(Y) \sum_{X} P(X \mid Y)[-\log P(X \mid Y)] = \sum_{X,Y} P(X,Y)[-\log P(X \mid Y)]$$

- Conditional Entropy: The *minimum average* number of bits to transmit to convey a symbol X, after symbol Y has already been conveyed
  - Averaged over all values of X and Y


And now
for something
completely different...

The statistical concept of correlatedness

- Two variables X and Y are correlated if If knowing X gives you an *expected* value of Y

- X and Y are uncorrelated if knowing X tells you nothing about the *expected* value of Y
  – Although it could give you other information
  – How?

Correlation vs. Causation

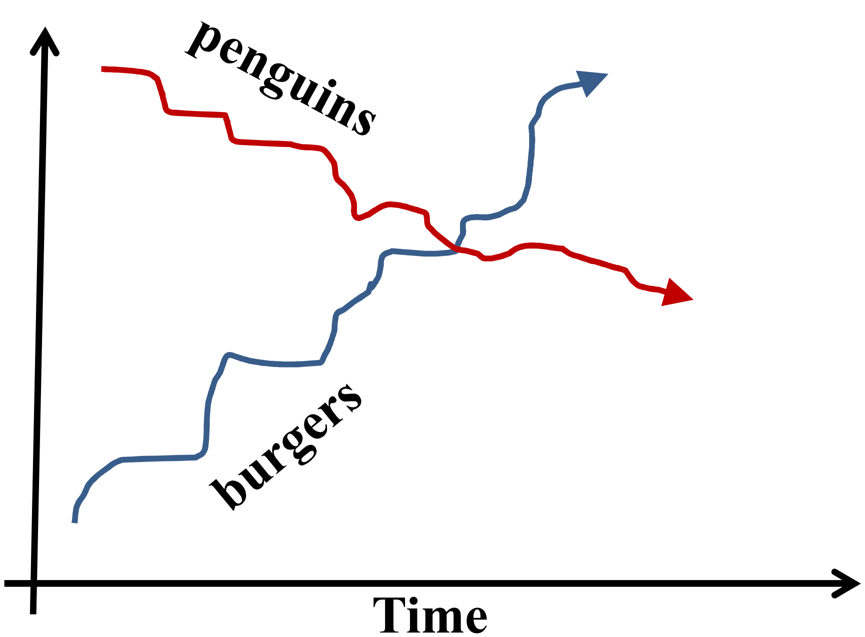
- The consumption of burgers has gone up steadily in the past decade

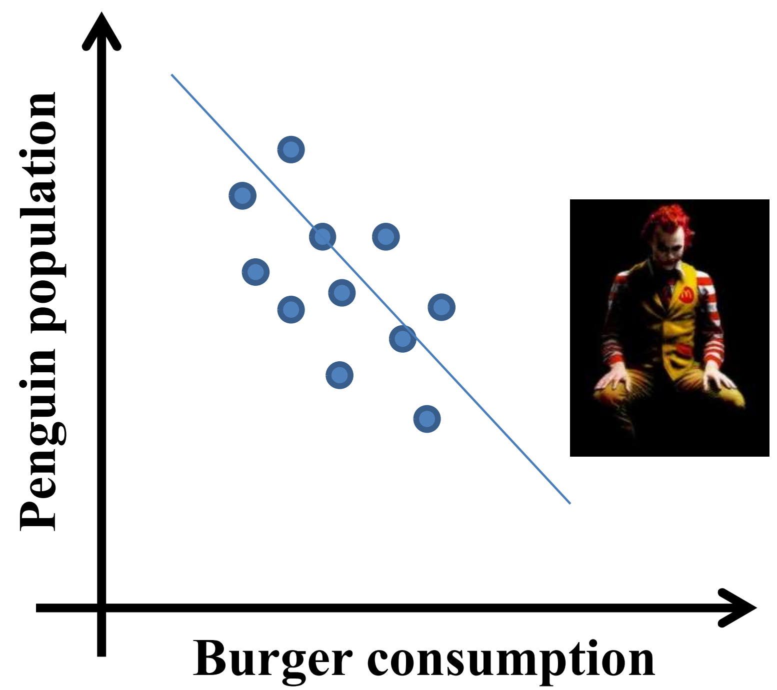
- In the same period, the penguin population of Antarctica has gone down

Correlation, not Causation
(unless McDonalds has a
top-secret Antarctica division)

The concept of *correlation*

- Two variables are correlated if knowing the value of one gives you information about the ***expected value*** of the other





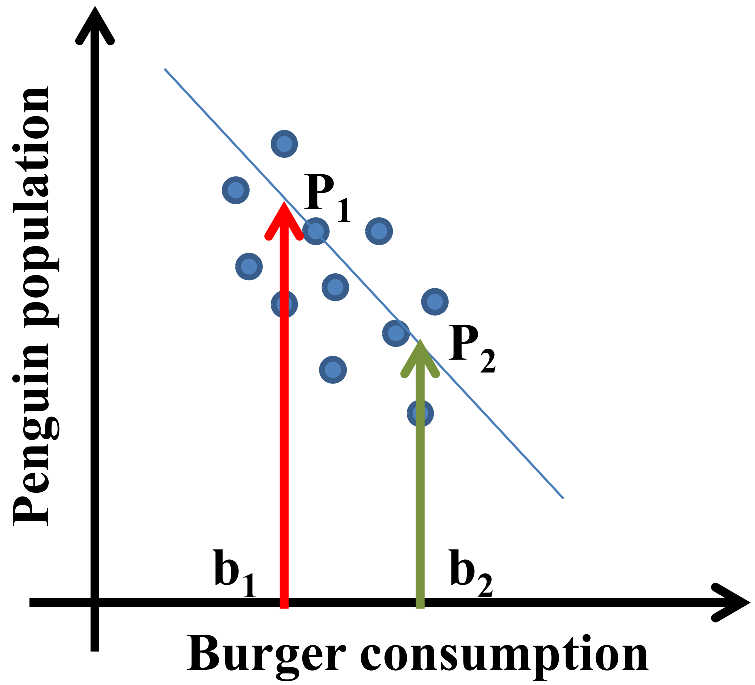
A brief review of basic probability

- *Uncorrelated:* Two random variables X and Y are uncorrelated iff:
  - The *average* value of the product of the variables equals the product of their individual averages

- Setup: Each draw produces one instance of X and one instance of Y
  - I.e one instance of (X, Y)
- $E[XY] = E[X]E[Y]$

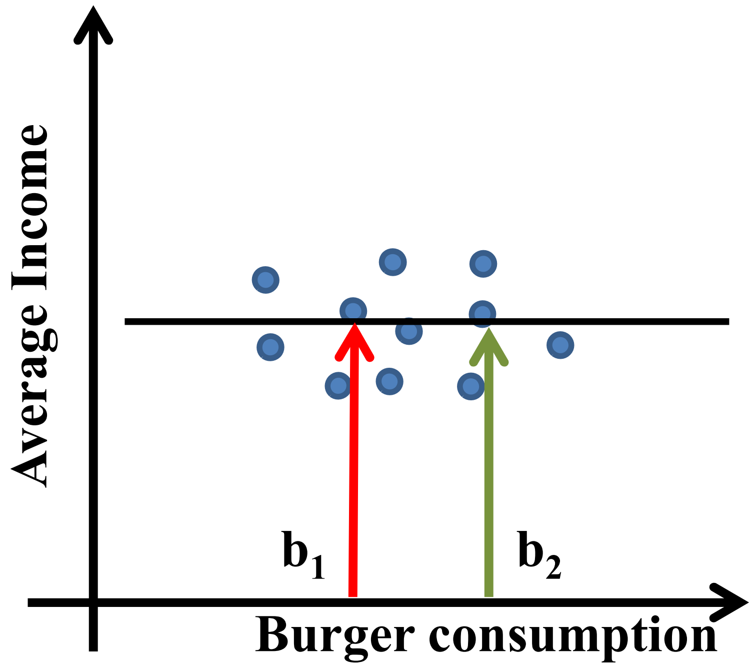
- The average value of Y is the same regardless of the value of X

Correlated Variables



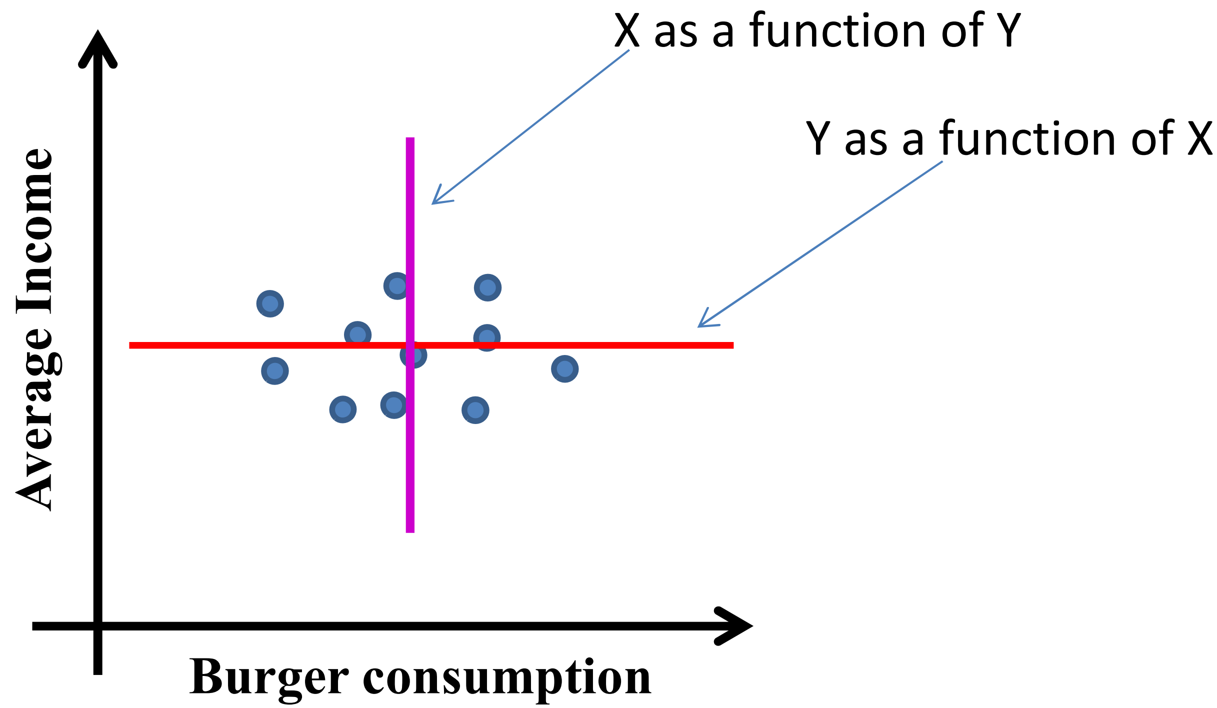
- Expected value of Y given X:
  - Find average of Y values of all samples at (or close) to the given X
  - If this is a function of X, X and Y are correlated

Uncorrelatedness



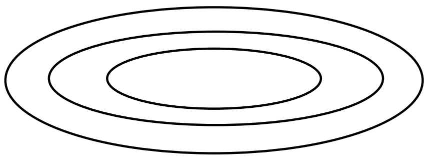
- Knowing X does not tell you what the *average* value of Y is
  - And vice versa

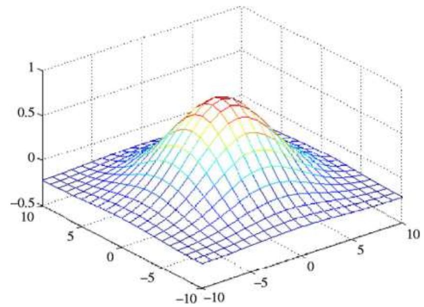
Uncorrelated Variables

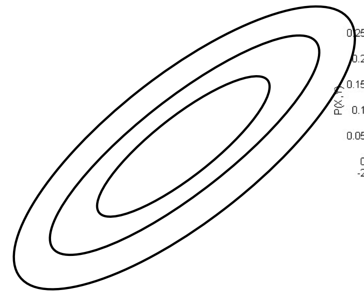


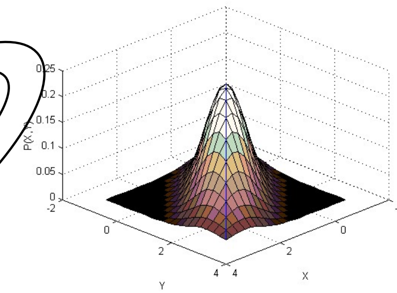
- The average value of Y is the same regardless of the value of X and vice versa

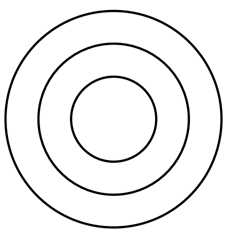
Uncorrelatedness in Random Variables

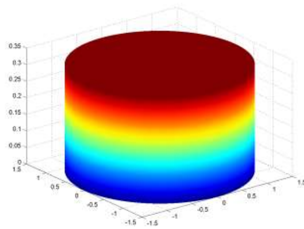


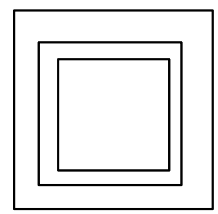


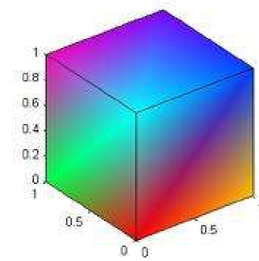












- Which of the above represent uncorrelated RVs?

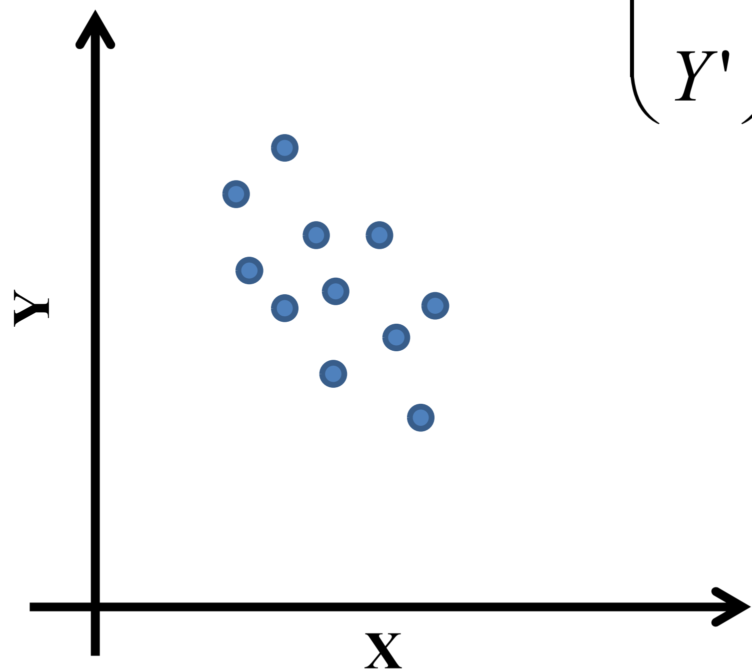
Benefits of uncorrelatedness..

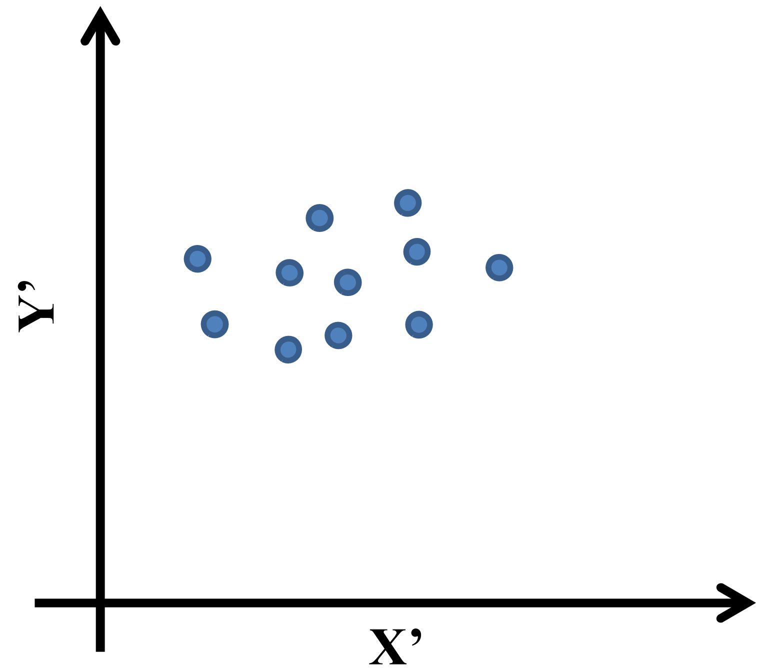
- Uncorrelatedness of variables is generally considered desirable for modelling and analyses
  - For Euclidean error based regression models and probabilistic models, uncorrelated variables can be separately handled
    - Since the value of one doesn't affect the average value of others
    - Greatly reduces the number of model parameters
  - Otherwise their interactions must be considered

- We will frequently transform correlated variables to make them uncorrelated
  - "Decorrelating" variables

The notion of *decorrelation*

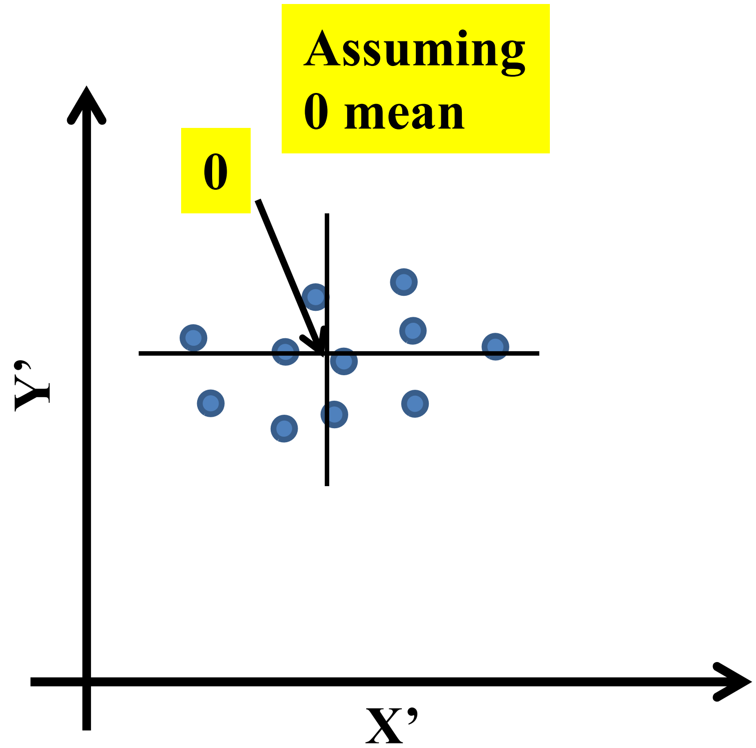
$$\begin{pmatrix} X' \\ Y' \end{pmatrix} = M \begin{pmatrix} X \\ Y \end{pmatrix}$$





- So how does one transform the correlated variables (X, Y) to the uncorrelated (X', Y')

What does “uncorrelated” mean



- $E[X'] = constant$
- $E[Y'] = constant$
- $E[Y'|X'] = constant$
- $E[X'Y'] = E[X']E[Y']$
- All will be 0 for centered data

$$E\left[\begin{pmatrix} X' \\ Y' \end{pmatrix}\begin{pmatrix} X' & Y' \end{pmatrix}\right] = E\begin{pmatrix} X'^2 & X'Y' \\ X'Y' & Y'^2 \end{pmatrix} = \begin{pmatrix} E[X'^2] & 0 \\ 0 & E[Y'^2] \end{pmatrix} = diagonal \quad matrix$$

- If $\mathbf{Y}$ is a matrix of vectors, $\mathbf{YY}^T$ = diagonal

Decorrelation

- Let $\mathbf{X}$ be the matrix of correlated data vectors
  - Each component of $\mathbf{X}$ informs us of the mean trend of other components

- Need a transform $\mathbf{M}$ such that if $\mathbf{Y} = \mathbf{MX}$ such that the covariance of $\mathbf{Y}$ is diagonal
  - $\mathbf{YY}^T$ is the covariance if $\mathbf{Y}$ is zero mean
  - For uncorrelated components, $\mathbf{YY}^T = \mathbf{Diagonal}$

$\Rightarrow \mathbf{MXX}^T\mathbf{M}^T = \mathbf{Diagonal}$

$\Rightarrow \mathbf{M}.\mathrm{Cov}(\mathbf{X}).\mathbf{M}^T = \mathbf{Diagonal}$

Decorrelation

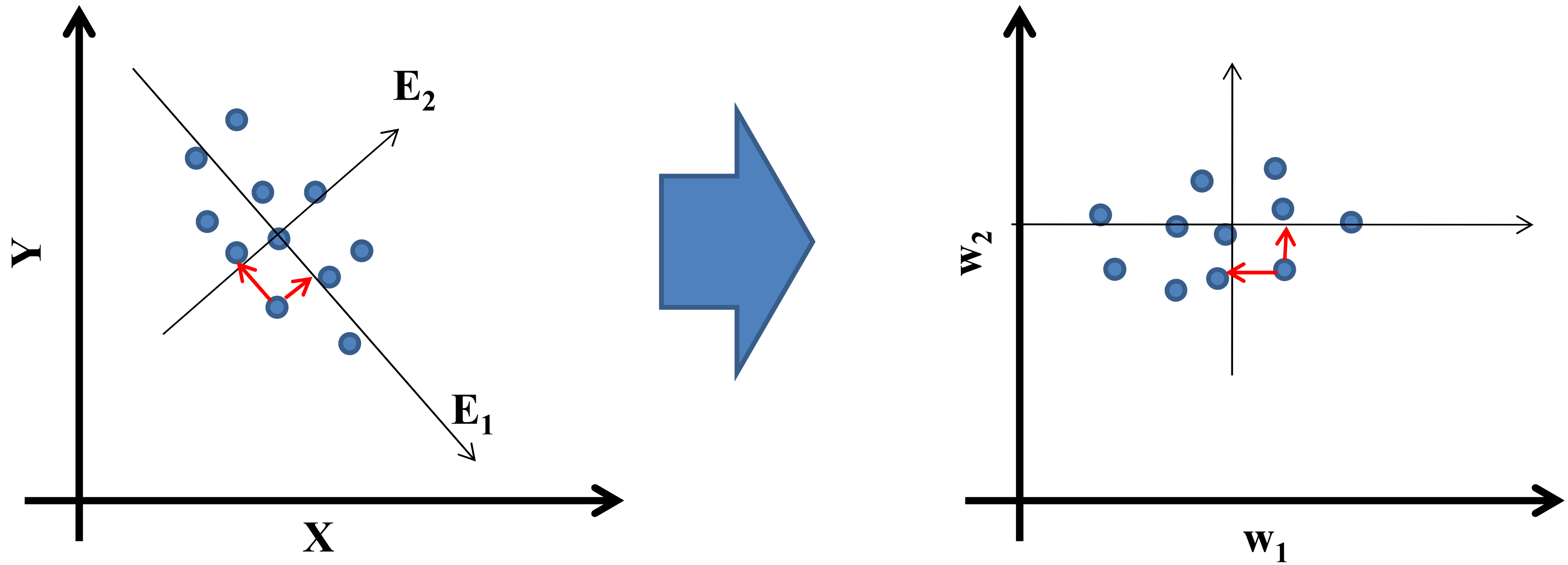
- Easy solution:
  - Eigen decomposition of $\mathrm{Cov}(\mathbf{X})$:

$$\mathrm{Cov}(\mathbf{X}) = \mathbf{E}\Lambda\mathbf{E}^{\mathrm{T}}$$

  - $\mathbf{E}\mathbf{E}^{\mathrm{T}} = \mathrm{I}$
- Let $\mathbf{M} = \mathbf{E}^{\mathrm{T}}$
- $\mathbf{M}\mathrm{Cov}(\mathbf{X})\mathbf{M}^{\mathrm{T}} = \mathbf{E}^{\mathrm{T}}\mathbf{E}\Lambda\mathbf{E}^{\mathrm{T}}\mathbf{E} = \Lambda = \text{diagonal}$
- PCA: $\mathbf{Y} = \mathbf{E}^{\mathrm{T}}\mathbf{X}$
  - Projects the data onto the Eigen vectors of the covariance matrix
  - *Diagonalizes* the covariance matrix
  - "Decorrelates" the data

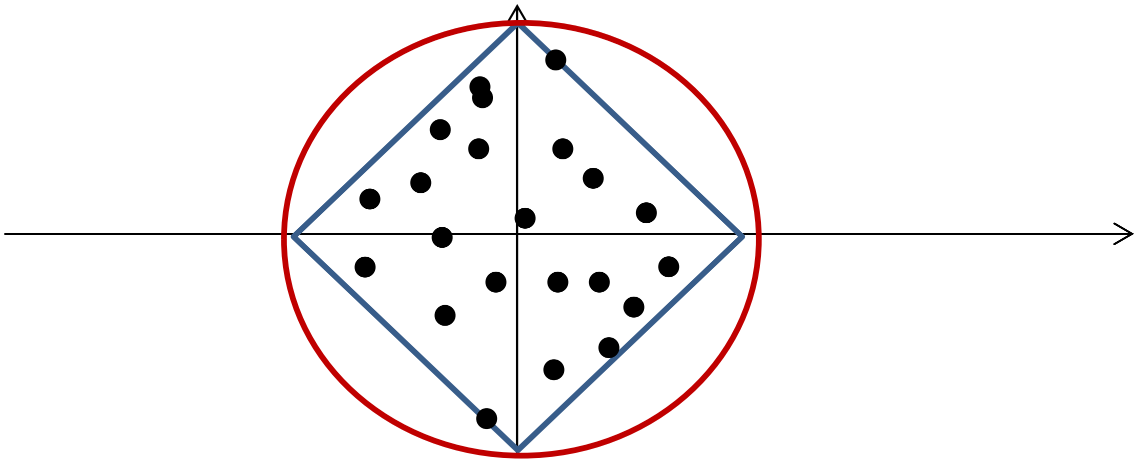
PCA

$$\mathbf{X} = w_1\mathbf{E}_1 + w_2\mathbf{E}_2$$



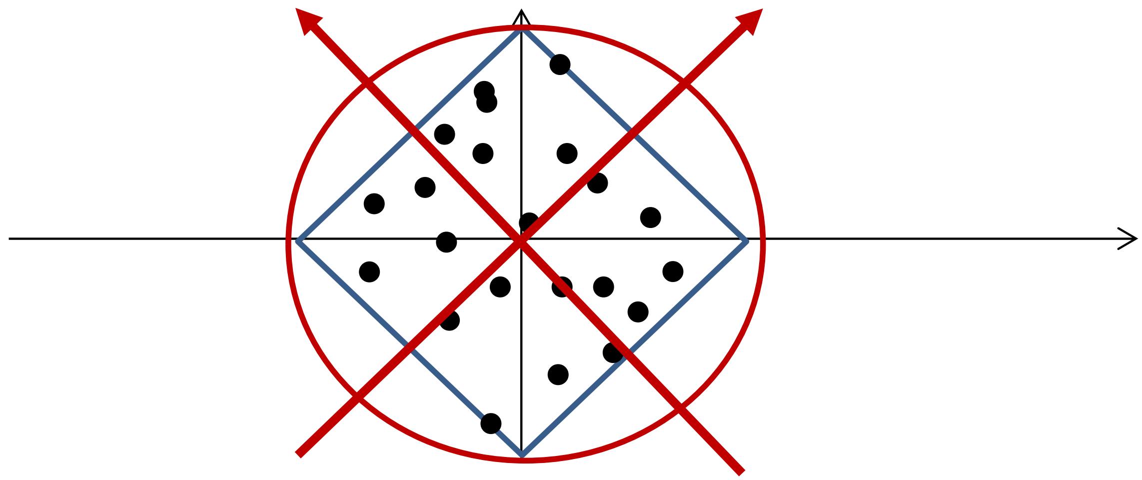
- PCA: $\mathbf{Y} = \mathbf{E}^{\mathrm{T}}\mathbf{X}$
  - Projects the data onto the Eigen vectors of the covariance matrix
    - Changes the coordinate system to the Eigen vectors of the covariance matrix
  - *Diagonalizes* the covariance matrix
  - “Decorrelates” the data

Decorrelating the data



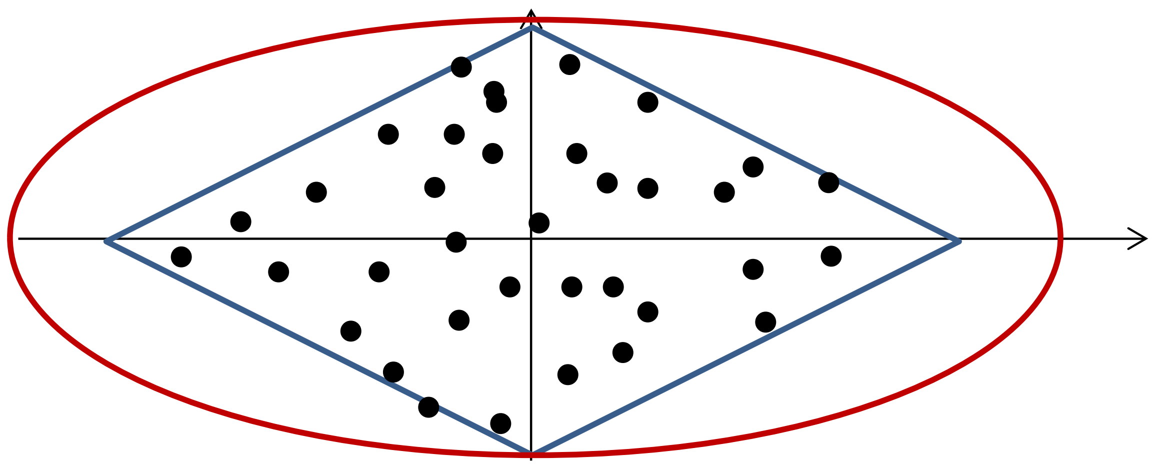
- Are there other decorrelating axes?

Decorrelating the data



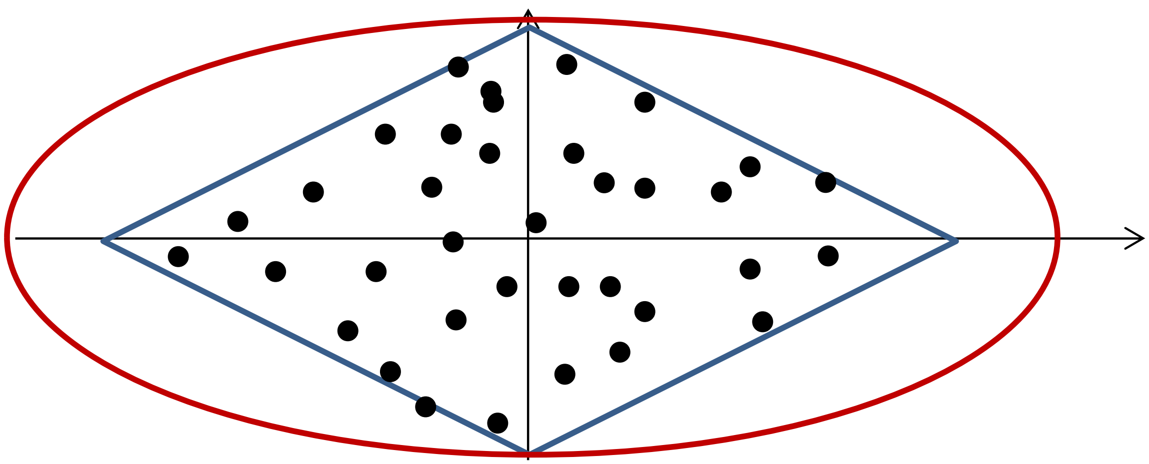
- Are there other decorrelating axes?

Decorrelating the data



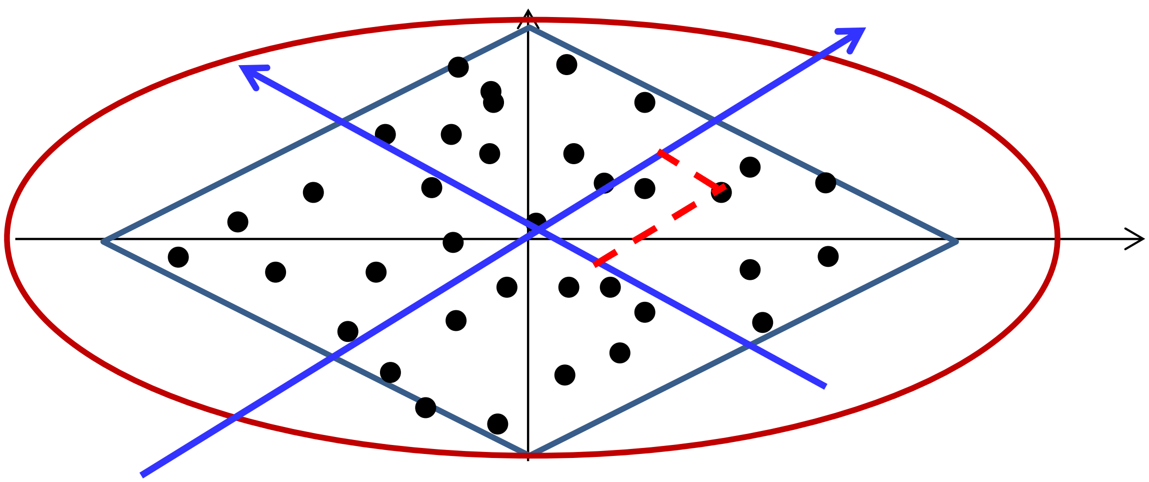
- Are there other decorrelating axes?

Decorrelating the data



- Are there other decorrelating axes?
- What about if we don't require them to be orthogonal?

Decorrelating the data



- Are there other decorrelating axes?
- What about if we don't require them to be orthogonal?
- What is special about these axes?

The statistical concept of *Independence*

- Two variables X and Y are *dependent* if If knowing X gives you *any information about* Y

- X and Y are *independent* if knowing X tells you nothing at all of Y

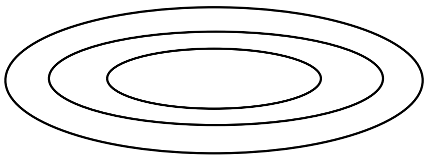
A brief review of basic probability

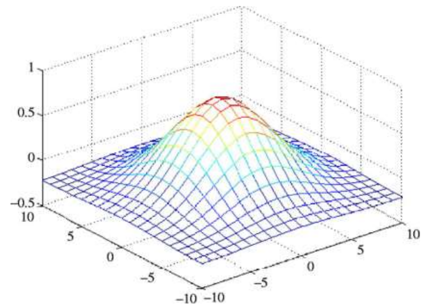
- ***Independence:*** Two random variables X and Y are independent iff:
  - Their joint probability equals the product of their individual probabilities
- $P(X,Y) = P(X)P(Y)$
- Independence implies uncorrelatedness
  - The average value of X is the same regardless of the value of Y
    - $E[X|Y] = E[X]$
  - But uncorrelatedness does not imply independence

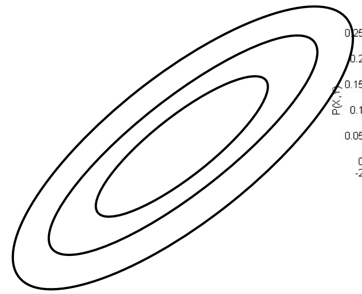
A brief review of basic probability

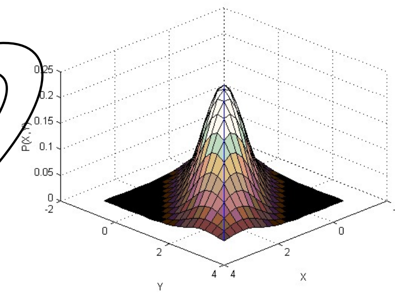
- *Independence:* Two random variables X and Y are independent iff:
- The average value of ***any function*** *of* X is the same regardless of the value of Y
  - Or any function of Y
- **E[f(X)g(Y)] = E[f(X)] E[g(Y)] for all f(), g()**

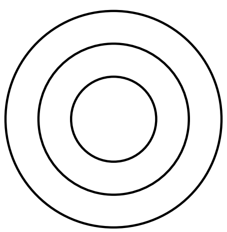
Independence

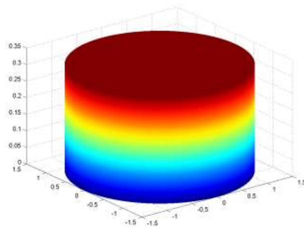


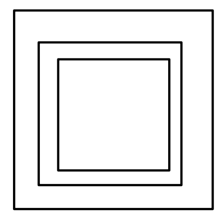


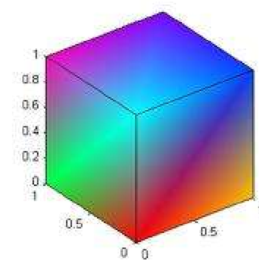






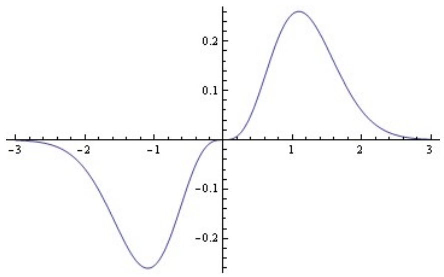


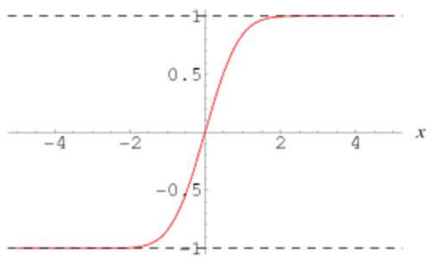




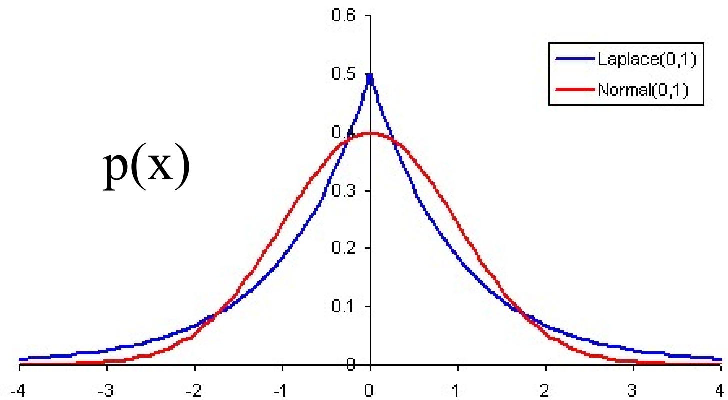
- Which of the above represent independent RVs?
- Which represent uncorrelated RVs?

A brief review of basic probability





y = f(x)



- The expected value of an odd function of an RV is 0 if
  - The RV is 0 mean
  - The PDF is of the RV is symmetric around 0
- **E[f(X)] = 0 if f(X) is odd symmetric**

A brief review of basic info. theory

- Conditional entropy of $X|Y = H(X)$ if X is independent of Y

$$H(X|Y) = \sum_Y P(Y) \sum_X P(X|Y)[-\log P(X|Y)] = \sum_Y P(Y) \sum_X P(X)[-\log P(X)] = H(X)$$

- Joint entropy of X and Y is the sum of the entropies of X and Y if they are independent

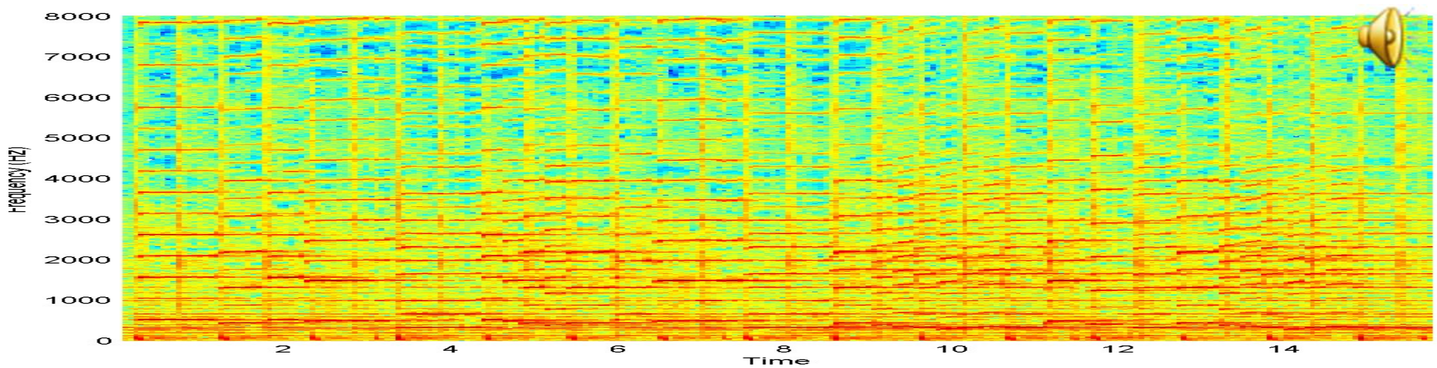
$$H(X,Y) = \sum_{X,Y} P(X,Y)[-\log P(X,Y)] = \sum_{X,Y} P(X,Y)[-\log P(X)P(Y)]$$

$$= -\sum_{X,Y} P(X,Y)\log P(X) - \sum_{X,Y} P(X,Y)\log P(Y) = H(X) + H(Y)$$

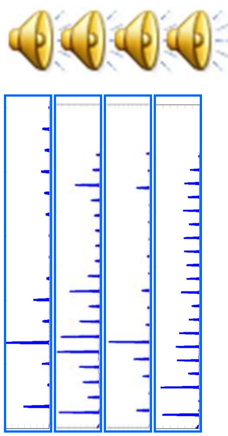
Onward..

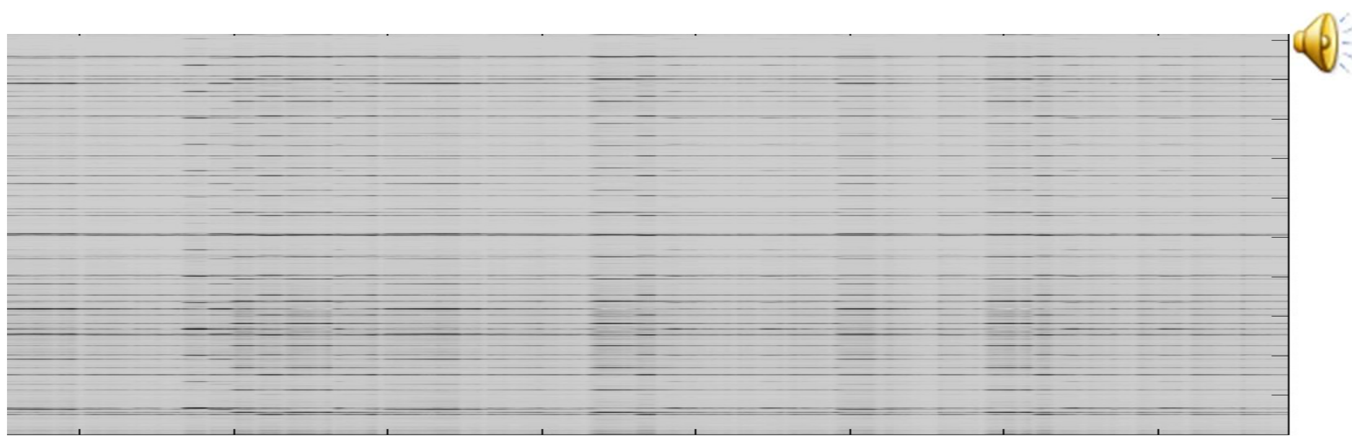
Projection: multiple notes

M =



W =

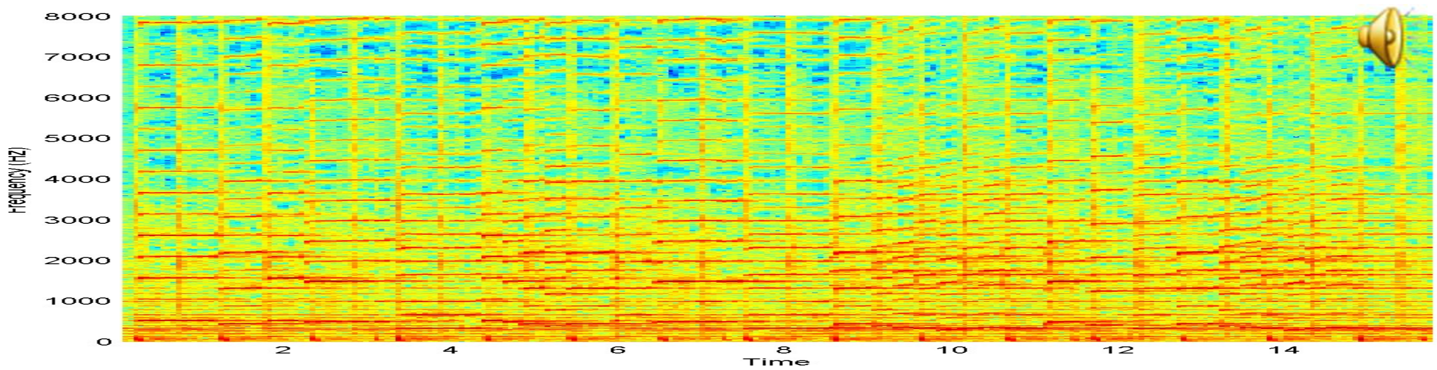




- $\mathbf{P} = \mathbf{W}\,(\mathbf{W}^T\mathbf{W})^{-1}\,\mathbf{W}^T$
- Projected Spectrogram = **PM**

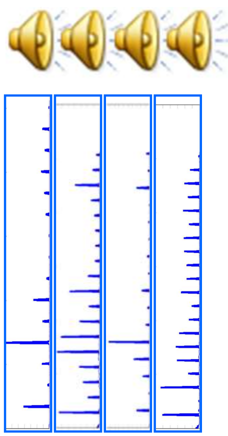
We’re actually computing a score

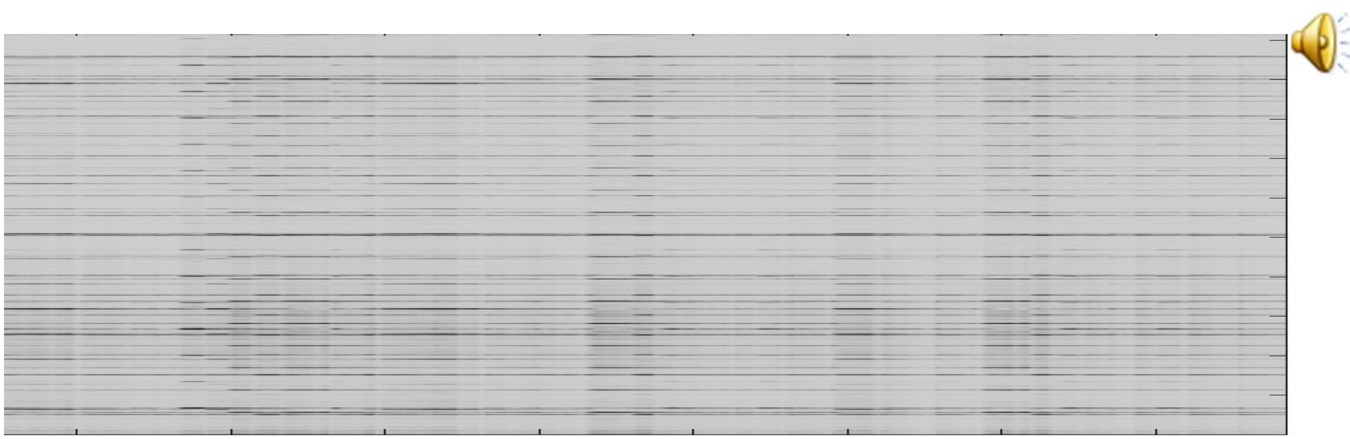
M =



H = ?

W =

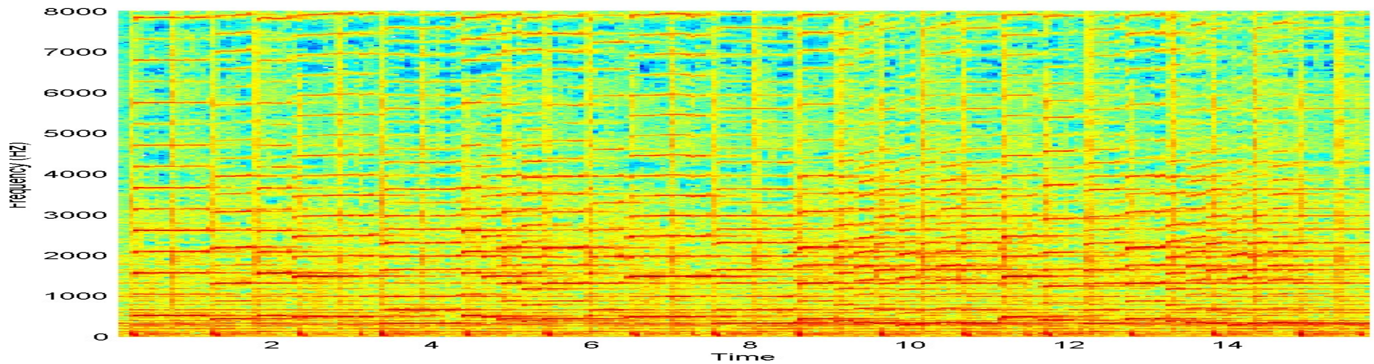




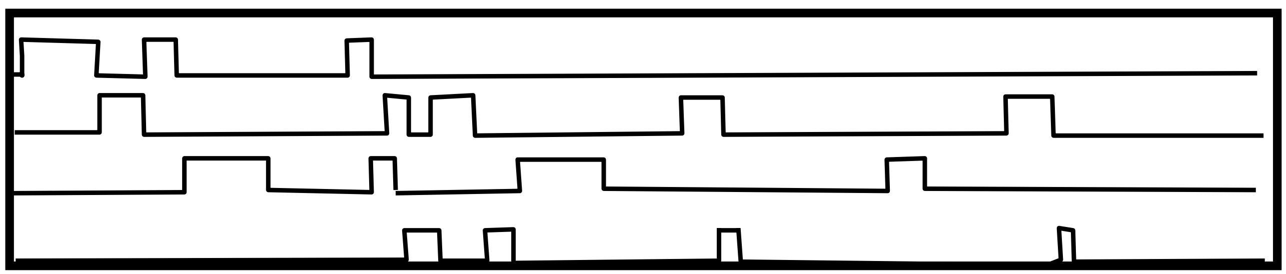
- $\mathbf{M} \sim \mathbf{W}\mathbf{H}$
- $\mathbf{H} = \text{pinv}(\mathbf{W})\mathbf{M}$

How about the other way?

M =



H =

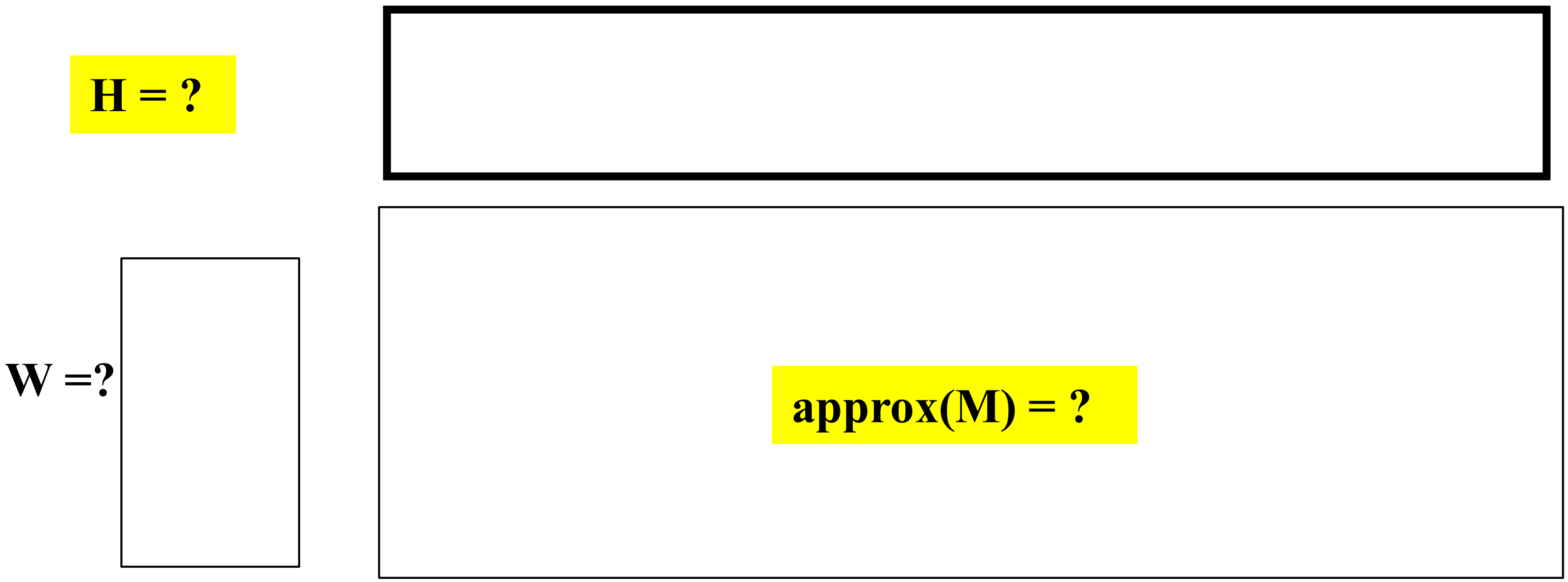


W = ?

U = ?

- M ~ WH   W = Mpinv(H)   U = WH

When both parameters are unknown



- Must estimate both $\mathbf{H}$ and $\mathbf{W}$ to best approximate $\mathbf{M}$
- Ideally, must learn *both* the *notes* and *their* transcription!

A least squares solution

$$\mathbf{W},\mathbf{H} = \arg\min_{\overline{\mathbf{W}},\overline{\mathbf{H}}} \| \mathbf{M} - \overline{\mathbf{W}}\overline{\mathbf{H}} \|_F^2 + \Lambda(\overline{\mathbf{W}}^T\overline{\mathbf{W}} - \mathbf{I})$$

- Constraint: $\mathbf{W}$ is orthogonal
  - $\mathbf{W}^T\mathbf{W} = \mathbf{I}$
- The solution: $\mathbf{W}$ are the Eigen vectors of $\mathbf{M}\mathbf{M}^T$
  - PCA!!

- $\mathbf{M} \sim \mathbf{W}\mathbf{H}$ is an approximation
- Also, the rows of $\mathbf{H}$ are *decorrelated*
  - Trivial to prove that $\mathbf{H}\mathbf{H}^T$ is diagonal

PCA

$$\mathbf{W}, \mathbf{H} = \arg\min_{\overline{\mathbf{W}}, \overline{\mathbf{H}}} \| \mathbf{M} - \overline{\mathbf{W}\mathbf{H}} \|_F^2$$

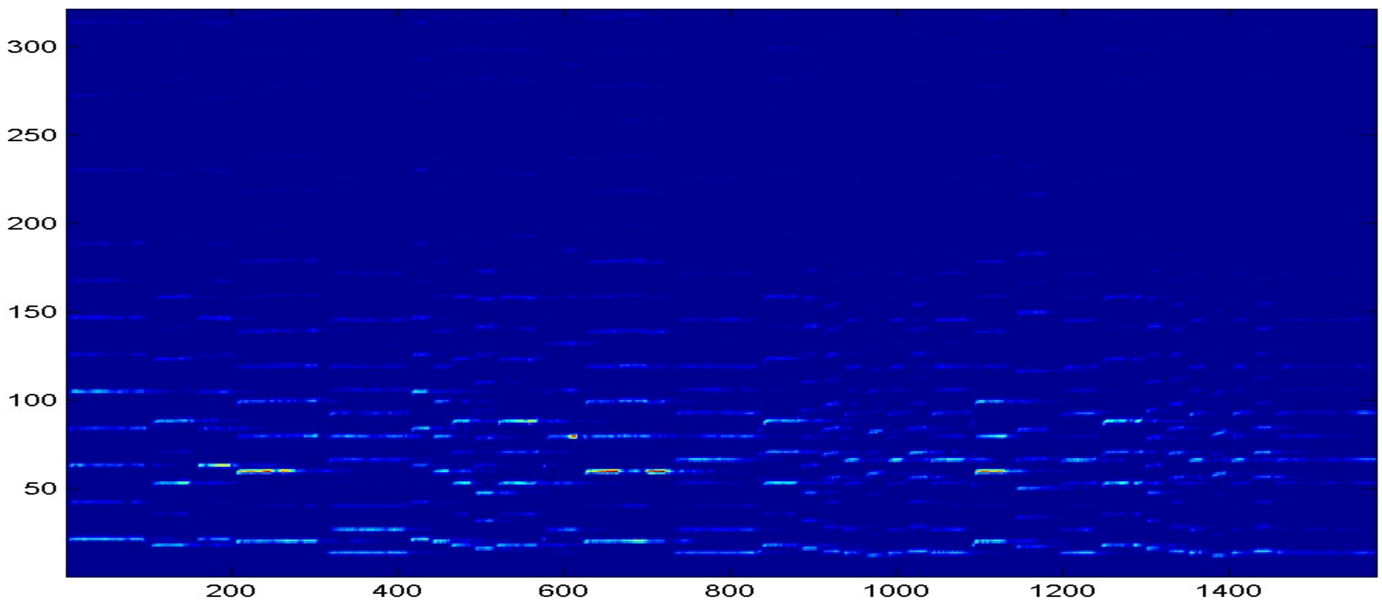
$$\mathbf{M} \approx \mathbf{W}\mathbf{H}$$

$$\mathbf{W}\mathbf{W}^{\mathbf{T}} = \text{Diagonal} \quad \text{OR} \quad \mathbf{H}\mathbf{H}^{\mathbf{T}} = \text{Diagonal}$$

The conditions are equivalent

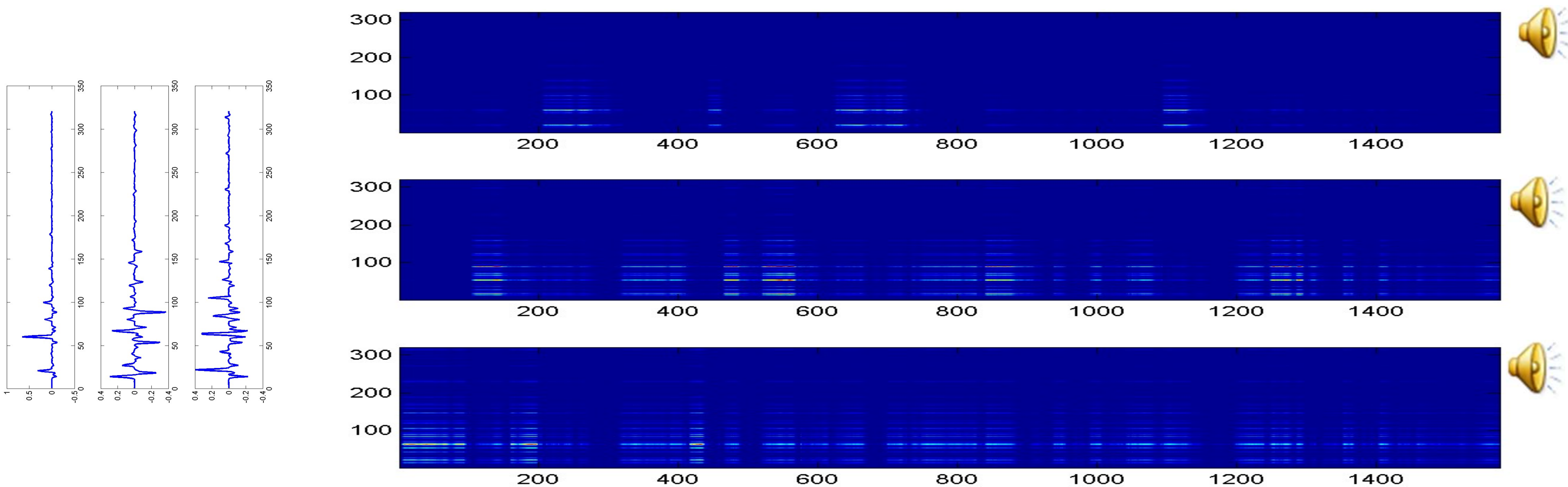
- The columns of $\mathbf{W}$ are the bases we have learned
  - The linear “building blocks” that compose the music
- They represent “learned” notes
  - $\mathbf{w}_i \mathbf{h}_i$ is the contribution of the ith note to the music
    - $\mathbf{w}_i$ is the ith column of $\mathbf{W}$
    - $\mathbf{h}_i$ is the ith row of $\mathbf{H}$

So how does that work?



- There are 12 notes in the segment, hence we try to estimate 12 notes..

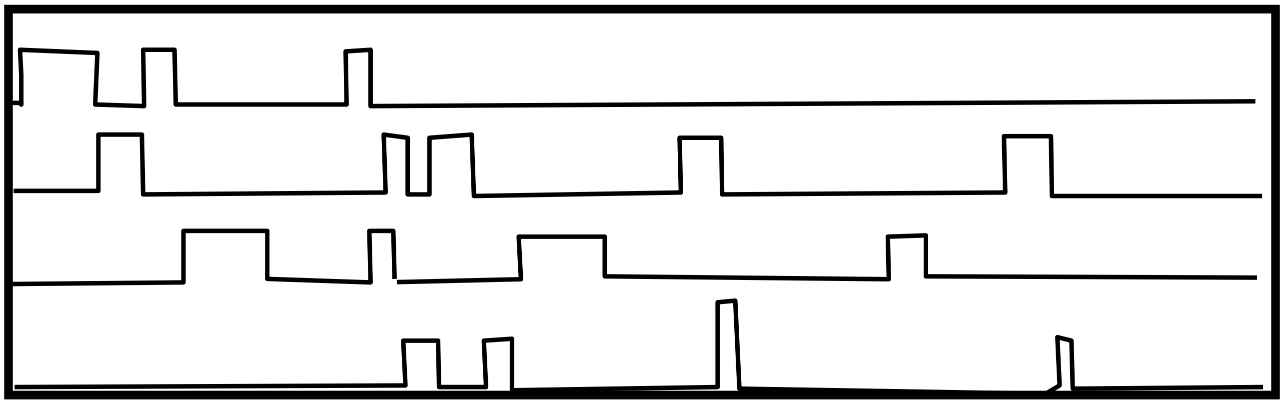
So how does that work?



- There are 12 notes in the segment, hence we try to estimate 12 notes..
- Results are not good

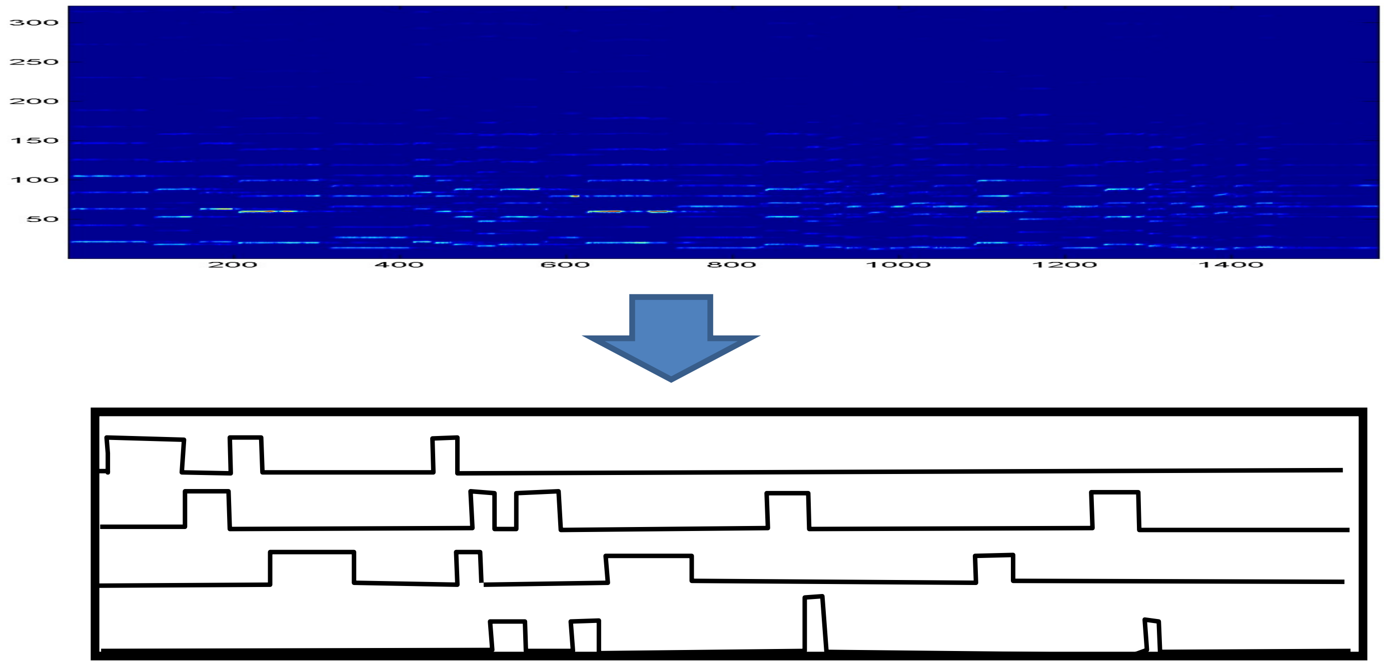
PCA through decorrelation of notes

$$\mathbf{W}, \mathbf{H} = \arg\min_{\overline{\mathbf{W}}, \overline{\mathbf{H}}} \| \mathbf{M} - \overline{\mathbf{H}} \|_F^2 + \Lambda(\overline{\mathbf{H}}\overline{\mathbf{H}}^T - \mathbf{D})$$



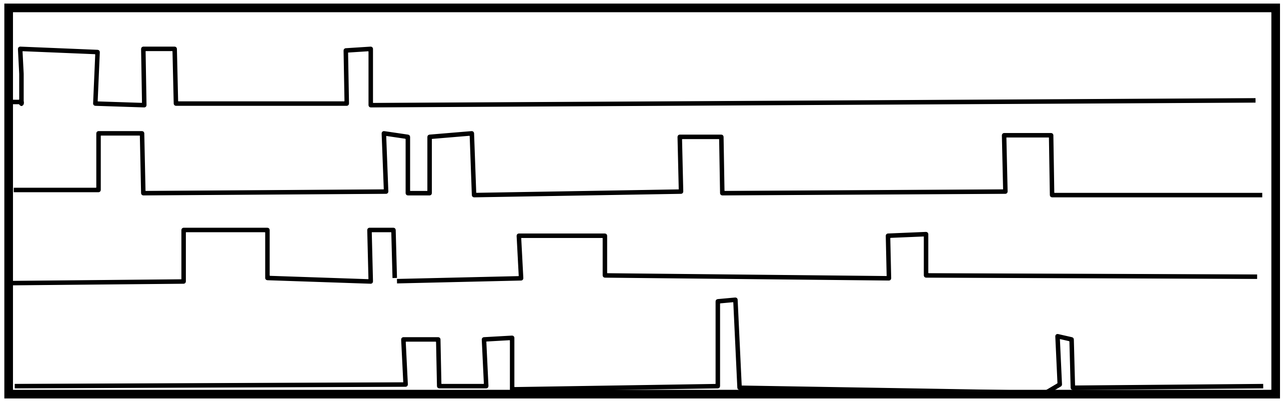
- Different constraint: Constraint $\mathbf{H}$ to be decorrelated
  - $\mathbf{H}\mathbf{H}^{\mathrm{T}} = \mathbf{D}$
- This will result exactly in PCA too
- Decorrelation of $\mathbf{H}$ Interpretation: What does this mean?

Decorrelation



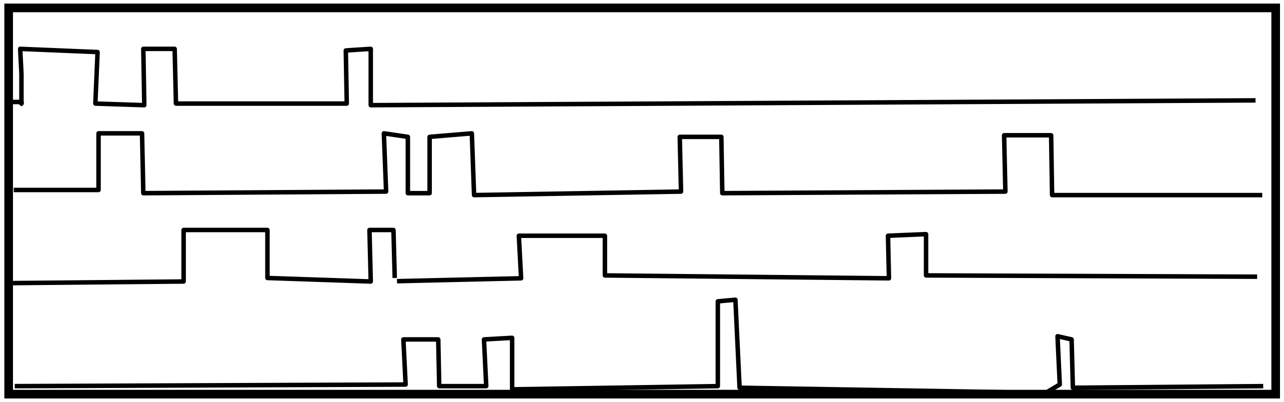
- Alternate view: Find a matrix $\mathbf{B}$ such that the rows of $\mathbf{H}=\mathbf{BM}$ are uncorrelated
- Will find $\mathbf{B} = \mathbf{W}^{\mathrm{T}}$
- $\mathbf{B}$ is the *decorrelating matrix* of **M**

What *else* can we look for?



- Assume: The “transcription” of one note does not depend on what else is playing
  - Or, in a multi-instrument piece, instruments are playing independently of one another
- Not strictly true, but still..

What *else* can we look for?



- Assume: The “transcription” of one note does not depend on what else is playing
  - Or, in a multi-instrument piece, instruments are playing independently of one another

- **Attempting to find statistically independent components of the mixed signal**
  - ***Independent Component Analysis***

Formulating it with Independence

$$\mathbf{W},\mathbf{H} = \arg\min_{\overline{\mathbf{W}},\overline{\mathbf{H}}} \|\mathbf{M} - \overline{\mathbf{W}}\overline{\mathbf{H}}\|_F^2 + \Lambda(\textit{rows of } \mathbf{H} \textit{ are independent})$$

- Impose statistical independence constraints on decomposition