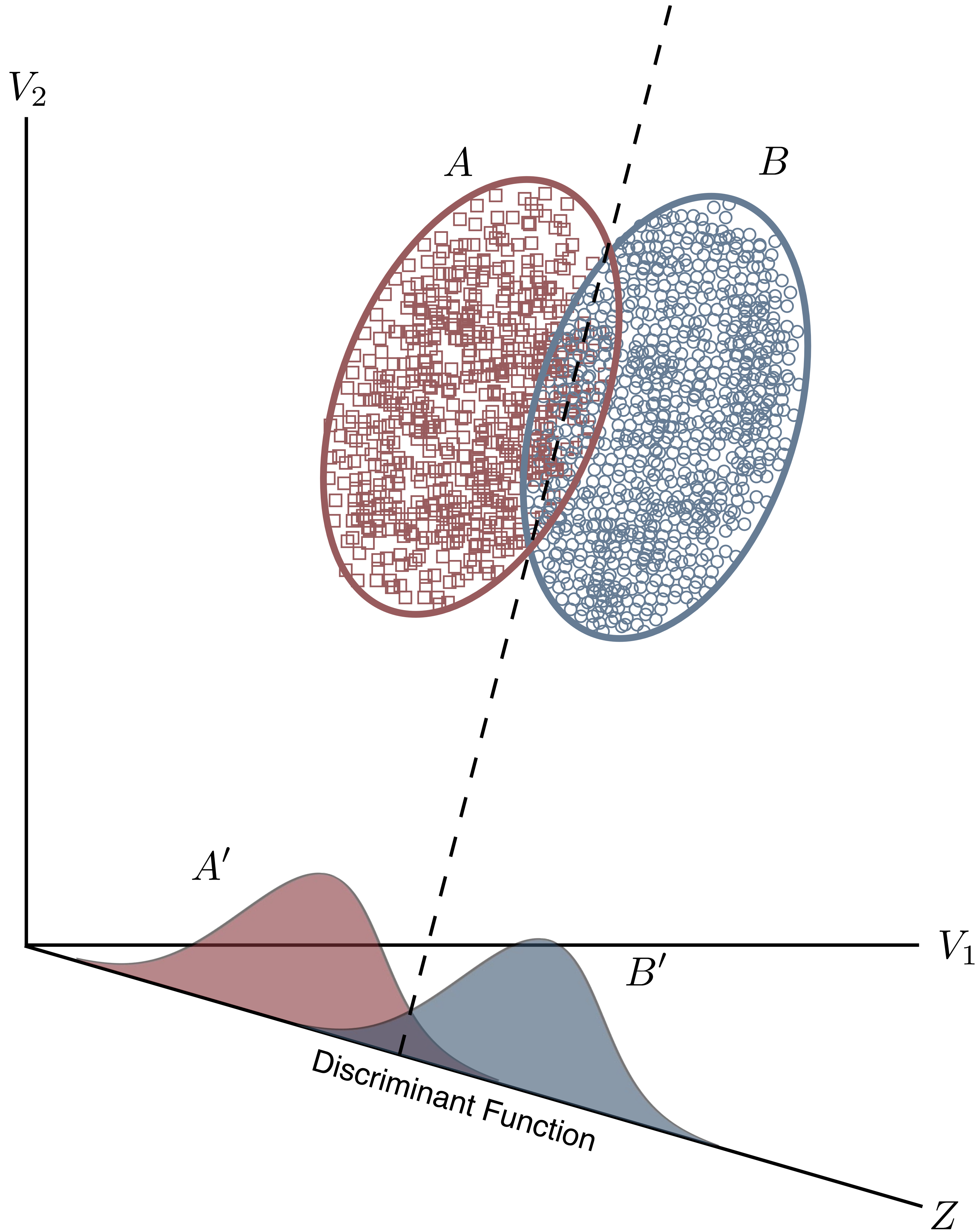
Admin

- Assignment 1 and Matrix Quiz:
  - We will post your marks on Blackboard as soon as we can.
- Assignment 2:
  - Due 13 May.


V_2
A
B
A'
V_1
B'
Discriminant Function
Z

A very brief intro to SDT (Signal Detection Theory)

If you have encountered SDT before, it was likely in the context of collecting formal data. We will discuss it as a representation of one type of decision problem, without the presumption that formal data is being collected.

Decision		**Reality**	
		Enemy Present	*Enemy Absent*
	Yes	Hit	False Alarm
	No	Miss	Correct Rejection

A very brief intro to SDT (Signal Detection Theory)

If you have encountered SDT before, it was likely in the context of collecting formal data. We will discuss it as a representation of one type of decision problem, without the presumption that formal data is being collected.

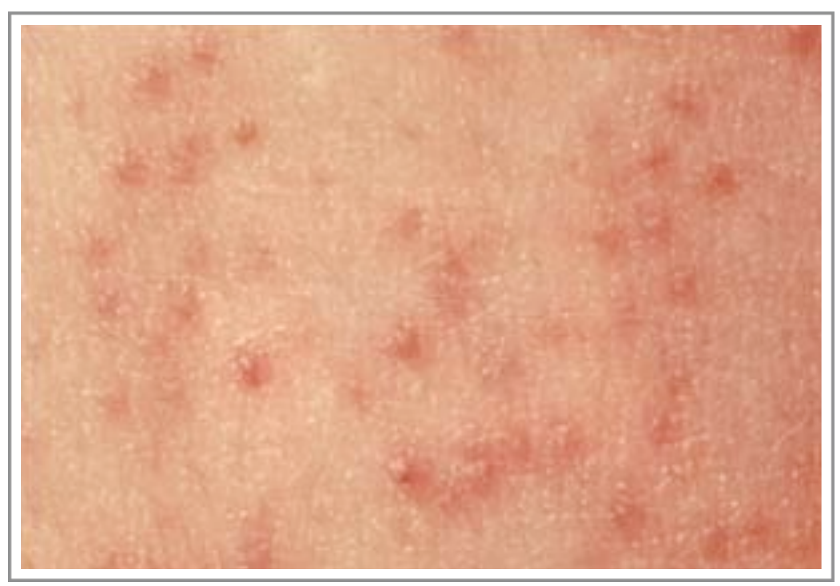
Reality

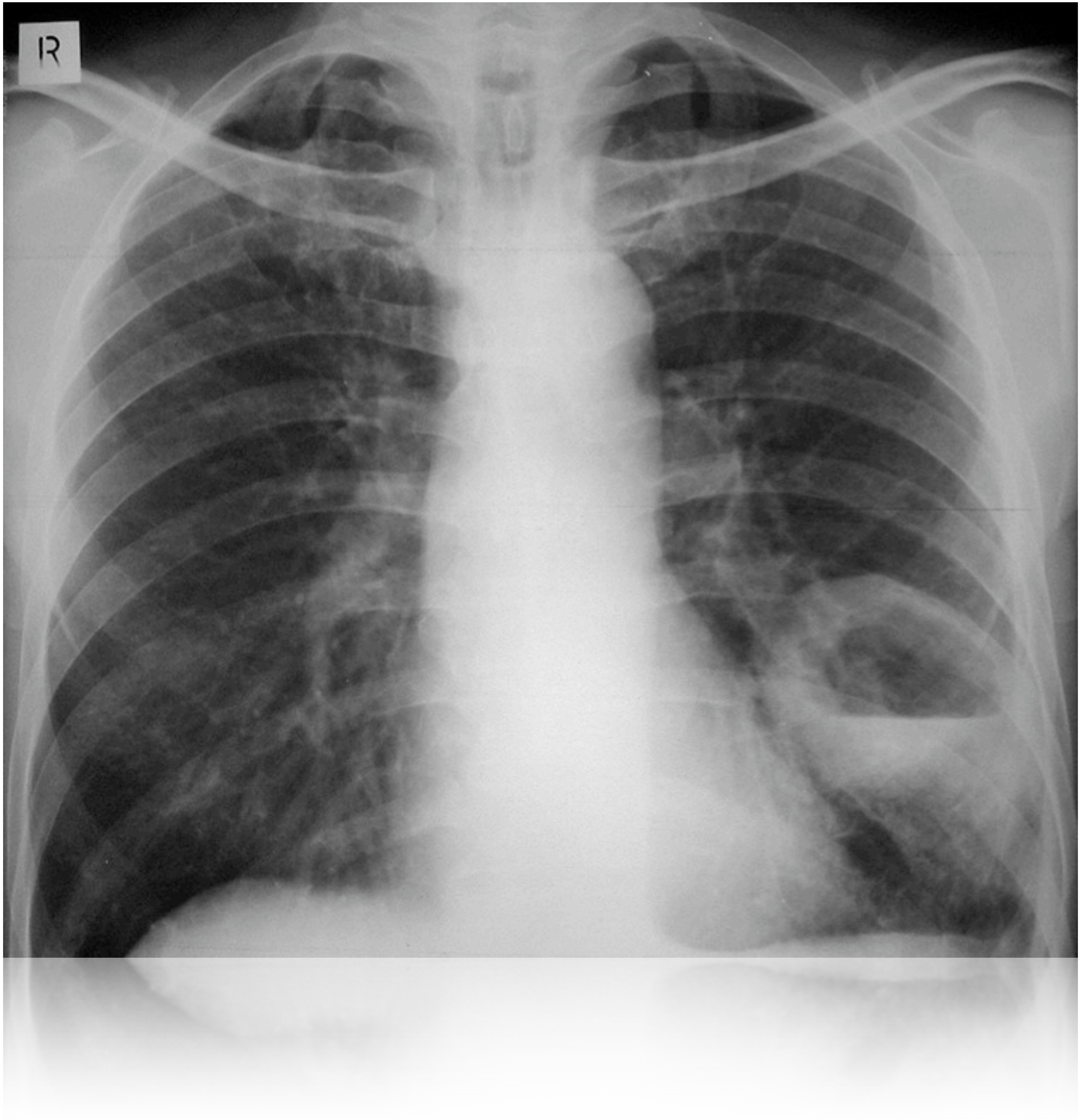
		Stimulus Present	*Stimulus Absent*
Decision	*Yes*	Hit	False Alarm
	No	Miss	Correct Rejection

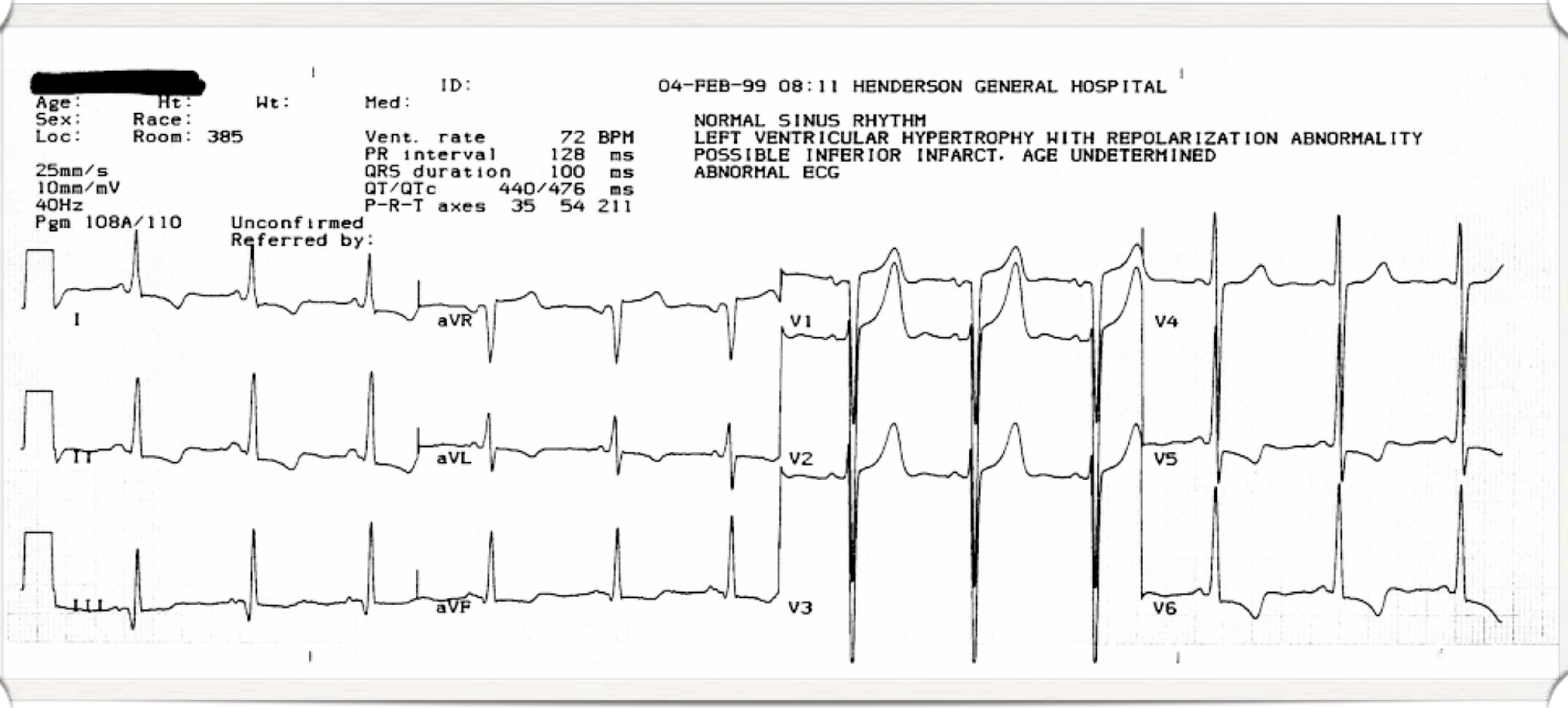


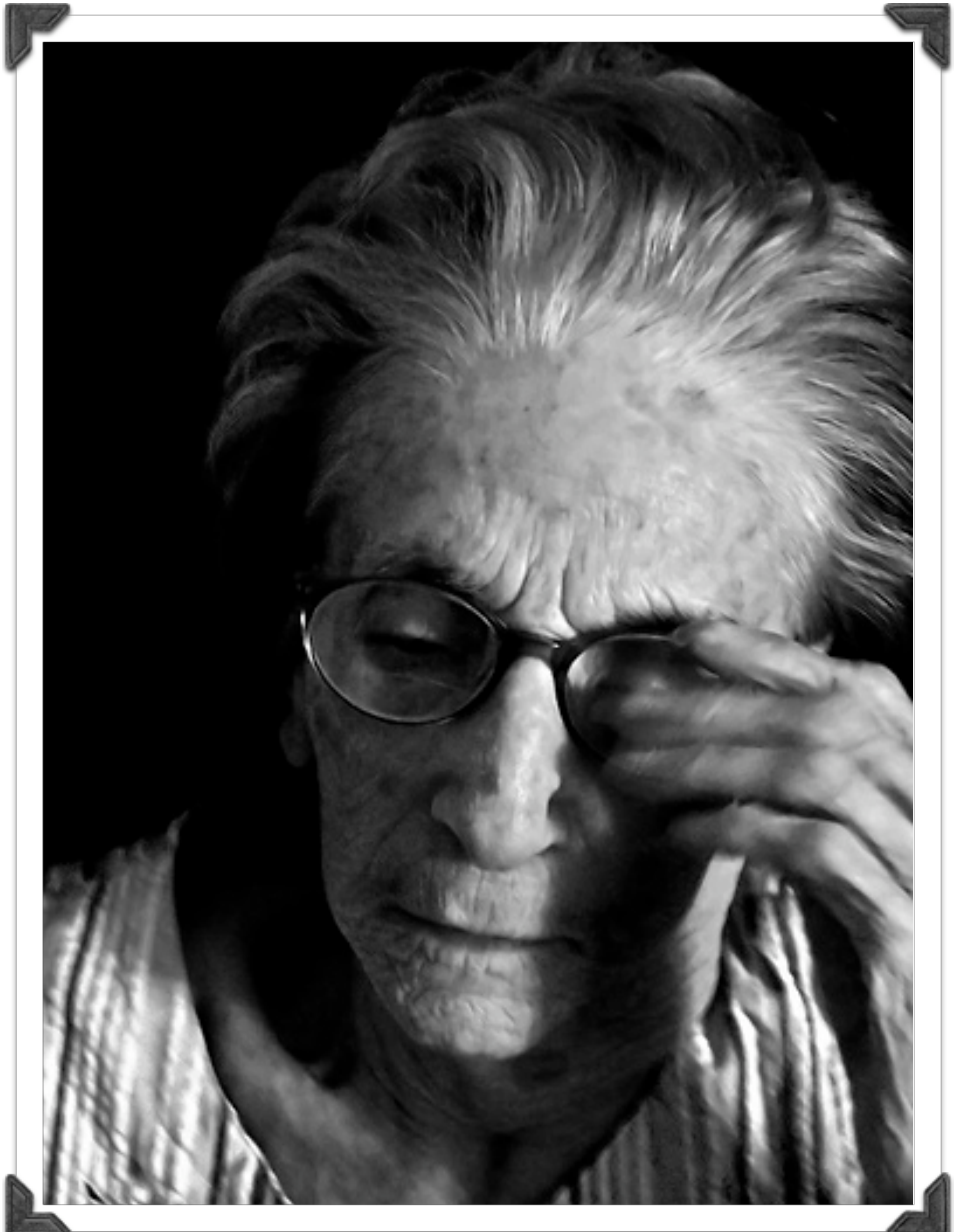






R


[redacted]
ID:
04-FEB-99 08:11 HENDERSON GENERAL HOSPITAL
Age: Ht: Wt: Med:
Sex: Race:
Loc: Room: 385
Vent. rate 72 BPM
PR interval 128 ms
QRS duration 100 ms
QT/QTc 440/476 ms
P-R-T axes 35 54 211
NORMAL SINUS RHYTHM
LEFT VENTRICULAR HYPERTROPHY WITH REPOLARIZATION ABNORMALITY
POSSIBLE INFERIOR INFARCT, AGE UNDETERMINED
ABNORMAL ECG
25mm/s
10mm/mV
40Hz
Pgm 108A/110
Unconfirmed
Referred by:
I
aVR
V1
V4
II
aVL
V2
V5
III
aVF
V3
V6

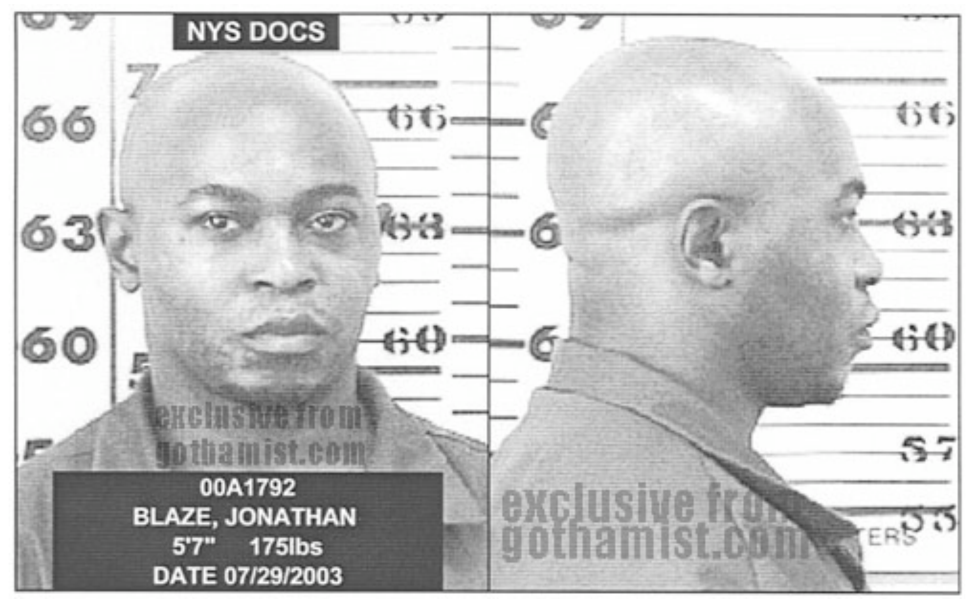



PL SUN A00
40

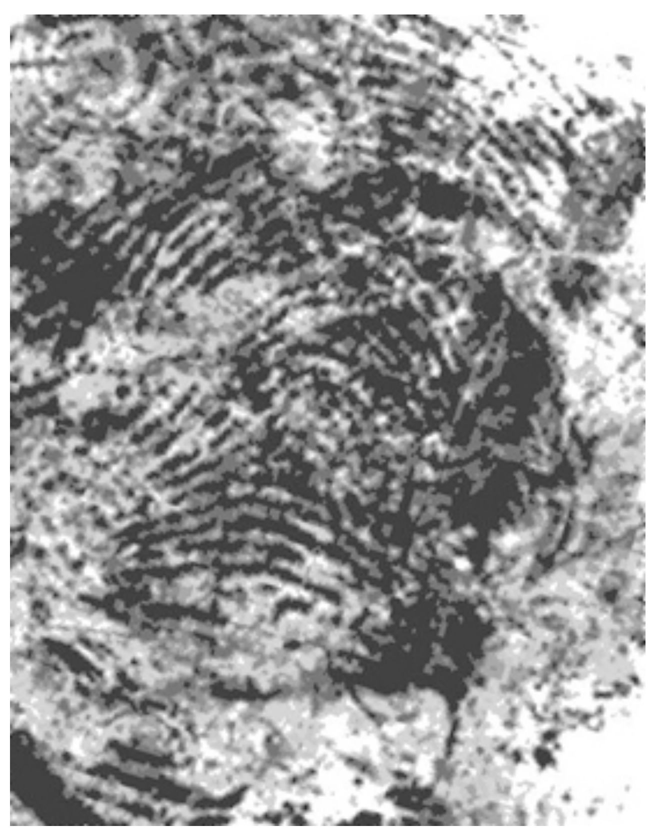

PL SUN A00
40

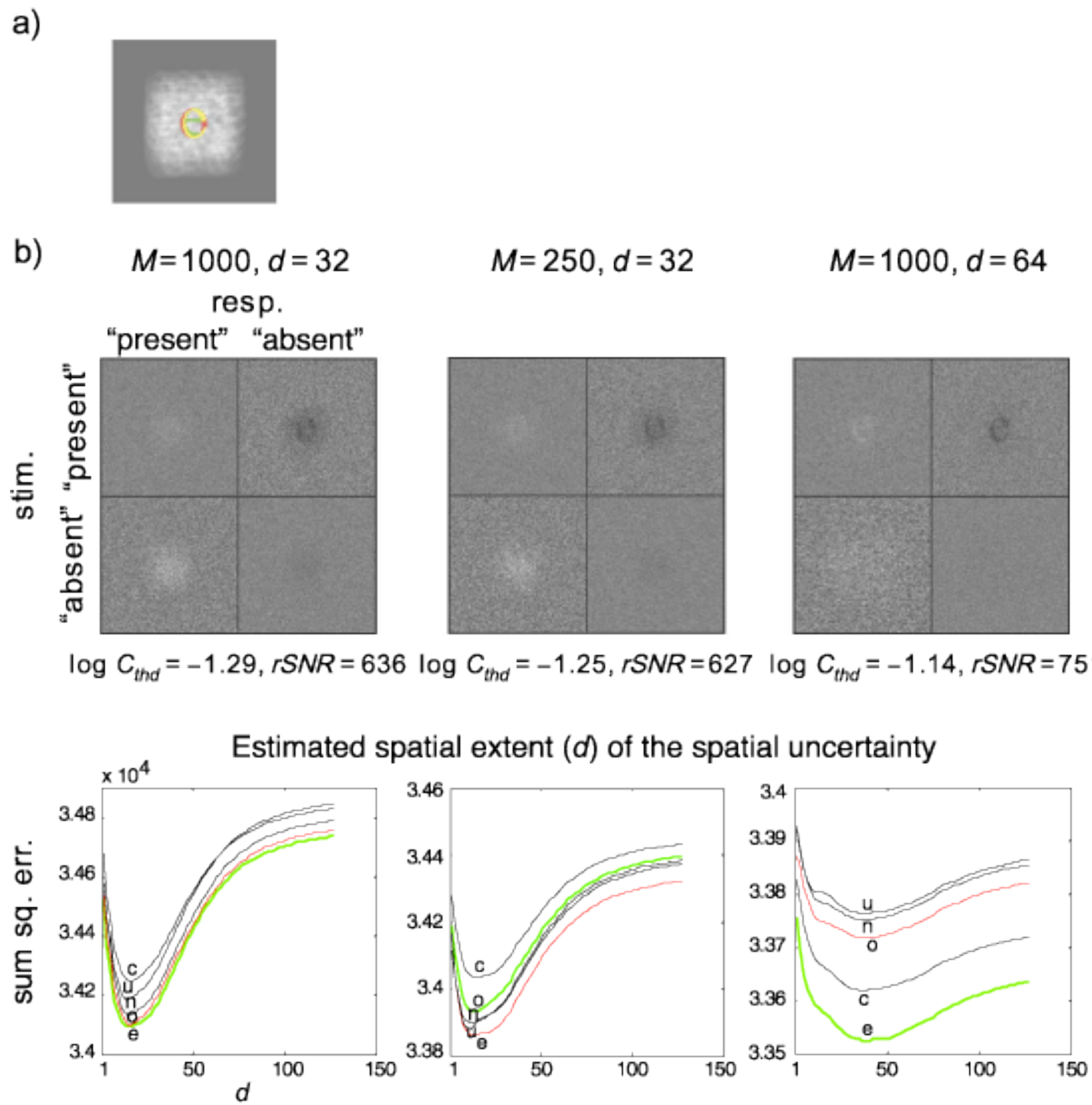

UN A00


N A00

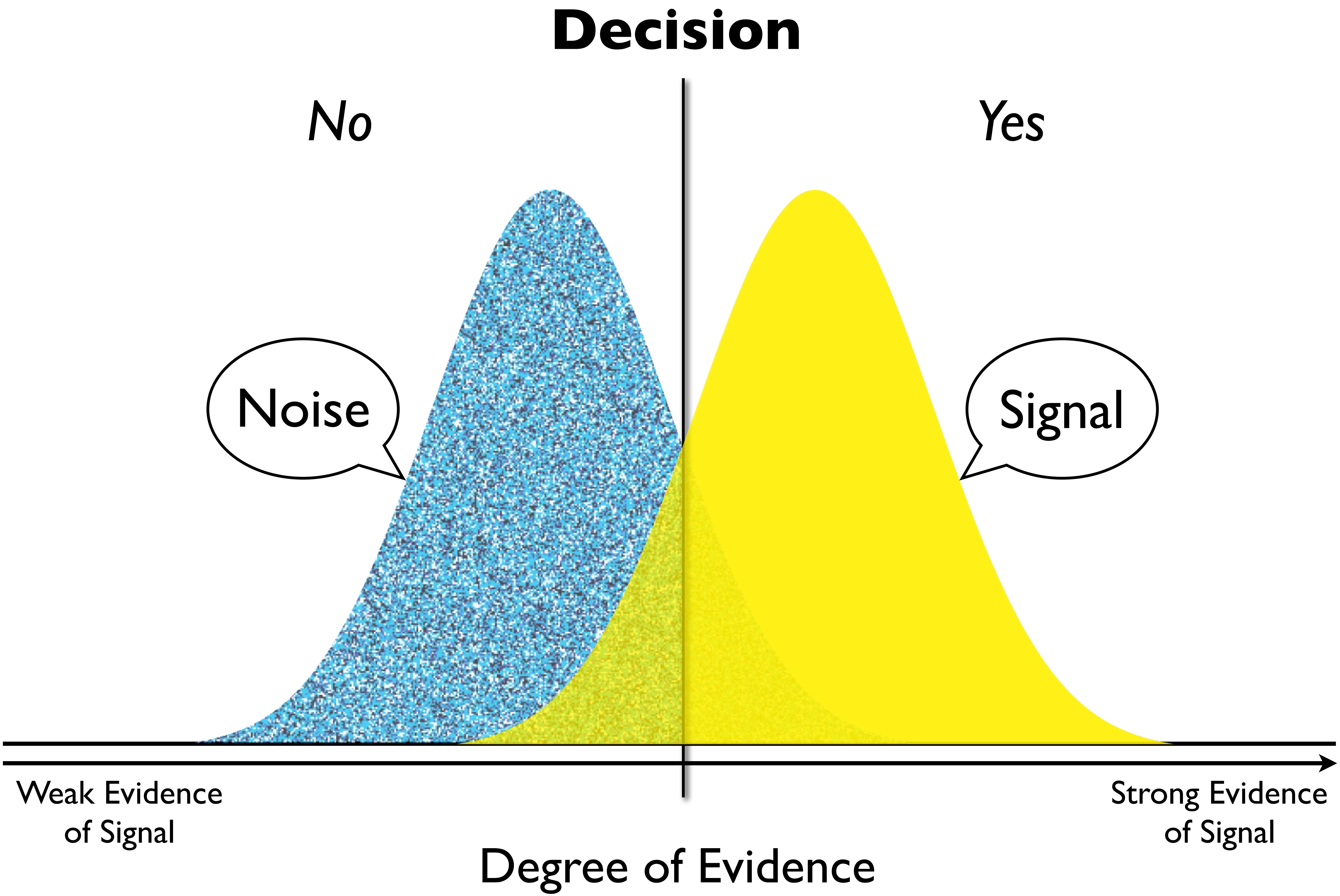

NYS DOCS
66
63
60
00A1792
BLAZE, JONATHAN
5'7" 175lbs
DATE 07/29/2003
exclusive from gothamist.com


PRESS

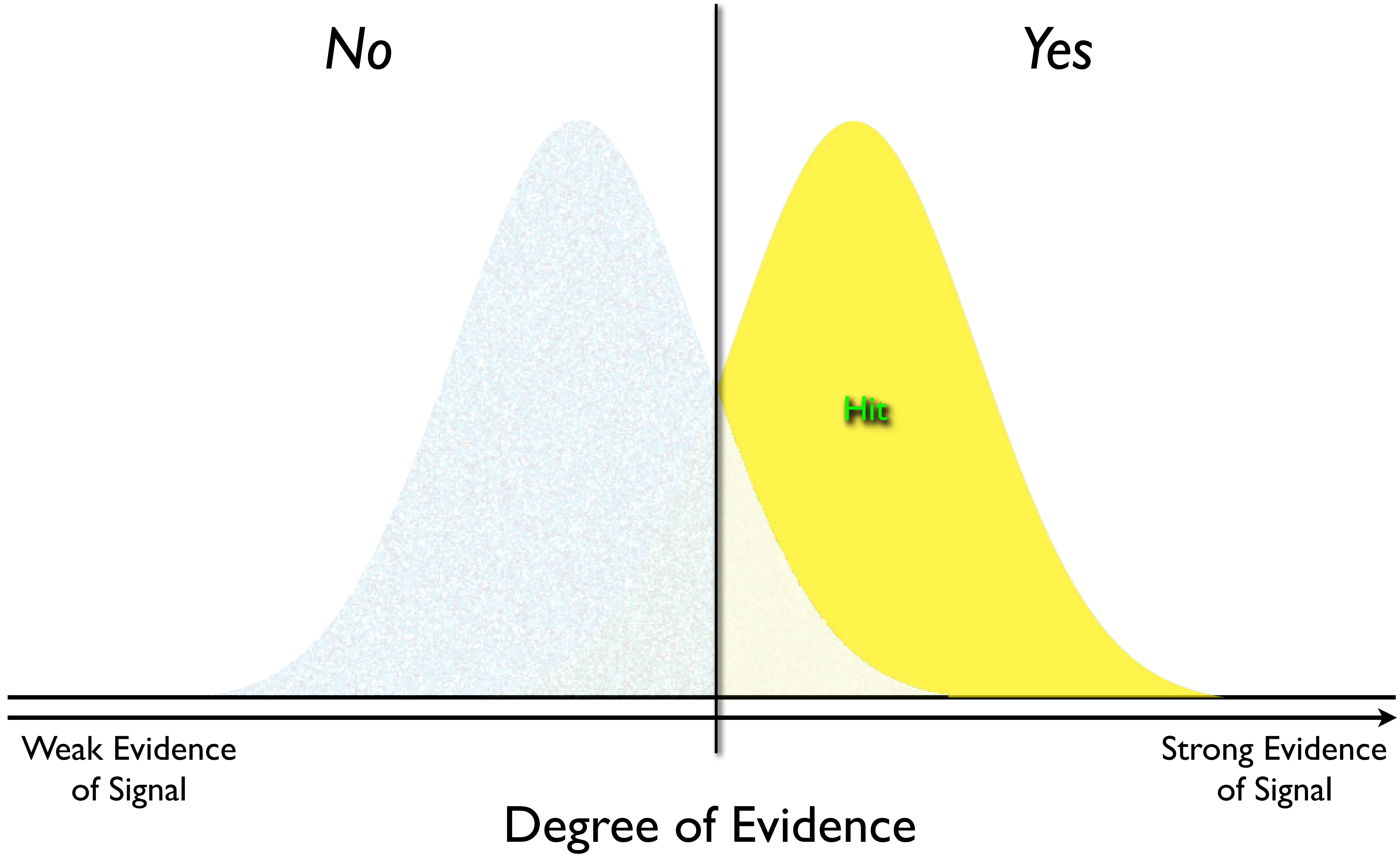



a)
b)
M=1000, d=32
M=250, d=32
M=1000, d=64
resp.
"present" "absent"
stim.
"absent" "present"
log C_thd = −1.29, rSNR = 636
log C_thd = −1.25, rSNR = 627
log C_thd = −1.14, rSNR = 751
Estimated spatial extent (d) of the spatial uncertainty
sum sq. err.
x 10^4
d

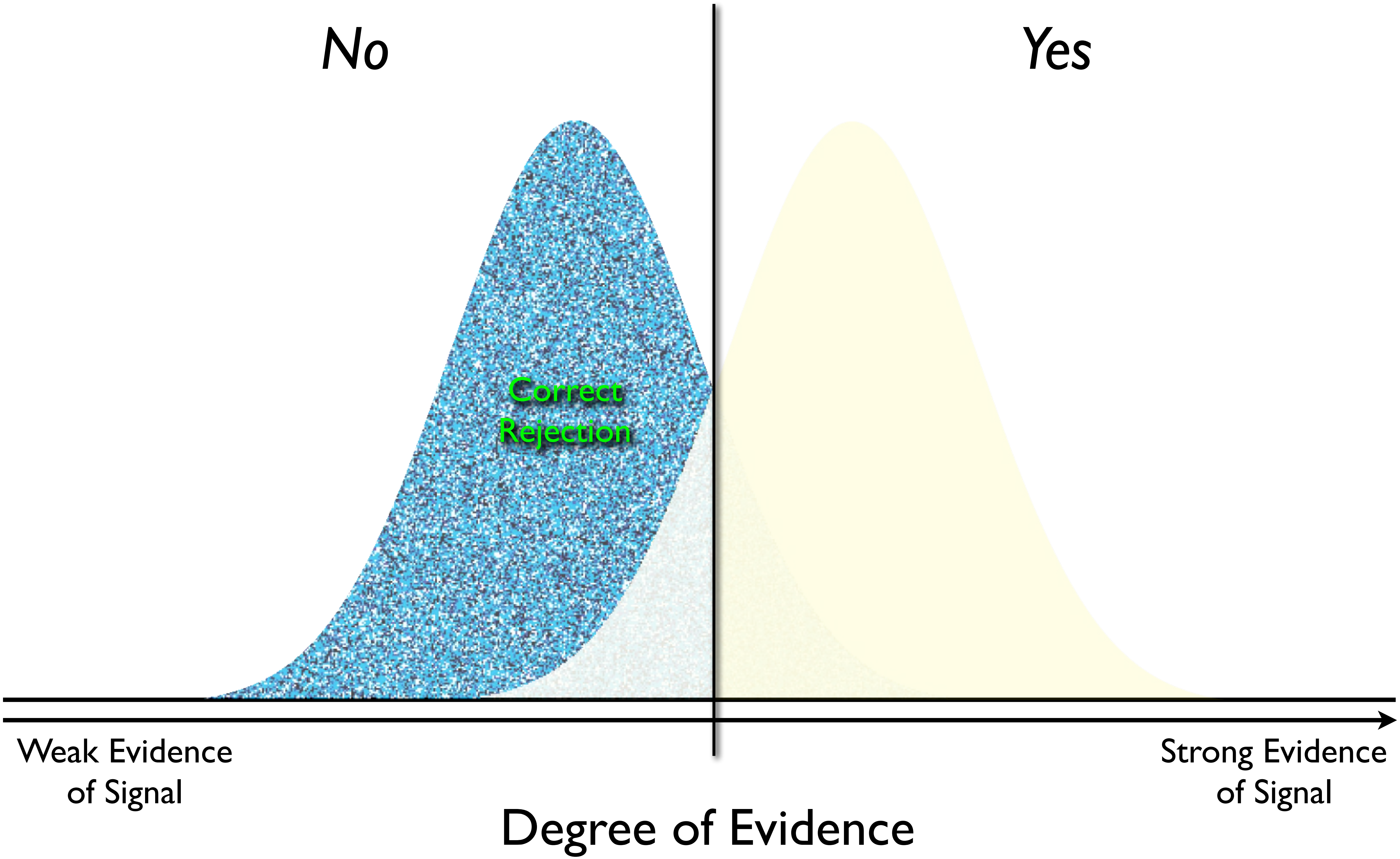
		Reality	
		+	**–**
Decision	*Yes*	Hit	False Alarm
	No	Miss	Correct Rejection

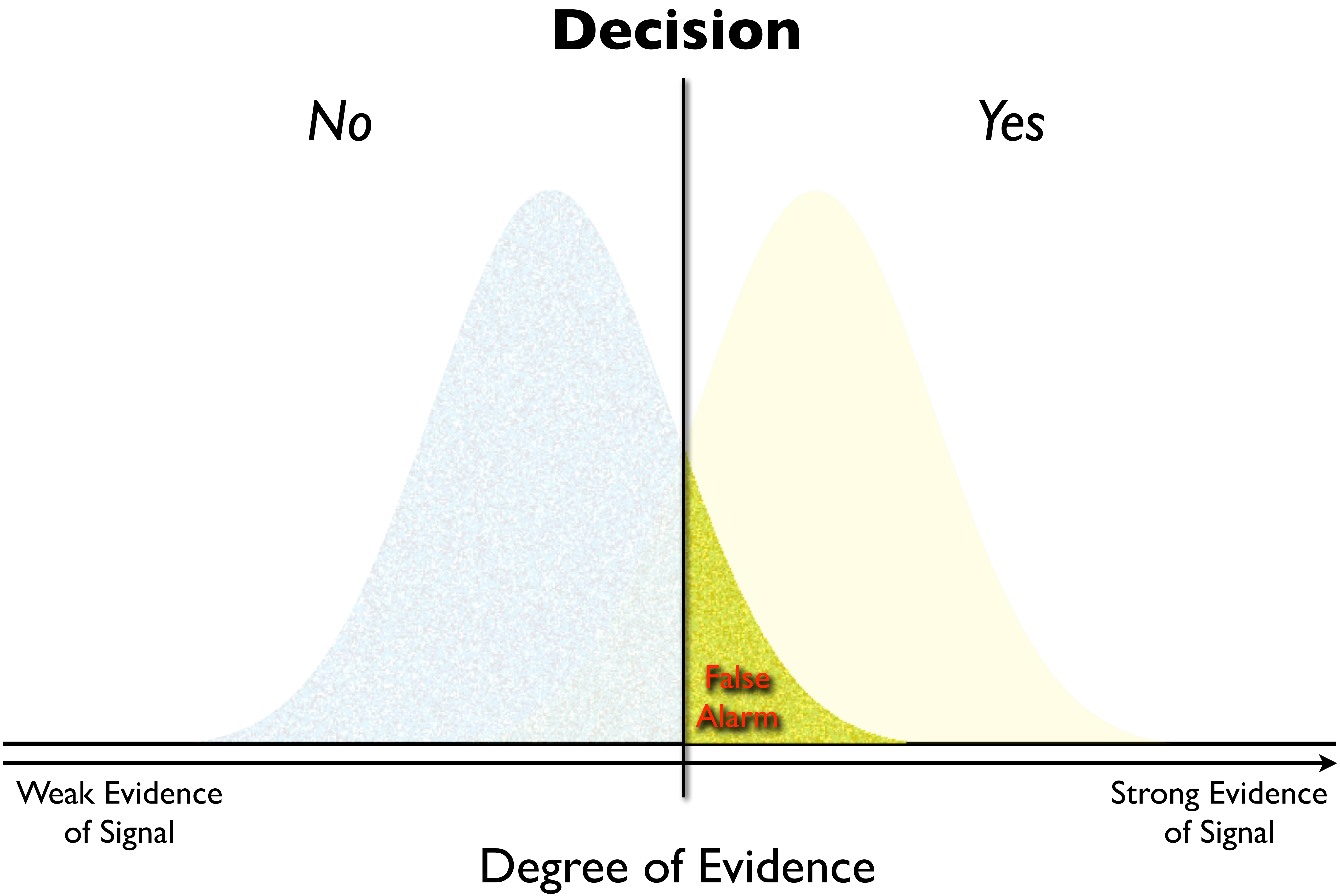

Decision
No
Yes
Noise
Signal
Weak Evidence of Signal
Strong Evidence of Signal
Degree of Evidence

Decision

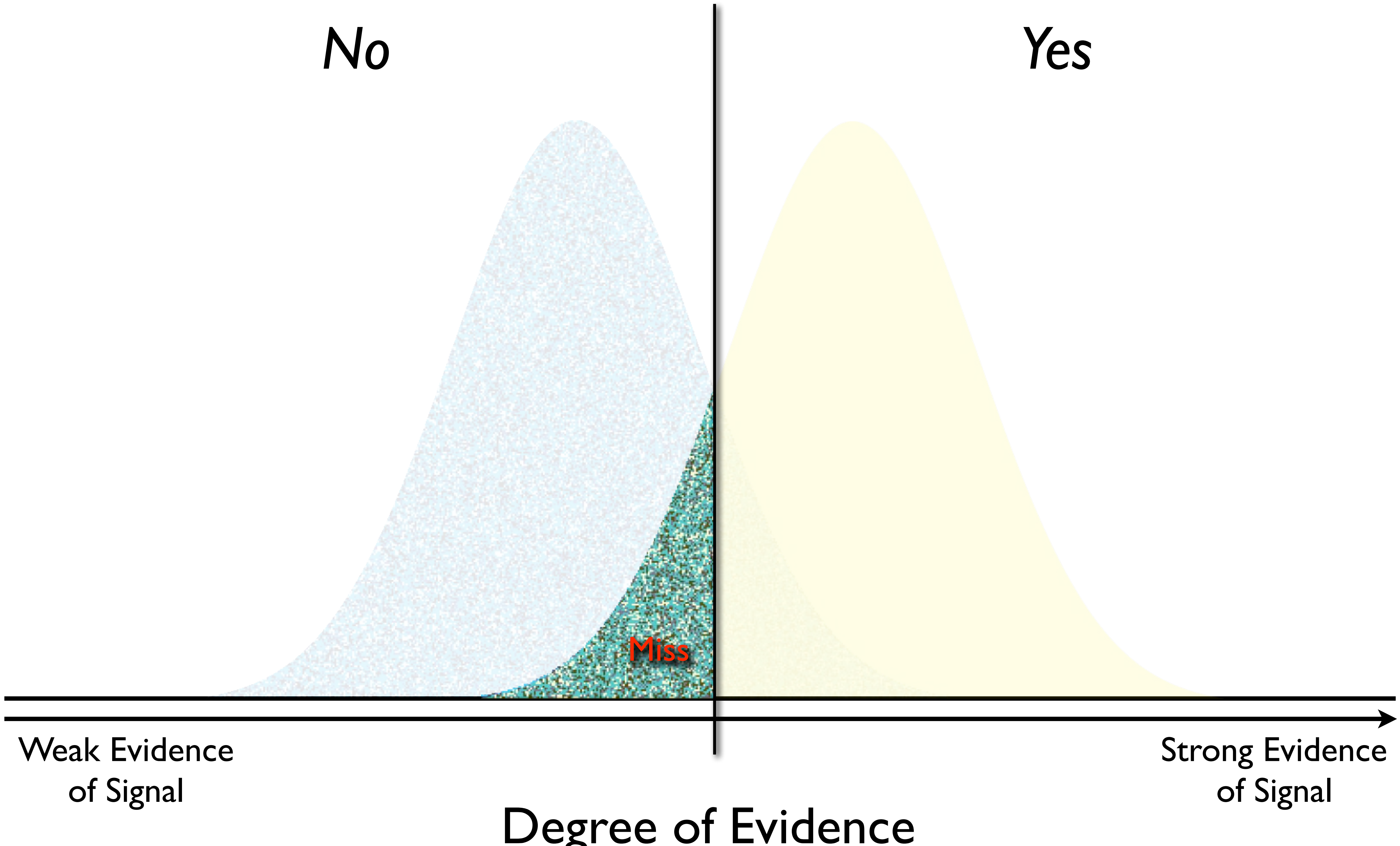

No
Yes
Hit
Weak Evidence of Signal
Strong Evidence of Signal
Degree of Evidence

Decision

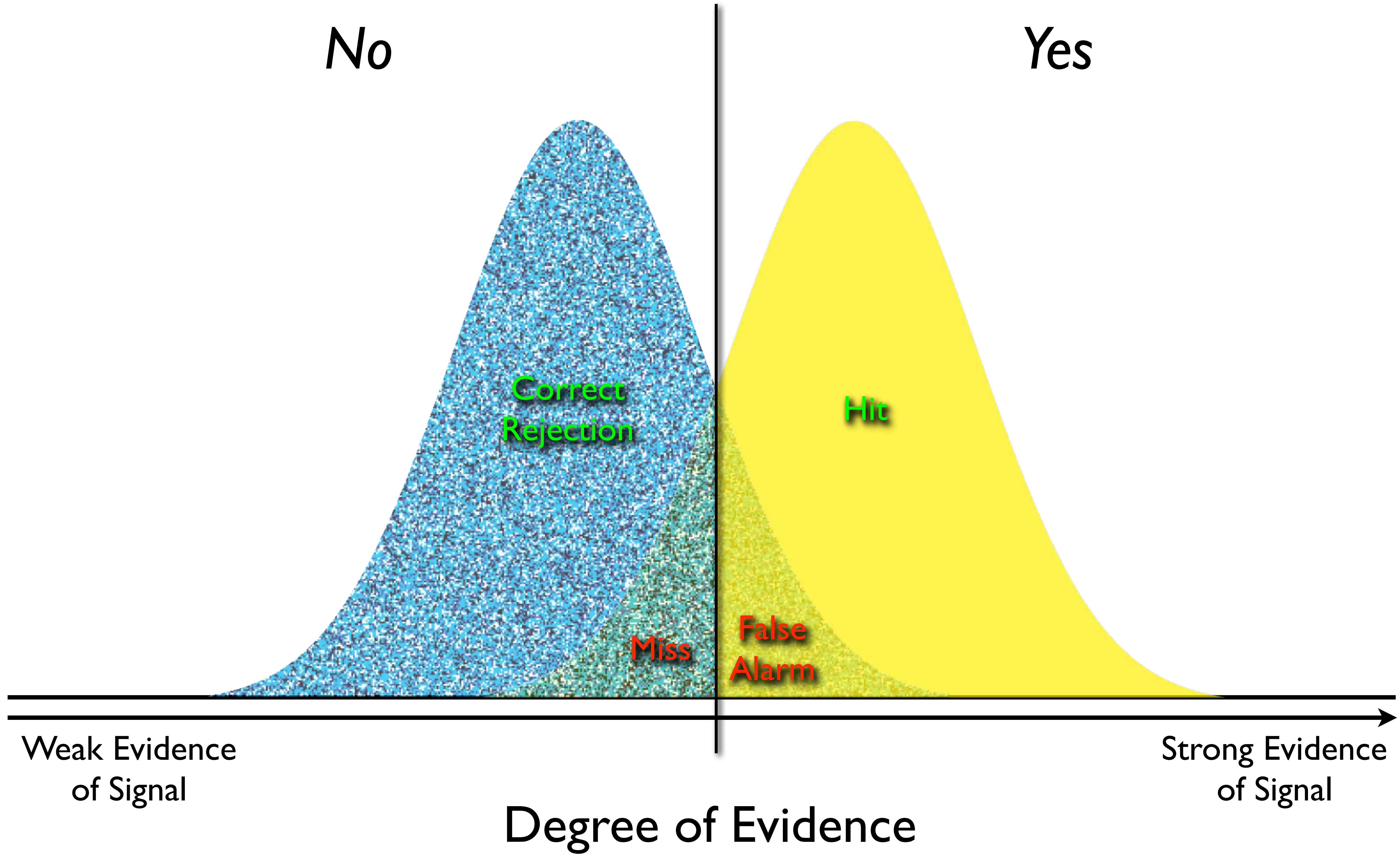

No
Yes
Correct
Rejection
Weak Evidence
of Signal
Strong Evidence
of Signal
Degree of Evidence


Decision
No
Yes
False
Alarm
Weak Evidence
of Signal
Strong Evidence
of Signal
Degree of Evidence

Decision


No
Yes
Miss
Weak Evidence of Signal
Strong Evidence of Signal
Degree of Evidence

Decision

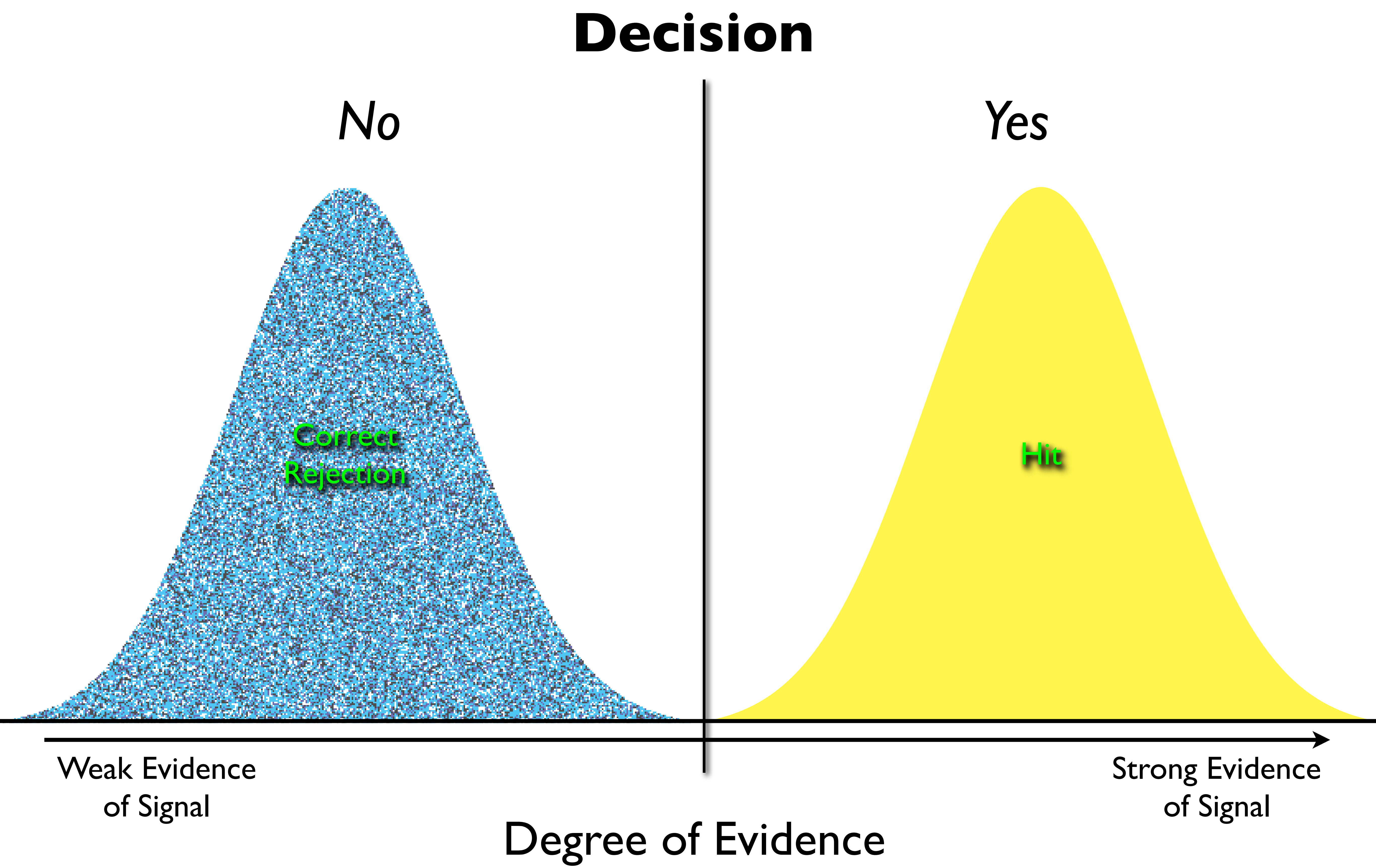

No
Yes
Correct
Rejection
Hit
Miss
False
Alarm
Weak Evidence
of Signal
Strong Evidence
of Signal
Degree of Evidence

So what can we vary?

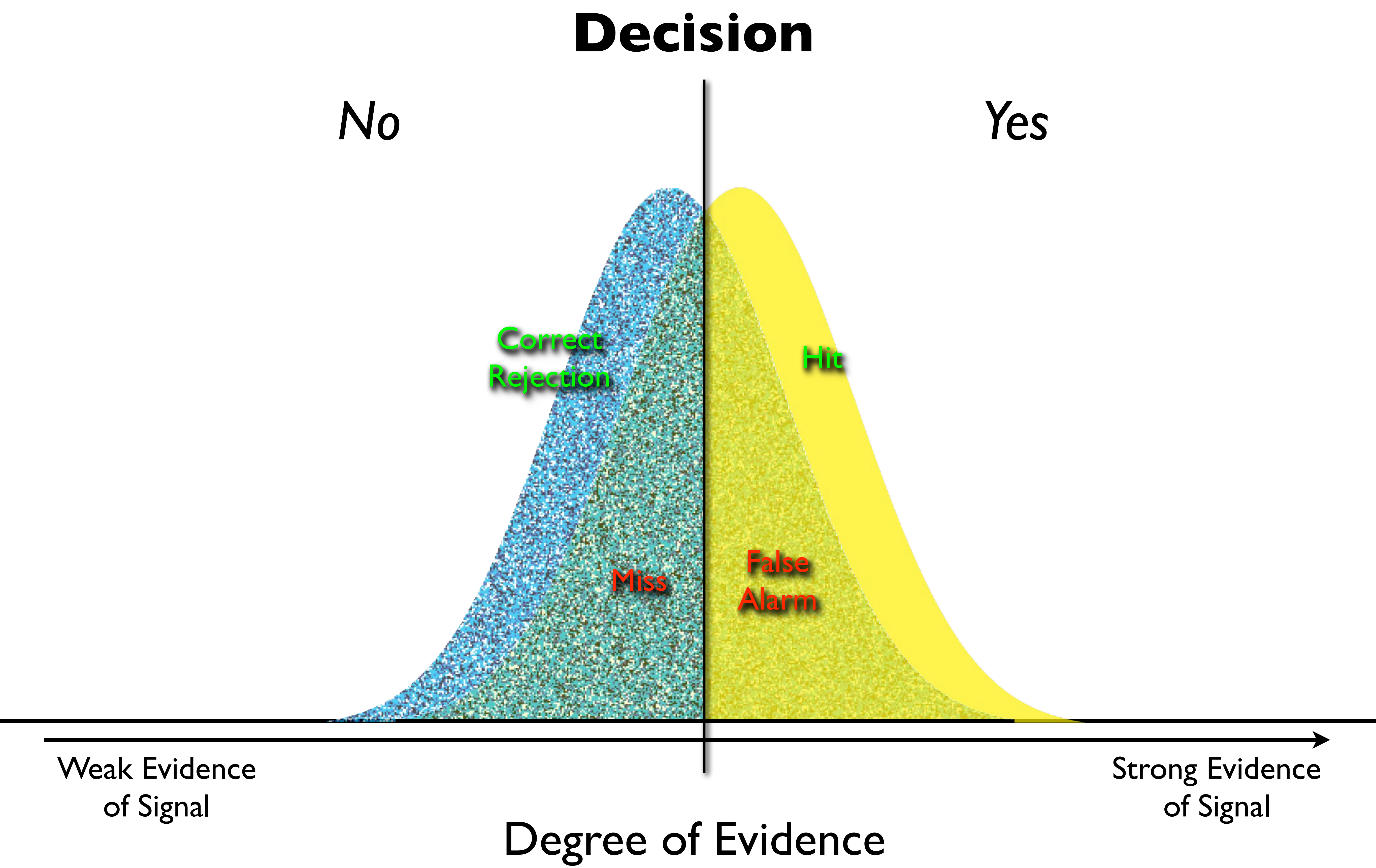
(to represent different decision problems)

1. Degree of overlap

1. Degree of overlap

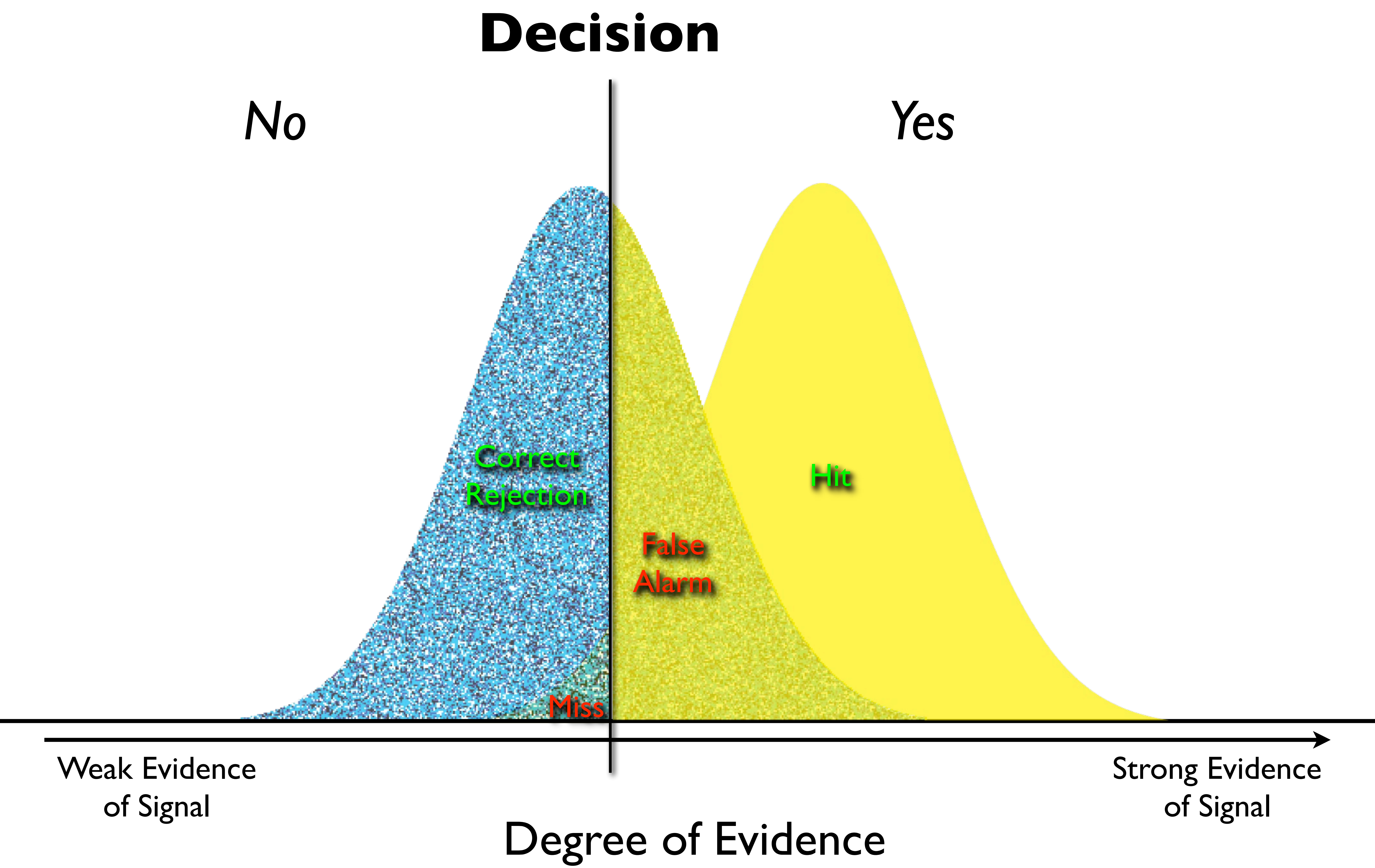


1. Degree of overlap

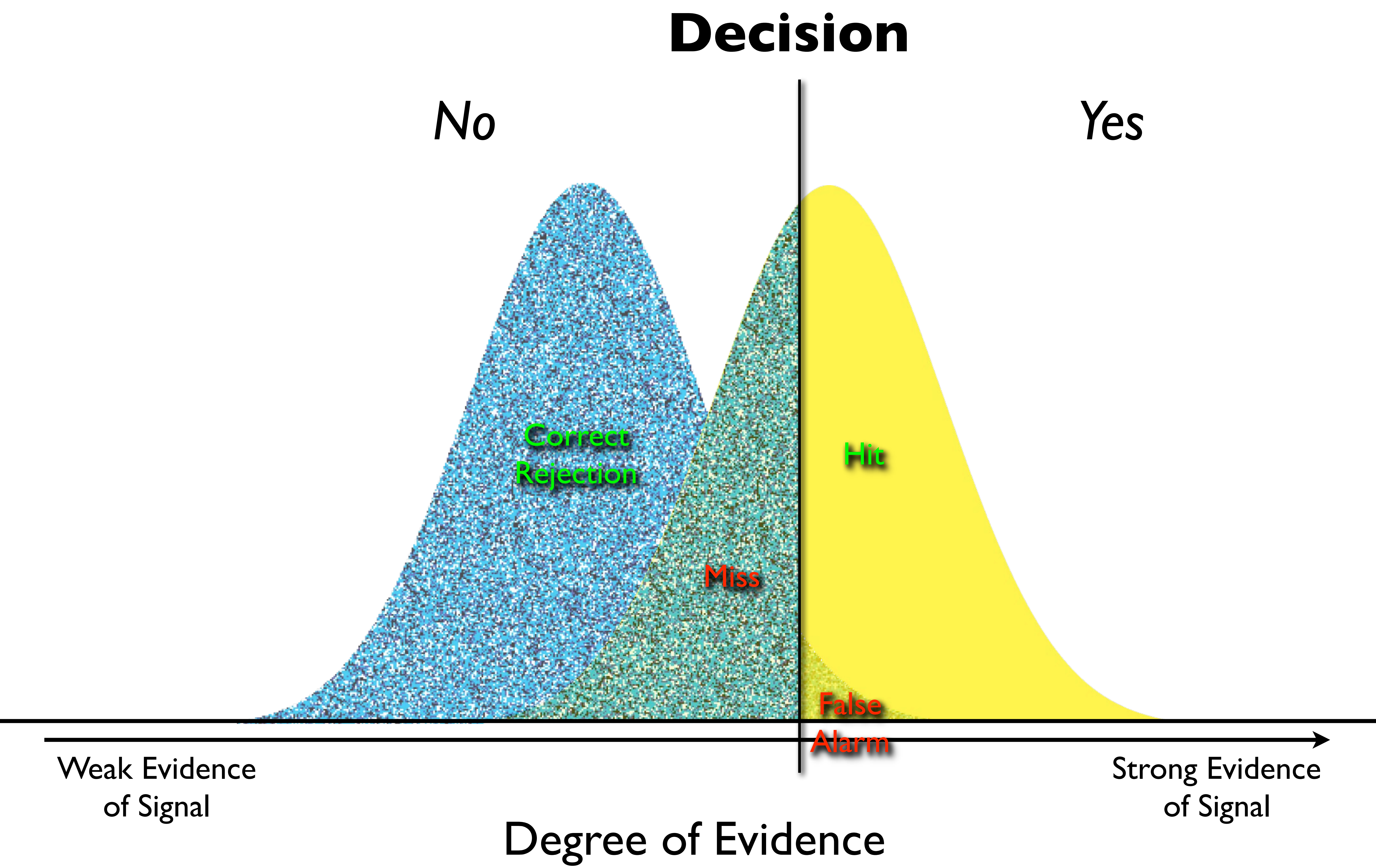


2. Location of the Criterion

2. Location of the Criterion

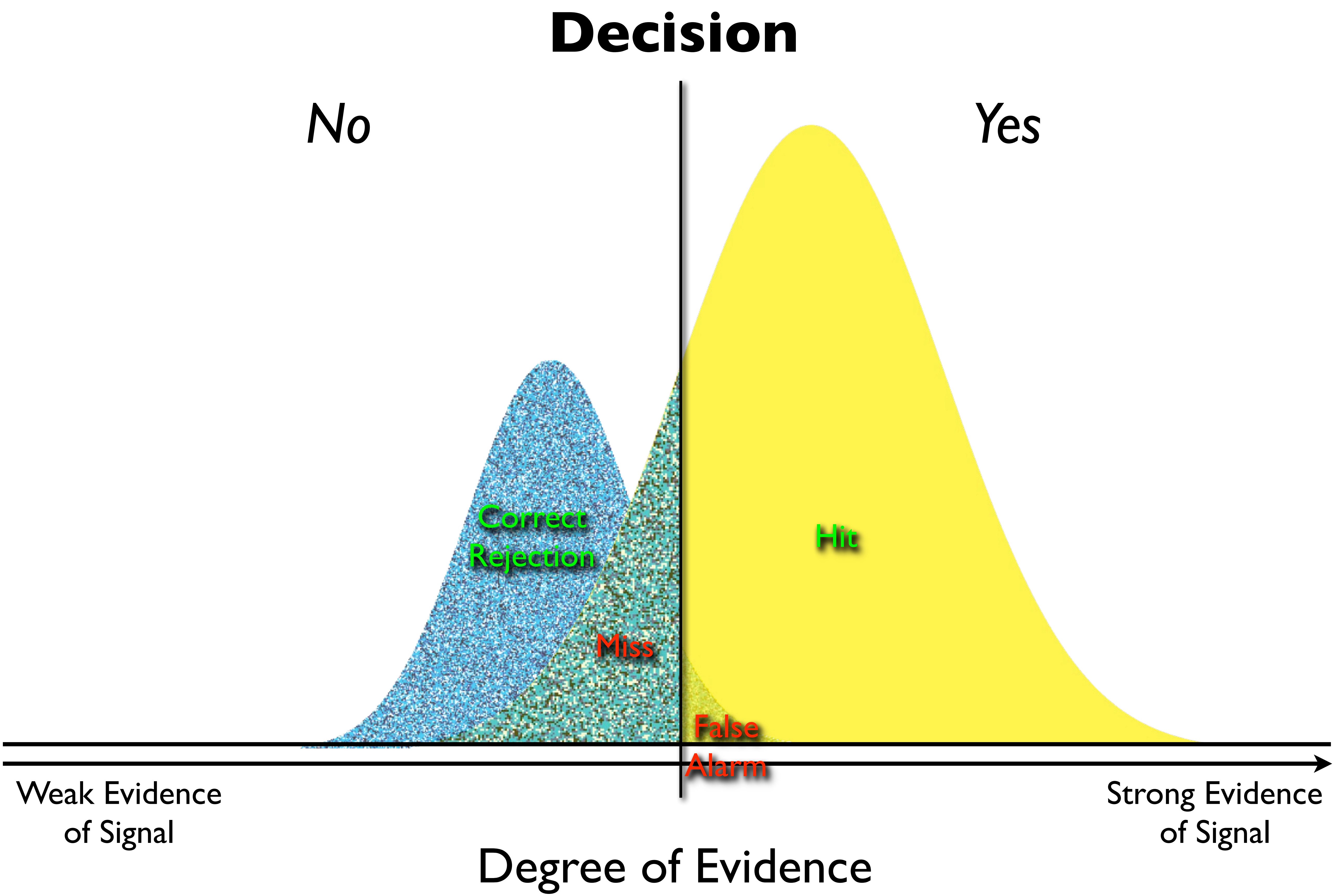


2. Location of the Criterion

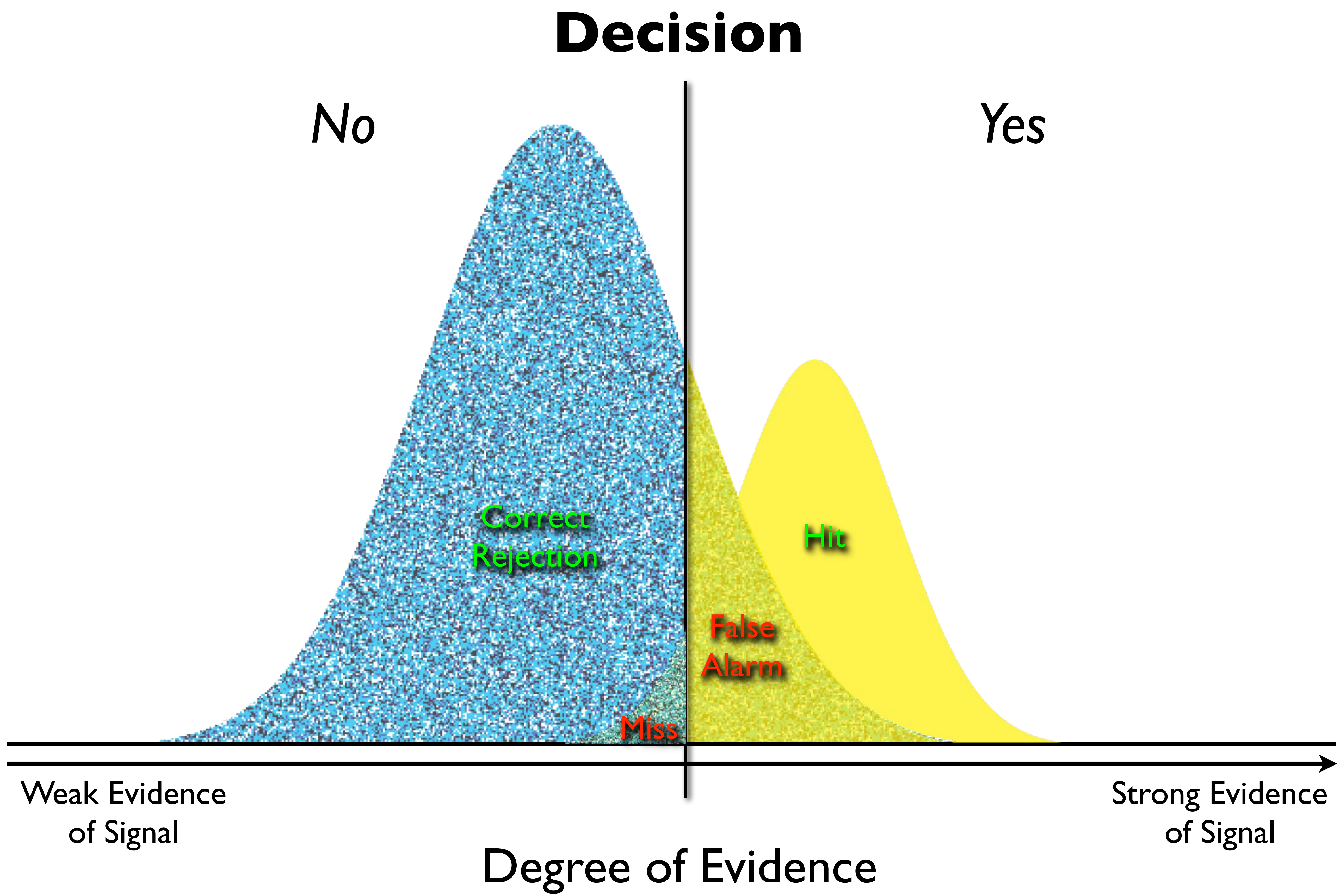


3. Size of curves: base rates

3. Size of curves: base rates

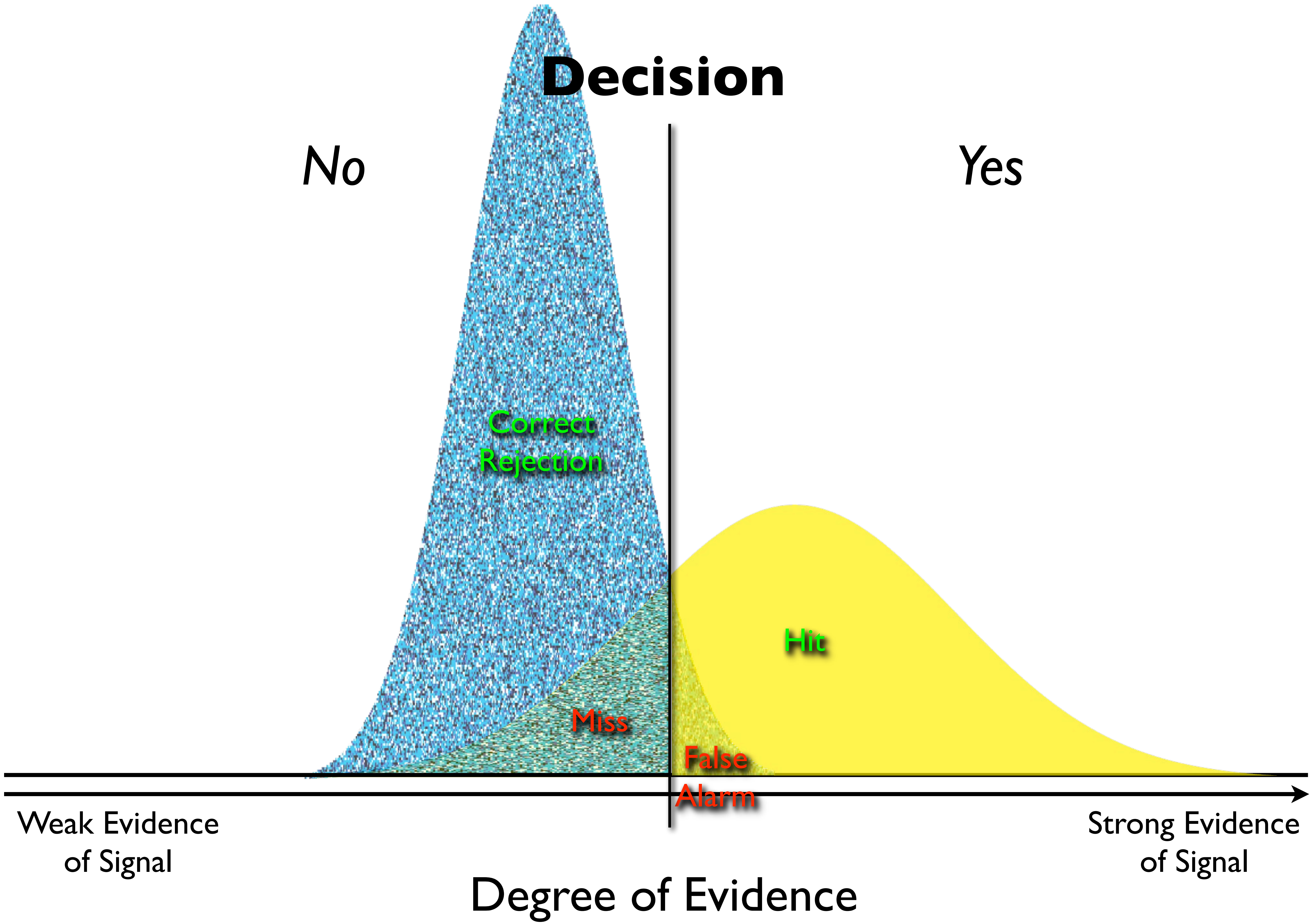


3. Size of curves: base rates

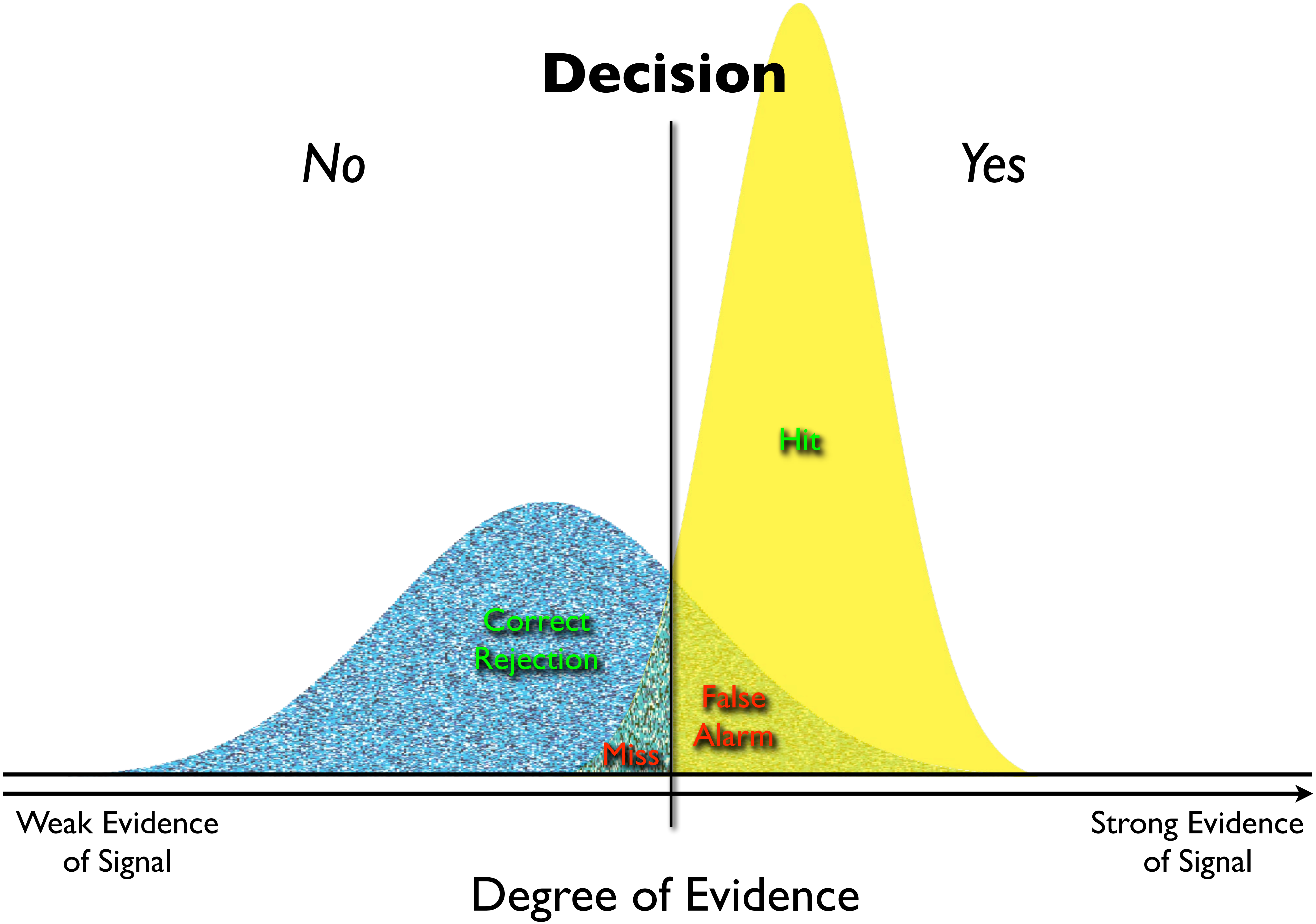


4. Spread (variance) of curves

4. Spread (variance) of curves



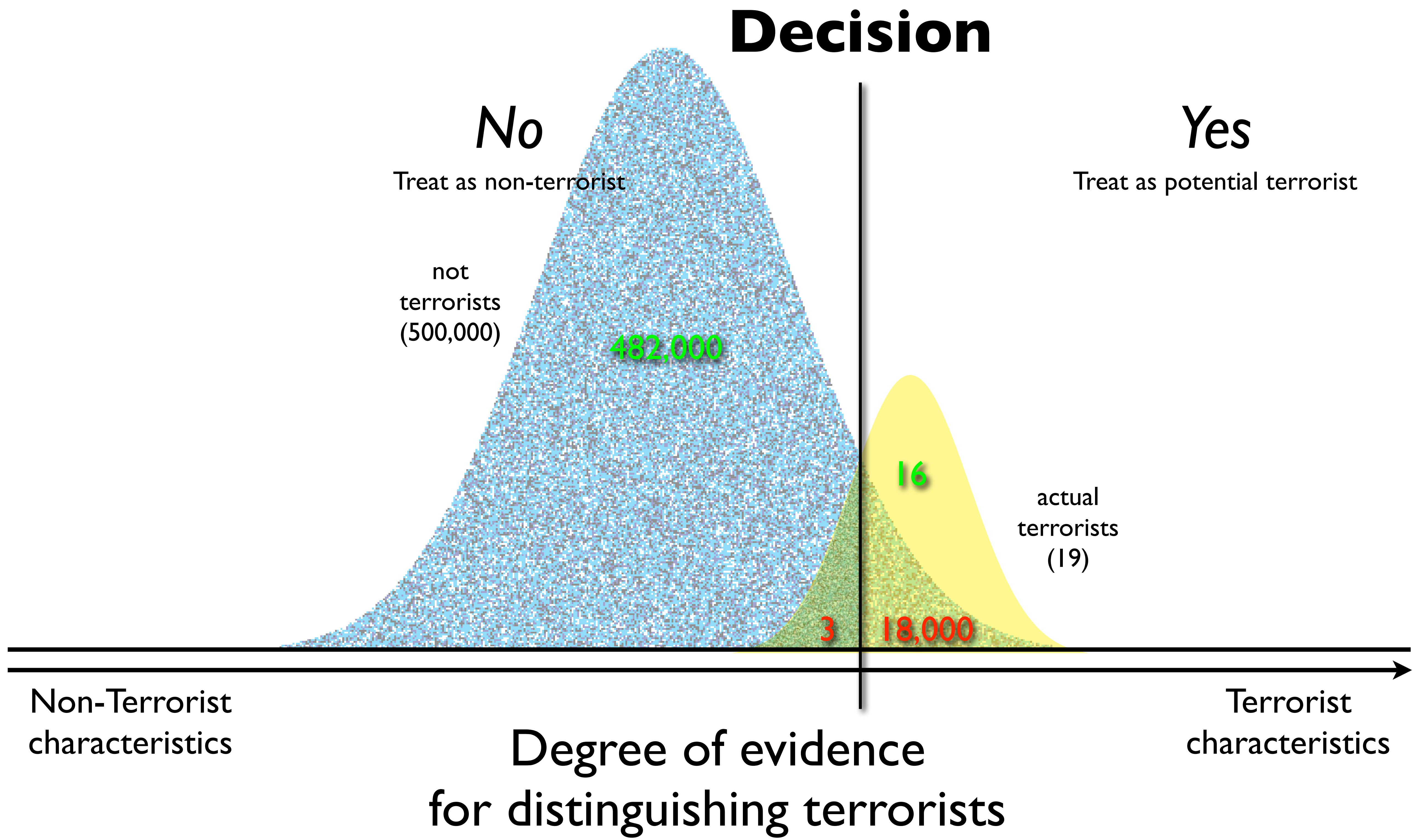
4. Spread (variance) of curves

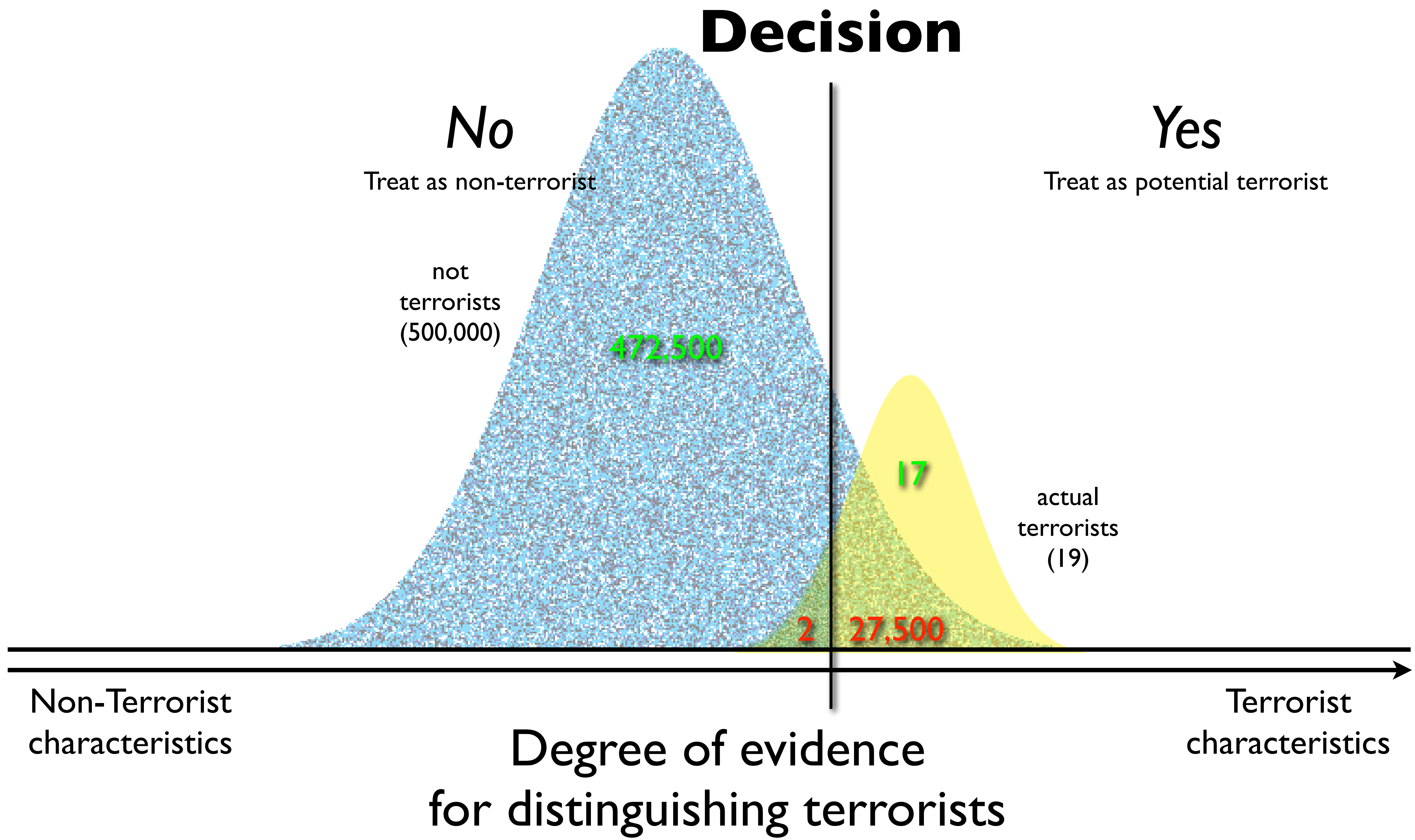


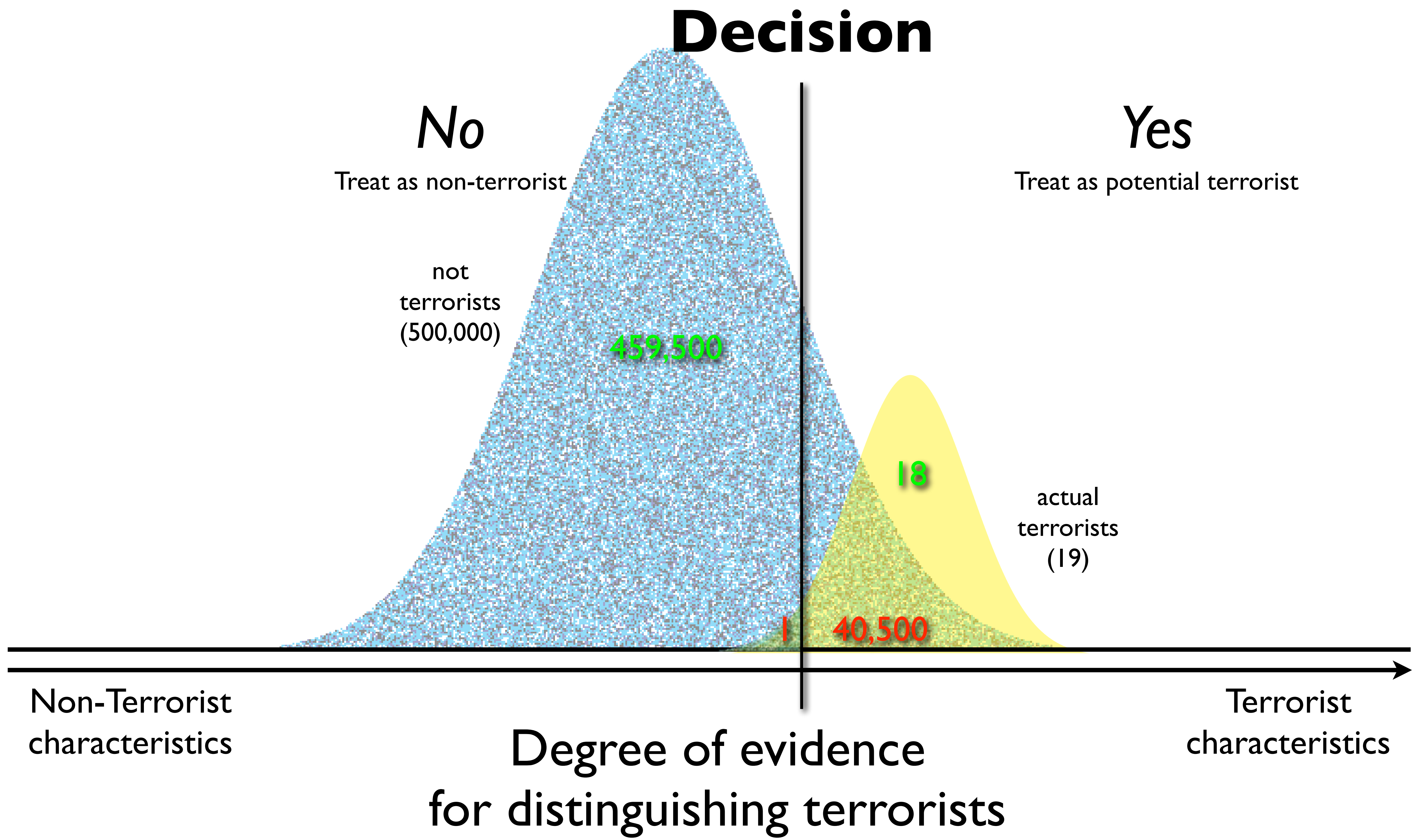
Why worry about this kind of representation?

To anticipate problems you are likely to have. For example: Is the effect you are looking for likely to be tough to detect; are you likely to get a lot of false alarms; can you anticipate having to set a very conservative criterion; is it likely you will have to look at a lot of cases?


Zone
E
O LIQUID
OR GELS
PERMITTED
BEYOND
SECURIT
NON, NEIN, GEEN, NÃO,
NO
ABSOLUTELY
NO
LIQUIDS
ALLOWED
TO INCLUDE:
ALL AEROSOLS, GELS,
TOILETRIES, SOAP, MOUTHWASH,
SHAVING CREAM, TOOTHPASTE,
PERFUME, COLOGNE, ETC.
POLICE


Decision
No
Treat as non-terrorist
Yes
Treat as potential terrorist
not
terrorists
(500,000)
482,000
16
actual
terrorists
(19)
3
18,000
Non-Terrorist
characteristics
Terrorist
characteristics
Degree of evidence
for distinguishing terrorists


Decision
No
Treat as non-terrorist
Yes
Treat as potential terrorist
not
terrorists
(500,000)
472,500
17
actual
terrorists
(19)
2
27,500
Non-Terrorist
characteristics
Terrorist
characteristics
Degree of evidence
for distinguishing terrorists



Decision
No
Treat as non-terrorist
Yes
Treat as potential terrorist
not
terrorists
(500,000)
459,500
18
actual
terrorists
(19)
1
40,500
Non-Terrorist
characteristics
Degree of evidence
for distinguishing terrorists
Terrorist
characteristics

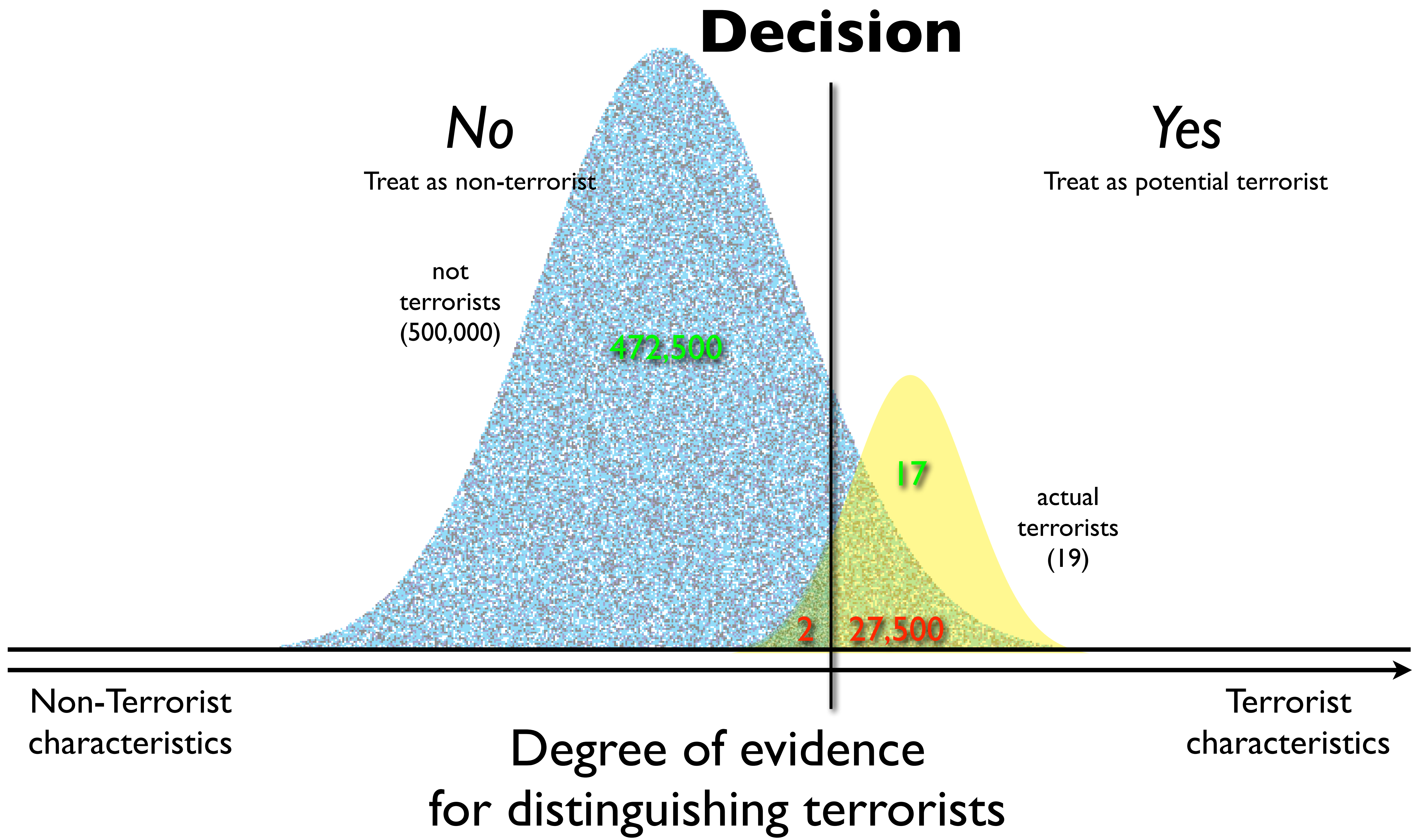
Anti-terror critics just don't get it, says Reid

Alan Travis, home affairs editor
Thursday August 10, 2006
The Guardian

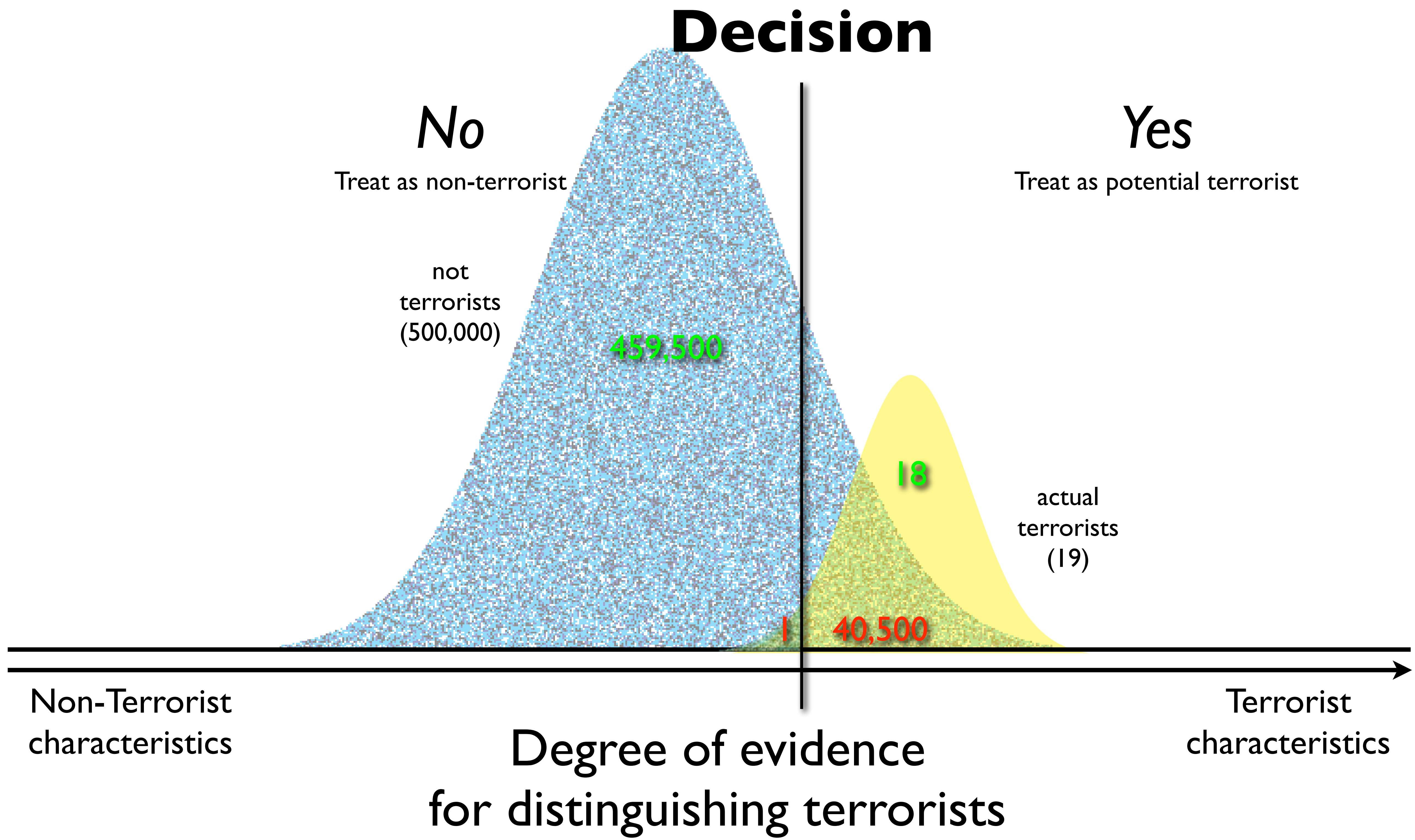
John Reid yesterday accused the government's anti-terror critics of putting national security at risk by their failure to recognise the serious nature of the threat facing Britain. "They just don't get it," he said.

The home secretary yesterday gave the thinktank Demos his strongest hint yet that a new round of anti-terror legislation is on the way this autumn by warning that traditional civil liberty arguments were not so much wrong as just made for another age.

"Sometimes we may have to modify some of our own freedoms in the short term in order to prevent their misuse and abuse by those who oppose our fundamental values and would destroy all of our freedoms in the modern world," he said.



Decision
No
Treat as non-terrorist
Yes
Treat as potential terrorist
not
terrorists
(500,000)
472,500
17
actual
terrorists
(19)
2
27,500
Non-Terrorist
characteristics
Terrorist
characteristics
Degree of evidence
for distinguishing terrorists



It's very easy to forget the high prior probability of a non-linear tradeoff: allowing at least a few errors of one kind will probably produce a big reduction in the number of the other type.

Research Questions

- Is the overall relationship statistically significant and how strong is the relationship?
- What variables are individually important in separating (discriminating) between the groups?

A simple example

2 group Discriminant Analysis

Two groups of inmates:

- Group 1 = convicted for murder
- Group 2 = convicted for fraud

Two measured variables:

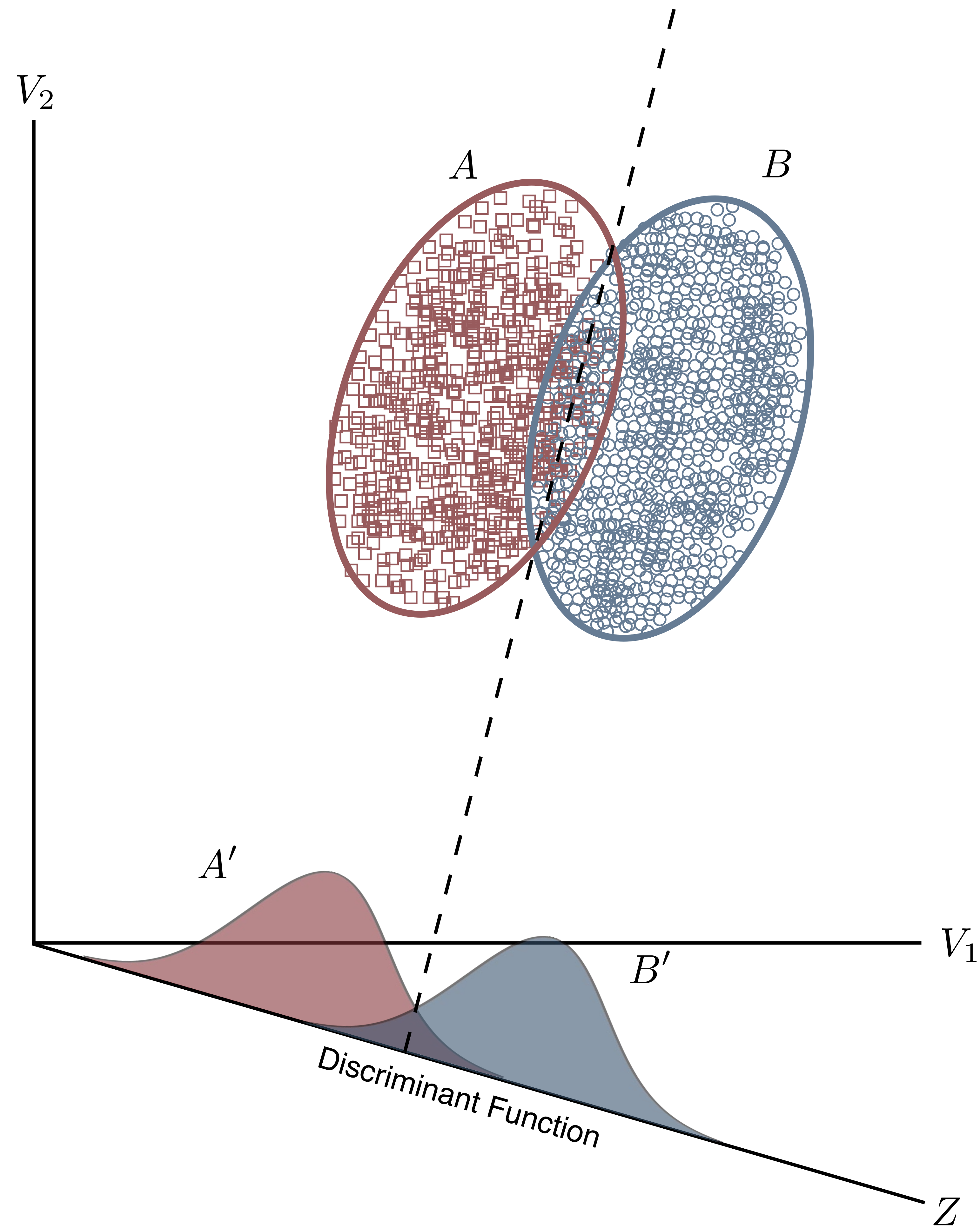
- a measure of intelligence (Y_1)
- a measure of aggression (Y_2)

$$Y_1 Y_2 \longleftarrow X$$

2 continuous variables

categorical 2 levels

		Group	
		Murderer	*Fraudster*
Classification	*“Murderer”*	Hit	False Alarm
	“Fraudster”	Miss	Correct Rejection



V_2
A
B
A'
V_1
B'
Discriminant Function
Z

Principal components analysis

- Purposes
- Motivational examples
- Design Issues
- Representing PCA
    - Logically: Euler Diagrams
    - Geometric: A vector representation
    - Schematic: A ‘boxes of data’ representation
    - Algebraic: A formulaic representation
    - Matrix: The Fundamental Equations
    - Schematic: The matrices linked

Purposes of principal components analysis and factor analysis

- To simplify a data set, by reducing multidimensional data to lower dimensions for analysis.
  - reduce a large number of variables to a smaller number with maximum spread among cases.
- To summarise patterns of intercorrelations among variables.
- To provide an operational definition for an unobserved, hypothetical construct using observed variables.
- To test a theory about the nature of the underlying variables.

What sort of questions are being investigated?

Distinctiveness, typicality, and recollective experience in face recognition: A principal components analysis

In this study, participants rated previously unseen faces on six dimensions: familiarity, distinctiveness, attractiveness, memorability, typicality, and resemblance to a familiar person. The faces were then presented again in a recognition test in which participants assigned their positive recognition decisions to either remember (R), know (K), or guess categories. On all dimensions except typicality, faces that were categorized as R responses were associated with significantly higher ratings than were faces categorized as K responses. Study ratings for R and K responses were then subjected to a principal components analysis. The factor loadings suggested that R responses were influenced primarily by the distinctiveness of faces, but K responses were influenced by moderate ratings on all six dimensions. These findings indicate that the structural features of a face influence the subjective experience of recognition.

Procrastination, a principal components analysis

The revised Eysenck Personality Questionnaire (EPQ), the Beck Depression Inventory, the Jenkins Activity Survey, and 3 time-usage measures constructed by the present authors were administered to 227 undergraduates who were chronic academic procrastinators. Three principal components were found, suggesting orthogonal personality variables associated with different types of procrastination (high EPQ psychoticism, neurotic extraverted, and depressed procrastination). Findings are discussed in terms of treatment for procrastinators.

What sort of questions are being investigated?

Perceived cognitive function is a major determinant of health related quality of life in a non-selected population of patients with coronary artery disease: A principal components analysis

Four independent principal factors representing perceived cognitive, physical, social and emotional functions underlying the patients' HRQL were found. Identical factors were recognized with an alternate technique. The major factor - explaining 43% of HRQL - was perceived cognitive function reflecting ability to concentrate, activity drive, memory and problem solving. Cognitive function correlated to EQ but not to CCS. Perceived physical function/general health explained 9% of HRQL and was as expected related both to EQ and CCS. Total CHP scores differed significantly to those of healthy controls. Conclusions: Perceived cognitive function seems to be a major determinant of HRQL in CAD patients. This, in addition to earlier reports of possible prognostic information of reduced cognitive function, would prompt us to propose that HRQL assessments should include questions aimed to assess cognitive function.

A new whole-mouth gustatory test procedure: Thresholds and principal components analysis in healthy men and women

Gustatory testing using the whole-mouth method was performed in 123 healthy young adult males and females. The average thresholds for detection and recognition of the 4 basic tastes were not greatly different from the normal thresholds previously reported in Japan. Results indicate that the whole-mouth gustatory test procedure employed in this study may be useful for evaluating gustatory function clinically. Principal components analysis confirmed that the sweet, salty, sour and bitter are indeed the four basic tastes and revealed that the sensation of taste is detected before the specific taste is identified.

A motivational example

Consider an investigation into the nature of intelligence. Data on six measures are collected:

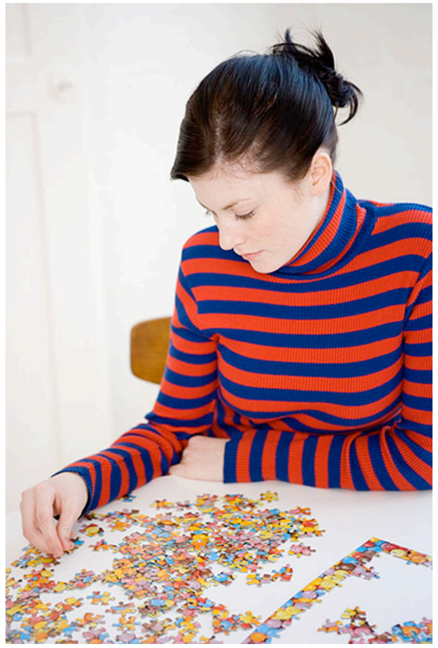
ability to recite song lyrics from memory

ability to hold two conversations at once

speed at completing crosswords

ability to assemble something from IKEA

ability to use a street directory



speed at completing jigsaw puzzles

What might be the 'underlying factors'?

Design Issues

- Selecting measures:
    - We want to sample to get a representative coverage of the conceptual domain. (e.g., a range of useful measures of what we mean by “intelligence”).
- Selecting participants:
    - We want to sample to get a representative coverage of participants. (e.g., a range of people that we would like to generalise to).
- Data collection method:
    - Self Report? Behavioural measures? Question wording? Response scales? (e.g., are these variables measuring what we expect them to measure?)

Design Issues

What makes a variable interesting or important?

1. What the variable is measuring; its meaning; the concept or construct it is pointing to.
2. That cases, (people), vary on that measure. If there is no variance then there is no information about differences between cases. Variance is a measure of the amount of information that the variable conveys.

By analogy, a factor will be important:

- If it is measuring something, and
- If it has a large variance.

Statistics and mathematics can do nothing about the first (1) because “data do not know where they come from”, but mathematics can work with variance and maximise the variance accounted for.

Variance is a big concept in principal components analysis and factor analysis.

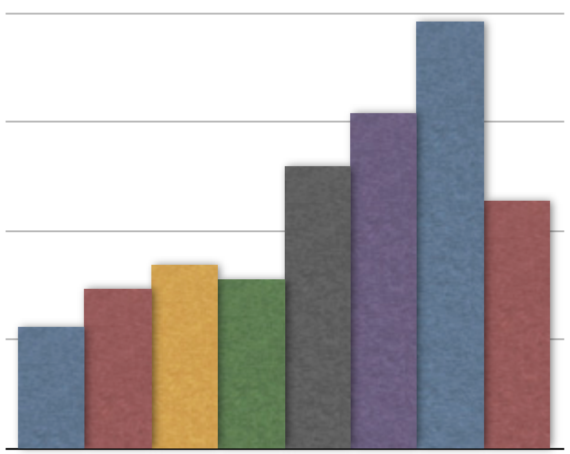
Measures that ‘define’ success...

Typing Speed | Emotional Stability | Chess Experience

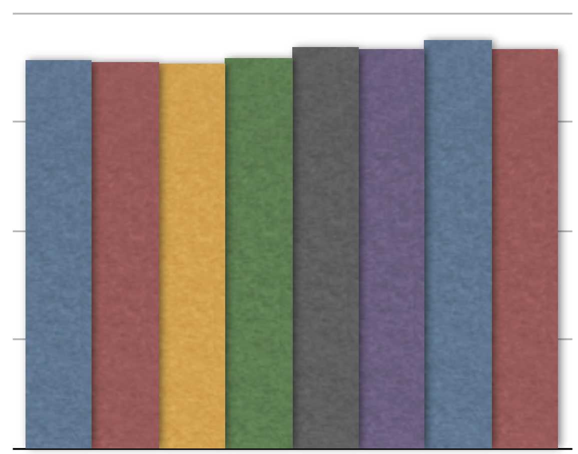
V_1 | V_2 | V_3

...but how do we know whether we have a ‘good’ measure?

One criterion for a ‘good’ variable is that is serves to distinguish between cases.



Good



Not so good

Recall from Lecture 2...

	Typing Speed	Emotional Stability	Chess Experience
	2	4	5
	1	7	2
	9	0	5
	6	2	4
	2	6	3
Variance	11.5	8.2	1.7

By computing the variance for each measure, the three measures may be correlated.

So the interpretations of the measures are not independent.

Another approach is to combine the three measures into a composite and compute the variance of the composite variable.

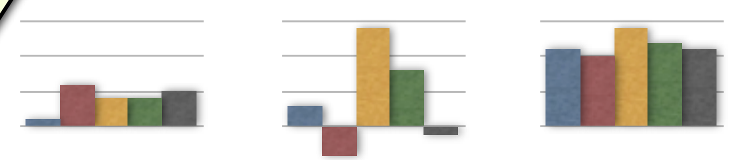
But how do we combine the scores?

	a_1	a_2	a_3
C_1	1	1	-1
C_2	1	-1	1
C_3	1	1	1

*Note that the variance of the linear composite can get large if we change the magnitude of the weights. So the weights are constrained so that their sums of squares are equal.

Recall from Lecture 2...

	Typing Speed	Emotional Stability	Chess Experience	C_1 $(1,1,-1)$	C_2 $(1,-1,1)$	C_3 $(1,1,1)$
	2	4	5	1	3	11
	1	7	2	6	-4	10
	9	0	5	4	14	14
	6	2	4	4	8	12
	2	6	3	5	-1	11
Variance	11.5	8.2	1.7	3.5	51.5	2.3



The goal here is to find the linear composite such that the scatter (spread) of the scores is a large as possible. That is, the linear composite has the largest possible variance. This gives the ‘most important factor’. The optimum weights depend essentially on the pattern of correlations among the variables.

Recall from Lecture 2...

Another motivational example

Consider an investigation into the nature of intelligence. Data on six measures are collected:

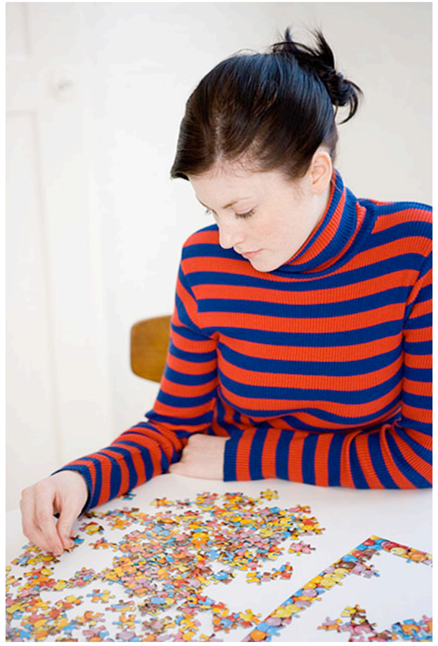
ability to recite song lyrics from memory

ability to hold two conversations at once

speed at completing crosswords

ability to assemble something from IKEA

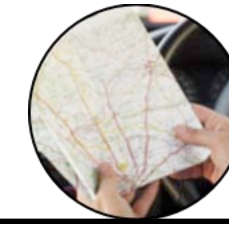
ability to use a street directory



speed at completing jigsaw puzzles

What might be the ‘underlying factors’?

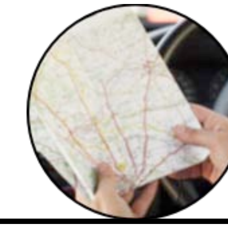
Correlations among six variables

						
	1.00					
	0.64	1.00				
	0.65	0.49	1.00			
	0.15	-0.04	-0.13	1.00		
	0.40	0.19	0.15	0.71	1.00	
	0.14	-0.01	-0.04	0.70	0.47	1.00

Look for patterns of high correlations that might reveal that there are not six independent ‘things’ being measured.

Note: Real data never look as clean as this!

Patterns in the correlations

						
	1.00					
	0.64	1.00				
	0.65	0.49	1.00			
	0.15	-0.04	-0.13	1.00		
	0.40	0.19	0.15	0.71	1.00	
	0.14	-0.01	-0.04	0.70	0.47	1.00

Another motivational example

Consider an investigation into the nature of intelligence. Data on six measures are collected:

ability to recite song lyrics from memory

ability to hold two conversations at once

speed at completing crosswords

ability to assemble something from IKEA

ability to use a street directory

speed at completing jigsaw puzzles

What might be the ‘underlying factors’?

Principal components analysis

- Purposes
- Motivational examples
- Design Issues
- Representing PCA
    - Logically: Euler Diagrams
    - Geometric: A vector representation
    - Schematic: A ‘boxes of data representation’
    - Algebraic: A formulaic representation
    - Matrix: The Fundamental Equations
    - Schematic: The matrices linked

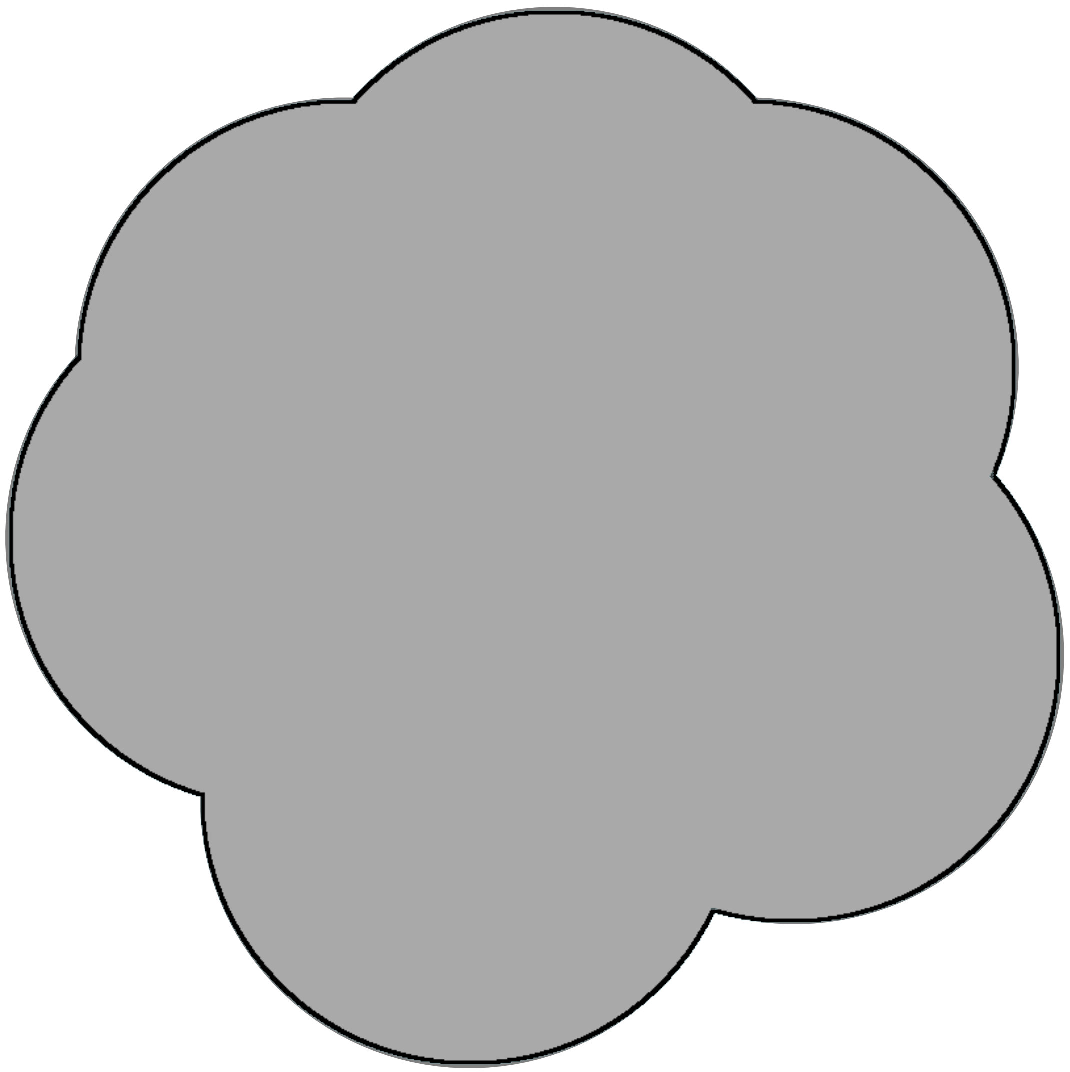
Representing PCA

Logically: Euler Diagrams

- Consider six variables that are intercorrelated.
  - Some more than others...
- The aim is to simplify our description of the information provided by the variables.
- A further aim may be to define the constructs which the variables describe.

Representing PCA

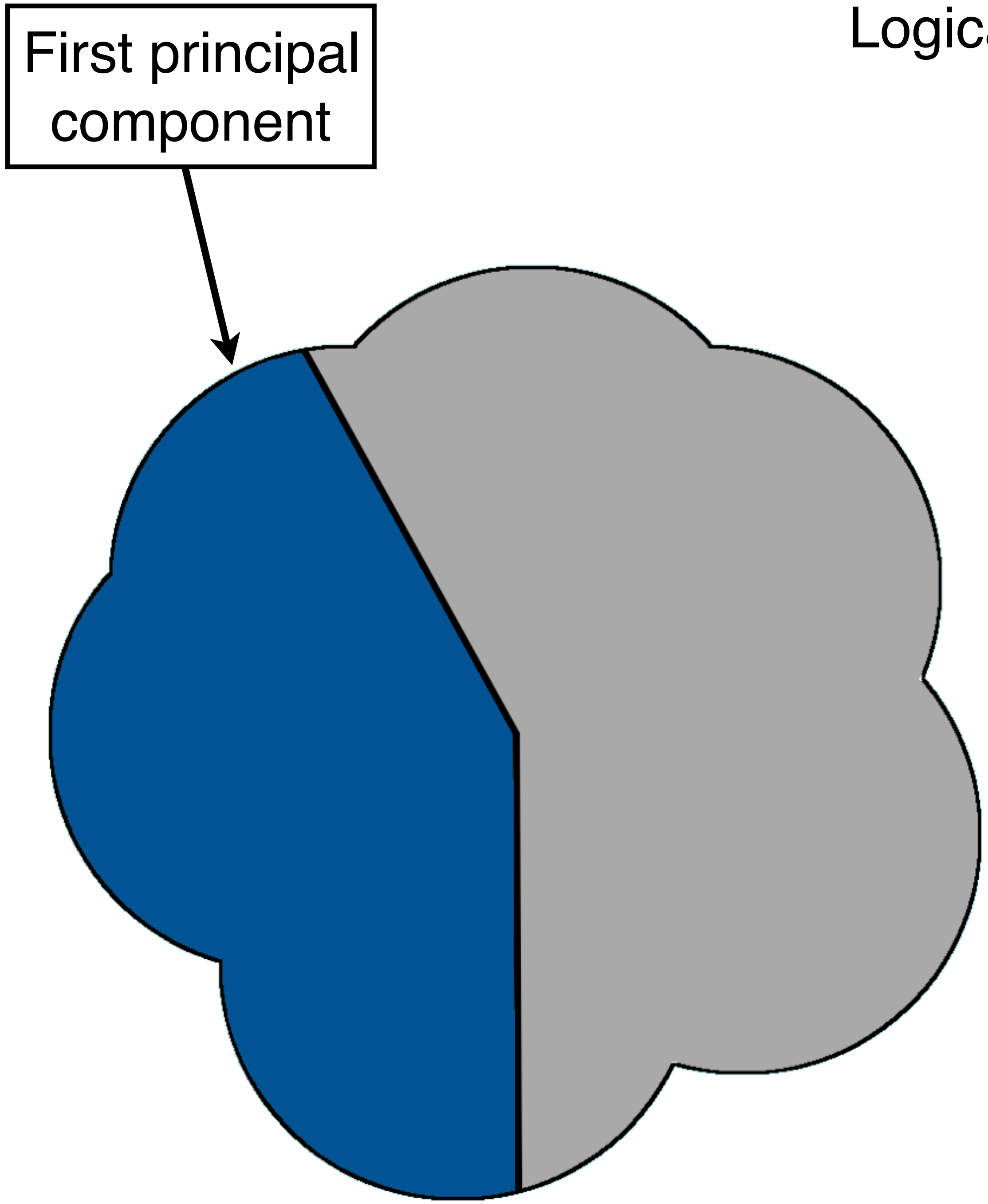
Logically: Euler Diagrams



- Each variable is set to have a variance of 1 (standardised), so the total variance of the six variables is 6.
- This total variance is subjected to a PCA.
- A linear composite is created...

Representing PCA

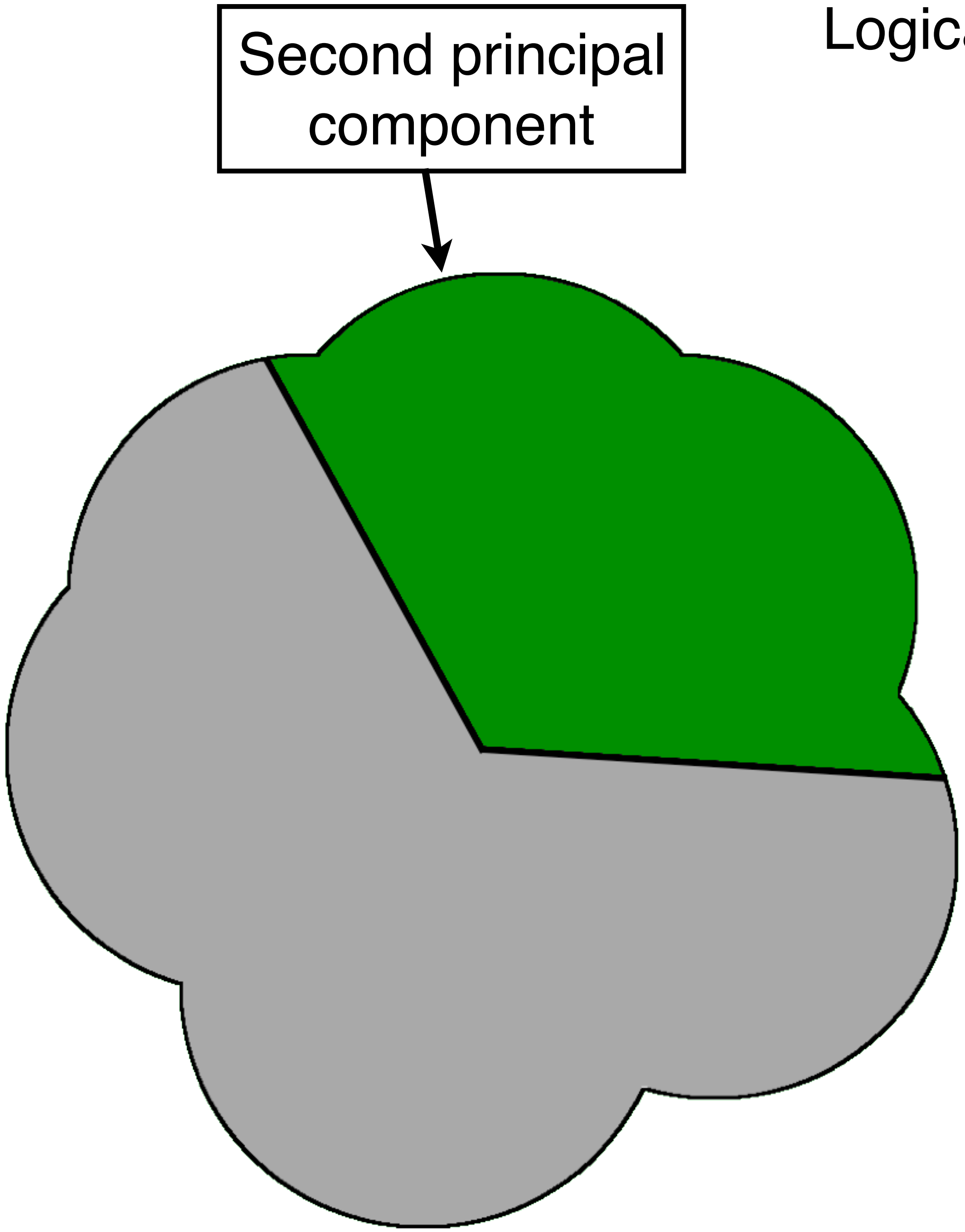
Logically: Euler Diagrams



- This first “principal component” is designed to account for as much as possible of the original variance
- It also represents the main “direction” that the variables are describing.

Representing PCA

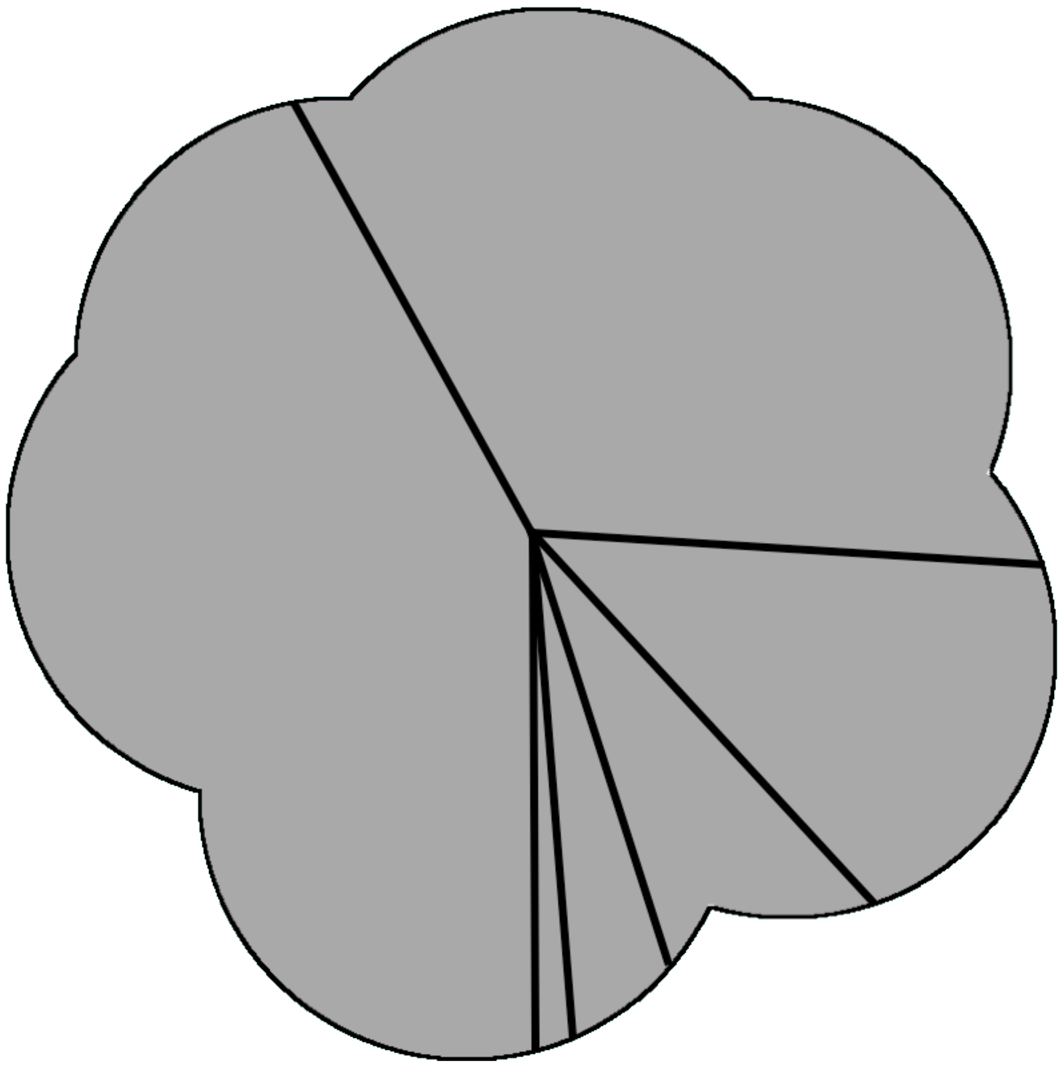
Logically: Euler Diagrams



- A second linear composite is created.
- This second principal component is designed to account for as much of the *remaining* variance as possible.
- This second principal component is uncorrelated with the first component.

Representing PCA

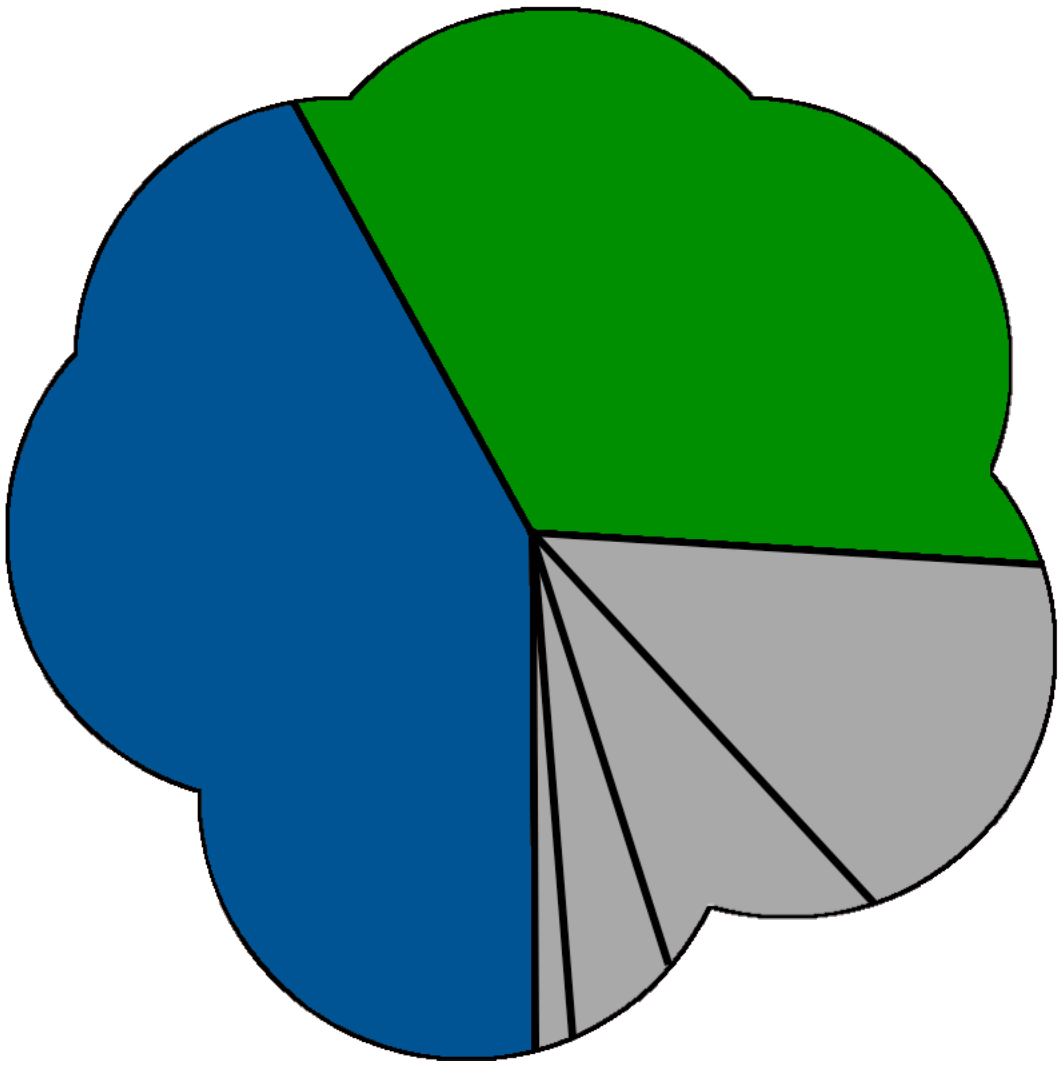
Logically: Euler Diagrams



- In total, six linear composites are created which, in combination, explain all of the original variance.
- The six correlated variables have been replaced with six uncorrelated linear composites.

Representing PCA

Logically: Euler Diagrams

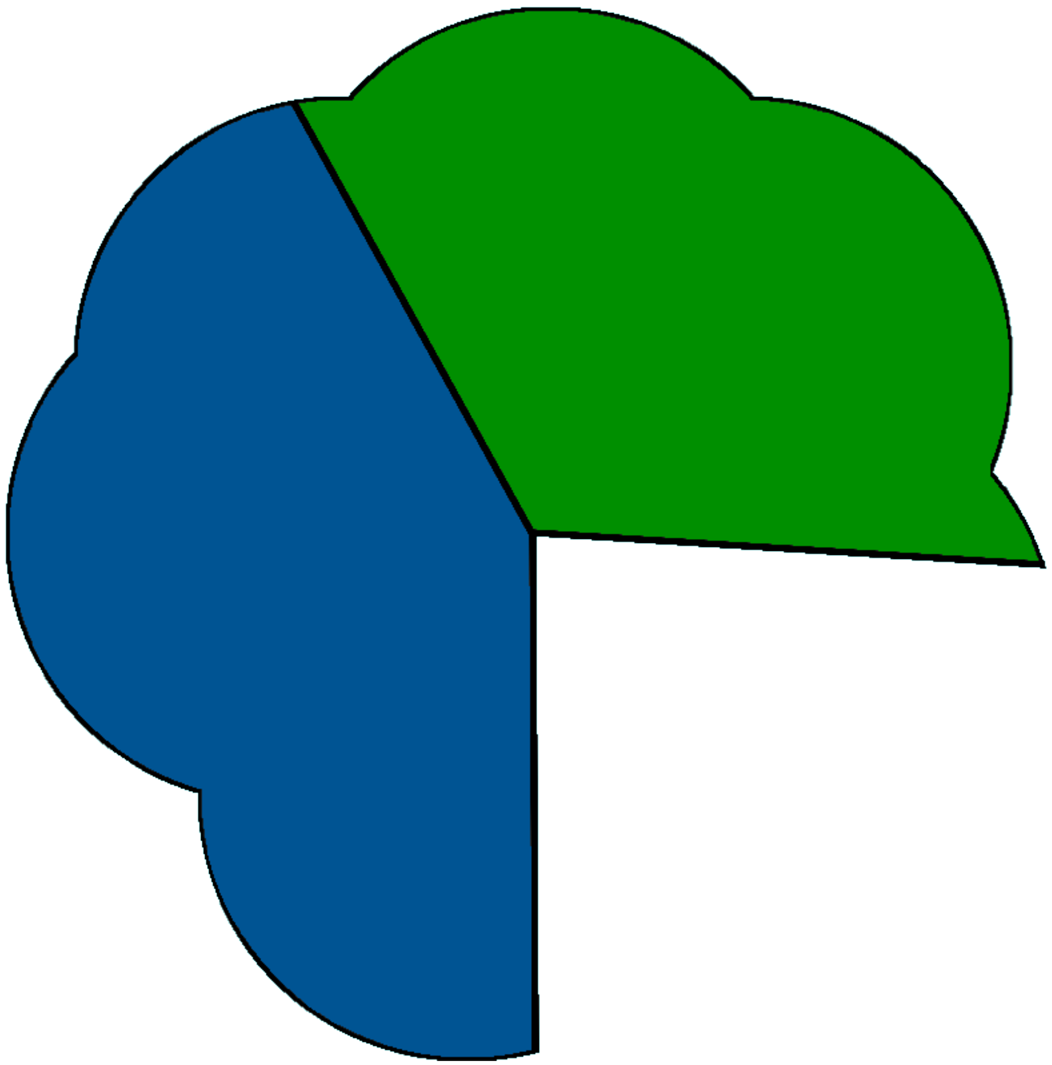


So what?

- Because of the way they are formed, earlier components contain more information than later components.
- In this example, the first two components explain 76% of the original variance.

Representing PCA

Logically: Euler Diagrams

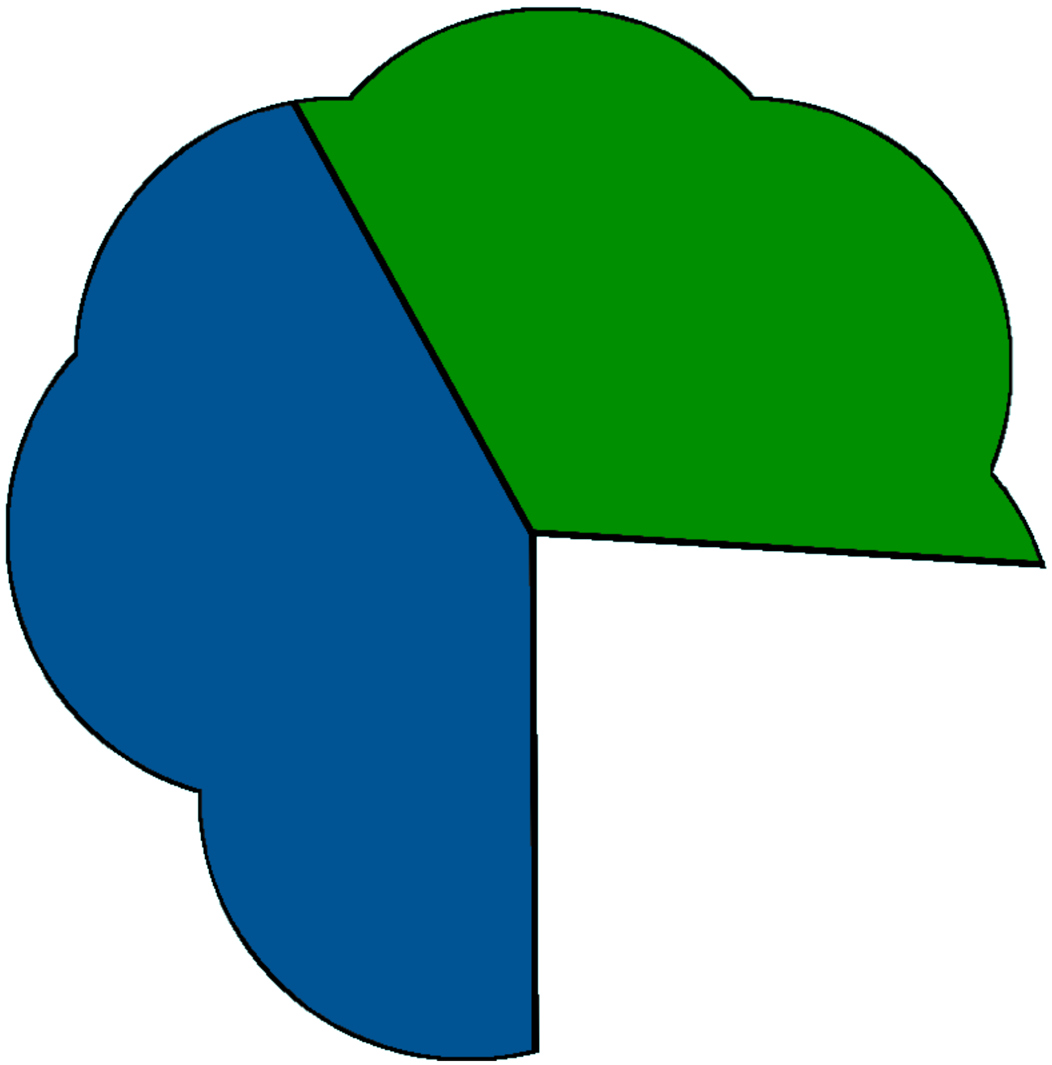


So what?

- So if we throw away the other four components, 24% of the original information is lost.
- But now we can discuss two variables instead of six.

Representing PCA

Logically: Euler Diagrams



So what?

- So we've simplified the description of the original variables (at the expense of some information).
- We've also defined two constructs (macro-variables) that describe (most of) the information in the original data.

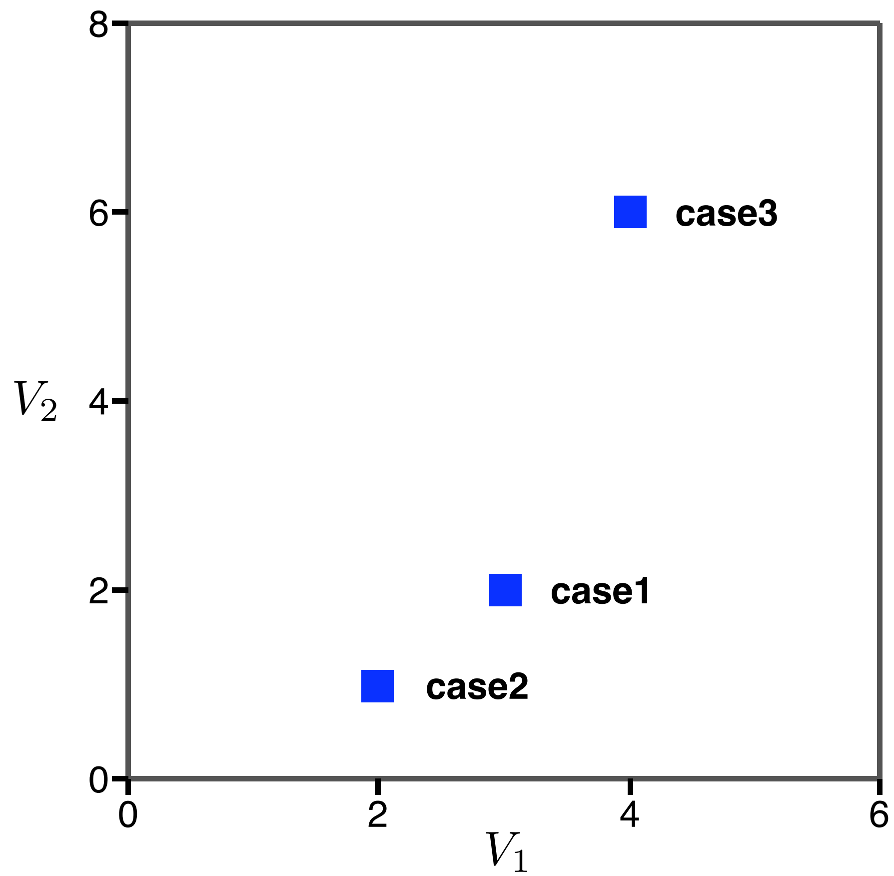
Principal components analysis

- Purposes
- Motivational examples
- Design Issues
- Representing PCA
    - Logically: Euler Diagrams
    - Geometric: A vector representation
    - Schematic: A 'boxes of data' representation
    - Algebraic: A formulaic representation
    - Matrix: The Fundamental Equations
    - Schematic: The matrices linked

Representing PCA

Geometric: A vector representation

Cases	V_1	V_2
1	3	2
2	2	1
3	4	6



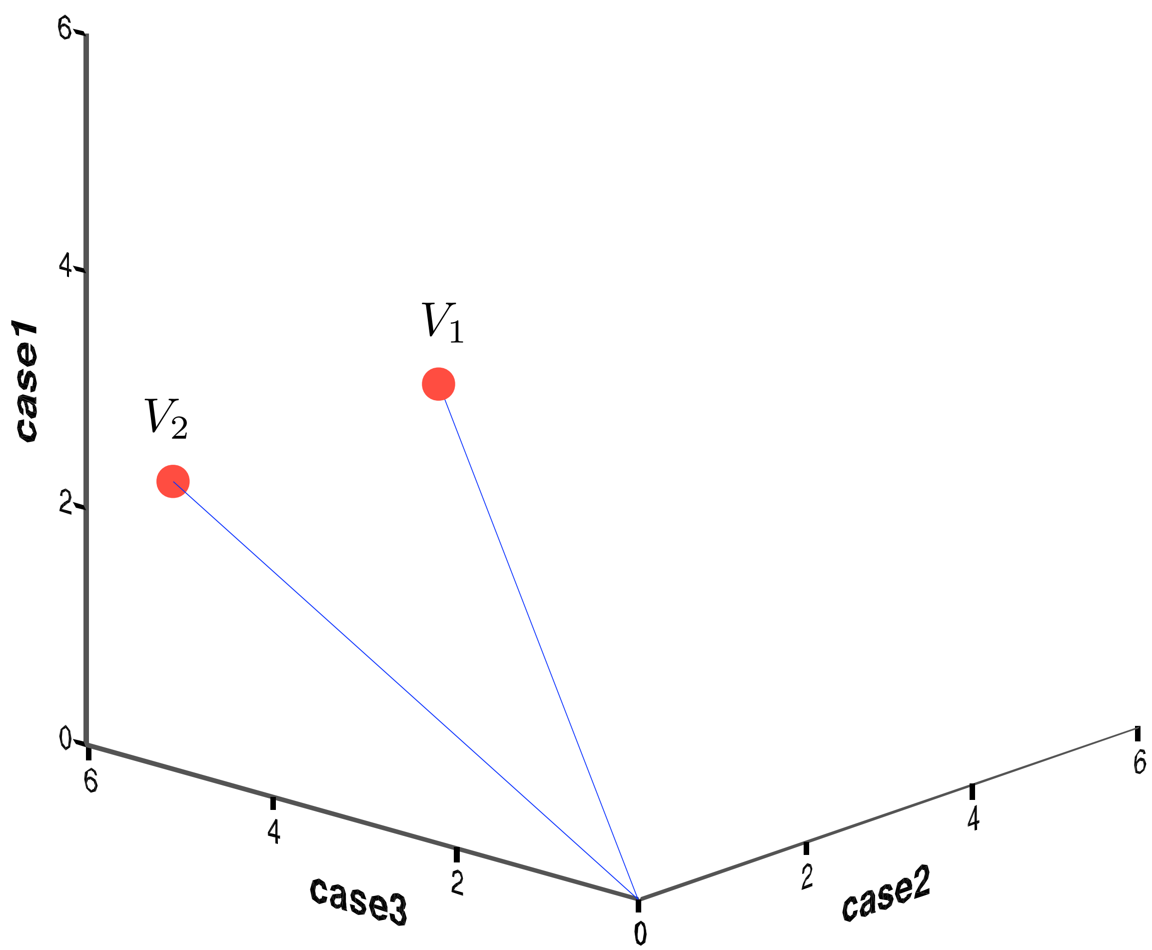
Cases plotted in 'variable space'

- Normally when we plot a scatterplot the variables define the axes and there is a point for each case. That is, we plot the cases in the space of the variables.

Representing PCA

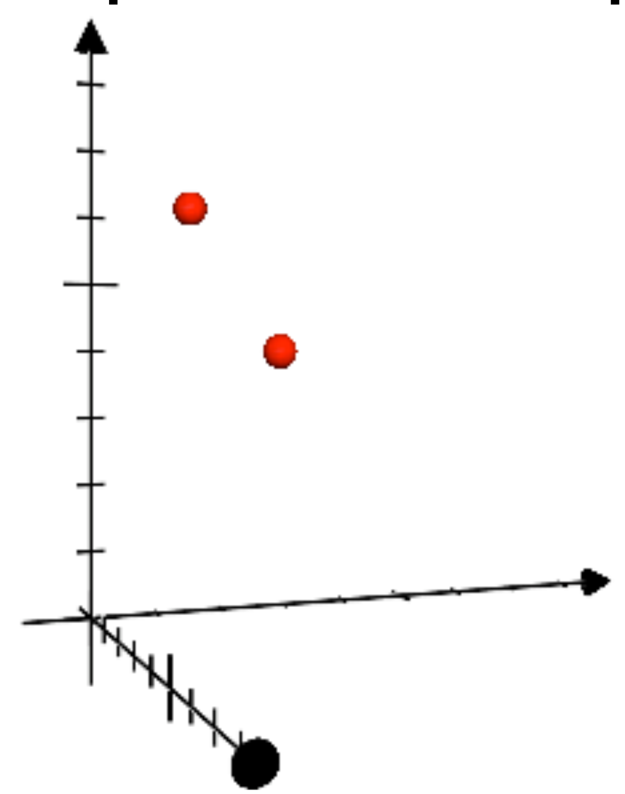
Geometric: A vector representation

Cases	V_1	V_2
1	3	2
2	2	1
3	4	6



Variables plotted in 'case space'

- We could also plot the variables in the space defined by cases. V_1 and V_2 are points in this space. The geometric definition of a vector is directional arrow of a given length coming from the origin of the space to the point.

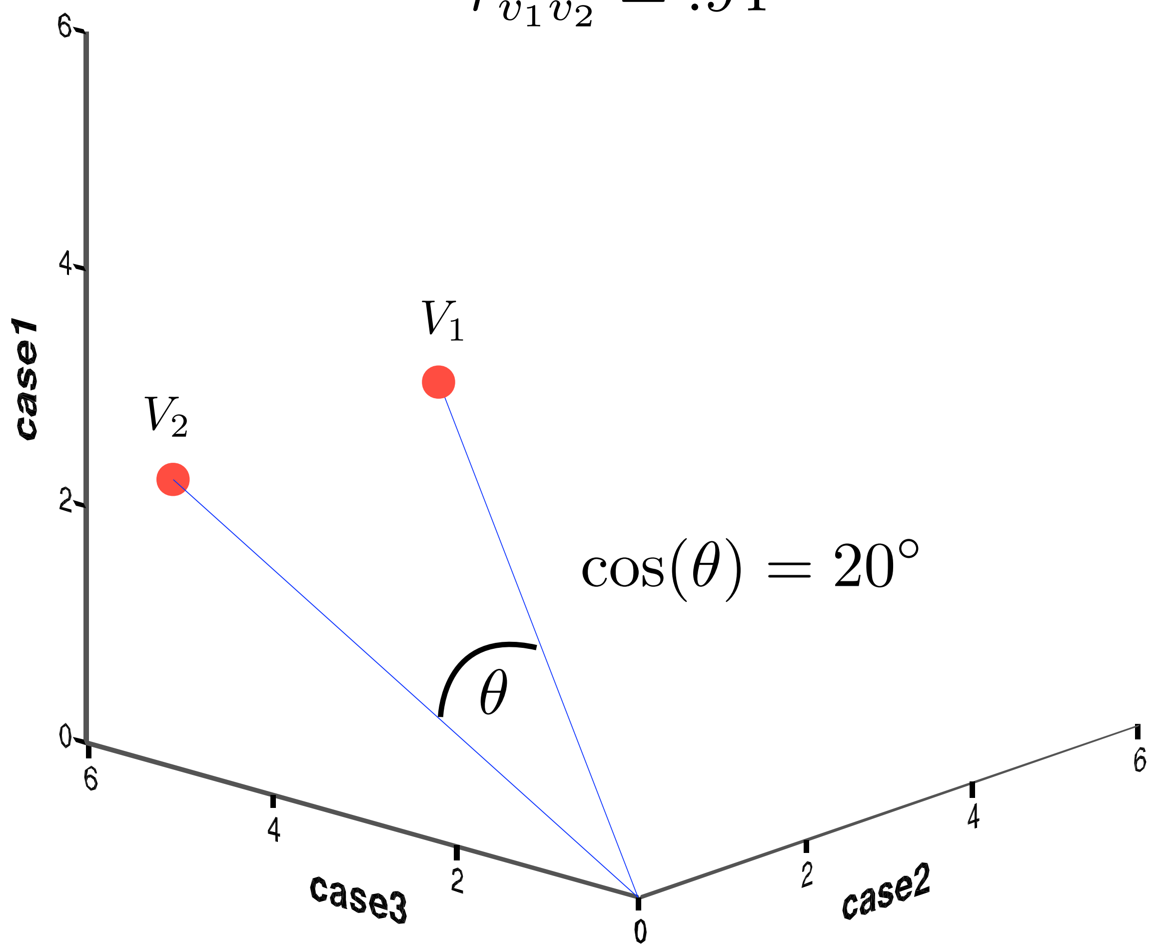


Representing PCA

Geometric: A vector representation

Cases	V_1	V_2
1	3	2
2	2	1
3	4	6

$r_{v_1 v_2} = .94$



- Even though the two vectors are plotted in three dimensional space only two dimensions are really needed to represent the two vectors.
  - Even if there were 100 cases and two variables, the vectors would be plotted in two dimensions.

$$\frac{180 \arccos(.94)}{\pi} = 20°$$

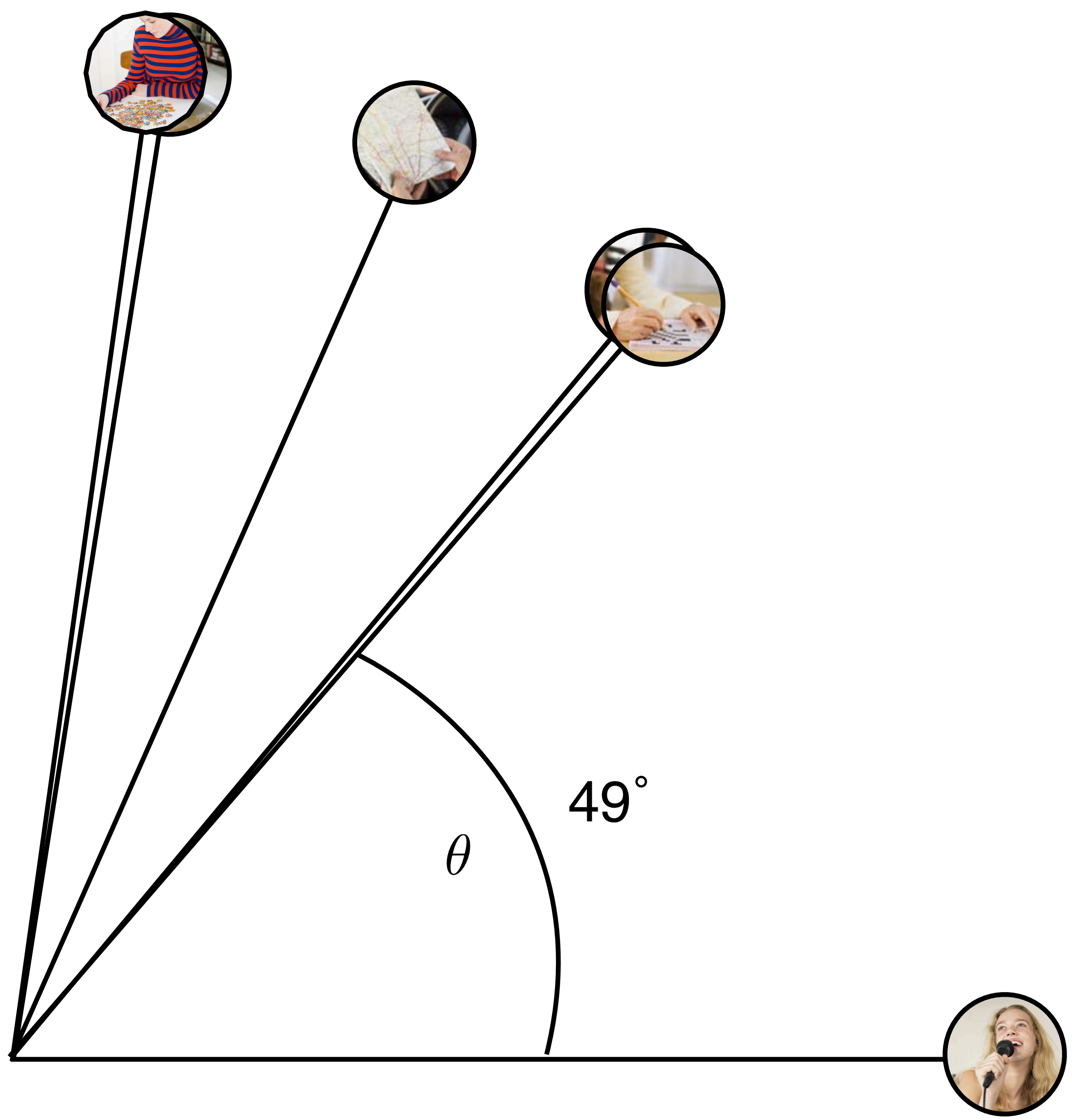
The angle between the two vectors 'measures' the correlation between the two variables. If two variables are perfectly correlated then they are coincident and the angle is 0. If the correlation is 0, the angle is 90° and the two vectors are orthogonal.

	1.00					
	0.64	1.00				
	0.65	0.49	1.00			
	0.15	-0.04	-0.13	1.00		
	0.40	0.19	0.15	0.71	1.00	
	0.14	-0.01	-0.04	0.70	0.47	1.00

r

	0					
	50	0				
	49	60	0			
	81	92	98	0		
	66	79	82	45	0	
	82	90	92	45	62	0

$\cos(\theta)$



There does appear to be two 'clusters' of variables. There seems to be two underlying factors in the variables. How can we find these factors? In this simple case we would be tempted to just draw them through the centres of the 'clusters' of variables. In real data, the patterns are not that clear and we need a technique to find the patterns. One such technique is principal components analysis.

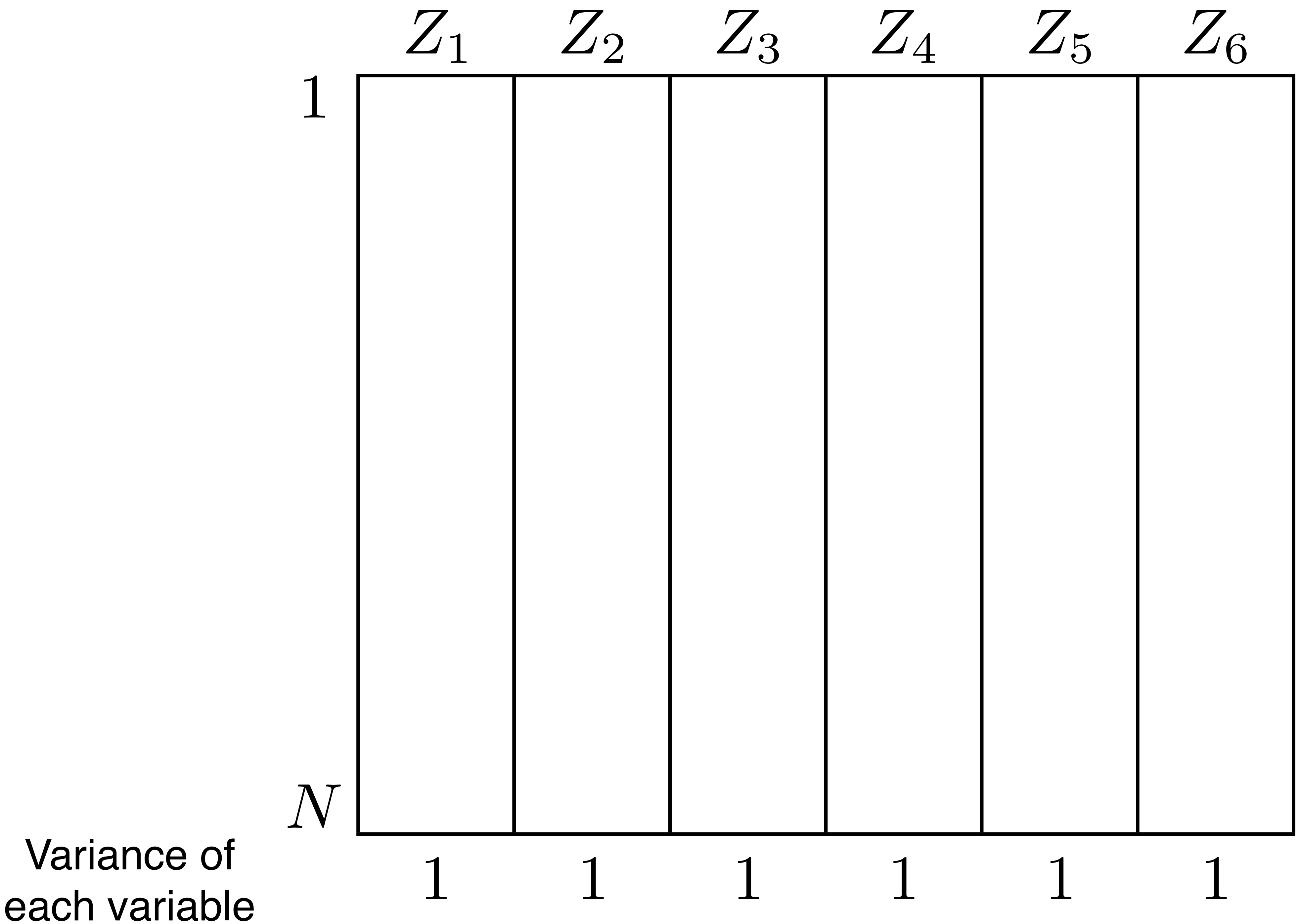
Principal components analysis

- Purposes
- Motivational examples
- Design Issues
- Representing PCA
    - Logically: Euler Diagrams
    - Geometric: A vector representation
    - Schematic: A ‘boxes of data’ representation
    - Algebraic: A formulaic representation
    - Matrix: The Fundamental Equations
    - Schematic: The matrices linked

Representing PCA

Schematic: A ‘boxes of data’ representation

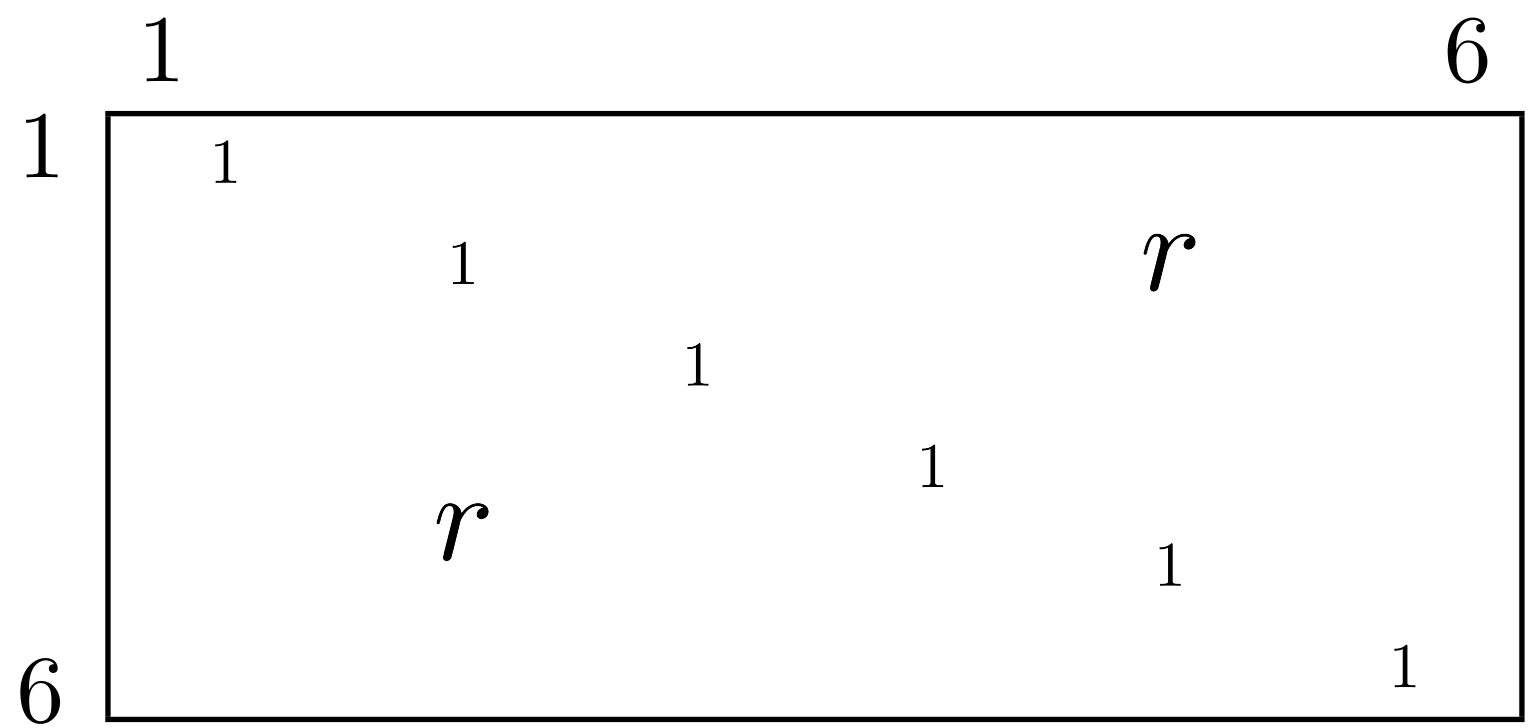
Original Correlated Variables (standardised)



Representing PCA

Schematic: A ‘boxes of data’ representation

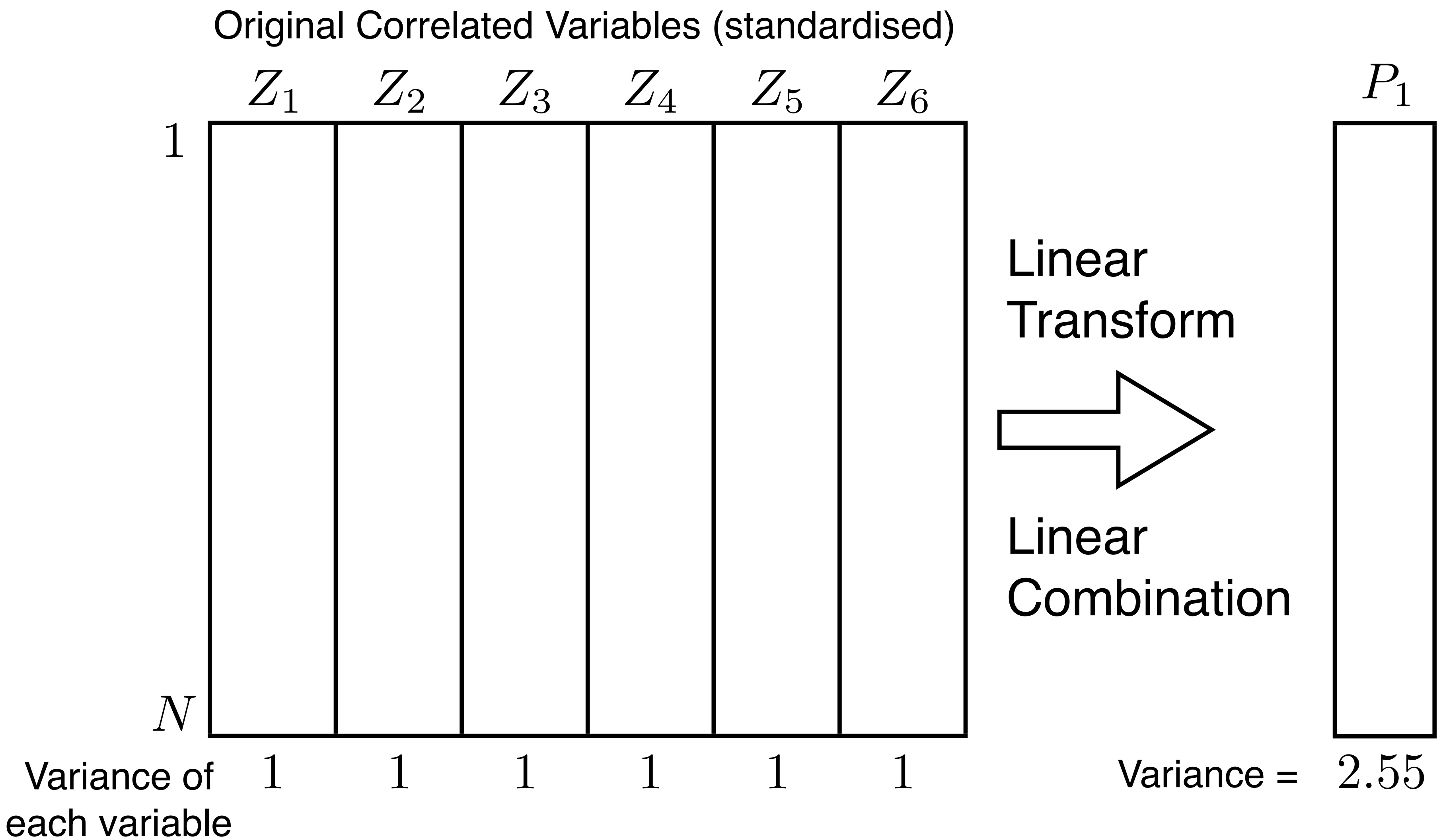
From raw data to correlation matrix
(variance-covariance matrix)



Total variance = 6

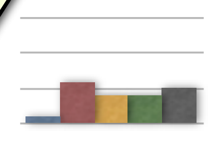
Representing PCA

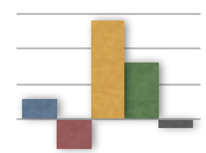
Schematic: A ‘boxes of data’ representation

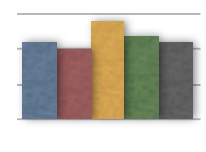


Find a linear combination of the six variables that has the maximum variance.

	Typing Speed	Emotional Stability	Chess Experience	C_1 $(1,1,-1)$	C_2 $(1,-1,1)$	C_3 $(1,1,1)$
	2	4	5	1	3	11
	1	7	2	6	-4	10
	9	0	5	4	14	14
	6	2	4	4	8	12
	2	6	3	5	-1	11
Variance	11.5	8.2	1.7	3.5	51.5	2.3





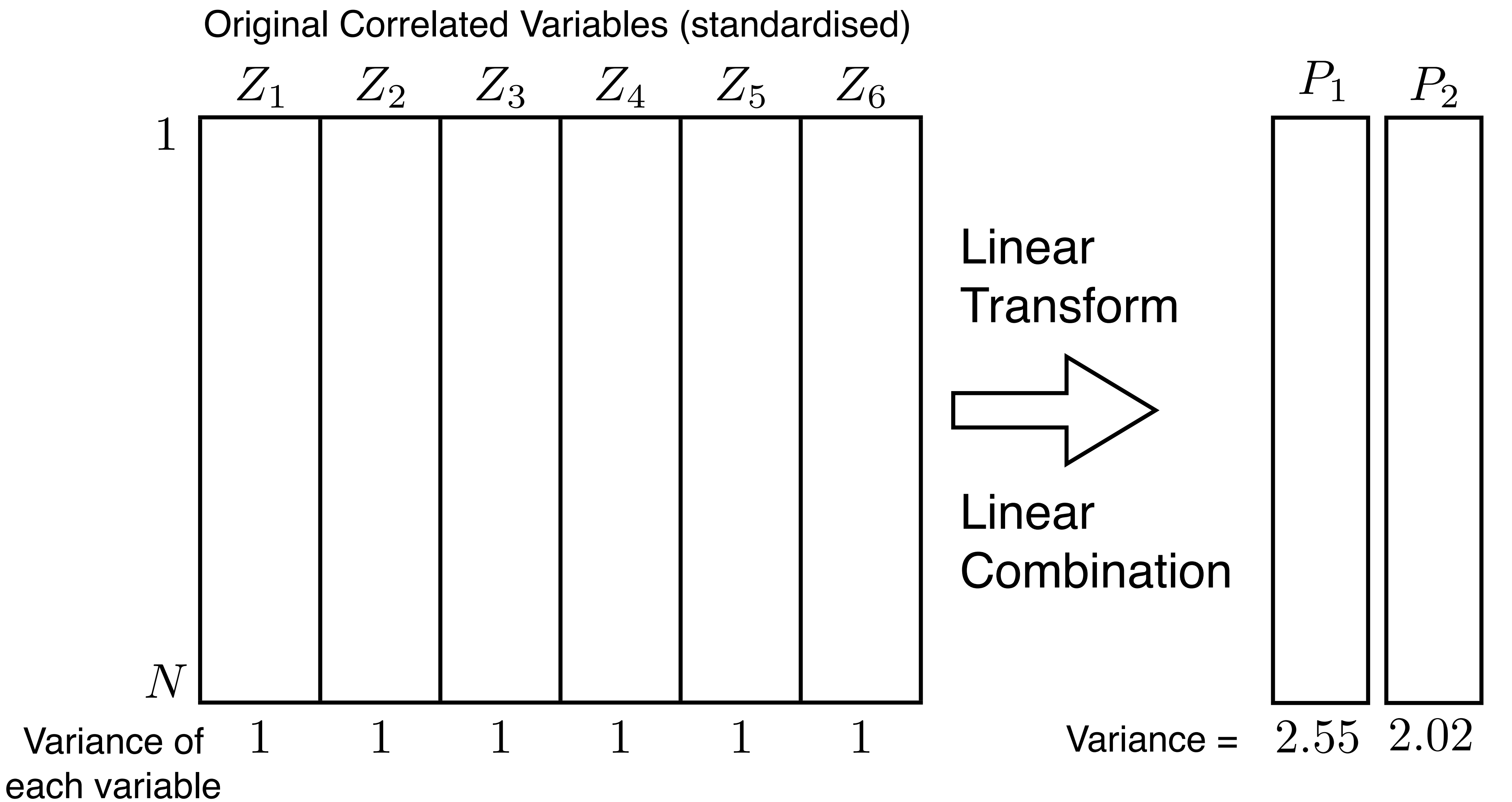


The goal here is to find the linear composite such that the scatter (spread) of the scores is a large as possible. That is, the linear composite has the largest possible variance. This gives the 'most important factor'. The optimum weights depend essentially on the pattern of correlations among the variables.

Recall from Lecture 2...

Representing PCA

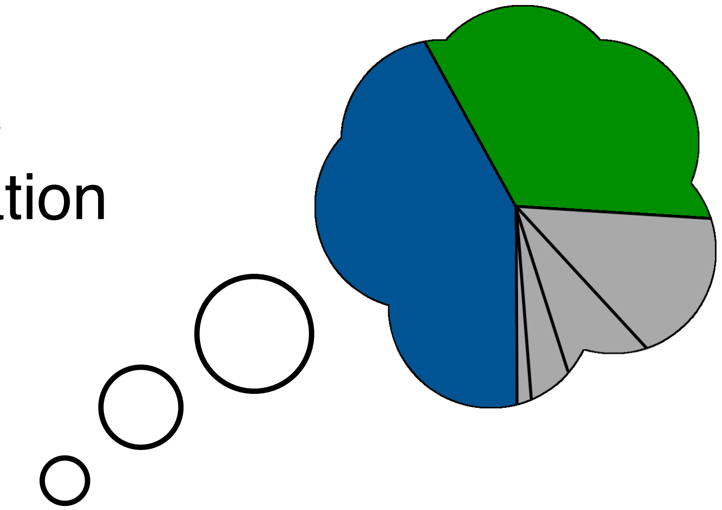
Schematic: A ‘boxes of data’ representation

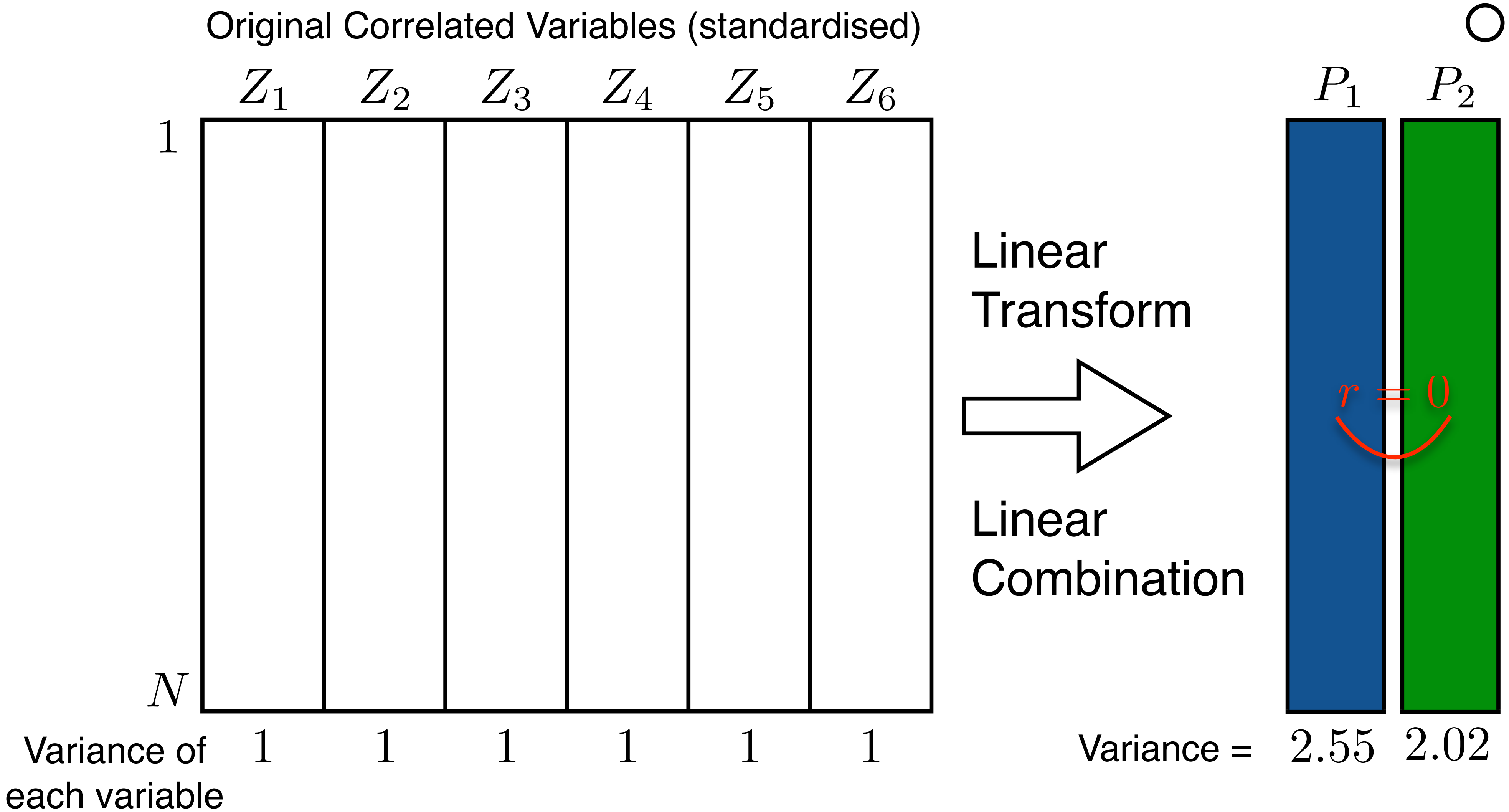


Find a second linear combination, uncorrelated (at right angles) with the first, that has a maximum of the residual variance.

Representing PCA

Schematic: A ‘boxes of data’ representation

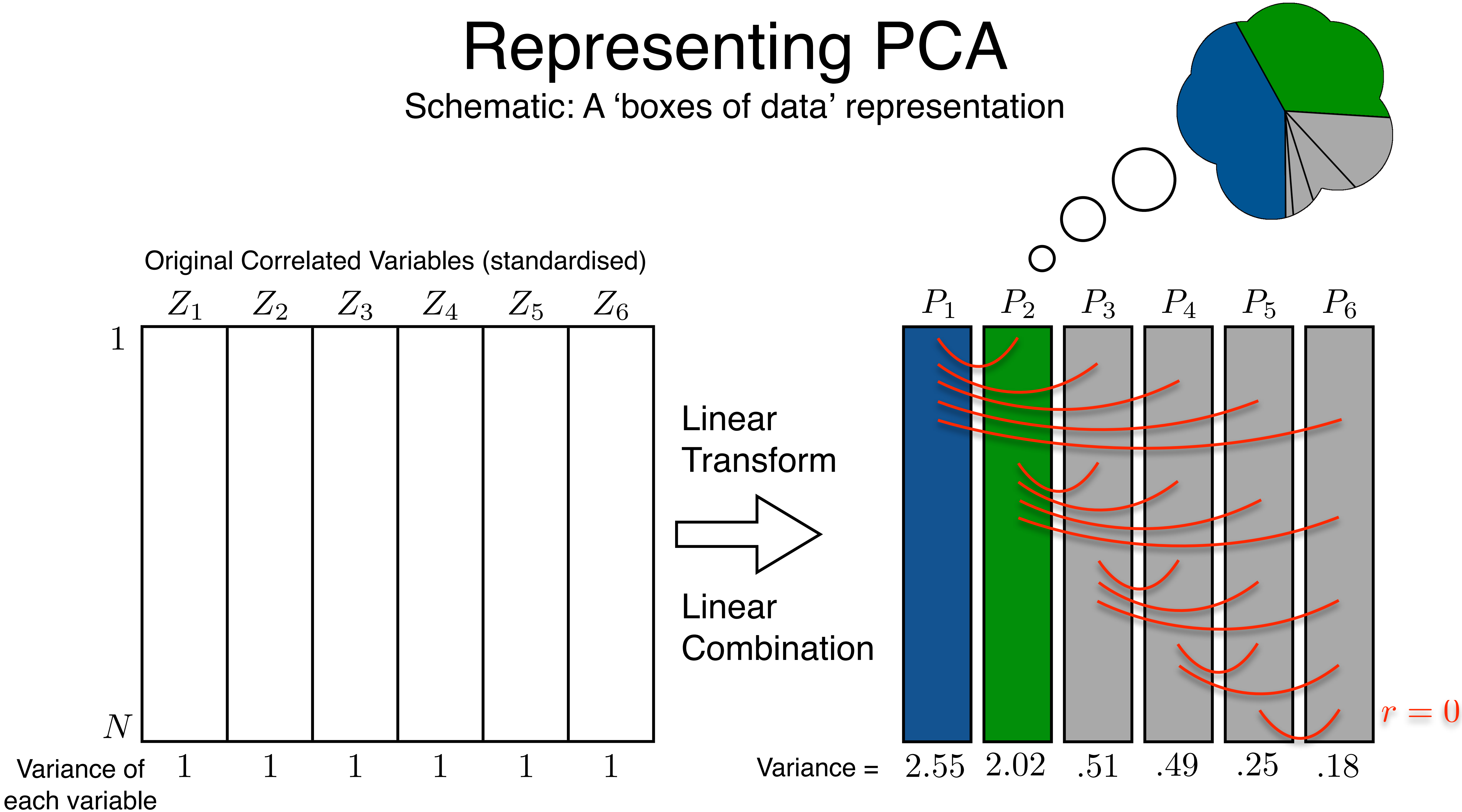




Find a second linear combination, uncorrelated (at right angles) with the first, that has a maximum of the residual variance.

Representing PCA

Schematic: A ‘boxes of data’ representation



Find a second linear combination,
uncorrelated (at right angles) with the first,
that has a maximum of the residual variance.

Representing PCA

Schematic: A ‘boxes of data’ representation

Variance-covariance matrix of the six new uncorrelated variables

$$\begin{array}{cc} & \begin{array}{cccccc}1 & & & & & 6\end{array} \\ \begin{array}{c}1\\ \\ \\ \\ \\ 6\end{array} & \begin{bmatrix}2.55 & & & & & \\ & 2.02 & & & zero & \\ & & .51 & & & \\ & & & .49 & & \\ & zero & & & .25 & \\ & & & & & .18\end{bmatrix}\end{array}$$

Total variance = 6

Representing PCA

Schematic: A ‘boxes of data’ representation

Summary

- A full PCA transforms a set of correlated measured variables into a set of uncorrelated variables (linear combinations).
- These are new composite scores or synthetic variables.
- We can use this if we know:
  - How many dimensions are needed to adequately represent the information in the original variables.
  - How to interpret the linear combination.

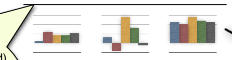
Principal components analysis

- Purposes
- Motivational examples
- Design Issues
- Representing PCA
  - Logically: Euler Diagrams
  - Geometric: A vector representation
  - Schematic: A ‘boxes of data’ representation
  - Algebraic: A formulaic representation
  - Matrix: The Fundamental Equations
  - Schematic: The matrices linked

Representing PCA

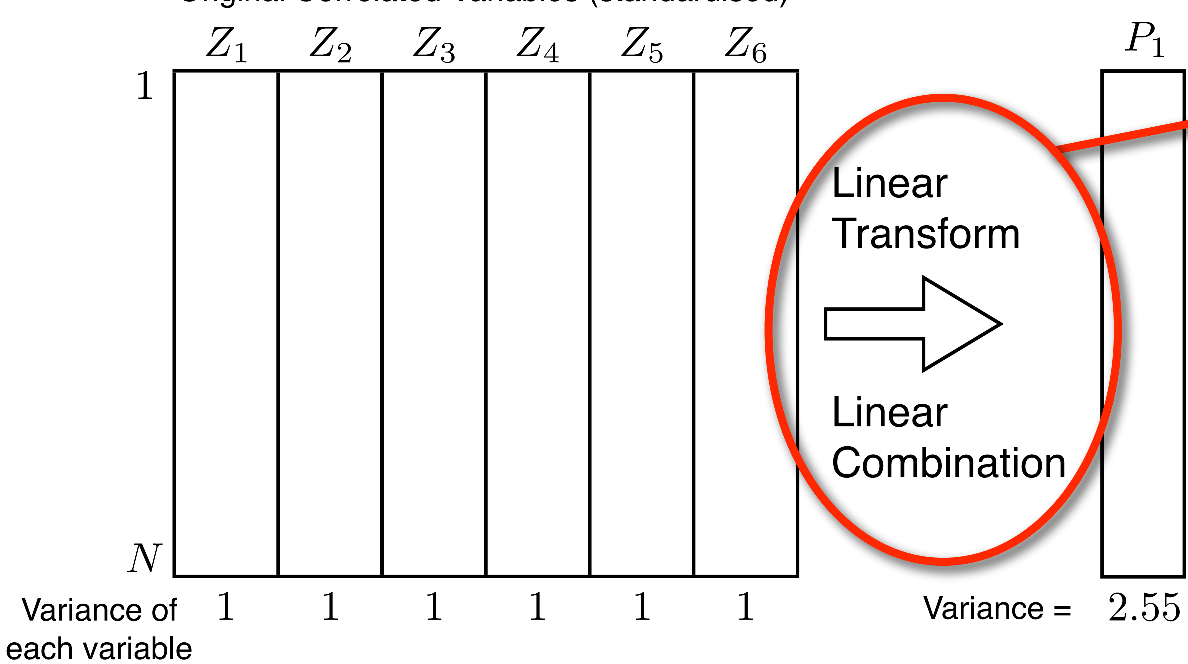
Algebraic: A formulaic representation

	Typing Speed	Emotional Stability	Chess Experience	C_1 $(1, 1, -1)$	C_2 $(1, -1, 1)$	C_3 $(1, 1, 1)$
	2	4	5	1	3	11
	1	7	2	6	-4	10
	9	0	5	4	14	14
	6	2	4	4	8	12
	2	6	3	5	-1	11
Variance	11.5	8.2	1.7	3.5	51.5	2.3



The goal here is to find the linear composite such that the scatter (spread) of the scores is a large as possible. That is, the linear composite has the largest possible variance. This gives the 'most important factor'. The optimum weights depend essentially on the pattern of correlations among the variables.

Recall from Lecture 2...



We need to fill this out a bit.

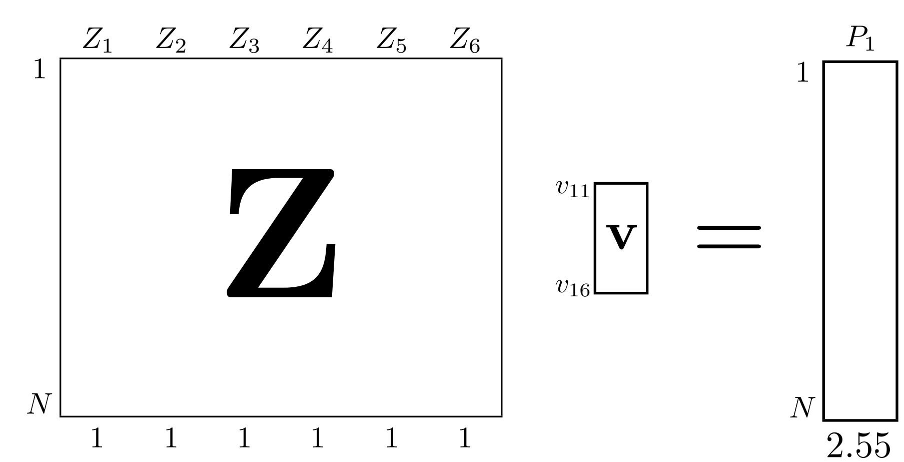
$$P_1 = v_{11}Z_{i1} + v_{21}Z_{i2} + \cdots + v_{61}Z_{i6}$$

$$P_1 = \sum v_{j1}Z_{ij}$$

$$P_1 = Zv_1$$

Find v_1 so that $var(P_1)$ is a maximum

$$var(P_1) = \mathbf{v}_1'\mathbf{R}\mathbf{v}_1$$



constraint: $\mathbf{v}_1'\mathbf{v}_1 = 1$ ← the weights are normalised so the variance can't be made arbitrarily large.

Representing PCA

Algebraic: A formulaic representation

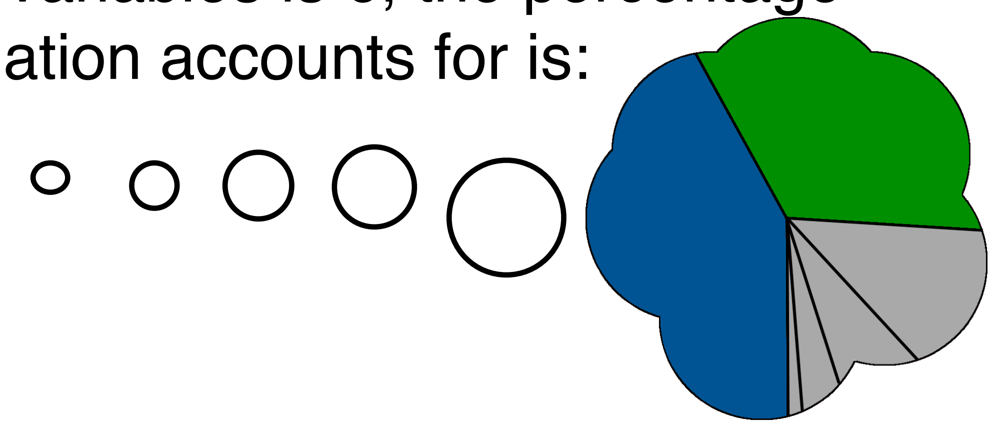
Goal: Find v_1 so that $var(P_1)$ is a maximum

This leads to an eigen equation (see later). In this case, the weights that maximise the variance of P_1 are:

$$\mathbf{v_1} = \begin{bmatrix} .48 \\ .35 \\ .32 \\ .41 \\ .50 \\ .37 \end{bmatrix}$$

This is the first eigenvector. The first eigenvalue is the $var(P_1) = 2.55$, and since the total variance in the six variables is 6, the percentage of variance that the first linear combination accounts for is:

$$\frac{100 \times 2.55}{6} = 42.5\%$$



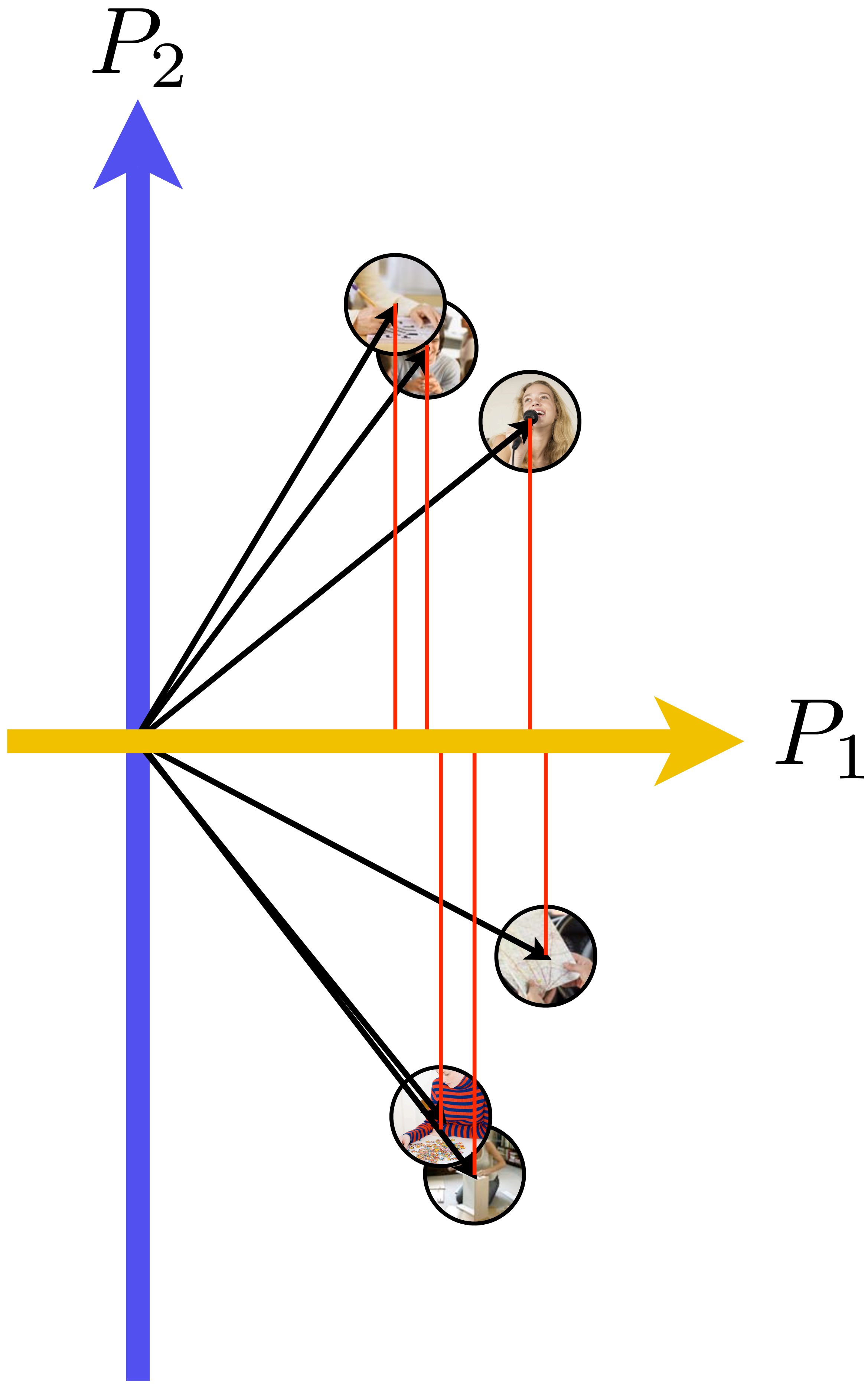
$$\mathbf{v_2} = \begin{bmatrix} -.35 \\ -.43 \\ -.48 \\ .48 \\ .24 \\ .42 \end{bmatrix}$$

This is the second eigenvector. The 2nd eigenvalue is $var(P_2) = 2.02$, and since the total variance in the six variables is 6, the percentage of variance that the second linear combination accounts for is:

$$\frac{100 \times 2.02}{6} = 33.7\%$$

The variance accounted for by both principal components is:

$$\frac{100 \times (2.55 + 2.02)}{6} = 76.2\%$$



- The first principal component finds the direction in which all the variables seem to be pointing.
  - The sums of squares of the projections of the endpoints of the vectors onto the principal direction is the amount of variance of the variables accounted for by that direction.

- The second component is at right angles (uncorrelated) with the first.
- From geometry we see that the linear composites are the new orthogonal directions in the space of the variables.

...a brief aside...

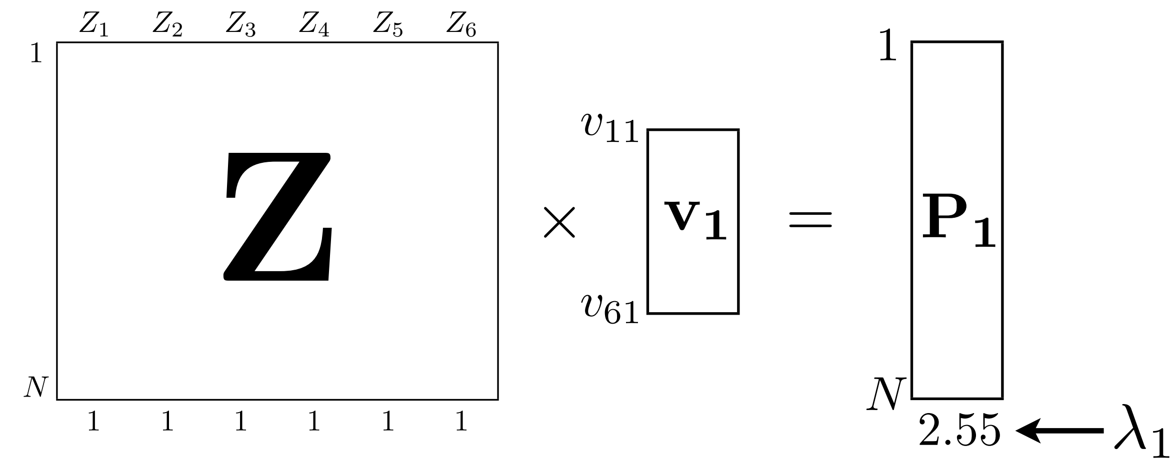
Principal components analysis

- Purposes
- Motivational examples
- Design Issues
- Representing PCA
  - Logically: Euler Diagrams
  - Geometric: A vector representation
  - Schematic: A ‘boxes of data’ representation
  - Algebraic: A formulaic representation
  - Matrix: The Fundamental Equations
  - Schematic: The matrices linked

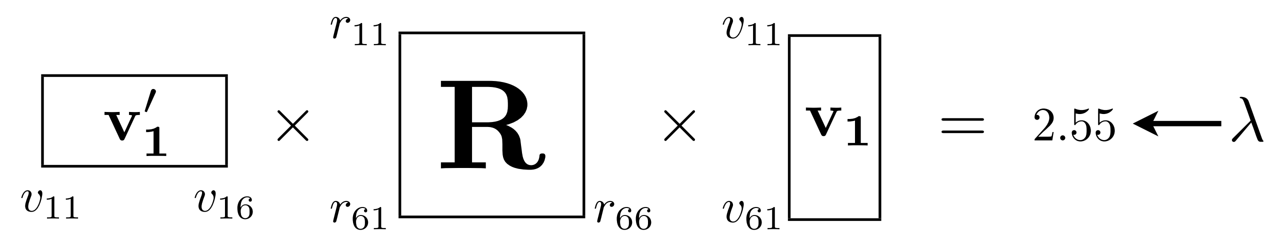
Representing PCA

Matrix: The Fundamental Equations

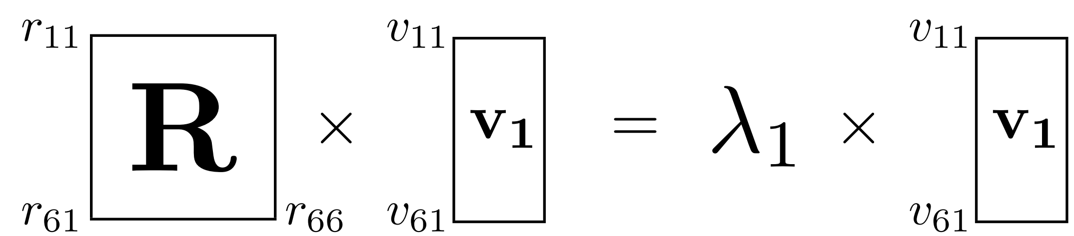
$$\mathbf{Z}\mathbf{v_1} = \mathbf{P_1}$$



$$\mathbf{v_1'}\mathbf{R}\mathbf{v_1} = \lambda_1$$



$$\mathbf{R}\mathbf{v_1} = \lambda_1\mathbf{v_1}$$

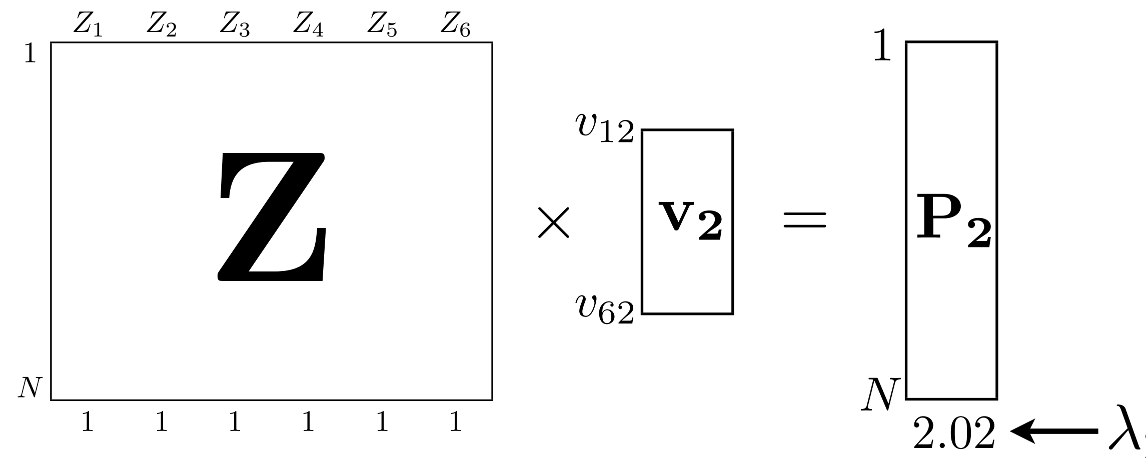


This is an 'eigen-equation' where λ_1 is the first eigenvalue and $\mathbf{v_1}$ is the first eigenvector. The weights, $\mathbf{v_1}$, specify the linear combination of the original variables that make the variance of the linear combination as large as possible. The variance of the linear composite is λ_1.

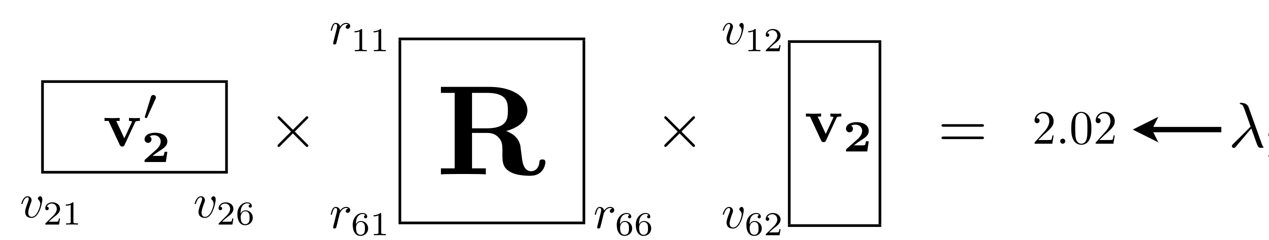
Representing PCA

Matrix: The Fundamental Equations

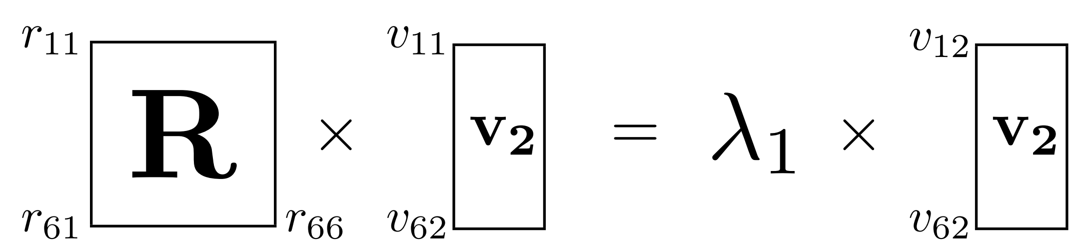
$$\mathbf{Z}\mathbf{v_2} = \mathbf{P_2}$$



$$\mathbf{v_2'}\mathbf{R}\mathbf{v_2} = \lambda_\mathbf{2}$$



$$\mathbf{R}\mathbf{v_2} = \lambda_\mathbf{2}\mathbf{v_2}$$



This is an 'eigen-equation' where λ_2 is the second eigenvalue and $\mathbf{v_2}$ is the second eigenvector. It's formed to be uncorrelated with the first and to have the maximum remaining variance.

Representing PCA

Matrix: The Fundamental Equations

There's an 'eigen-equation' for each principal component. Putting all the eigen-equations together gives us:

$$\mathbf{L} = \mathbf{V}'\mathbf{R}\mathbf{V}$$

$$\mathbf{RV} = \mathbf{VL} \qquad \begin{bmatrix} r_{11} & & \\ & \mathbf{R} & \\ r_{61} & & r_{66} \end{bmatrix} \times \begin{bmatrix} v_{11} & & \\ & \mathbf{V} & \\ v_{61} & & v_{66} \end{bmatrix} = \begin{bmatrix} v_{11} & & \\ & \mathbf{V} & \\ v_{61} & & v_{66} \end{bmatrix} \times \begin{bmatrix} l_{11} & & \\ & \mathbf{L} & \\ l_{61} & & l_{66} \end{bmatrix}$$

$$\begin{bmatrix} \lambda_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda_4 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda_5 & 0 \\ 0 & 0 & 0 & 0 & 0 & \lambda_6 \end{bmatrix}$$

The matrices $\mathbf{V}$ and $\mathbf{L}$ are special. Each of the columns in $\mathbf{V}$ is orthogonal to all the others and $\mathbf{V}'\mathbf{V} = \mathbf{I} = \mathbf{V}^{-\mathbf{1}}\mathbf{V}$, and therefore $\mathbf{V}^{-\mathbf{1}} = \mathbf{V}'$. $\mathbf{L}$ is a diagonal matrix with the eigenvalues down the diagonal.

Given these formulas and due to the special form of $\mathbf{V}$, the correlation matrix $\mathbf{R}$ can be expressed as:

$$\mathbf{R} = \mathbf{V}\mathbf{L}\mathbf{V}'$$

This is known as the singlular value descomposition (SVD) of the correlation matrix $\mathbf{R}$.

Representing PCA

Matrix: The Fundamental Equations

$\mathbf{R} = \mathbf{V}\mathbf{L}\mathbf{V}'$ can be rewritten as $\mathbf{R} = \mathbf{V}\sqrt{\mathbf{L}}\sqrt{\mathbf{L}}\mathbf{V}'$

Now if we let $\mathbf{A} = \mathbf{V}\sqrt{\mathbf{L}}$ and thus $\mathbf{A}' = \sqrt{\mathbf{L}}\mathbf{V}'$ then:

$$\mathbf{R} = \mathbf{A}\mathbf{A}'$$

This is the fundamental equation of Principal Components Analysis.

What this means is that all the information in $\mathbf{R}$ is reexpressed in $\mathbf{A}$, (the loading matrix) which gives the relationships (correlations) between the variables and components). So $\mathbf{A}$ contains all the information that's in $\mathbf{R}$.

Representing PCA

Matrix: The Fundamental Equations

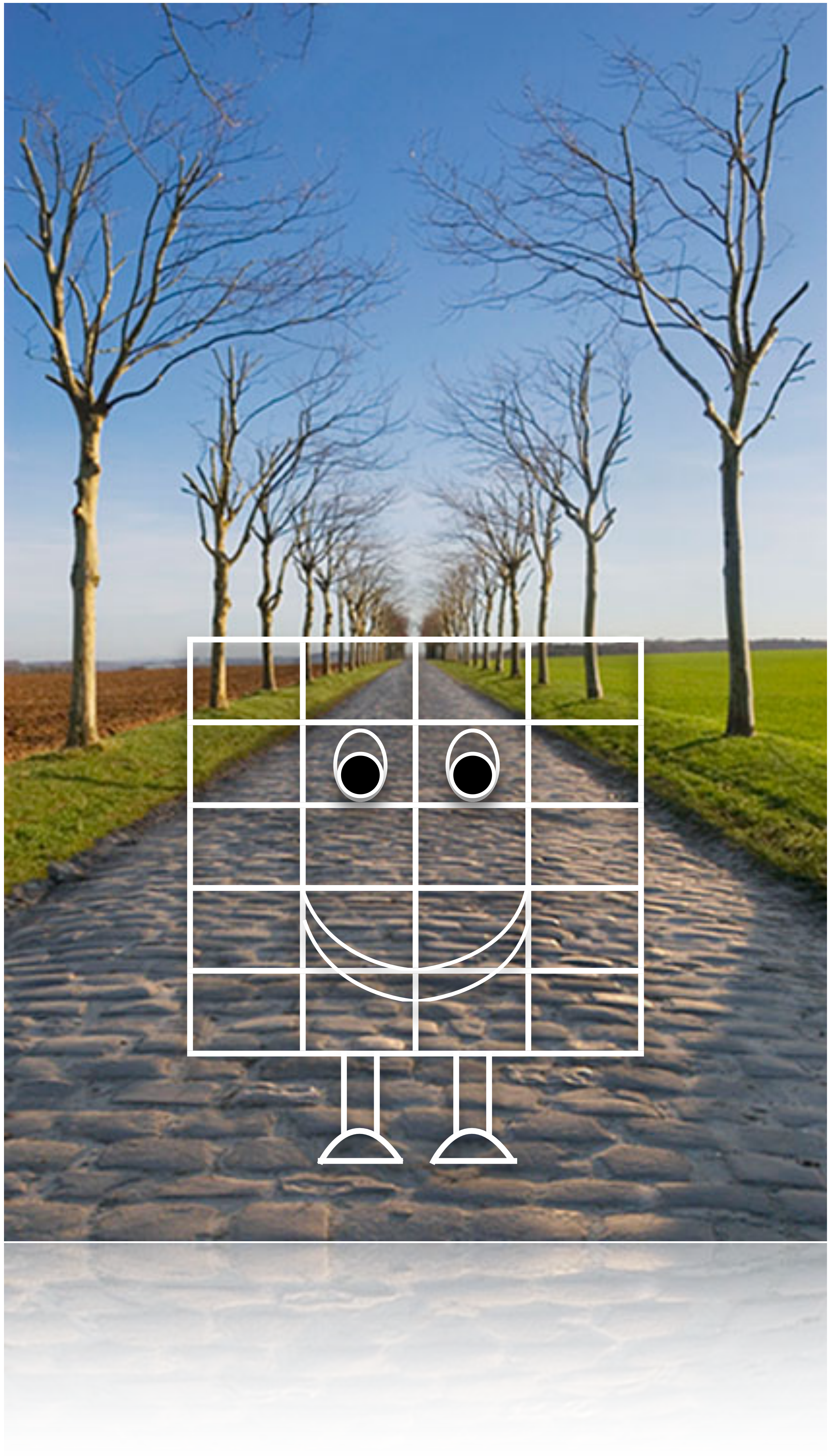
$$\mathbf{R} = \mathbf{A}\mathbf{A}'$$

$$\overset{\mathbf{R}}{\begin{bmatrix} 1.00 & 0.64 & 0.65 & 0.15 & 0.40 & 0.14 \\ 0.64 & 1.00 & 0.49 & -0.04 & 0.19 & -0.01 \\ 0.65 & 0.49 & 1.00 & -0.13 & 0.15 & -0.04 \\ 0.15 & -0.04 & -0.13 & 1.00 & 0.71 & 0.70 \\ 0.40 & 0.19 & 0.15 & 0.71 & 1.00 & 0.47 \\ 0.14 & -0.01 & -0.04 & 0.70 & 0.47 & 1.00 \end{bmatrix}} = \overset{\mathbf{A}}{\begin{bmatrix} 0.76 & 0.50 & 0.02 & -0.05 & -0.40 & -0.01 \\ 0.56 & 0.61 & -0.42 & 0.33 & 0.17 & 0.02 \\ 0.50 & 0.68 & 0.47 & -0.12 & 0.22 & 0.07 \\ 0.65 & -0.68 & -0.06 & -0.05 & 0.01 & 0.33 \\ 0.79 & -0.34 & -0.18 & -0.40 & 0.12 & -0.21 \\ 0.59 & -0.59 & 0.28 & 0.45 & 0.00 & -0.15 \end{bmatrix}} \times \overset{\mathbf{A}'}{\begin{bmatrix} 0.76 & 0.56 & 0.50 & 0.65 & 0.79 & 0.59 \\ 0.50 & 0.61 & 0.68 & -0.68 & -0.34 & -0.59 \\ 0.02 & -0.42 & 0.47 & -0.06 & -0.18 & 0.28 \\ -0.05 & 0.33 & -0.12 & -0.05 & -0.40 & 0.45 \\ -0.40 & 0.17 & 0.22 & 0.01 & 0.12 & 0.00 \\ -0.01 & 0.02 & 0.07 & 0.33 & -0.21 & -0.15 \end{bmatrix}}$$

The elements of $\mathbf{A}$ also turn out to be the correlations of each variable with each principal component. These correlations are called ‘loadings’ and indicate the relationship between each variable and each component. $\mathbf{A}$ is thus called the loading, pattern, or structure matrix.

Principal components analysis

- Purposes
- Motivational examples
- Design Issues
- Representing PCA
  - Logically: Euler Diagrams
  - Geometric: A vector representation
  - Schematic: A 'boxes of data' representation
  - Algebraic: A formulaic representation
  - Matrix: The Fundamental Equations
  - Schematic: The matrices linked



One matrix's journey of transformation: From real to synthetic

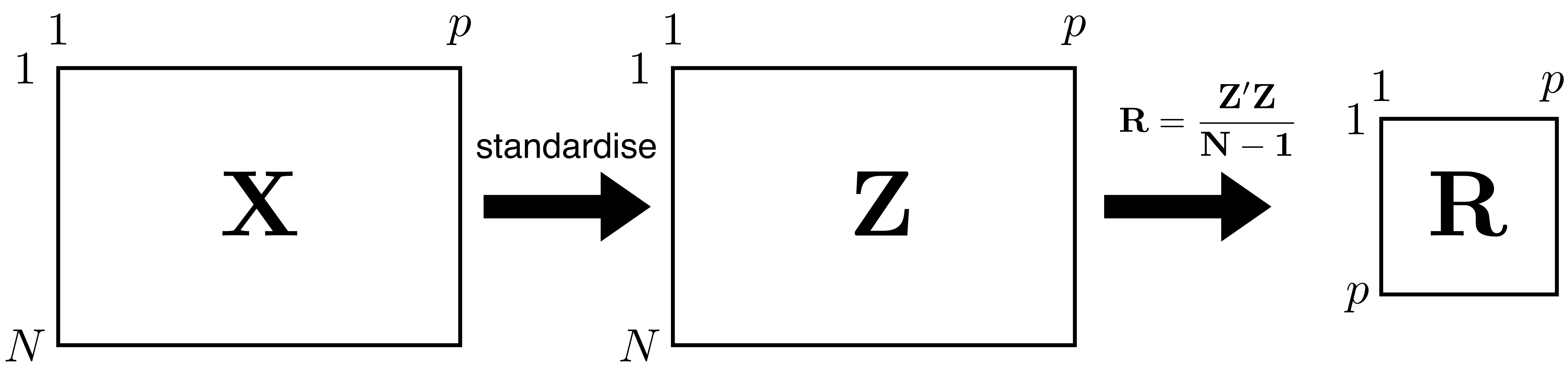
Representing PCA

Schematic: The matrices linked

T&F Example (page 615)

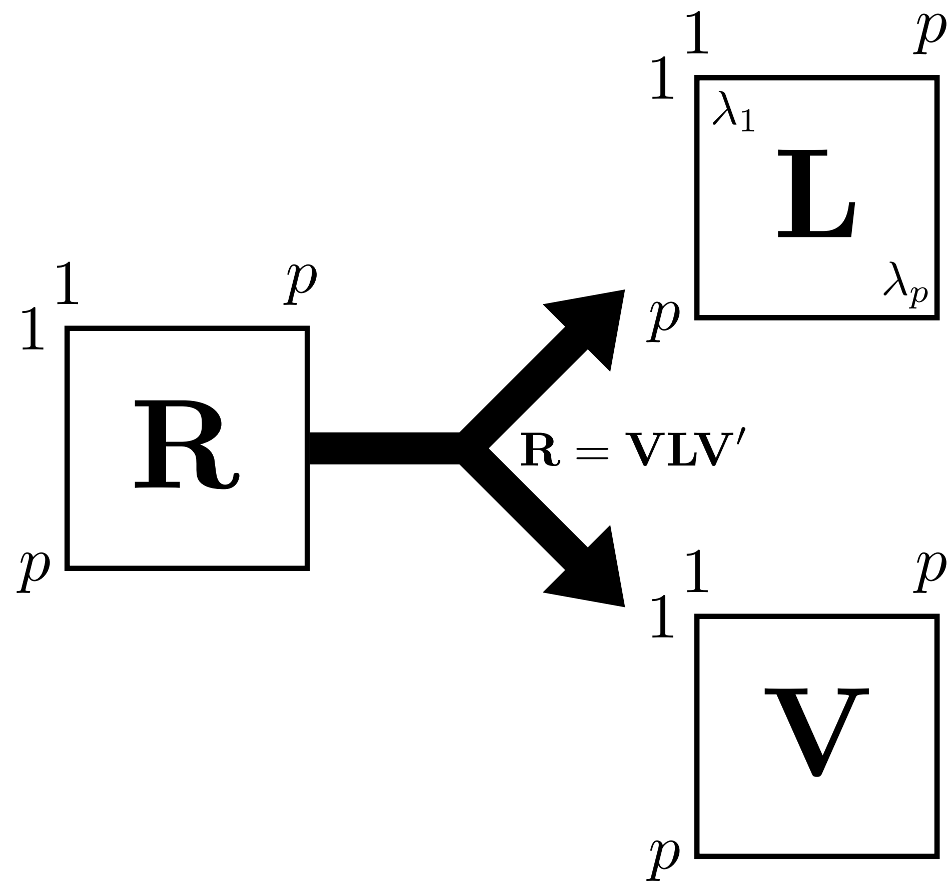
“Five subjects who were trying on ski boots late on a Friday night in January were asked about the importance of each of four variables to their selection of a ski resort. The variables were cost of ski ticket (COST), speed of ski lift (LIFT), depth of snow (DEPTH), and moisture of snow (POWDER). Larger numbers indicate greater importance. The researcher wanted to investigate the pattern of relationships among the variables in an effort to understand better the dimensions underlying choice of ski area.”

	Variables			
Skiers	*COST*	*LIFT*	*DEPTH*	*POWDER*
S_1	32	64	65	67
S_2	61	37	62	65
S_3	59	40	45	43
S_4	36	62	34	35
S_5	62	46	43	40



$$\mathbf{X} = \begin{bmatrix} 32 & 64 & 65 & 67 \\ 61 & 37 & 62 & 65 \\ 59 & 40 & 45 & 43 \\ 36 & 62 & 34 & 35 \\ 62 & 46 & 43 & 40 \end{bmatrix} \qquad \mathbf{Z} = \begin{bmatrix} -1.223 & 1.136 & 1.150 & 1.141 \\ 0.748 & -1.024 & 0.923 & 1.007 \\ 0.612 & -0.784 & -0.363 & -0.470 \\ -0.952 & 0.976 & -1.195 & -1.007 \\ 0.816 & -0.304 & -0.515 & -0.671 \end{bmatrix} \qquad \mathbf{R} = \begin{bmatrix} 1.000 & -0.953 & -0.055 & -0.130 \\ -0.953 & 1.000 & -0.091 & -0.036 \\ -0.055 & -0.091 & 1.000 & 0.990 \\ -0.130 & -0.036 & 0.990 & 1.000 \end{bmatrix}$$

Take X, standardise, get Z, calculate $\frac{\mathbf{Z'Z}}{\mathbf{N-1}}$, get R.



$$\mathbf{R} = \begin{bmatrix} 1.000 & -0.953 & -0.055 & -0.130 \\ -0.953 & 1.000 & -0.091 & -0.036 \\ -0.055 & -0.091 & 1.000 & 0.990 \\ -0.130 & -0.036 & 0.990 & 1.000 \end{bmatrix}$$

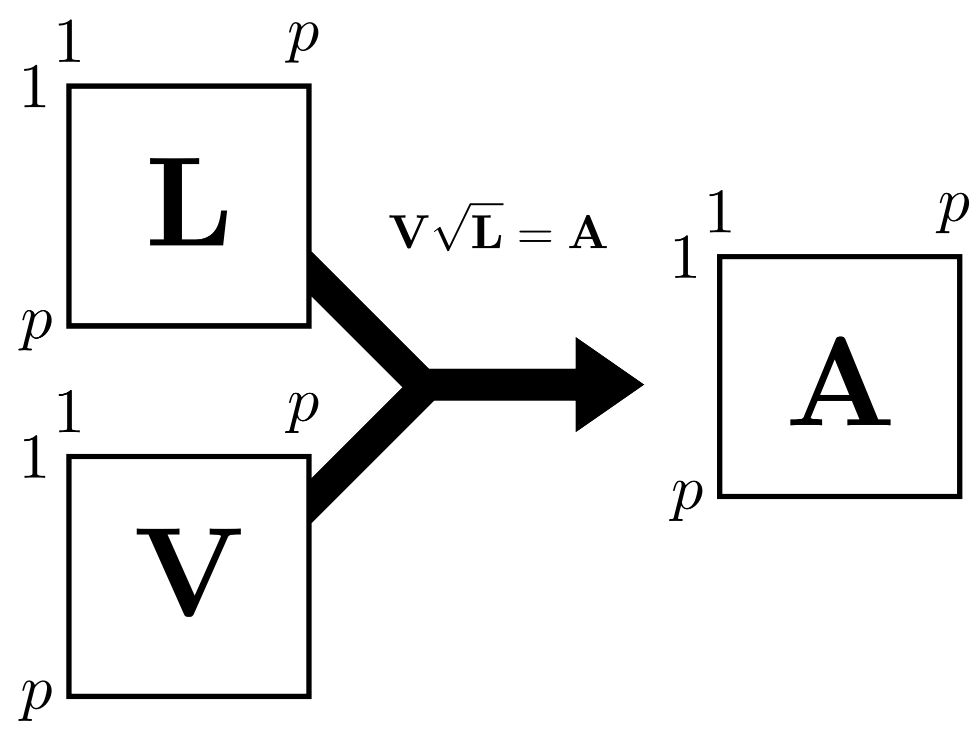
$$\mathbf{L} = \begin{bmatrix} 2.016 & 0 & 0 & 0 \\ 0 & 1.942 & 0 & 0 \\ 0 & 0 & 0.038 & 0 \\ 0 & 0 & 0 & 0.004 \end{bmatrix}$$

$$\mathbf{V} = \begin{bmatrix} 0.352 & -0.614 & 0.663 & -0.244 \\ -0.251 & 0.664 & 0.676 & -0.199 \\ -0.627 & -0.322 & 0.276 & 0.653 \\ -0.647 & -0.280 & -0.169 & -0.689 \end{bmatrix}$$

Take R,

decompose R (do a SVD)

get L and V.

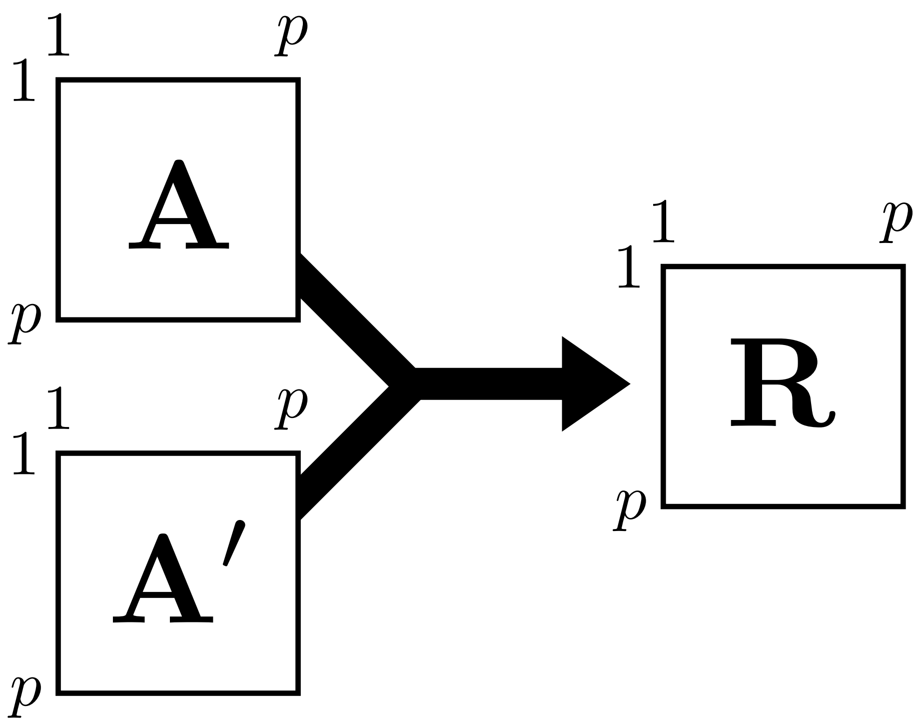


$$\mathbf{L} = \begin{bmatrix} 2.016 & 0 & 0 & 0 \\ 0 & 1.942 & 0 & 0 \\ 0 & 0 & 0.038 & 0 \\ 0 & 0 & 0 & 0.004 \end{bmatrix}$$

$$\mathbf{A} = \begin{bmatrix} 0.500 & -0.856 & 0.129 & -0.016 \\ -0.357 & 0.925 & 0.131 & -0.013 \\ -0.891 & -0.449 & 0.054 & 0.043 \\ -0.919 & -0.390 & -0.033 & -0.046 \end{bmatrix}$$

$$\mathbf{V} = \begin{bmatrix} 0.352 & -0.614 & 0.663 & -0.244 \\ -0.251 & 0.664 & 0.676 & -0.199 \\ -0.627 & -0.322 & 0.276 & 0.653 \\ -0.647 & -0.280 & -0.169 & -0.689 \end{bmatrix}$$

Take L and V, normalise columns, get A.



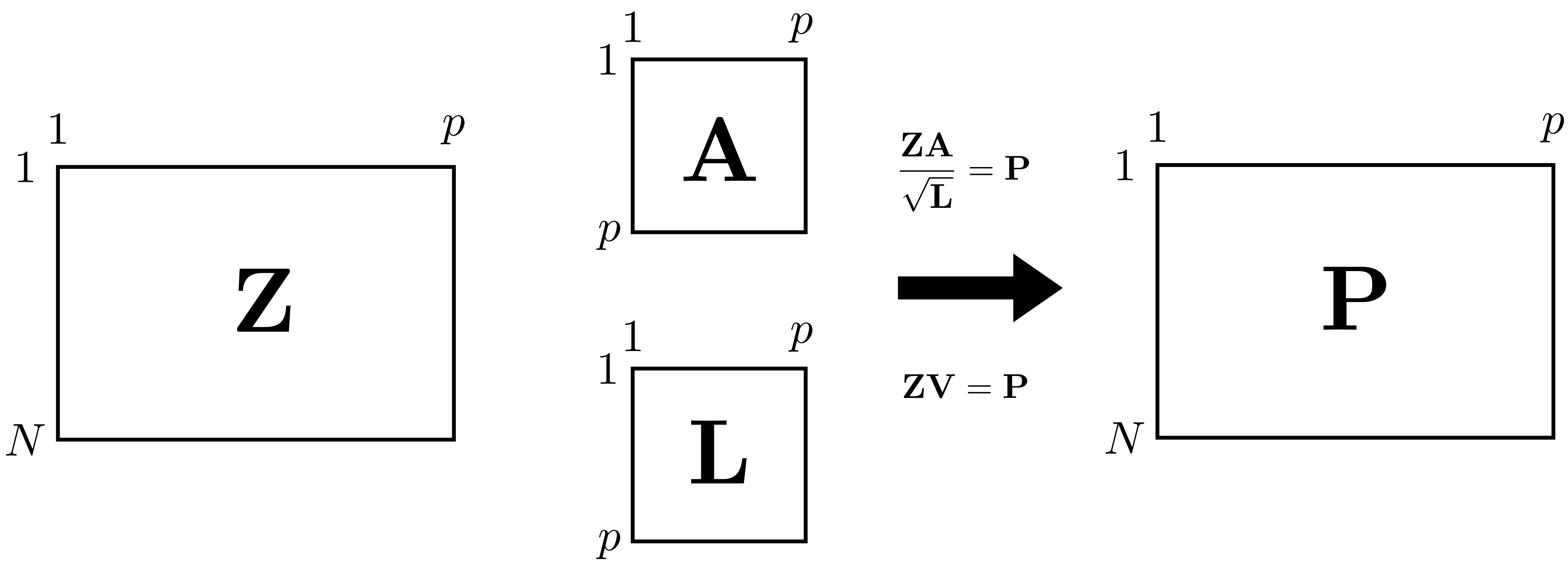
$$\mathbf{A}\mathbf{A}' = \mathbf{R}$$

$$\mathbf{A} = \begin{bmatrix} 0.500 & -0.856 & 0.129 & -0.016 \\ -0.357 & 0.925 & 0.131 & -0.013 \\ -0.891 & -0.449 & 0.054 & 0.043 \\ -0.919 & -0.390 & -0.033 & -0.046 \end{bmatrix}$$

$$\mathbf{A}' = \begin{bmatrix} 0.500 & -0.357 & -0.891 & -0.919 \\ -0.856 & 0.925 & -0.449 & -0.390 \\ 0.129 & 0.131 & 0.054 & -0.033 \\ -0.016 & -0.013 & 0.043 & -0.046 \end{bmatrix}$$

$$\mathbf{R} = \begin{bmatrix} 1.000 & -0.953 & -0.055 & -0.130 \\ -0.953 & 1.000 & -0.091 & -0.036 \\ -0.055 & -0.091 & 1.000 & 0.990 \\ -0.130 & -0.036 & 0.990 & 1.000 \end{bmatrix}$$

Take A, calculate $\mathbf{A}'\mathbf{A}$, recover R.



$$\mathbf{Z} = \begin{bmatrix} -1.223 & 1.136 & 1.150 & 1.141 \\ 0.748 & -1.024 & 0.923 & 1.007 \\ 0.612 & -0.784 & -0.363 & -0.470 \\ -0.952 & 0.976 & -1.195 & -1.007 \\ 0.816 & -0.304 & -0.515 & -0.671 \end{bmatrix}$$

$$\mathbf{A} = \begin{bmatrix} 0.500 & -0.856 & 0.129 & -0.016 \\ -0.357 & 0.925 & 0.131 & -0.013 \\ -0.891 & -0.449 & 0.054 & 0.043 \\ -0.919 & -0.390 & -0.033 & -0.046 \end{bmatrix}$$

$$\mathbf{L} = \begin{bmatrix} 2.016 & 0 & 0 & 0 \\ 0 & 1.942 & 0 & 0 \\ 0 & 0 & 0.038 & 0 \\ 0 & 0 & 0 & 0.004 \end{bmatrix}$$

$$\mathbf{P} = \begin{bmatrix} -2.177 & 0.816 & 0.082 & 0.038 \\ -0.710 & -1.718 & -0.112 & -0.069 \\ 0.945 & -0.648 & -0.146 & 0.093 \\ 0.821 & 1.899 & -0.130 & -0.049 \\ 1.121 & -0.349 & 0.306 & -0.012 \end{bmatrix}$$

$$var = \begin{matrix} 2.016 & 1.942 & 0.038 & 0.004 \end{matrix}$$

Take Z, A, and L, calculate $\frac{\mathbf{ZA}}{\sqrt{\mathbf{L}}} = \mathbf{P}$, get P.

Representing PCA

Schematic: The matrices linked

Summary (again)

- A full PCA transforms a set of correlated measured variables into a set of uncorrelated variables (linear combinations).
- These are new composite scores or synthetic variables.
- We can use this if we know:
  - How many dimensions are needed to adequately represent the information in the original variables.
  - How to interpret the linear combination.