

Introduction

Computational Social Intelligence - Lecture 01

Prof. Alessandro Vinciarelli
School of Computing Science &
Institute of Neuroscience and Psychology

<http://www.dcs.gla.ac.uk/vincia>
alessandro.Vinciarelli@glasgow.ac.uk



University
of Glasgow

EPSRC
Engineering and Physical Sciences
Research Council

FNSNF

Outline

- Introduction
- The Theory of Multiple Intelligences
- Types of Intelligence
- Social Intelligence and Computing
- Conclusions

Outline

- Introduction
- The Theory of Multiple Intelligences
- Types of Intelligence
- Social Intelligence and Computing
- Conclusions

Course Teacher

- Prof. Alessandro Vinciarelli
- web: <http://www.dcs.gla.ac.uk/vincia>
- e-mail: Alessandro.Vinciarelli@glasgow.ac.uk
- Twitter: @alevincia
- Phone: 0141-3302795
- Office: S111 (School of Computing Science)

Timetable

SEPTEMBER 2018						
SUN	MON	TUE	WED	THU	FRI	SAT
						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						

www.theprintablecalendar.com

OCTOBER 2018						
SUN	MON	TUE	WED	THU	FRI	SAT
	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			

www.theprintablecalendar.com

NOVEMBER 2018						
SUN	MON	TUE	WED	THU	FRI	SAT
				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	

www.theprintablecalendar.com

- The course will last for 10 weeks
- Thursday: 9.00-11.00
- Thursday:14.00-15.00

Course Material

- All the course material (slides, texts, etc.) is available on the Moodle page of the course:
<http://website>
- The slides are not a textbook, you need to study the texts that will be provided during the course.

Evaluation

- The course involves an Assessed Exercise that accounts for 20% of the final mark;
- The course includes 10 hours (out of 30) that will be dedicated to the Assessed Exercise;
- The exam accounts for 80% of the final mark.

Interdisciplinarity

- The course is interdisciplinary and it includes not only computing science, but also social psychology;
- The acquisition of social psychology and psychometric notions is crucial towards the successful completion of the course.

Methodology

- The course is methodologically oriented, it requires one to solve problems and to understand the theory behind the solutions;
- The course adopts mathematical and statistical methodologies like those taught, e.g., in the course “Machine Learning”.

Syllabus

- The course is research oriented and the syllabus includes research papers (available on Moodle) that need to be studied in view of the exam;
- The research papers to be studied are explicitly mentioned at the beginning of every lecture.

Role of Programming

- The course requires the use of programming, but it will not improve your knowledge of programming, it will teach you how to address new problems with programming;
- Programming contributes only to a limited extent to final mark;
- No coding examples will be provided.

Recap

- The course does not target the mere acquisition of technical skills, it requires the development of a scientific attitude;
- Every lecture requires one to study different types of material (including scientific articles);
- Code and coding are a tool and not a goal, the course is not based on coding examples and does not target software development.

Part I: Behaviour Observation

Design and organise the collection of behavioural data in view of the application of statistical and computational methodologies for human behaviour understanding.

- Methodology: Statistical Testing;
- Psychology: Behaviour Observation;
- Practical: Nonverbal Cues in Conversation.

Part II: Psychometrics

Measure social and psychological constructs - in quantitative terms - through the adoption of standard psychometric questionnaires.

- Methodology: Correlation and Associations
- Psychology: Psychometric Questionnaires
- Practical: Speech and Personality

Part III: Behaviour Understanding

Apply basic statistical methodologies (e.g., k-Means and Naïve Bayes Classifier) to automatically map behavioural observations into social and psychological constructs.

- Methodology: Bayesian Decision Theory
- Psychology: Human-Human Communication
- Practical: Smile Detection

Outline

- Introduction
- The Theory of Multiple Intelligences
- Social Intelligence and Computing
- Conclusions

This lecture is based on the following text
(available on Moodle):

- Davis et al., “The theory of multiple intelligences”, in “Cambridge Handbook of Intelligence”, Sternberg & Kaufman (eds.), 485-503, Cambridge University Press, 2011.

The Theory of Multiple Intelligences

“[...] individuals possess eight or more relatively autonomous intelligences [and] draw on these intelligences [to] solve problems that are relevant to the societies in which they live.”

Davis, Christodoulou, Seider & Gardner, “The theory of multiple intelligences”,
in “Cambridge Handbook of Intelligence”, Sternberg & Kaufman (eds.),
485-503, Cambridge University Press, 2011.

1. Isolation

“[...] certain individuals should demonstrate particularly high or low levels of a particular capacity in contrast to other capacities.”

Davis, Christodoulou, Seider & Gardner, “The theory of multiple intelligences”,
in “Cambridge Handbook of Intelligence”, Sternberg & Kaufman (eds.),
485-503, Cambridge University Press, 2011.

2. Specific Neural Structure

“[...] its neural structure and functioning
should be distinguishable from that of
other major human faculties.”

Davis, Christodoulou, Seider & Gardner, “The theory of multiple intelligences”,
in “Cambridge Handbook of Intelligence”, Sternberg & Kaufman (eds.),
485-503, Cambridge University Press, 2011.

3. Distinct Development

“It should have a distinct developmental trajectory [...] different intelligences should develop at different rates and along paths which are distinctive.”

Davis, Christodoulou, Seider & Gardner, “The theory of multiple intelligences”,
in “Cambridge Handbook of Intelligence”, Sternberg & Kaufman (eds.),
485-503, Cambridge University Press, 2011.

4. Evolutionary Basis

“[...] an intelligence ought to have a previous instantiation in primate or other species and putative survival value.”

Davis, Christodoulou, Seider & Gardner, “The theory of multiple intelligences”,
in “Cambridge Handbook of Intelligence”, Sternberg & Kaufman (eds.),
485-503, Cambridge University Press, 2011.

5. Symbolic Systems

“It should be susceptible to capture in symbol systems, of the sort used in formal or informal education.”

Davis, Christodoulou, Seider & Gardner, “The theory of multiple intelligences”,
in “Cambridge Handbook of Intelligence”, Sternberg & Kaufman (eds.),
485-503, Cambridge University Press, 2011.

6. Measurability

"It should be supported by evidence from psychometric tests of intelligence."

Davis, Christodoulou, Seider & Gardner, "The theory of multiple intelligences",
in "Cambridge Handbook of Intelligence", Sternberg & Kaufman (eds.),
485-503, Cambridge University Press, 2011.

7.Experimental

“It should be distinguishable from other intelligences through experimental psychological tasks.”

Davis, Christodoulou, Seider & Gardner, “The theory of multiple intelligences”,
in “Cambridge Handbook of Intelligence”, Sternberg & Kaufman (eds.),
485-503, Cambridge University Press, 2011.

8. Specific Processes

“[...] there should be identifiable mental processes that handle information related to each intelligence.”

Davis, Christodoulou, Seider & Gardner, “The theory of multiple intelligences”,
in “Cambridge Handbook of Intelligence”, Sternberg & Kaufman (eds.),
485-503, Cambridge University Press, 2011.

Recap

- There are at least eight criteria that must be respected for a particular ability to be accepted as a type of intelligence;
- The criteria address observable aspects of intelligence (evolution, development, brain, psychology and behaviour);
- The criteria ensure that the different types of intelligence are independent.

Outline

- Introduction
- The Theory of Multiple Intelligences
- **Types of Intelligence**
- Social Intelligence and Computing
- Conclusions

1.Linguistic

“An ability to analyse information and create products involving oral and written language such as speeches, books, and memos.”

Davis, Christodoulou, Seider & Gardner, “The theory of multiple intelligences”, in “Cambridge Handbook of Intelligence”, Sternberg & Kaufman (eds.), 485-503, Cambridge University Press, 2011.

2. Logical-Mathematical

“An ability to develop equations and proofs, make calculations, and solve abstract problems.”

Davis, Christodoulou, Seider & Gardner, “The theory of multiple intelligences”,
in “Cambridge Handbook of Intelligence”, Sternberg & Kaufman (eds.),
485-503, Cambridge University Press, 2011.

3.Spatial

“An ability to recognise and manipulate
large-scale and fine-grained spatial
images.”

Davis, Christodoulou, Seider & Gardner, “The theory of multiple intelligences”,
in “Cambridge Handbook of Intelligence”, Sternberg & Kaufman (eds.),
485-503, Cambridge University Press, 2011.

4.Musical

“An ability to produce, remember, and
make meaning of different patterns of
sound.”

Davis, Christodoulou, Seider & Gardner, “The theory of multiple intelligences”,
in “Cambridge Handbook of Intelligence”, Sternberg & Kaufman (eds.),
485-503, Cambridge University Press, 2011.

5. Naturalist

“An ability to identify and distinguish among different types of plants, animals, and weather formations that are found in the natural world.”

Davis, Christodoulou, Seider & Gardner, “The theory of multiple intelligences”, in “Cambridge Handbook of Intelligence”, Sternberg & Kaufman (eds.), 485-503, Cambridge University Press, 2011.

6. Bodily-Kinesthetic

“An ability to use one’s own body to create products or solve problems.”

Davis, Christodoulou, Seider & Gardner, “The theory of multiple intelligences”,
in “Cambridge Handbook of Intelligence”, Sternberg & Kaufman (eds.),
485-503, Cambridge University Press, 2011.

7. Interpersonal

“An ability to recognise and understand
other people's moods, desires,
motivations, and intentions.”

Davis, Christodoulou, Seider & Gardner, “The theory of multiple intelligences”,
in “Cambridge Handbook of Intelligence”, Sternberg & Kaufman (eds.),
485-503, Cambridge University Press, 2011.

8.Intrapersonal

“An ability to recognise and understand one's own moods, desires, motivations, and intentions.”

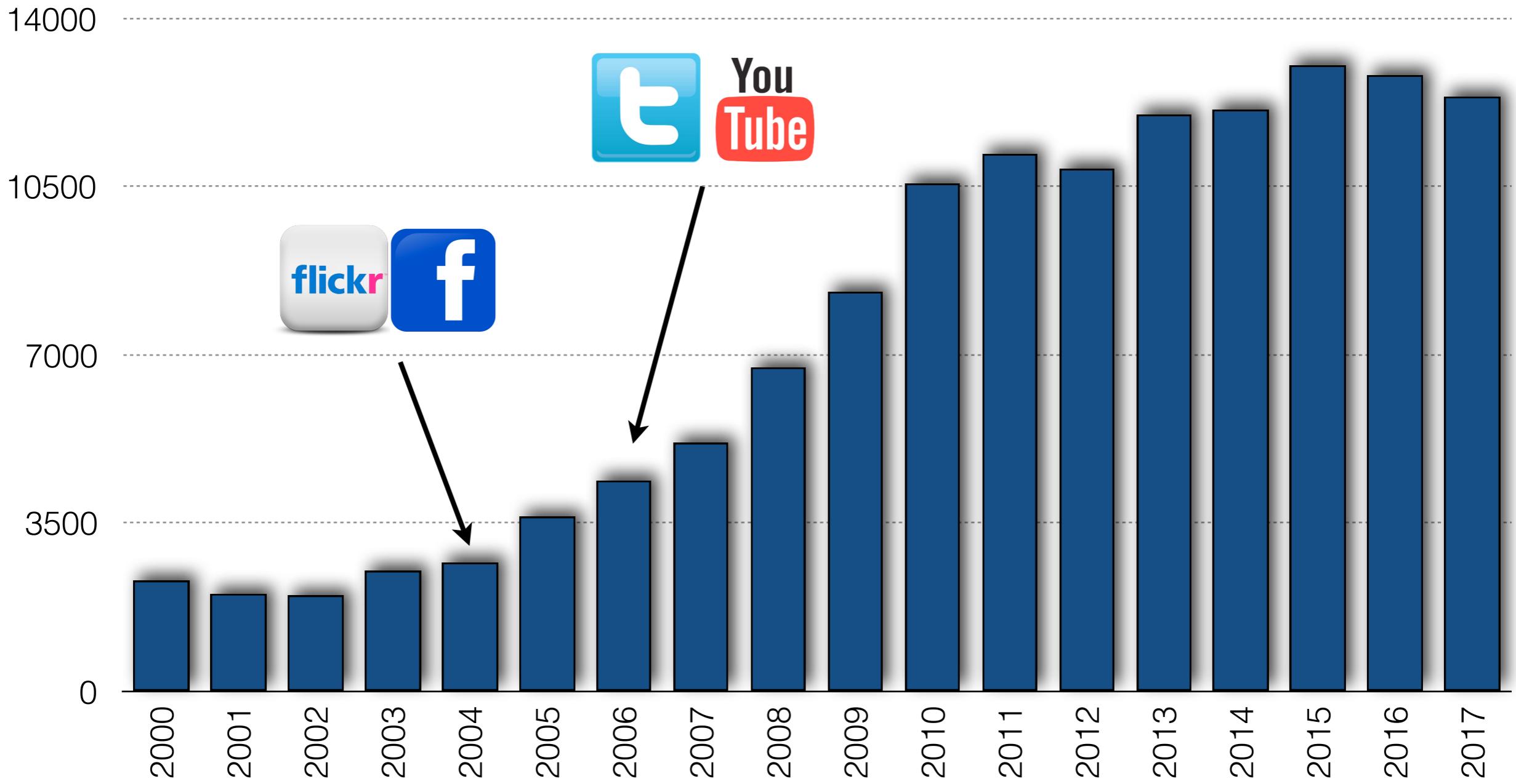
Davis, Christodoulou, Seider & Gardner, “The theory of multiple intelligences”,
in “Cambridge Handbook of Intelligence”, Sternberg & Kaufman (eds.),
485-503, Cambridge University Press, 2011.

Recap

- The application of the eight criteria leads to identification of eight types of intelligence;
- Social intelligence is the name the literature adopts for what the Theory of Multiple Intelligences calls interpersonal intelligence;
- Social intelligence is the skill underlying effective social interactions.

Outline

- Introduction
- The Theory of Multiple Intelligences
- Types of Intelligence
- Social Intelligence and Computing
- Conclusions



Number of publications that contain the word "social" in the largest publication repositories (ACM Digital Library and IEEEXplore).

Social Intelligence and Computing?

“If AI systems are indeed ever to walk among us, they’ll have to be able to understand that each of us has thoughts and feelings and expectations [and] they’ll have to adjust their behaviour accordingly.”

Hintze, “The four types of AI: what you need to know”, 2016, <https://www.weforum.org/agenda/2016/11/the-four-types-of-ai-what-you-need-to-know>

Social Intelligence and Computing?

“And new research into social robots - that know how to collaborate and build working alliances with humans - means that a future where robots and humans work together, each to do what it does best - is a strong likelihood.”

Meyerson, “Top 10 emerging technologies of 2015”, 2015, <https://www.weforum.org/agenda/2015/03/top-10-emerging-technologies-of-2015-2/#next-robotics>

Social Intelligence and Computing?

“tasks that are difficult to automate [...] will require [...] social intelligence.”

UK Government Office for Science, “Artificial intelligence: opportunities and implications for the future of decision making”, 2015, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/566075/gs-16-19-artificial-intelligence-ai-report.pdf

Social Intelligence and Computing?

“[...] the [next] wave of computerisation will depend on overcoming the engineering bottlenecks related to creative and social intelligence.”

Frey and Osborne, “Oxford Martin Programme on the Impacts of Future Technology”, 2013, https://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf

Social Intelligence and Computing?

“[...] machines that completely mimic, or even improve upon the human physical form, but are socially as limited as theme park animatronics [or, vice versa,] devices that have no physical dexterity, but are very social”

KPMG, “Social Robots”, 2016, <https://assets.kpmg.com/content/dam/kpmg/pdf/2016/06/social-robots.pdf>

Information Retrieval
Natural Language Processing
  

Expert Systems
Theorem Proving
 

Self Driving Cars
 

Music Retrieval
  

Computer Vision


Robotics
 

Artificial Agents
Social Media
 

Affective Computing
:) 

Linguistic

Logical-Mathematical

Spatial

Musical

Naturalist

Bodily-Kinaesthetic

Interpersonal

Intrapersonal

Recap

- The ultimate goal of Computational Social Intelligence (CSI) is to build socially intelligent machines;
- CSI is a subfield of Artificial Intelligence characterised by specific scientific and technological goals;
- It responds to economic and industrial needs stated by think tanks, companies and strategic institutions.

Outline

- Introduction
- The Theory of Multiple Intelligences
- Types of Intelligence
- Social Intelligence and Computing
- Conclusions

Conclusions

- Computational Social Intelligence encompasses computing technologies dealing with human-human and human-machine interactions;
- It is an interdisciplinary domain that requires both psychology and computing;
- It includes the analysis of social / psychological phenomena in observational data and the synthesis of social behaviour.

Thank You!

Basic Statistics

Computational Social Intelligence - Lecture 02

Prof. Alessandro Vinciarelli
School of Computing Science &
Institute of Neuroscience and Psychology

<http://www.dcs.gla.ac.uk/vincia>
alessandro.Vinciarelli@glasgow.ac.uk



University
of Glasgow

EPSRC
Engineering and Physical Sciences
Research Council

FNSNF

Texts (see Moodle)

This lecture is based on the following text
(available on Moodle):

- Appendix A of F.Camastra and A.Vinciarelli,
“Machine Learning for Audio, Image and Video
Processing”, Springer Verlag, 2008 (Pages 525
to 532 included).

Outline

- Definition of Probability
- Basic Laws of Probability
- Random Variables
- Conclusions

Outline

- Definition of Probability
- Basic Laws of Probability
- Random Variables
- Conclusions

Flipping Coins

- Flipping a coin is a simple statistical experiment and the outcome can be only Head (H) or Tail (T);
- Out of a sufficiently large number of attempts, it is possible to say that both H and T appear roughly half of the times;
- However, it is not possible to know what will be the outcome of the next attempt.

Number of times we observe outcome “Head” when flipping a coin

$$\frac{n(H)}{n} \approx \frac{1}{2}$$

Total number of times the coin is flipped

There can be small variations with respect to the expectations

Fraction of the times the outcome “Head” is observed

A set of K mutually exclusive events

Events are mutually exclusive when the occurrence of one implies the non-occurrence of the others (Head or Tail)

$$A_1, A_2, \dots, A_K$$

$$p(A_i) = \lim_{n \rightarrow \infty} \frac{n(A_i)}{n}$$

The probability of event "i"

The fraction of times the outcome is event "i" when n tends to infinity

The ratio is smaller or equal than 1 because $n(A)$ cannot be larger than n



$$0 \leq \frac{n(A)}{n} \leq 1 \Rightarrow 0 \leq p(A) \leq 1$$



The value of $p(A)$ is the value of the ratio when n tends to infinity



$n(A)$ cannot be lower than 0, then the ratio cannot be lower than 0

The ratio is always smaller than 1, then $p(A)$ is always smaller than 1

Outline

- Definition of Probability
- **Basic Laws of Probability**
- Random Variables
- Conclusions

The Sample Space

- The sample space is a set of L mutually exclusive events;
- An event A is the outcome of a statistical experiment that corresponds to one, several or all elements of the sample space;
- The main advantage of the sample space is that it allows one to explain the probability laws with the set theory.

The sample space contains L mutually exclusive events

The sample space contains L mutually exclusive events



$$\Omega = \{\omega_1, \omega_2, \dots, \omega_L\}$$

$$A = \{\omega_i, \omega_j, \dots, \omega_k\} \subseteq \Omega$$

The event A corresponds to one, several or all the events of the sample space

The occurrence of A depends on the occurrence of the events of the sample space

The sample space contains the six faces of a dice



$$\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$$

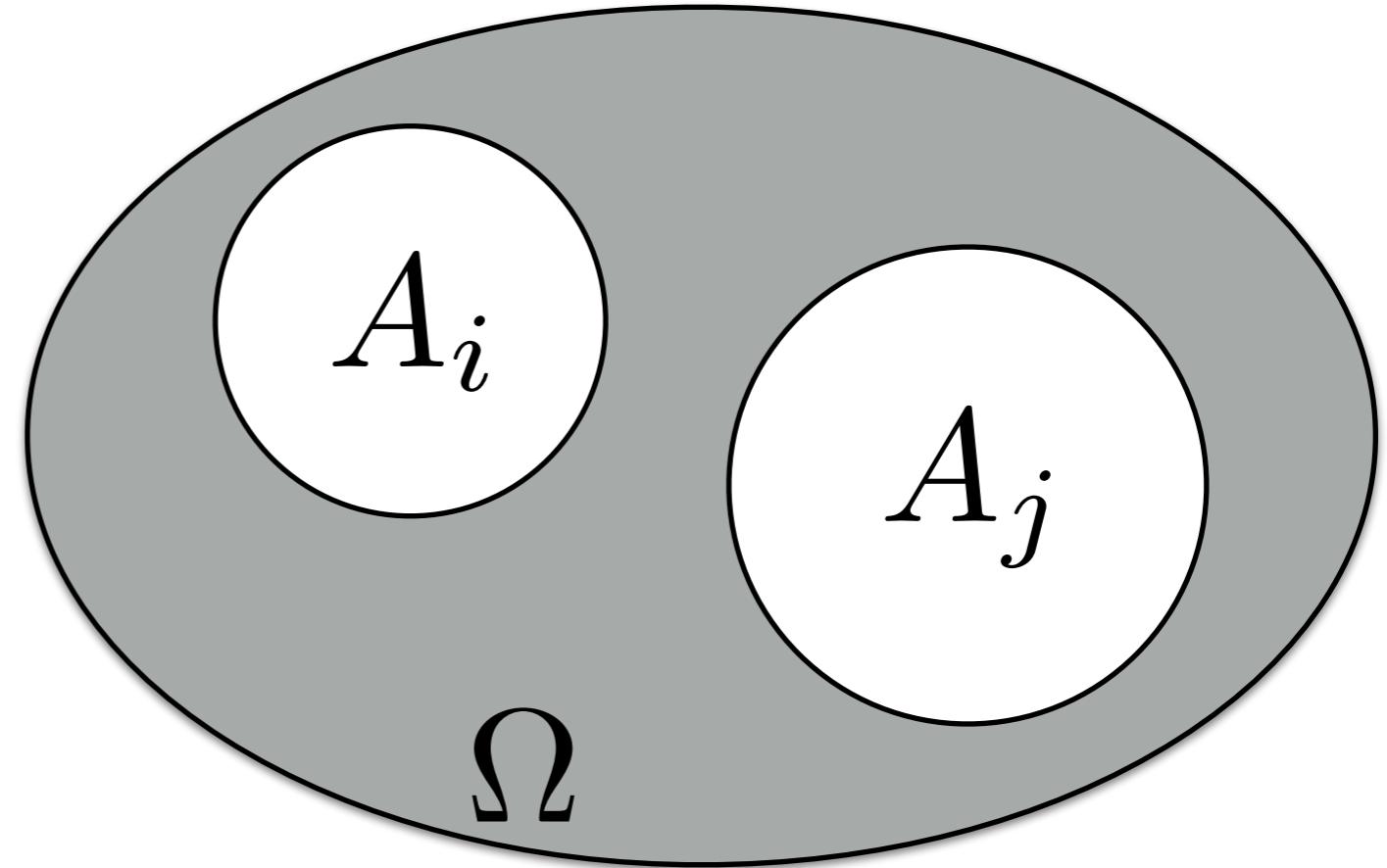
$$A = \{\omega_2, \omega_4, \omega_6\}$$

The event A corresponds to the occurrence of an even number

The event A is a subset of the sample space

Mutually Exclusive Events

No shared element ω_k
between the events
 A_i and A_j

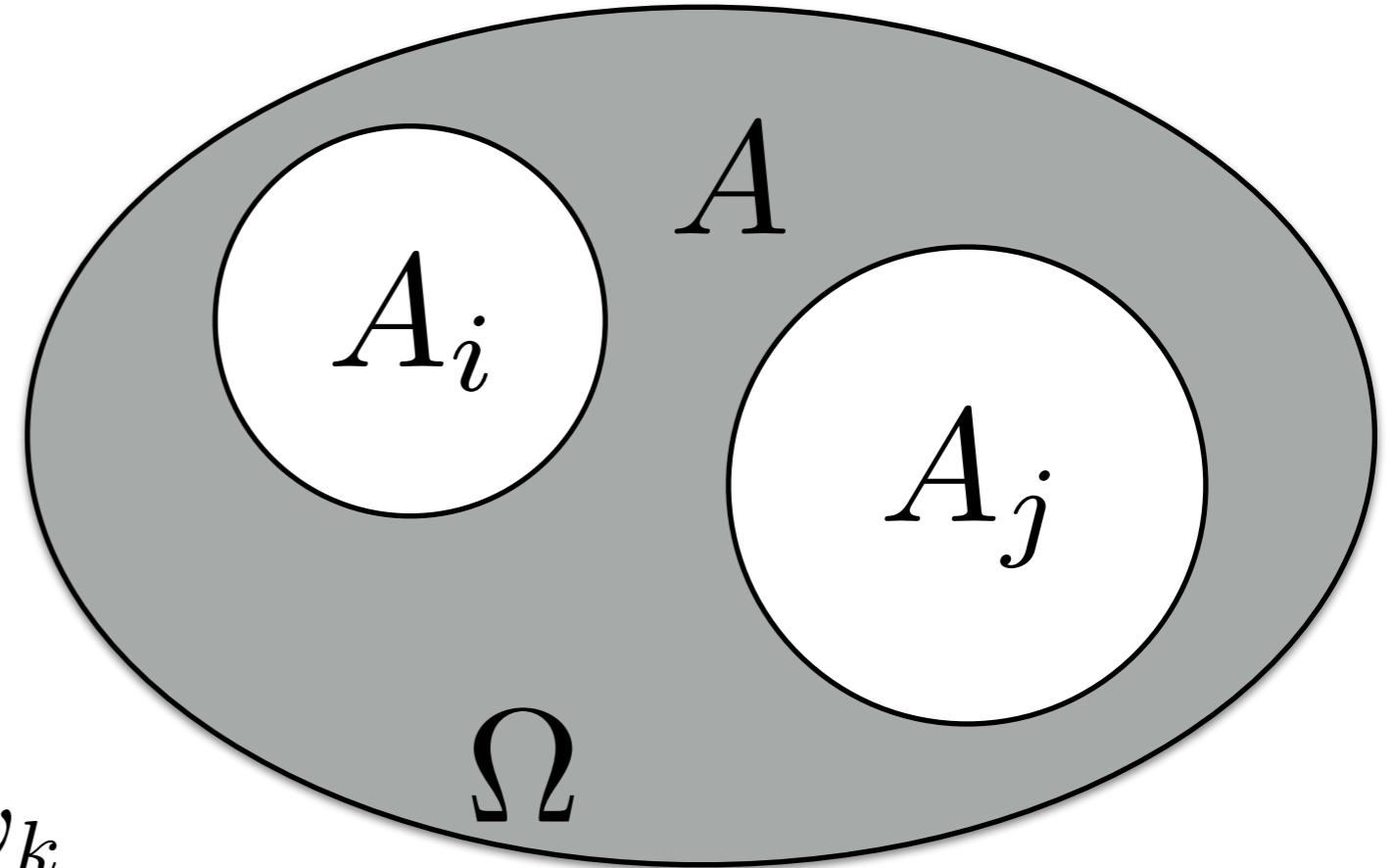


$$A_i \cap A_j = \emptyset$$

A_i and A_j
are said to be
mutually exclusive

Union of Mutually Exclusive Events

$$A = A_i \cup A_j$$



A includes all elements ω_k that belong to A_i or A_j

A is the union of A_i and A_j

The probability of the union of two mutually exclusive events

Addition Law (I)

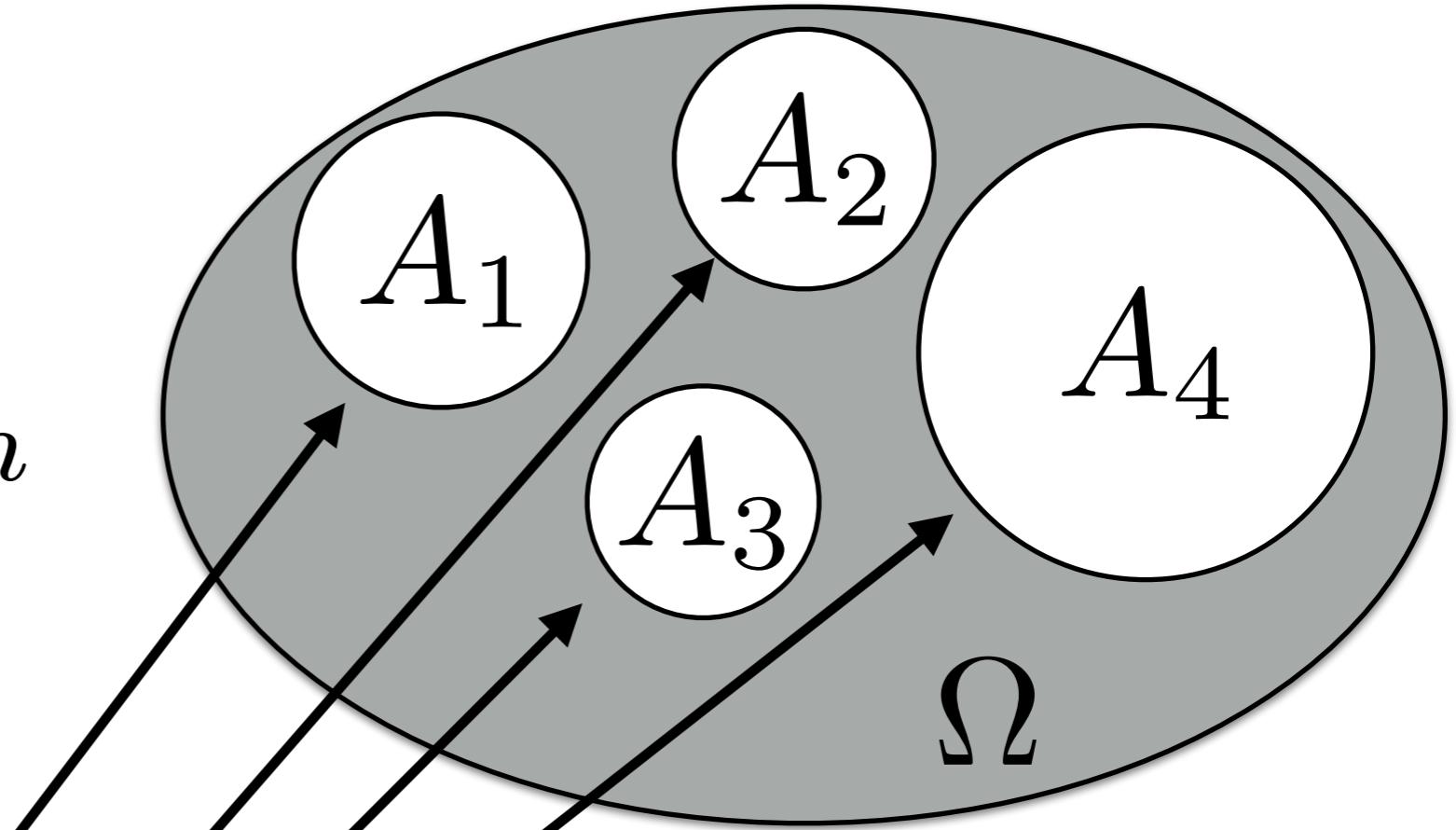
$$p(A) = \lim_{n \rightarrow \infty} \left[\frac{n(A_i)}{n} + \frac{n(A_j)}{n} \right] = p(A_i) + p(A_j)$$

The sum of the probabilities of the two mutually exclusive events

A diagram illustrates the derivation of the Addition Law (I). It shows a large bracket grouping the terms $\frac{n(A_i)}{n}$ and $\frac{n(A_j)}{n}$. Two arrows point from the text "The sum of the probabilities of the two mutually exclusive events" to the "+" sign between the terms and to the final equals sign.

Union of Mutually Exclusive Events

$$A = \bigcup_{n=1}^N A_n$$



An event A can be thought of as the union of N mutually exclusive events

A is the union of N
mutually exclusive
events

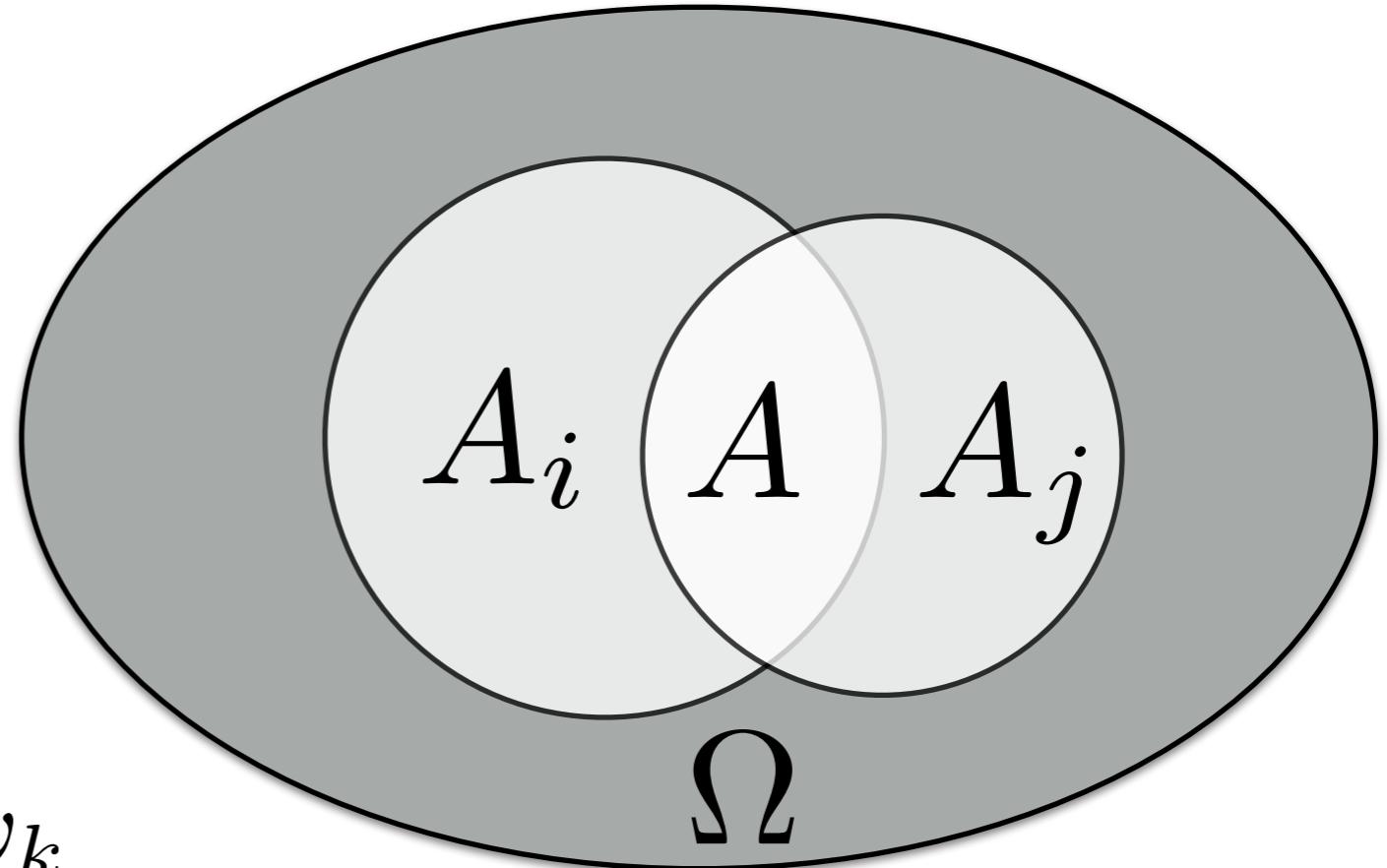
Addition Law (II)

$$p(A) = p\left(\bigcup_{n=1}^N A_n\right) = \sum_{n=1}^N p(A_n)$$

p(A) is the sum of the
probabilities of the
mutually exclusive
events

Intersection of Multiple Events

$$A = A_i \cap A_j$$



A includes all elements ω_k
that belong to both

A_i and A_j

A is the intersection
of A_i and A_j

Conditional Probability

The probability of A
given B

$$p(A|B)$$

Probability of the
intersection of A and B

$$= \frac{p(A \cap B)}{p(B)}$$

Probability of B

Product Law

The probability of the intersection

$$p(A \cap B) = p(A, B) = p(A|B)p(B)$$

The joint probability of A and B

The product comes from the definition of the conditional probability

The intersection is a subset of the two sets



$$A \cap B \subseteq B \Rightarrow 0 \leq p(A|B) \leq 1$$

$$B \subset A \Rightarrow p(A|B) = 1$$

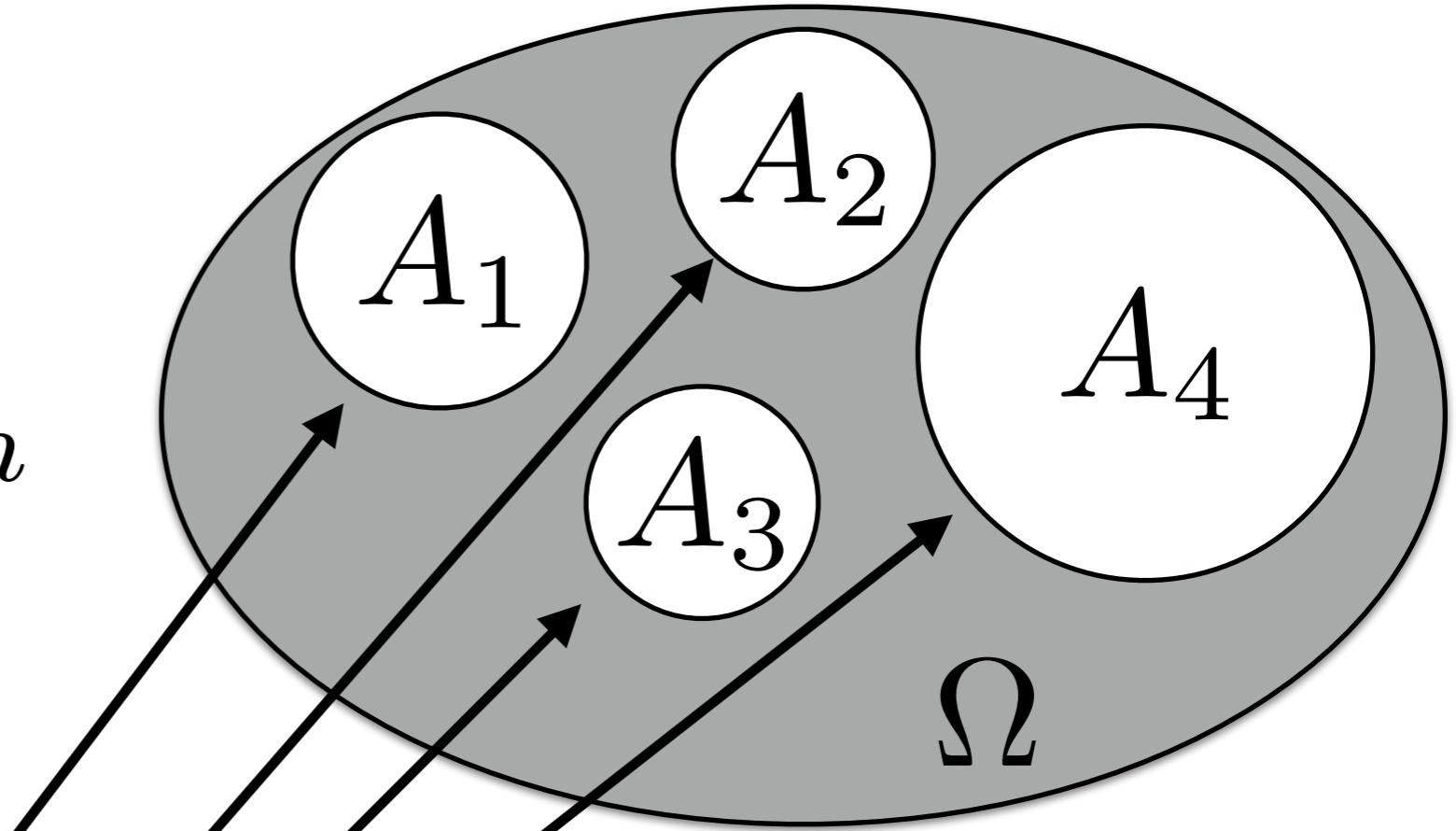


$$A \cap B = \emptyset \Rightarrow p(A|B) = 0$$

The intersection is empty

Union of Mutually Exclusive Events

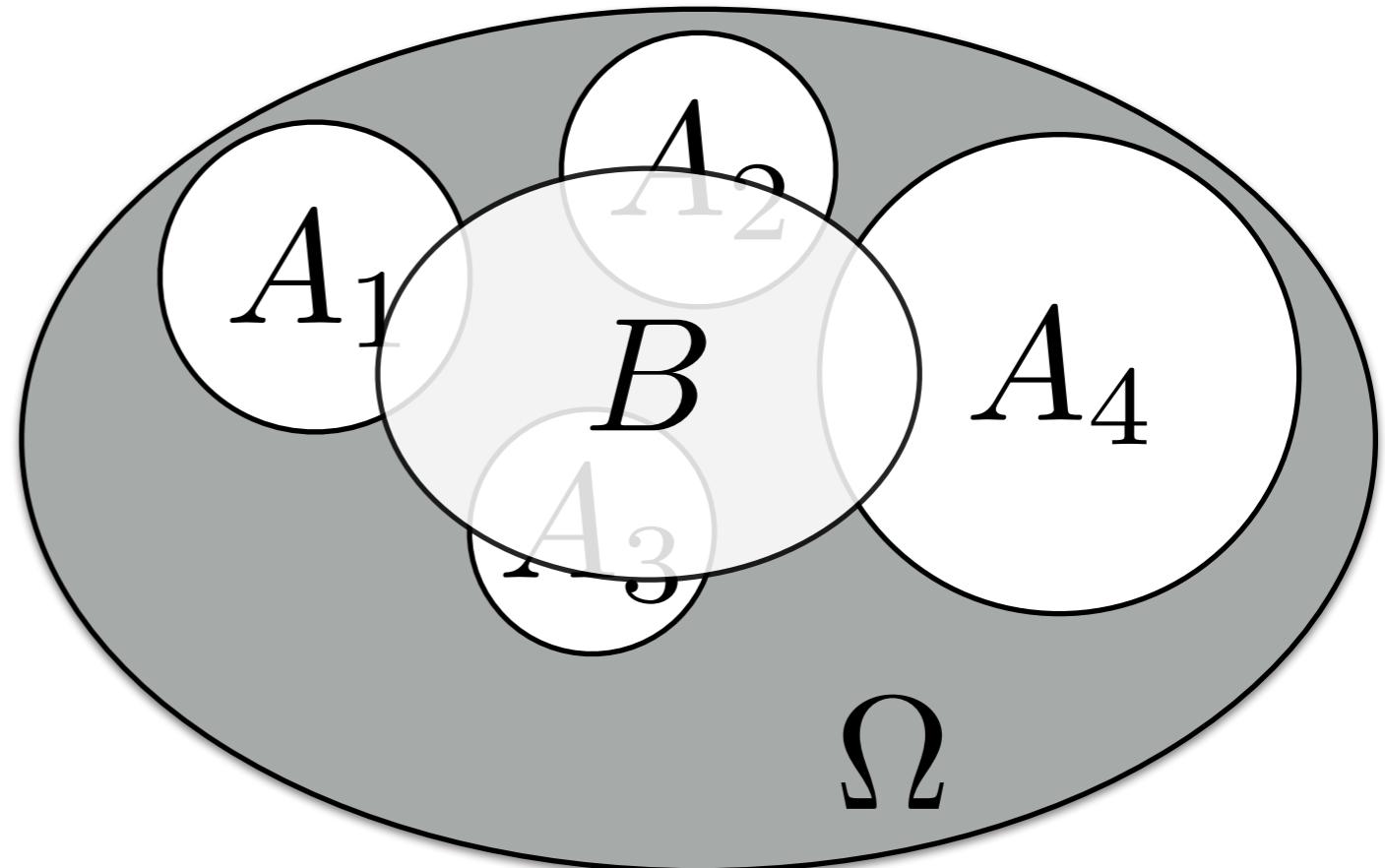
$$A = \bigcup_{n=1}^N A_n$$



An event A can be thought of as the union of N mutually exclusive events

Union of Mutually Exclusive Events

The intersection of A and B is the union of the intersections



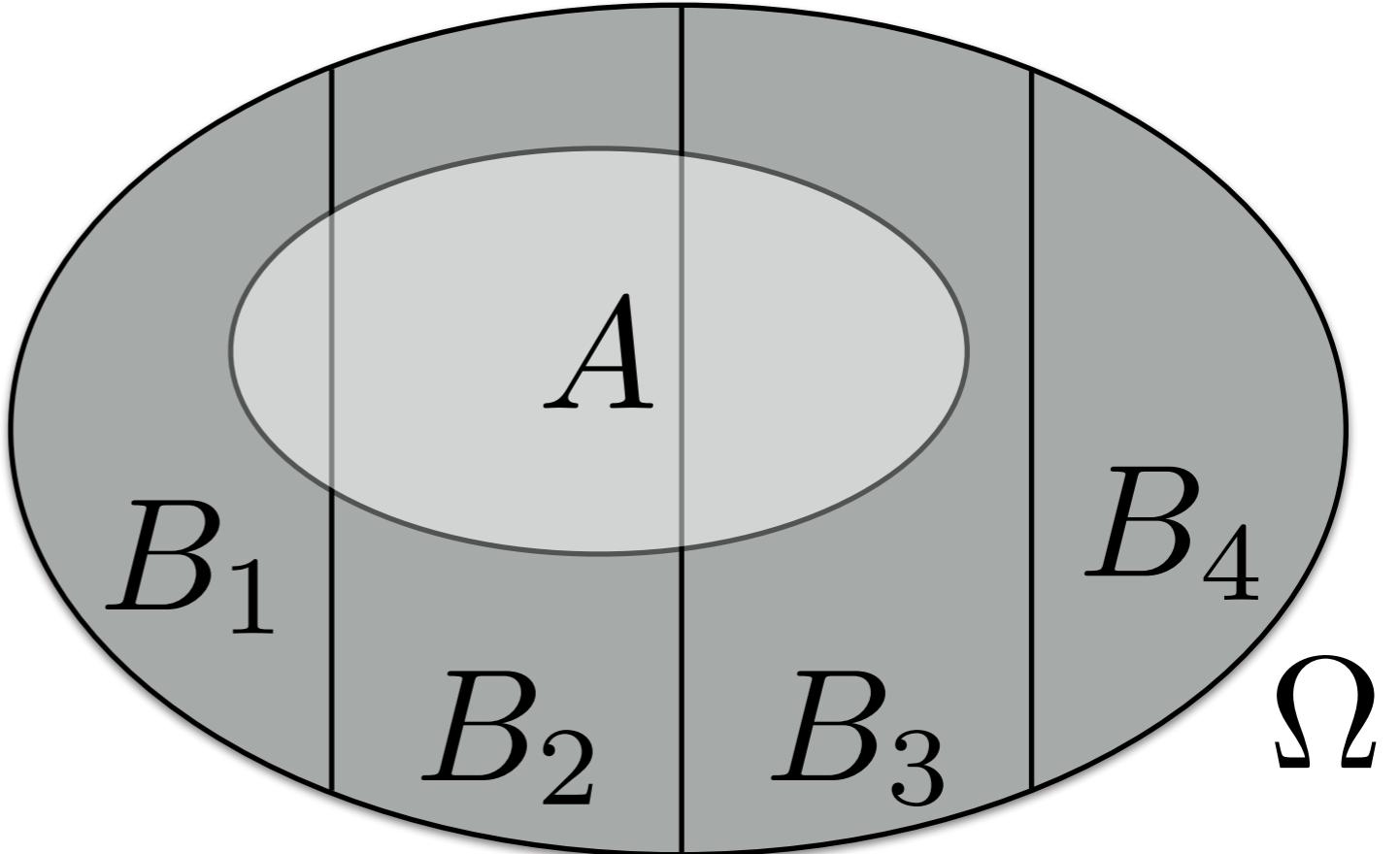
$$A \cap B = \bigcup_{n=1}^N A_n \cap B$$

Addition Law for Conditional Probability (I)

$$\begin{aligned} p(A|B) &= \sum_{n=1}^N \frac{p(A_n \cap B)}{p(B)} = \\ &= \sum_{n=1}^N p(A_n|B) \end{aligned}$$

Union of Mutually Exclusive Events

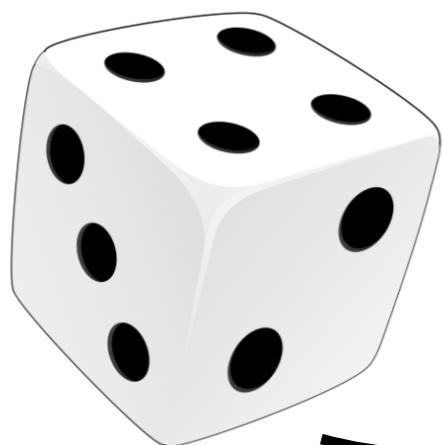
$$\Omega = \bigcup_{n=1}^N B_n$$



The sample space is the union of N mutually exclusive events

Addition Law for Conditional Probability (II)

$$\begin{aligned} p(A) &= \sum_{n=1}^N p(A \cap B_n) = \\ &= \sum_{n=1}^N p(A|B_n)p(B_n) \end{aligned}$$



When there are two dices, the sample spaces include the same events, but they are different

$$\Omega^{(1)} = \{\omega_1^{(1)}, \dots, \omega_6^{(1)}\}$$
$$\Omega^{(2)} = \{\omega_1^{(2)}, \dots, \omega_6^{(2)}\}$$
$$A_1 = \{\omega_i^{(1)}, \dots, \omega_j^{(1)}\}$$
$$A_2 = \{\omega_m^{(2)}, \dots, \omega_n^{(2)}\}$$

The events build upon different sample spaces

The definition of probability

$$p(A_1, A_2) = \lim_{n \rightarrow \infty} \frac{n(A_1, A_2)}{n} \approx$$
$$\approx \lim_{n \rightarrow \infty} \frac{n(A_1, A_2)}{n(A_2)} \approx p(A_1)$$

The number of times A2 occurs tends to infinite when n does

If the occurrence of A2 does not influence the occurrence of A1

Statistical Independence

$$p(A_1, A_2) = \lim_{n \rightarrow \infty} \frac{n(A_1, A_2)}{n} \simeq$$
$$\simeq \lim_{n \rightarrow \infty} \frac{n(A_1, A_2)}{n(A_2)} \frac{n(A_2)}{n} \simeq p(A_1)p(A_2)$$

Multiply and divide by $n(A_2)$

Only if the occurrence of A_2 does not influence the occurrence of A_1 and vice versa

Recap

- The probability is an estimate of how frequently an event occurs as an outcome of a statistical experiment;
- Set theory allows one to demonstrate the basic laws of probability (addition and product);
- Statistical independence allows one to write the joint probability of several events as the product of the individual event probabilities.

Outline

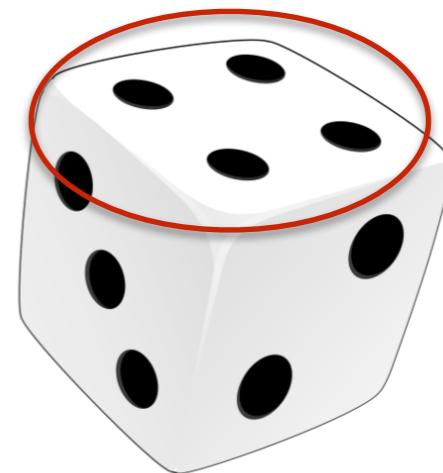
- Definition of Probability
- Basic Laws of Probability
- **Random Variables**
- Conclusions

A variable is said to be random when it is related to a statistical experiment

The value of a random variable depends on the outcome of the statistical experiment

$$\xi = \xi(\omega)$$

In the case of a dice, the outcome is the face and the value is the number “written” on it



A variable is discrete when its value belongs to a finite set X

If a value is not in X , its probability is null

$$\xi \in X = \{x_1, x_2, \dots, x_T\}$$

$$P(\xi = x_j) = p(x_j)$$

It is possible to estimate the probability that the variable takes a certain value in X

The probability distribution is a function that associates value and probability

The sum goes over all values of the discrete random variable

This is a constraint that must be respect for p to be a distribution

$$\sum_{k=1}^T p(x_k) = 1$$

The probability of a variable being lower than a certain value

$$P(\xi \leq x_j) = \sum_{x_k \leq x_j} p(x_k)$$

$$F(x_j) = P(\xi \leq x_j)$$

The cumulative probability function

The Addition Law allows the following estimate

↓

$$\sum_{x_k \leq x_j} p(x_k)$$

↑

The value of the cumulative probability function is a probability

A variable is continuous when its value belongs to a continuous interval



$$\xi \in [a, b]$$

$$P(x_1 \leq \xi \leq x_2) = \int_{x_1}^{x_2} p(x) dx$$

It is possible to estimate the probability of the value falling between two extremes

The function $p(x)$ is called probability density function

The integral goes over
all values of the
continuous random
variable

$$\int_{-\infty}^{\infty}$$

This is a constraint that
must be respect for p to
be a probability density
function

$$p(x)dx = 1$$

The probability of a variable being lower than a certain value

$$F(x_1) = \int_{-\infty}^{x_1} p(x) dx$$

$$F(x_1) = P(\xi \leq x_1)$$

The cumulative probability function

The probability density function allows the following estimate

$$F(x_1) = P(\xi \leq x_1)$$

The value of the cumulative probability function is a probability

Recap

- The value of the random variables depends on an underlying event;
- When a random variable is discrete, it is possible to define a probability distribution;
- When a random variable is continuous, it is necessary to adopt a probability density function.

Outline

- Definition of Probability
- Basic Laws of Probability
- Random Variables
- Conclusions

Conclusions

- The probability is an estimate of how frequently an event takes place over a sufficient number of attempts;
- Discrete variables have a distribution, continuous variables have a probability density function.

Thank You!

Hypothesis Testing

Computational Social Intelligence - Lecture 03

Prof. Alessandro Vinciarelli
School of Computing Science &
Institute of Neuroscience and Psychology

<http://www.dcs.gla.ac.uk/vincia>
alessandro.Vinciarelli@glasgow.ac.uk



University
of Glasgow

EPSRC
Engineering and Physical Sciences
Research Council

FNSNF

This lecture is based on the following text
(available on Moodle):

- D.C.Howell, "Statistical Methods for Psychology", Chapter 4, pp. 92-109, Cengage Learning, 2009.

Outline

- Sampling Distributions
- Hypothesis Testing
- The Null Hypothesis
- Error Types
- Directionality

Outline

- Sampling Distributions
- Hypothesis Testing
- The Null Hypothesis
- Error Types
- Directionality

The height of
individual i in the
population

$$\bar{h} = \frac{1}{N} \sum_{i=1}^N h_i$$

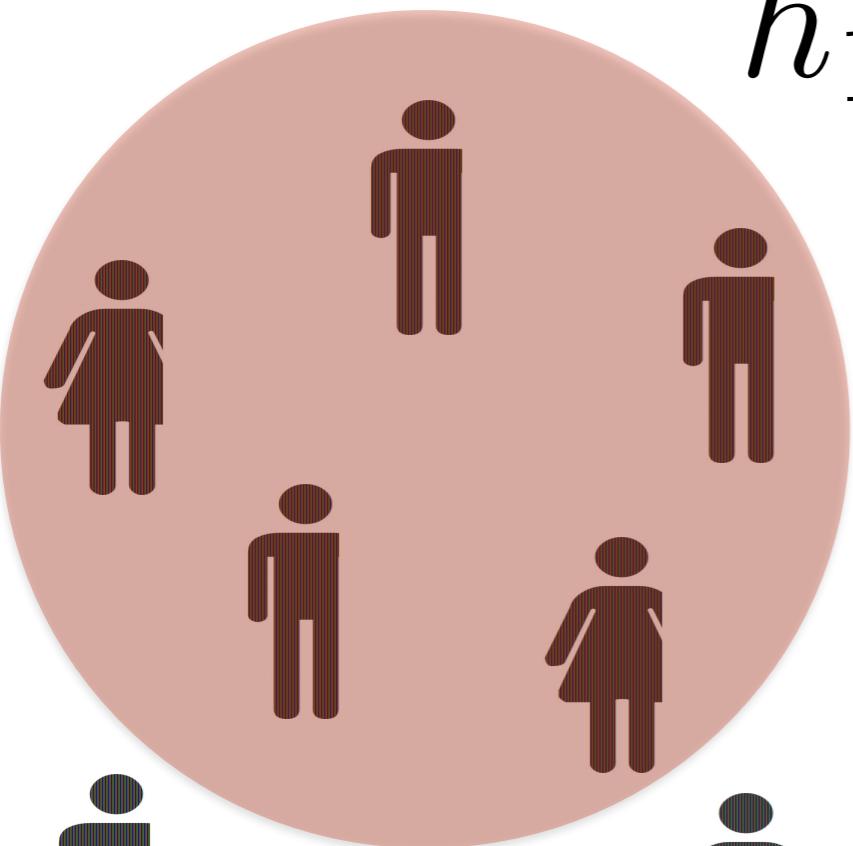
The average height of
the individuals in the
population



Sampling Distributions

“[...] the distribution of values obtained for [a] statistic over repeated sampling (i.e. running the experiment, or drawing samples, an unlimited number of times).”

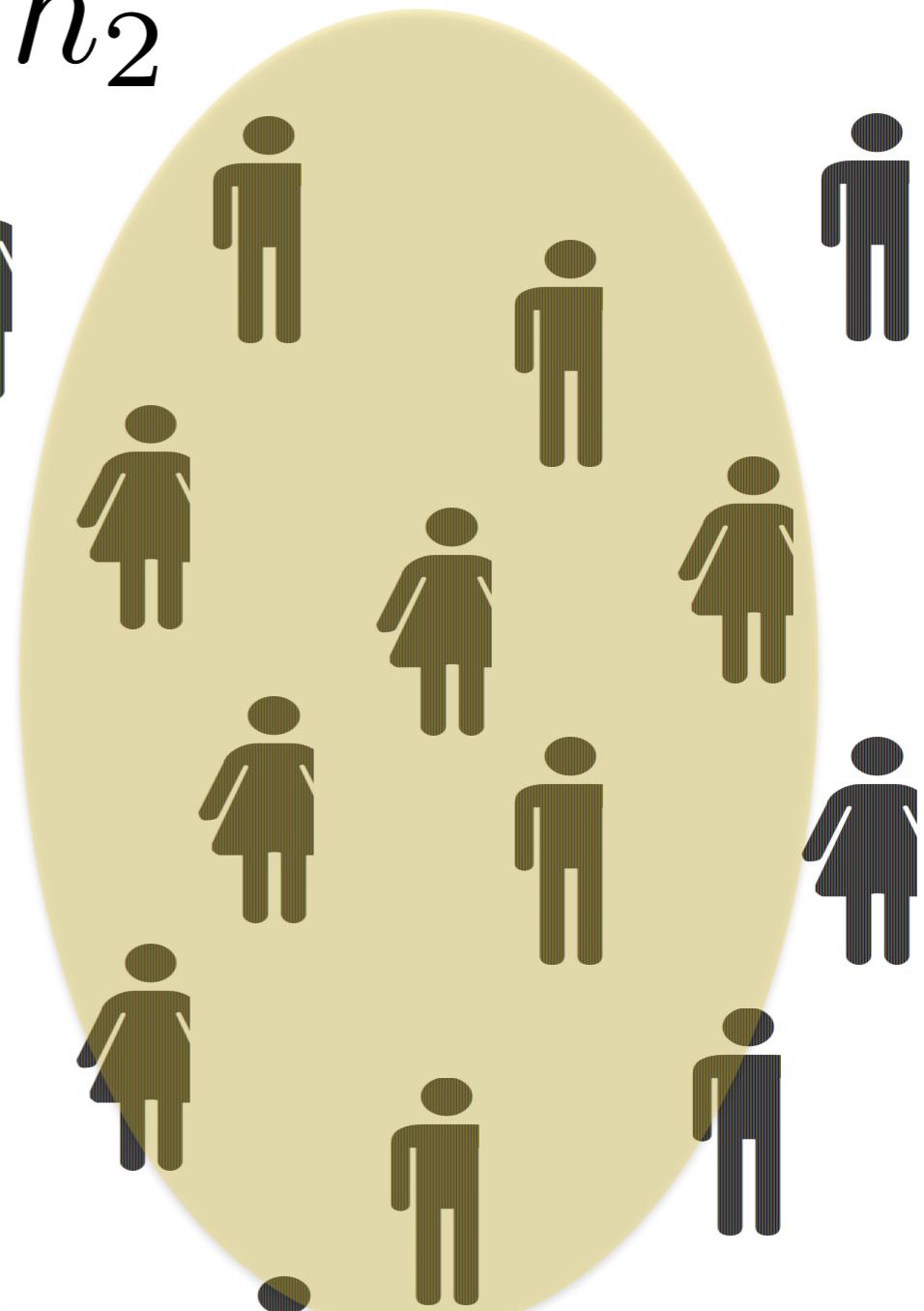
D.C.Howell, “Statistical Methods for Psychology”, Chapter 4,
Cengage Learning, 2009.



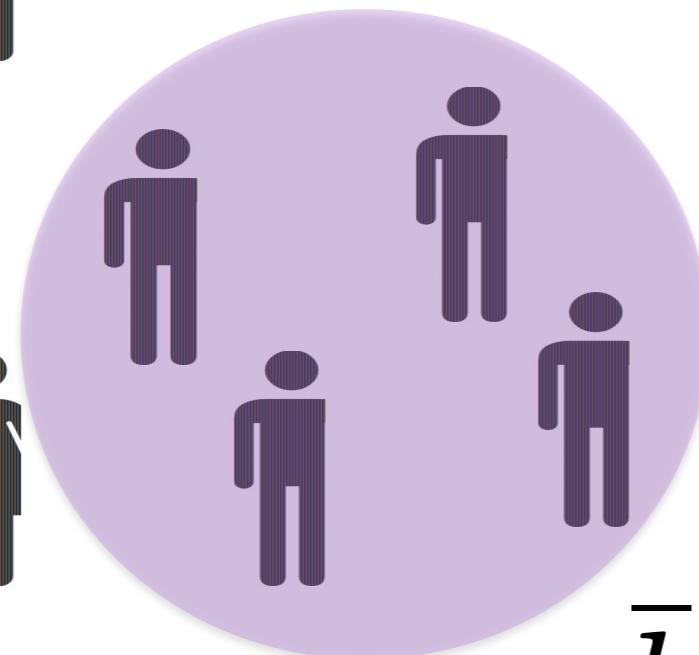
\bar{h}_1

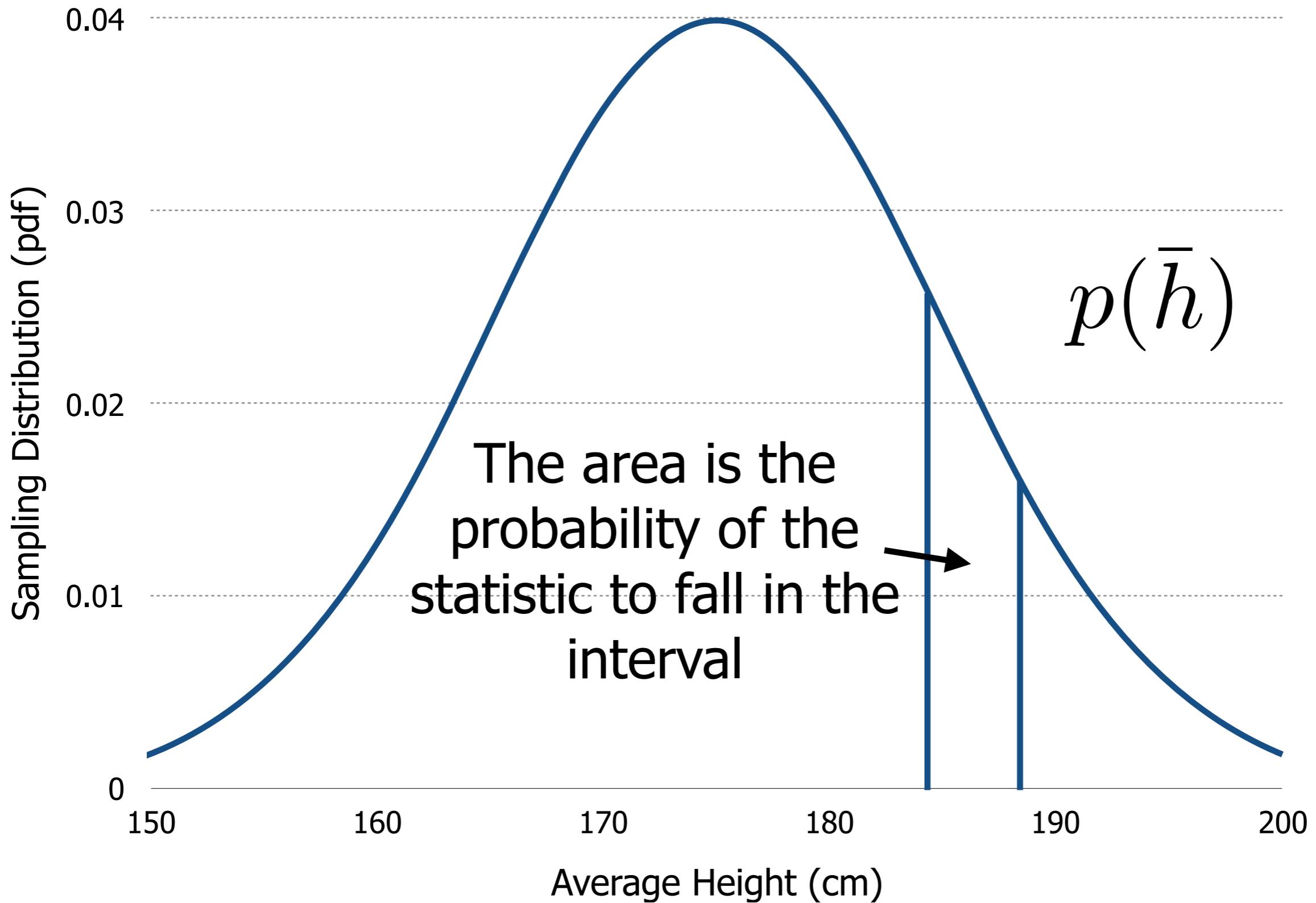


\bar{h}_2



\bar{h}_3



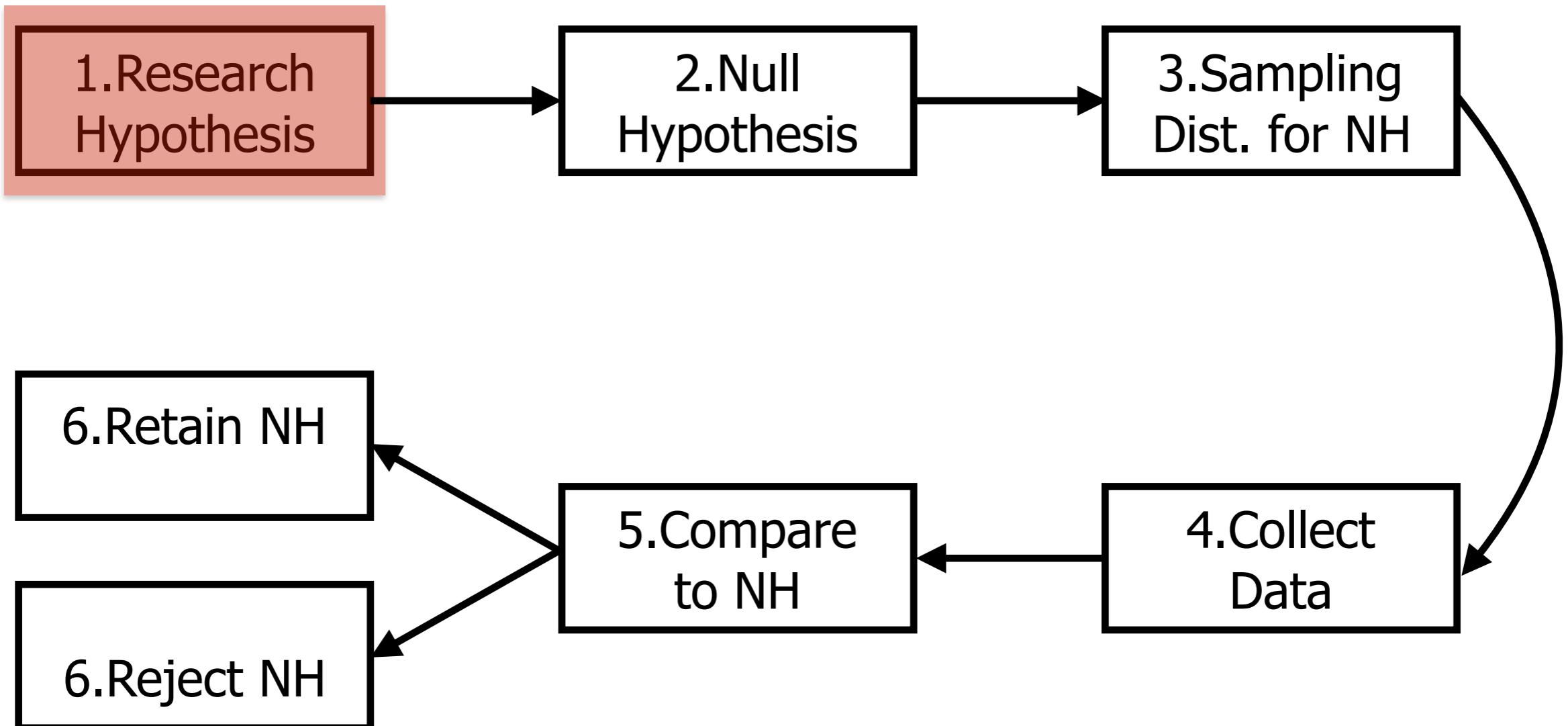


Outline

- Sampling Distributions
- Hypothesis Testing
- The Null Hypothesis
- Error Types
- Directionality

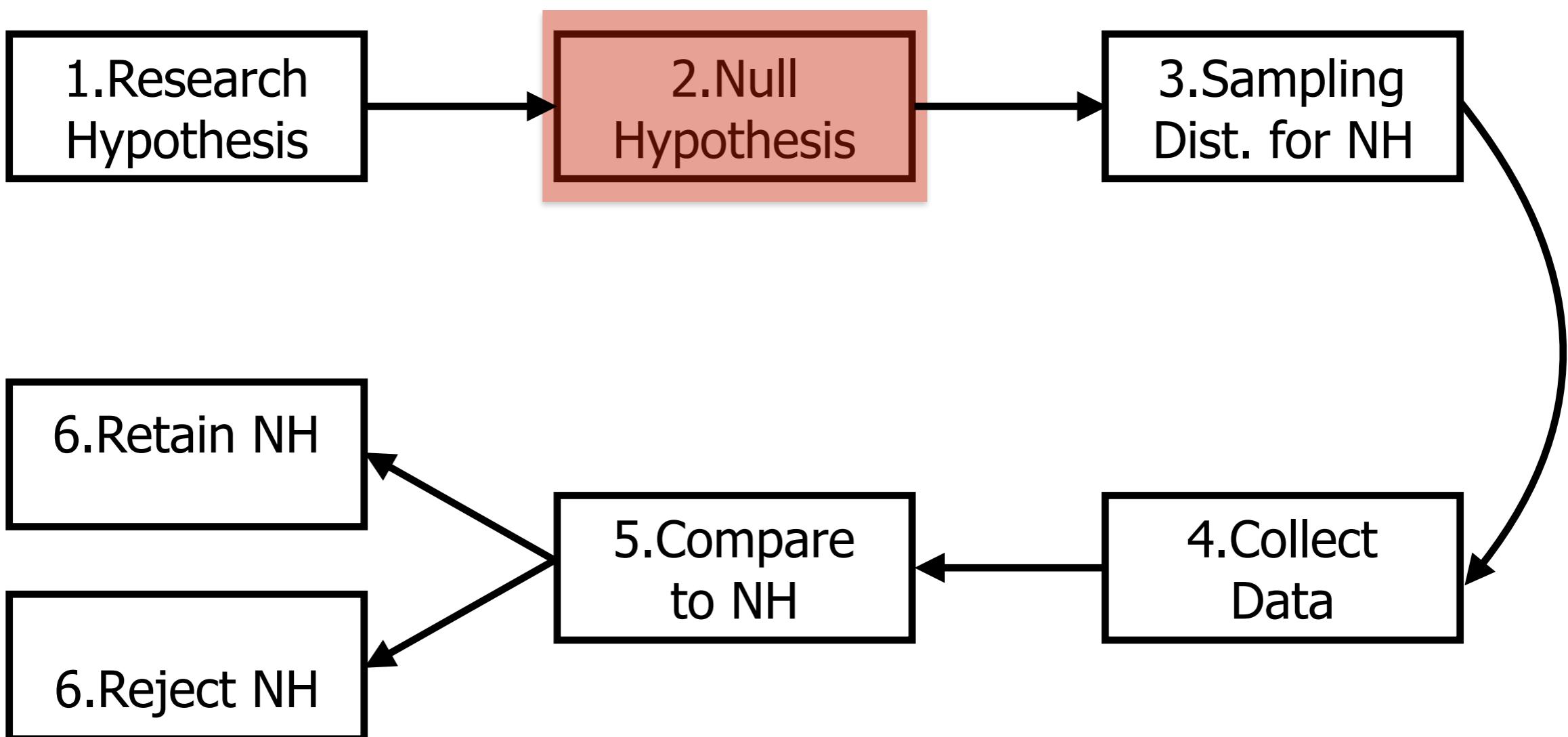
Hypothesis Testing

- State a research hypothesis;
- Setup the null hypothesis (NH);
- Construct the sampling distribution of the statistic when the null hypothesis is true;
- Collect data;
- Compare sample statistic to its distribution when the null hypothesis is true;
- Retain or reject the null hypothesis.



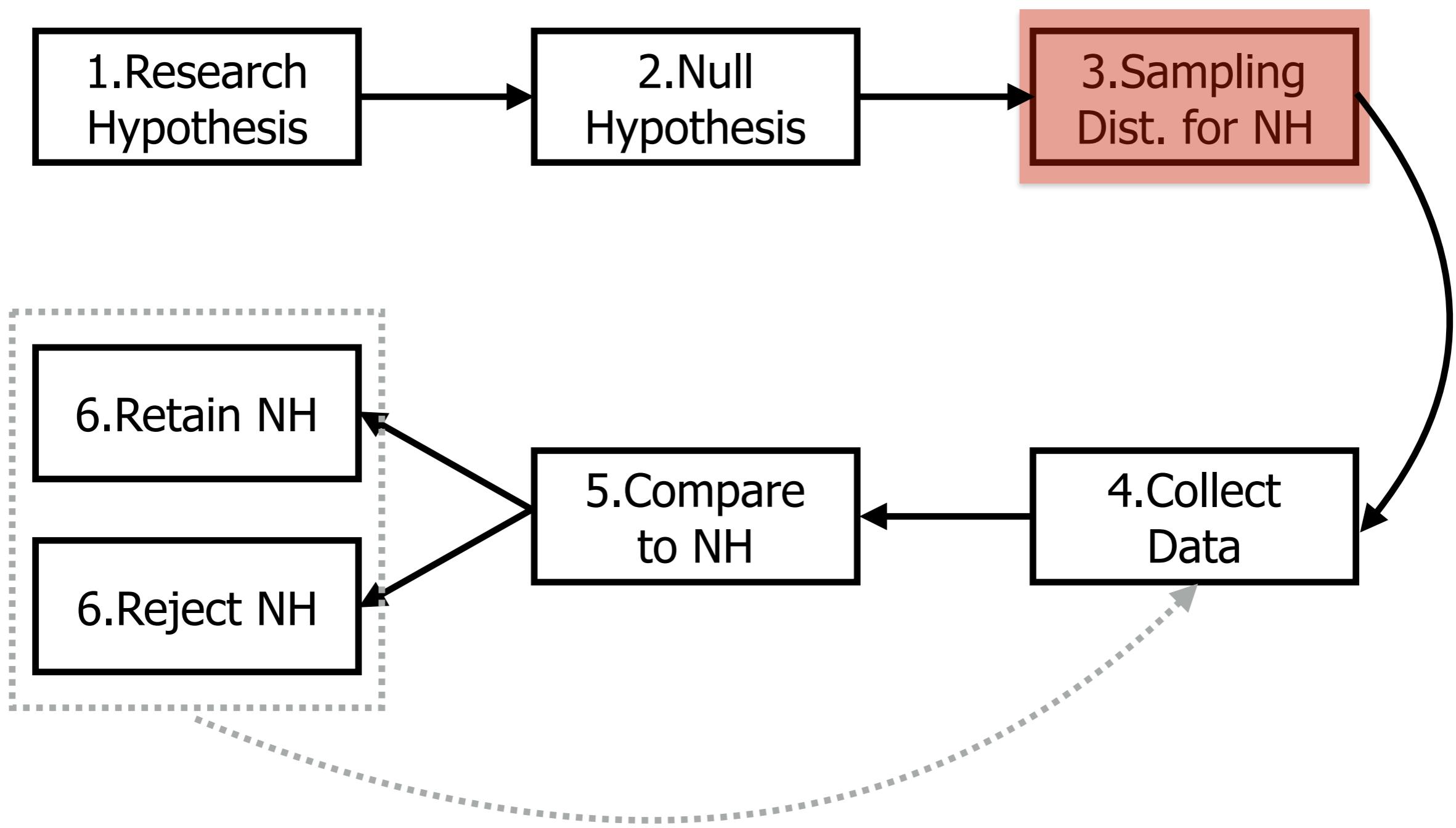
Research Hypothesis

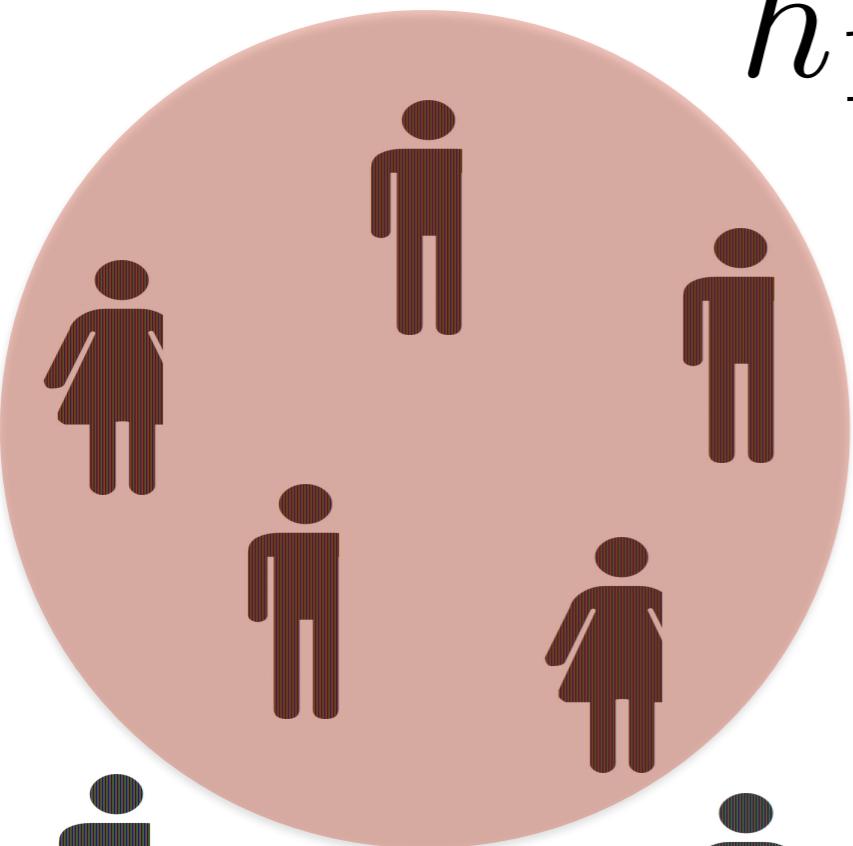
- Basketball players tend to be taller than the rest of the population;
- Primary school children tend to be smaller than the rest of the population.



Null Hypothesis

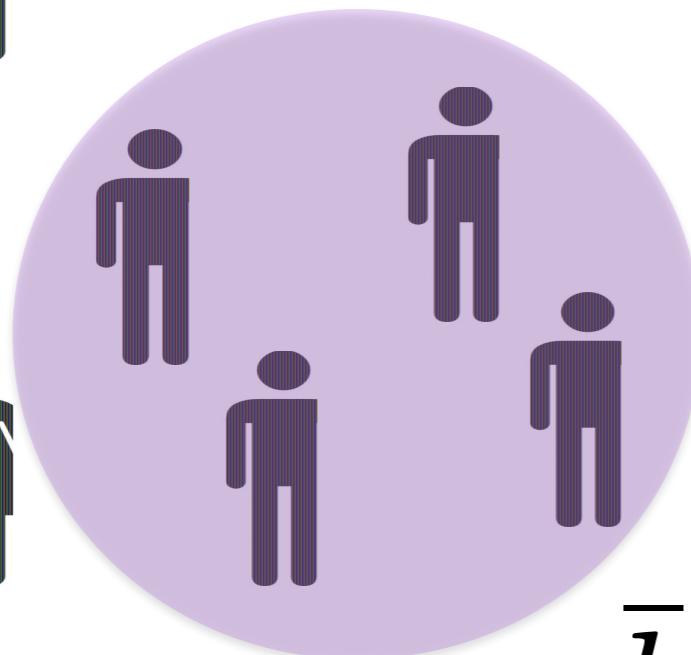
- Basketball players do not tend to be taller than the rest of the population;
- Primary school children do not tend to be smaller than the rest of the population.



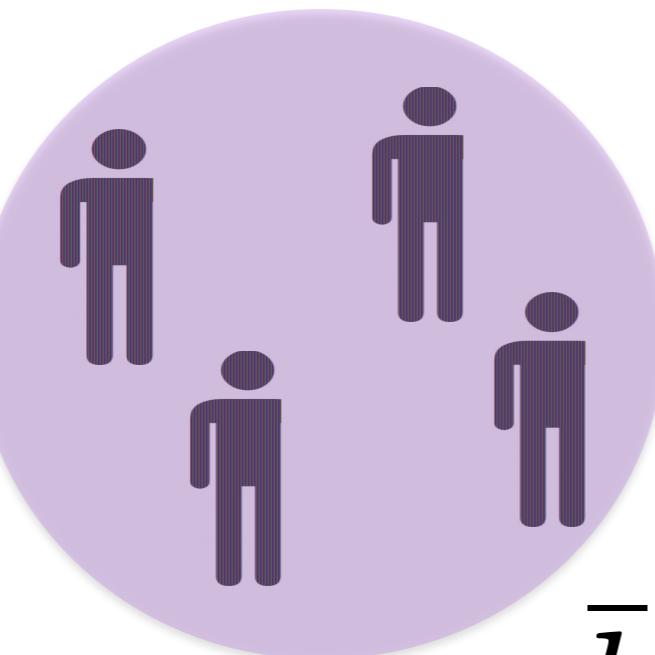
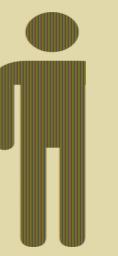


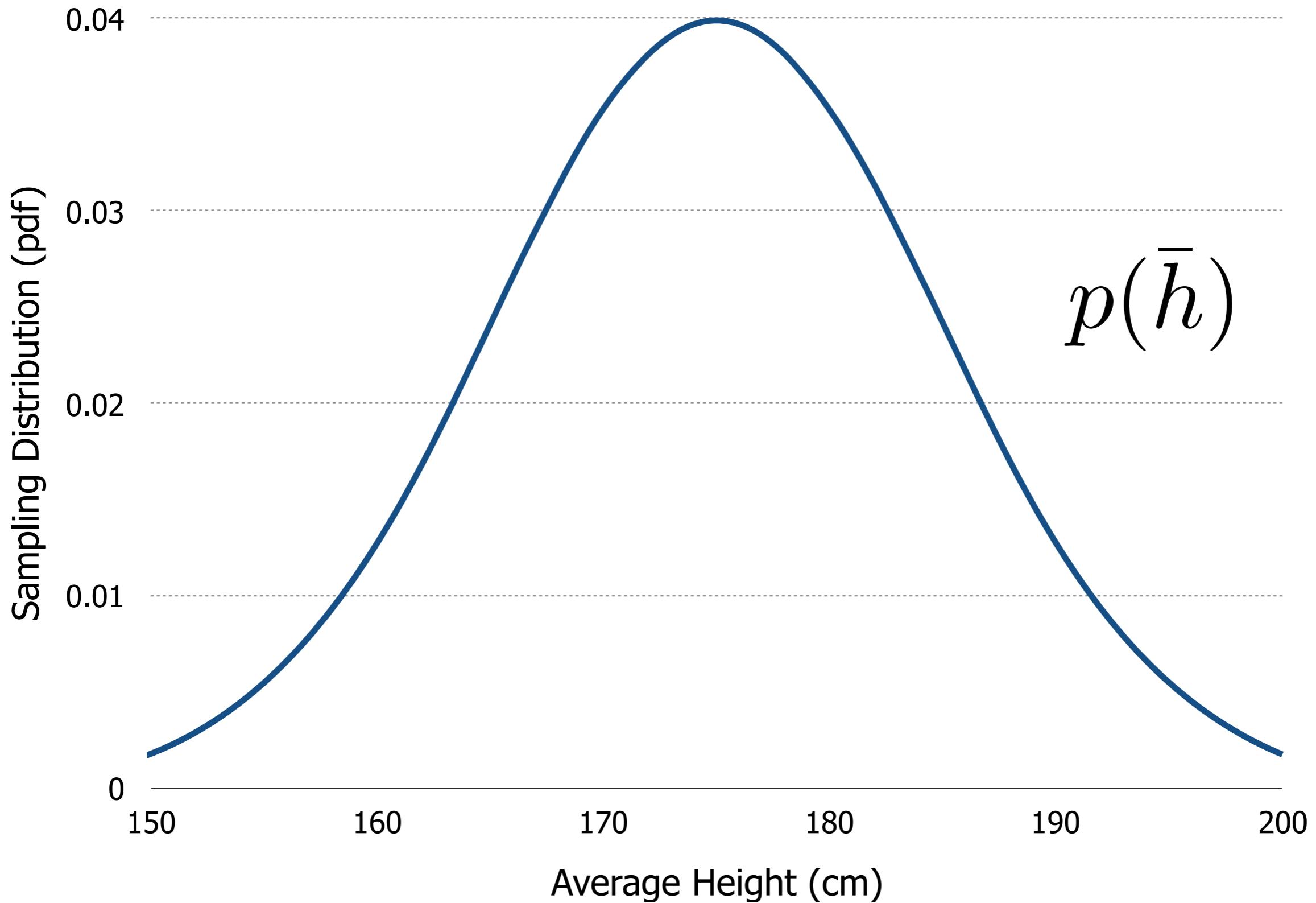
\bar{h}_1

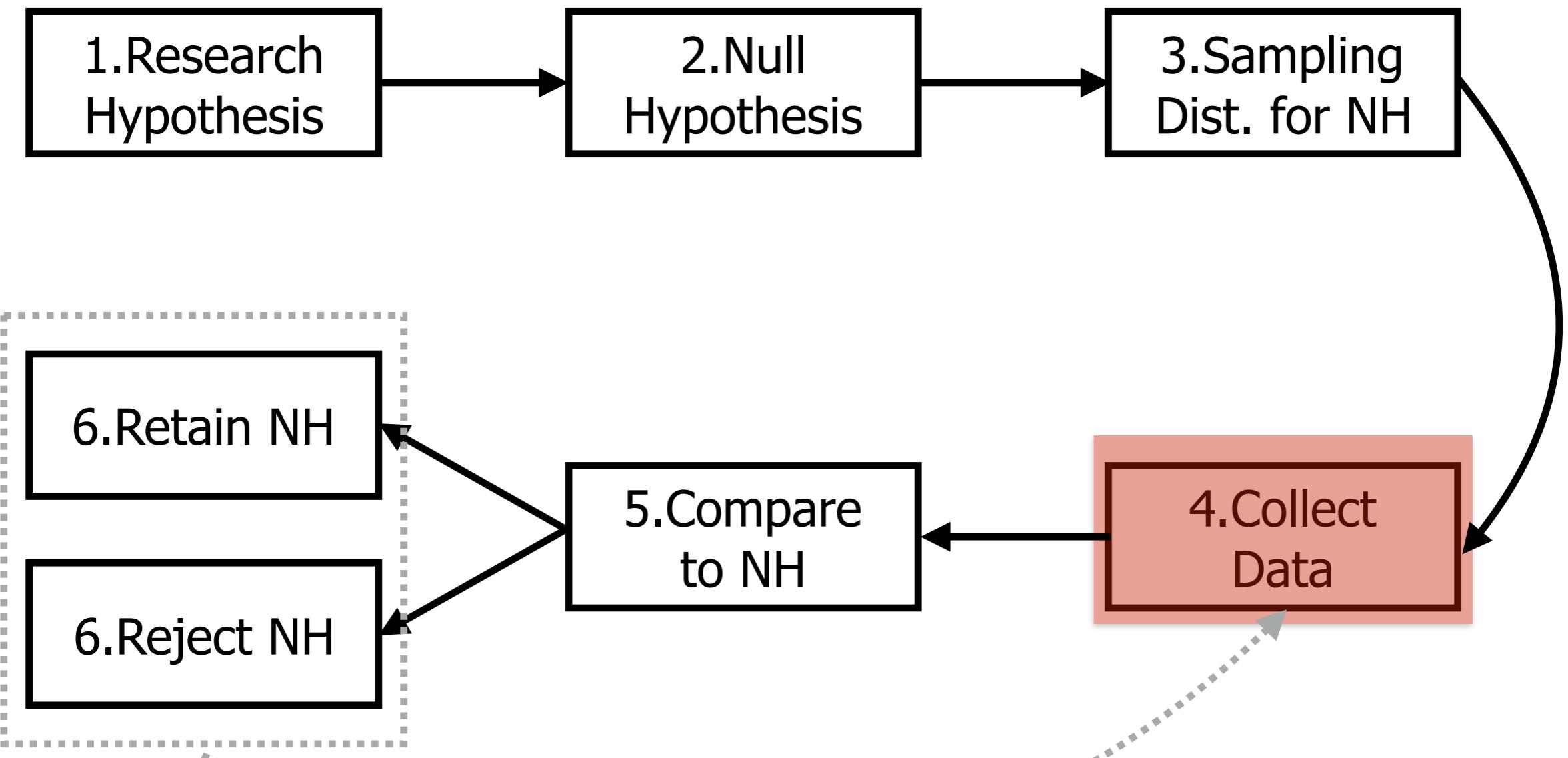
\bar{h}_2



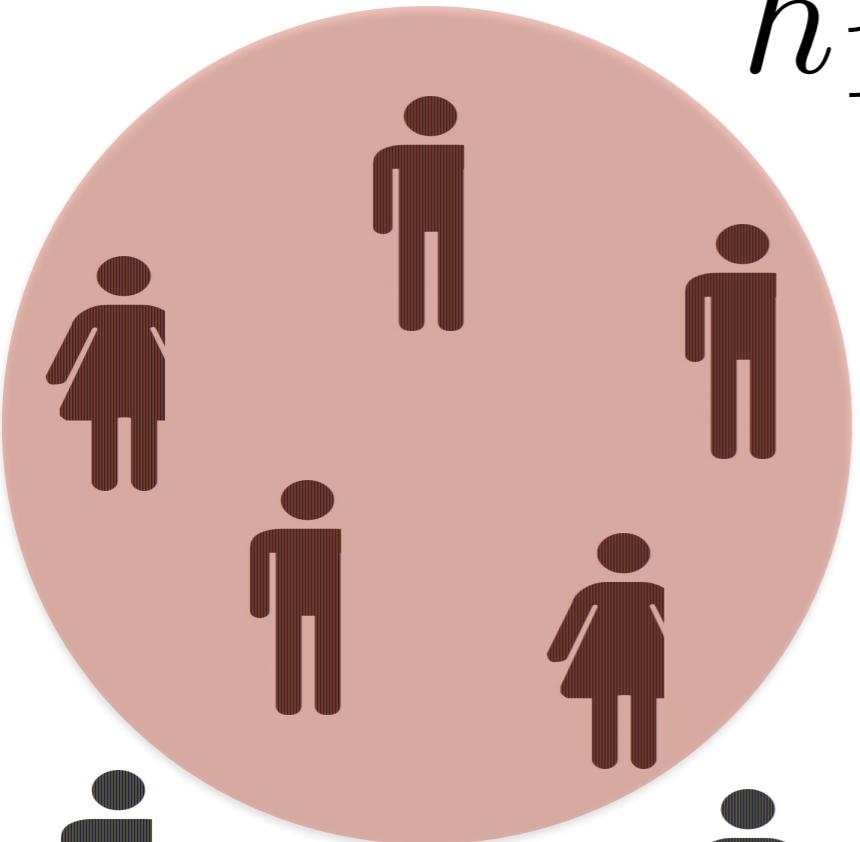
\bar{h}_3

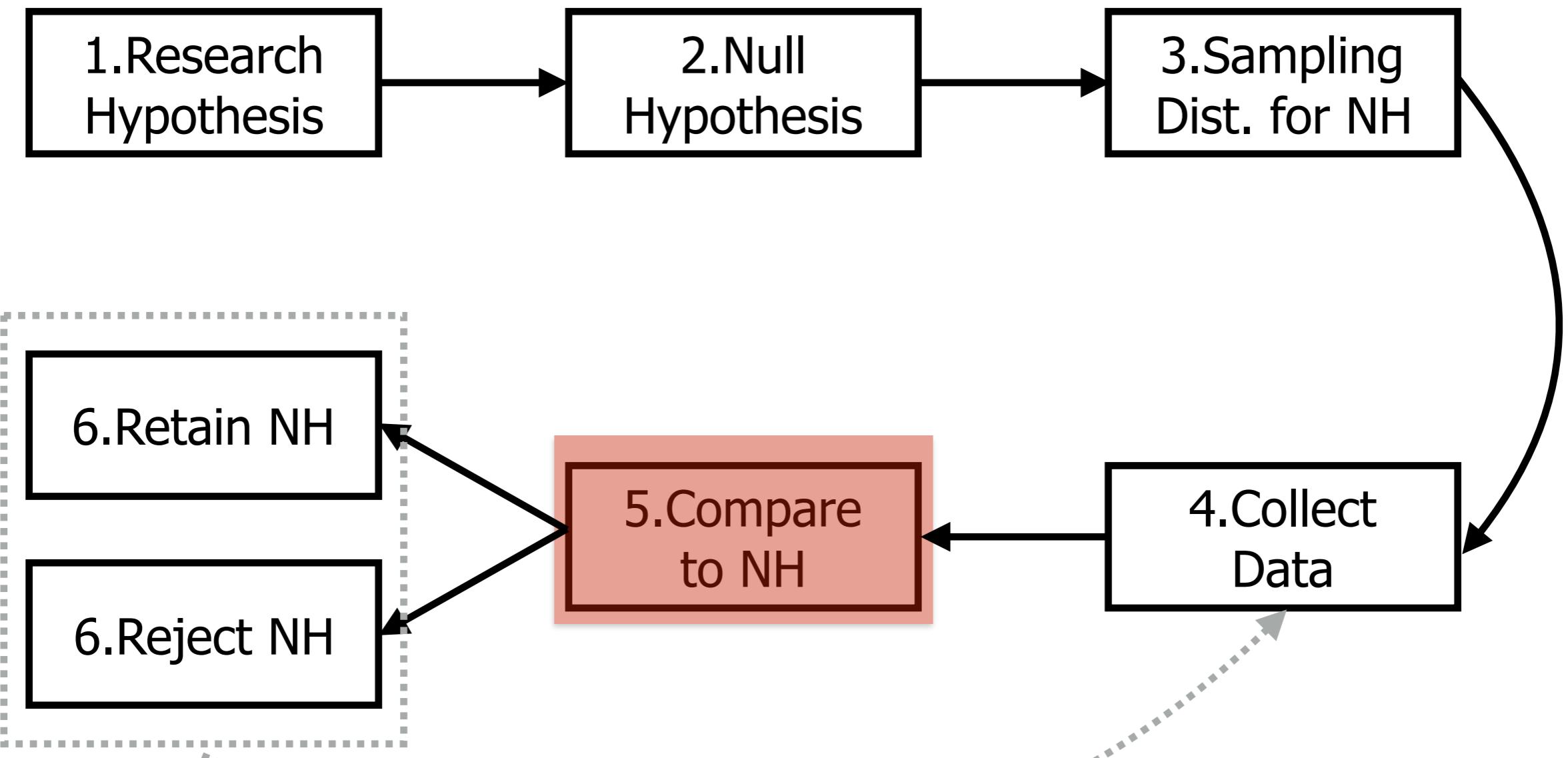


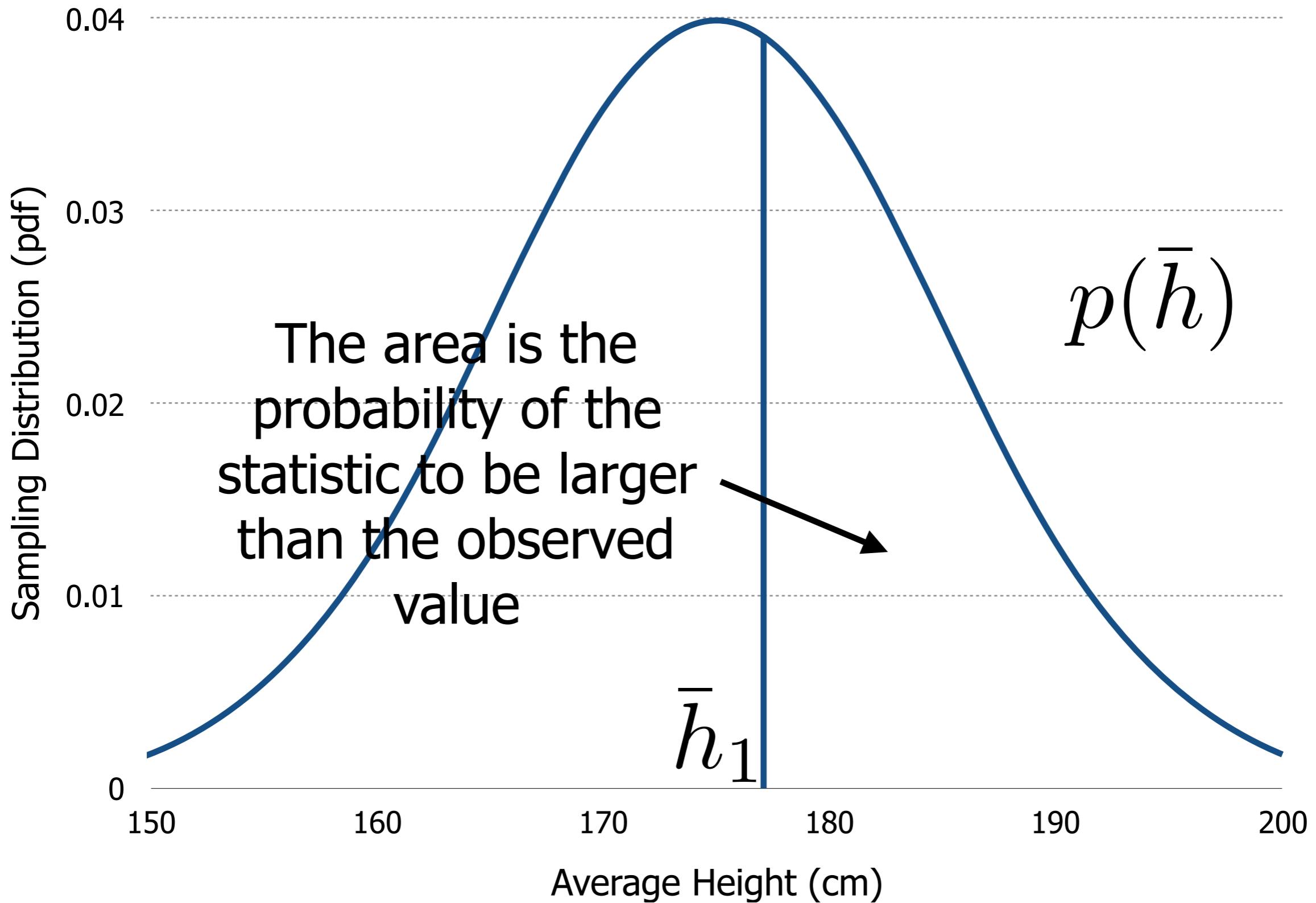


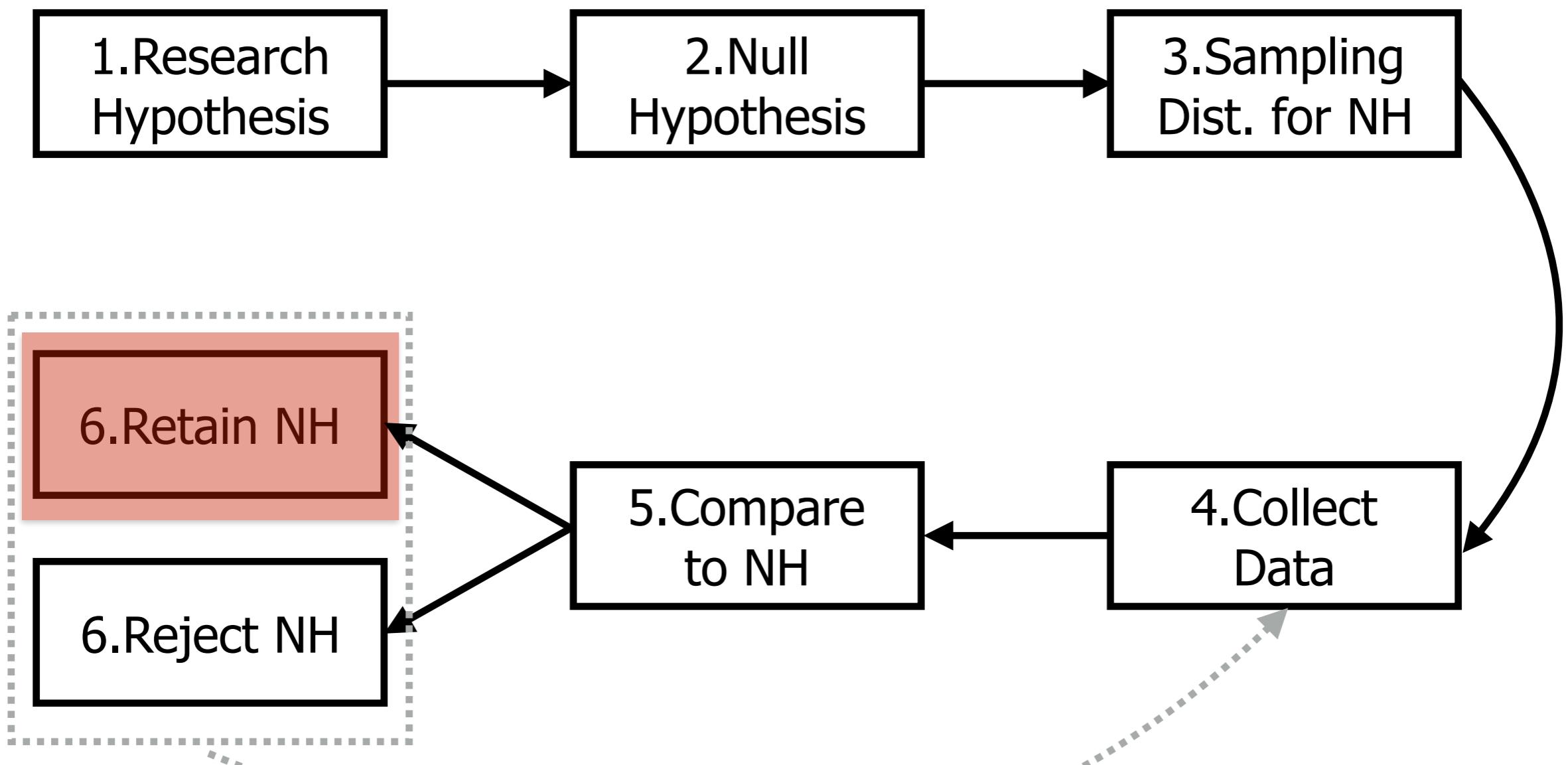


\bar{h}_1



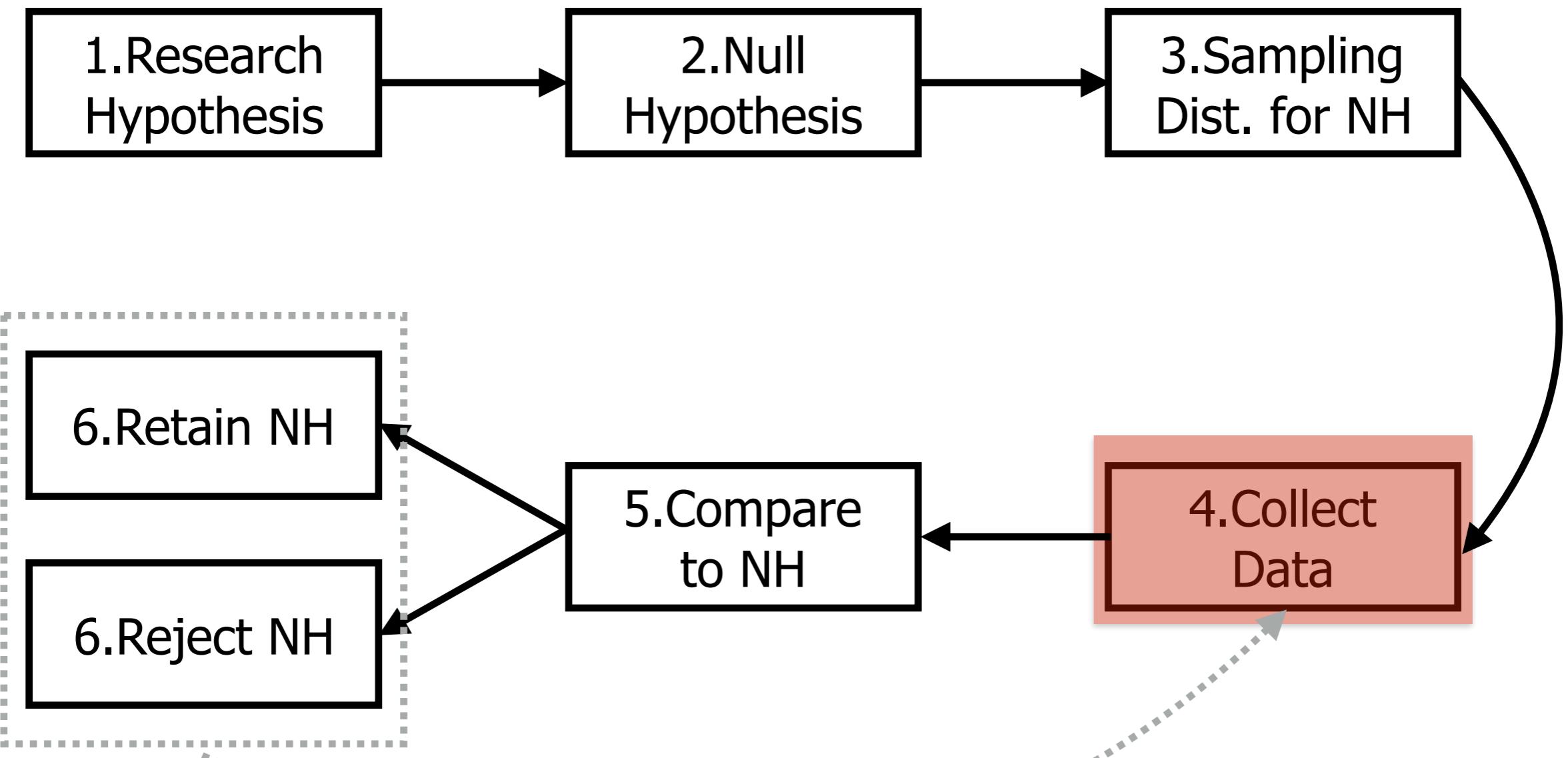




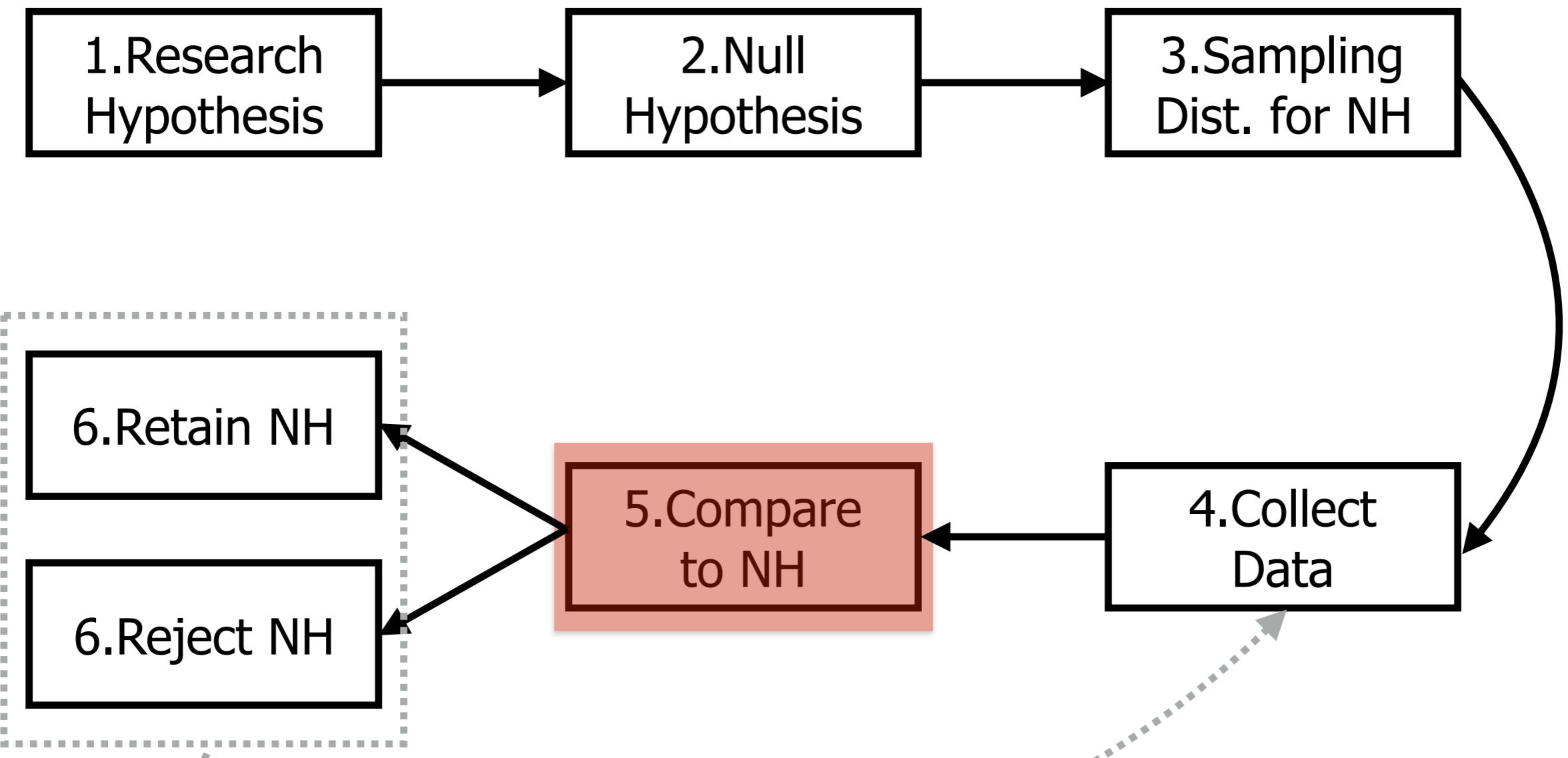


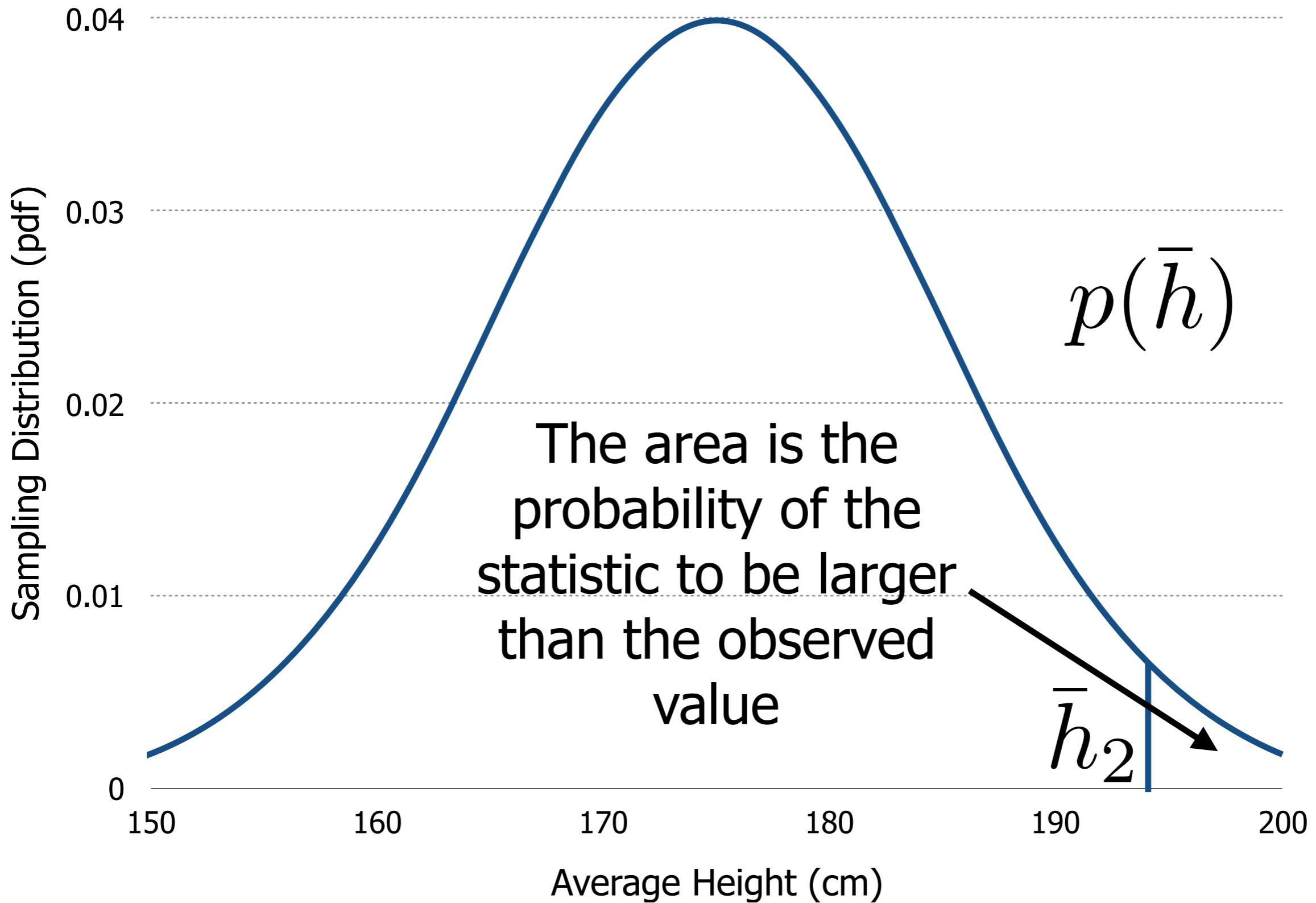
Fail to Reject the Null Hypothesis

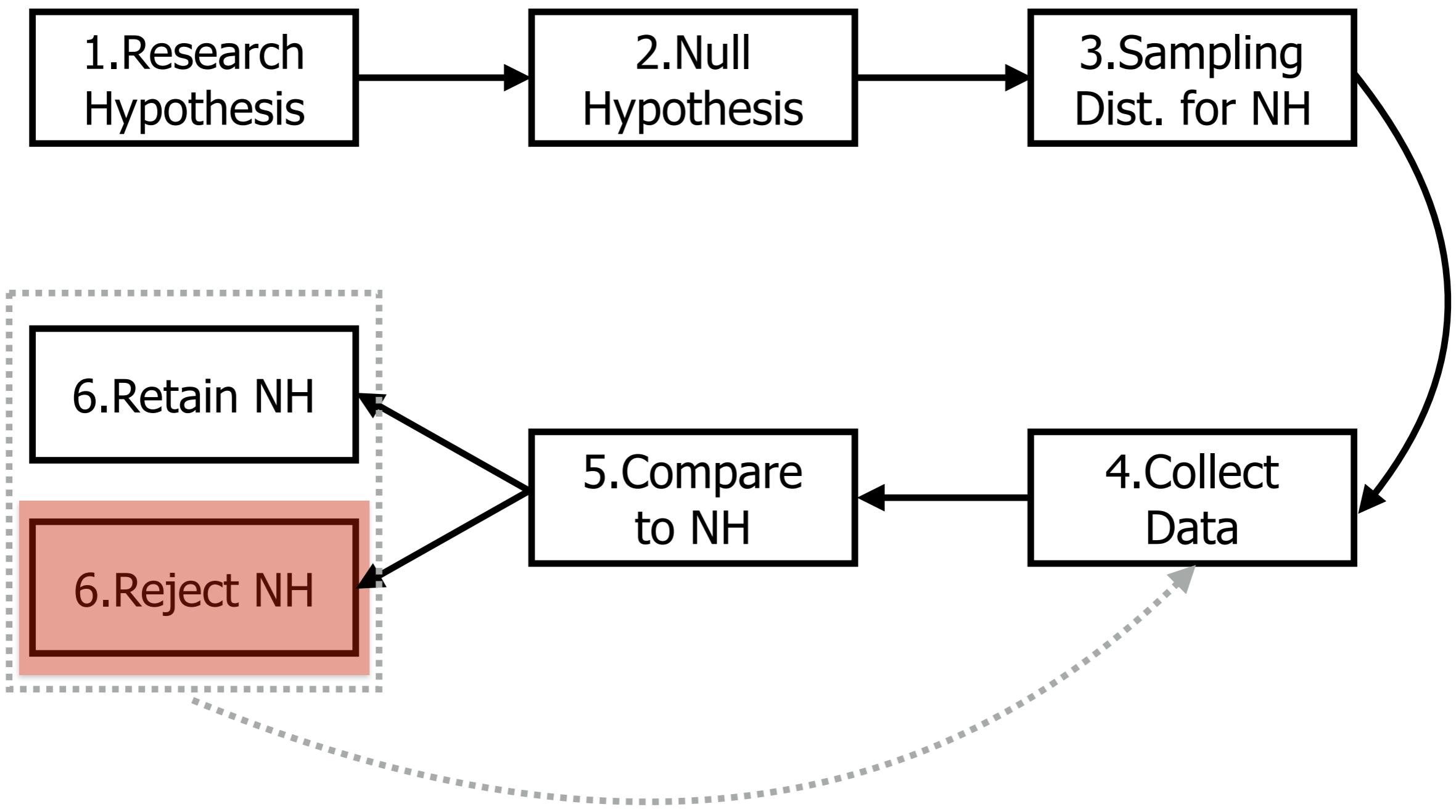
- When the probability to observe a average height higher or equal to the one I observe is large enough, the null hypothesis fails to be rejected;
- What “large enough” means exactly will be explained later.



\bar{h}_2 

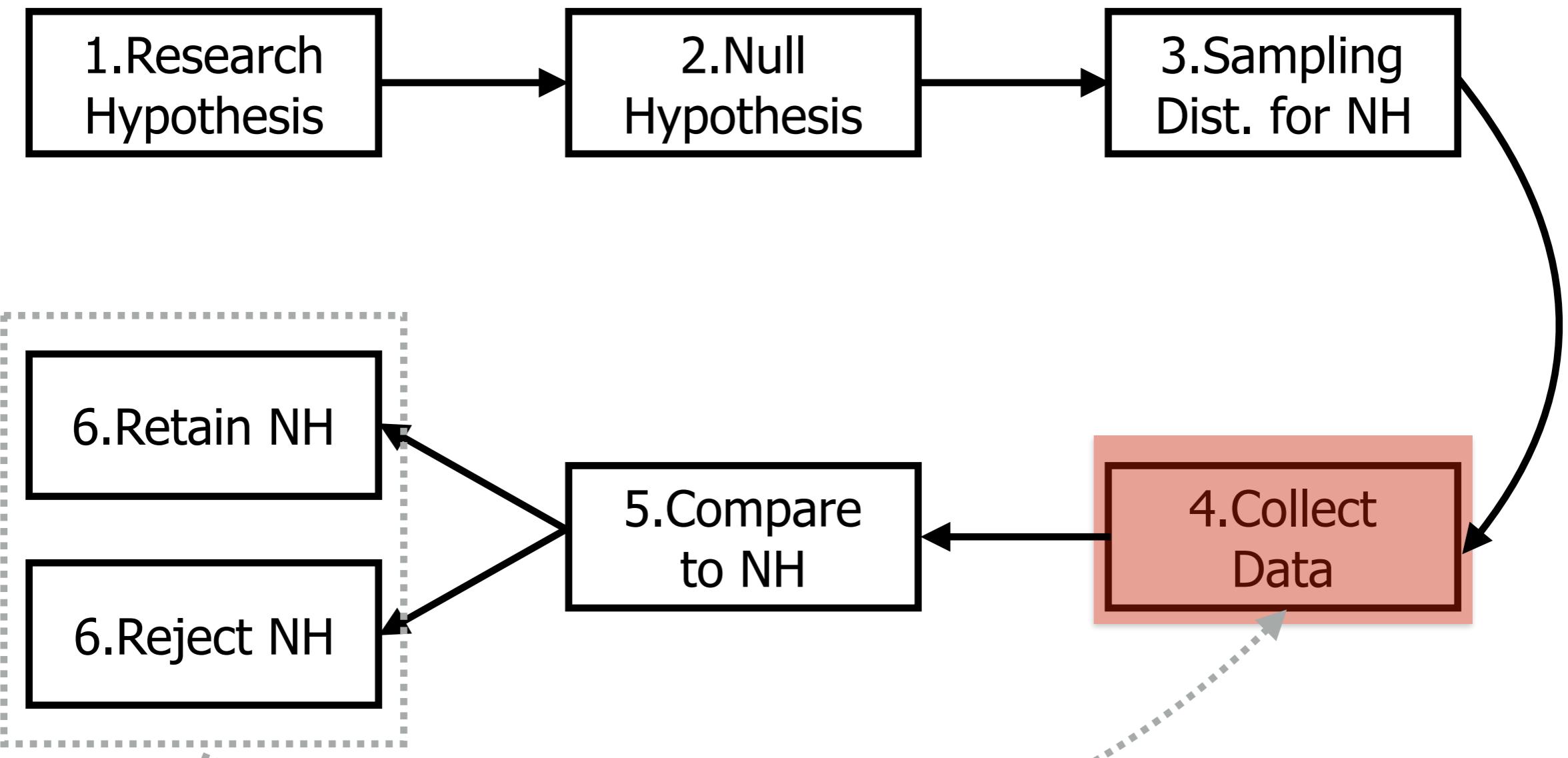


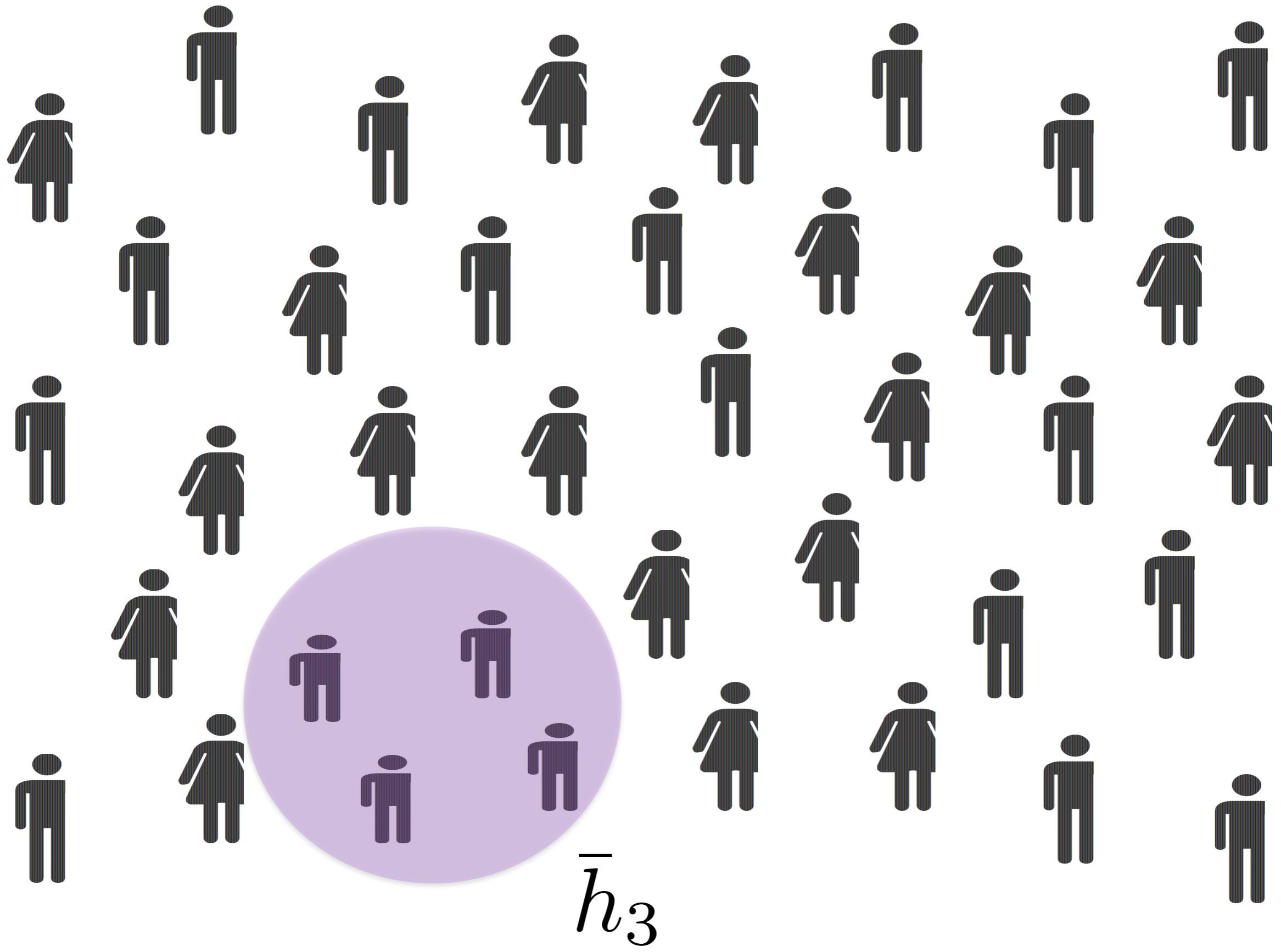


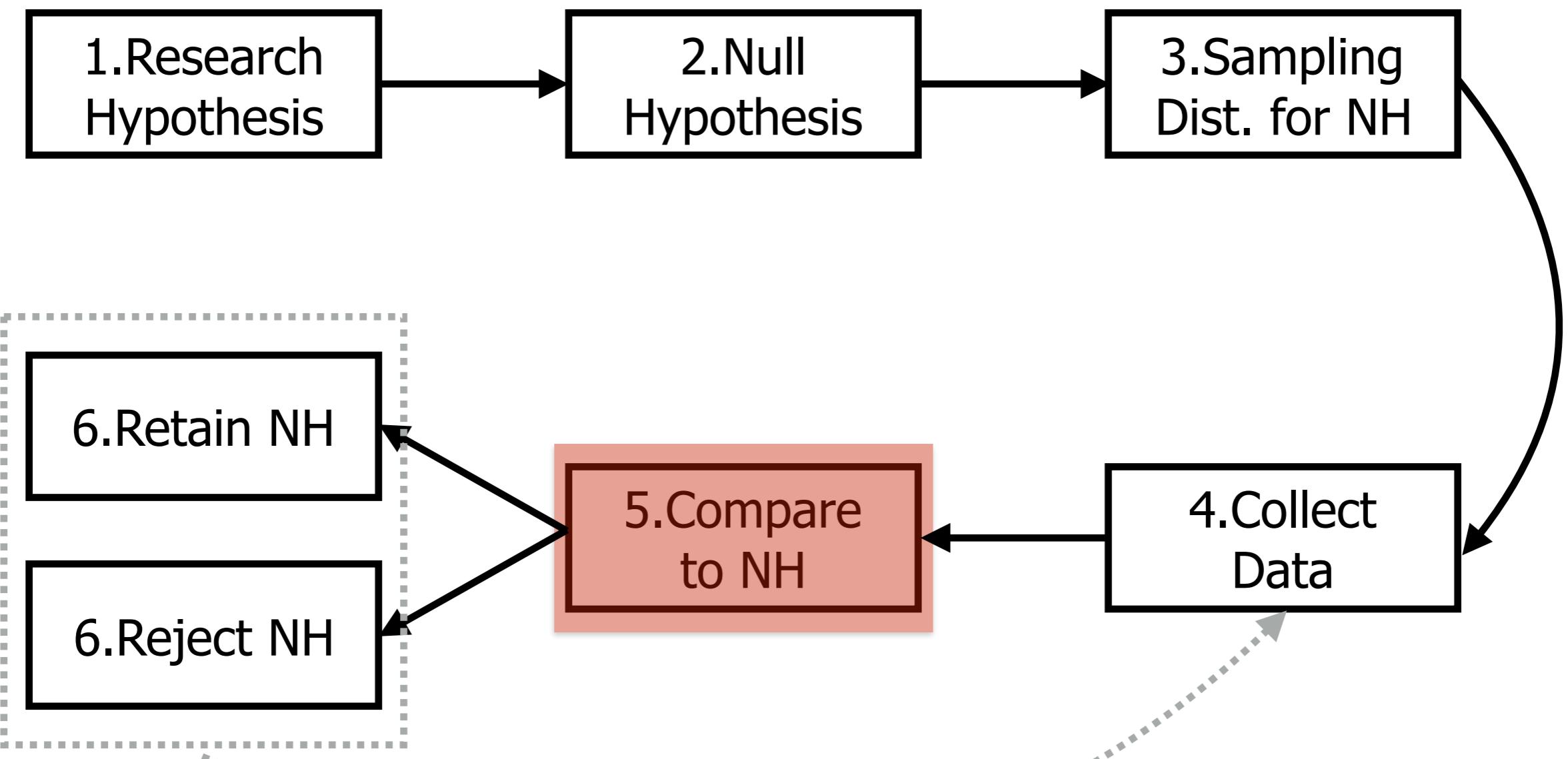


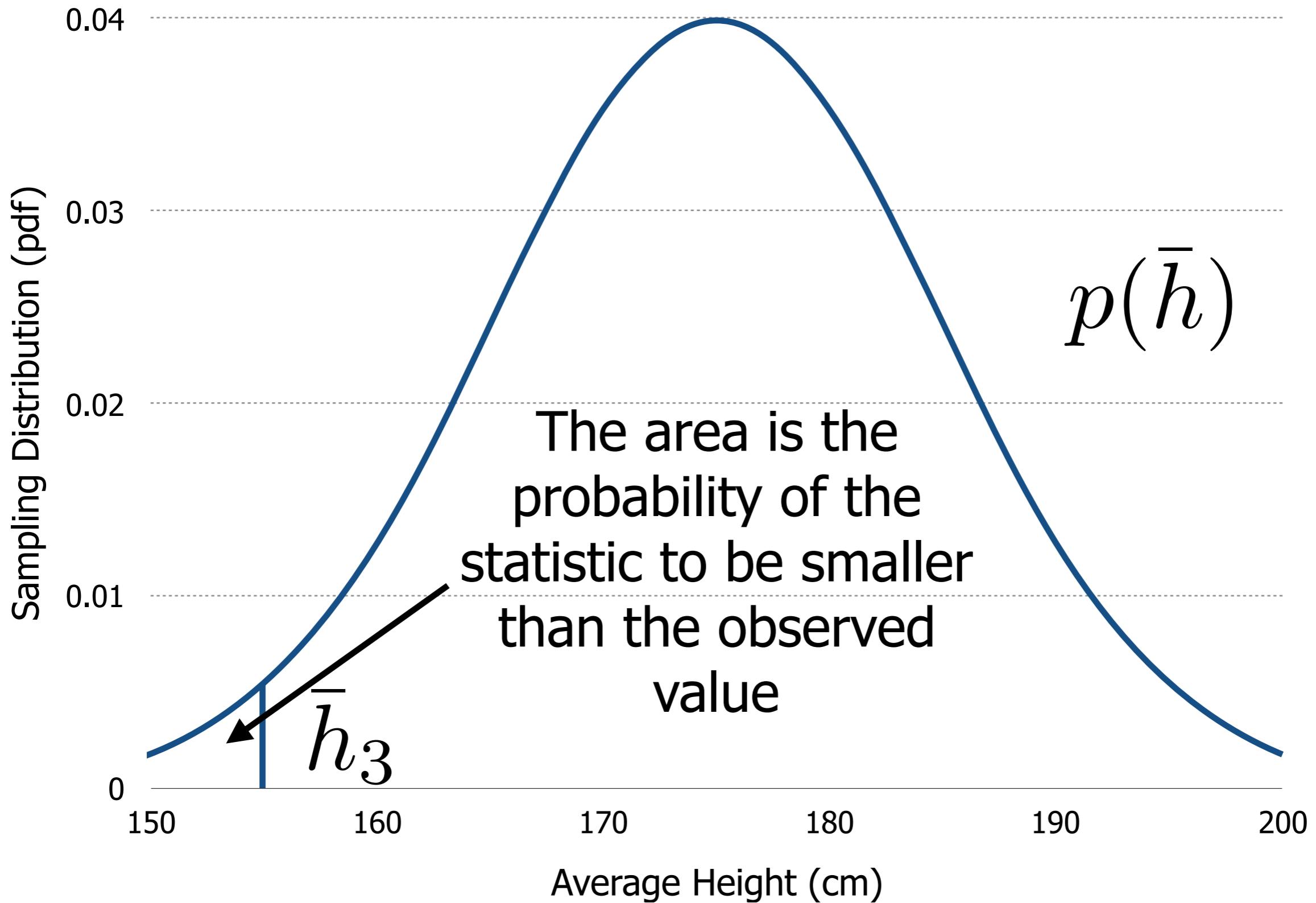
Reject the Null Hypothesis

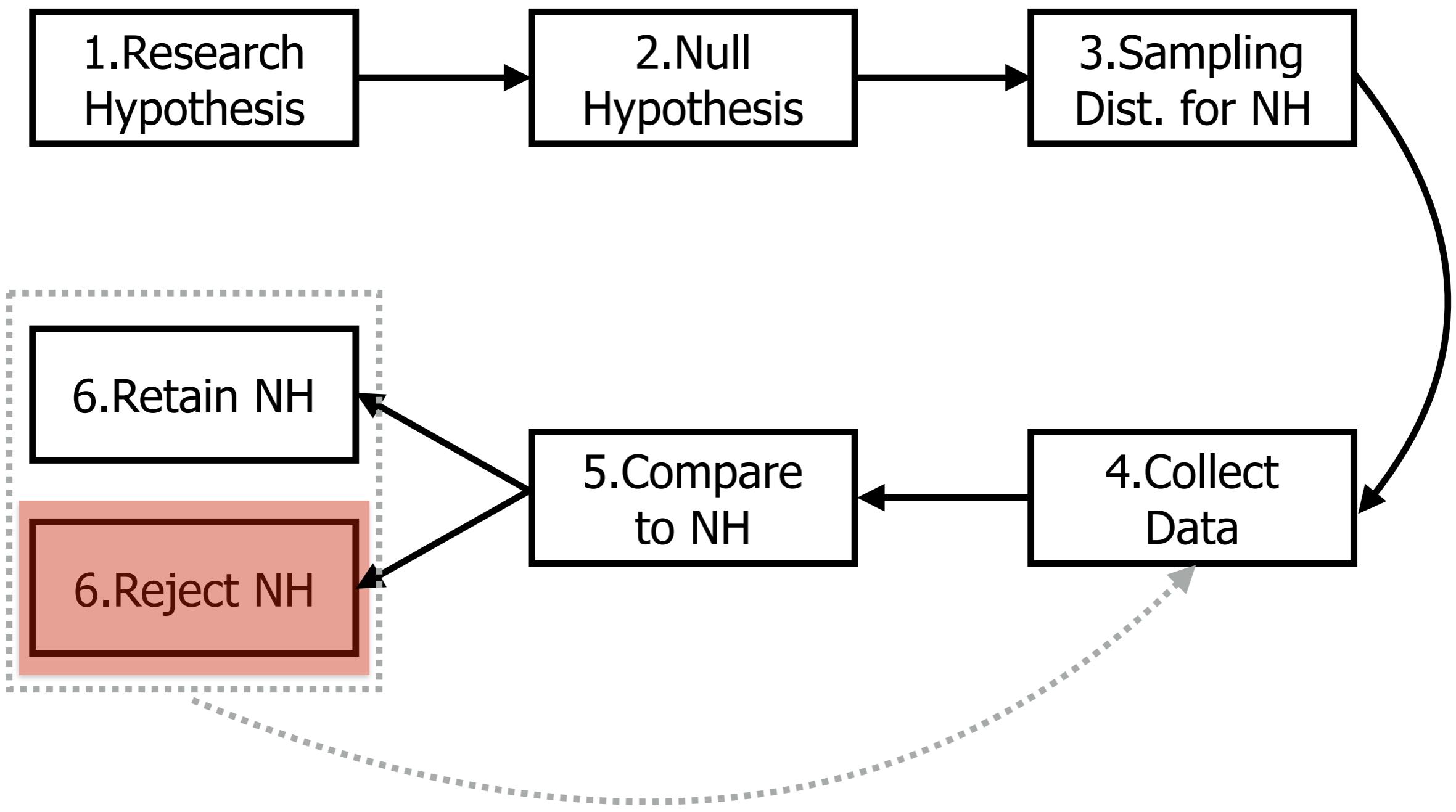
- When the probability to observe a average height higher than or equal to the one I observe is small enough, the null hypothesis can be rejected;
- What “small enough” means exactly will be explained later.











Reject the Null Hypothesis

- When the probability to observe a average height lower than or equal to the one I observe is small enough, the null hypothesis can be rejected;
- What “small enough” means exactly will be explained later.

Outline

- Sampling Distributions
- Hypothesis Testing
- The Null Hypothesis
- Error Types
- Directionality

The Null Hypothesis?

“[...] we can never prove something to be true, but we can prove something to be false. Observing 3000 people with two arms does not prove the statement ‘Everyone has two arms’. However, finding one person with three arms does disprove the original statement [...]”

D.C.Howell, “Statistical Methods for Psychology”, Chapter 4,
Cengage Learning, 2009.

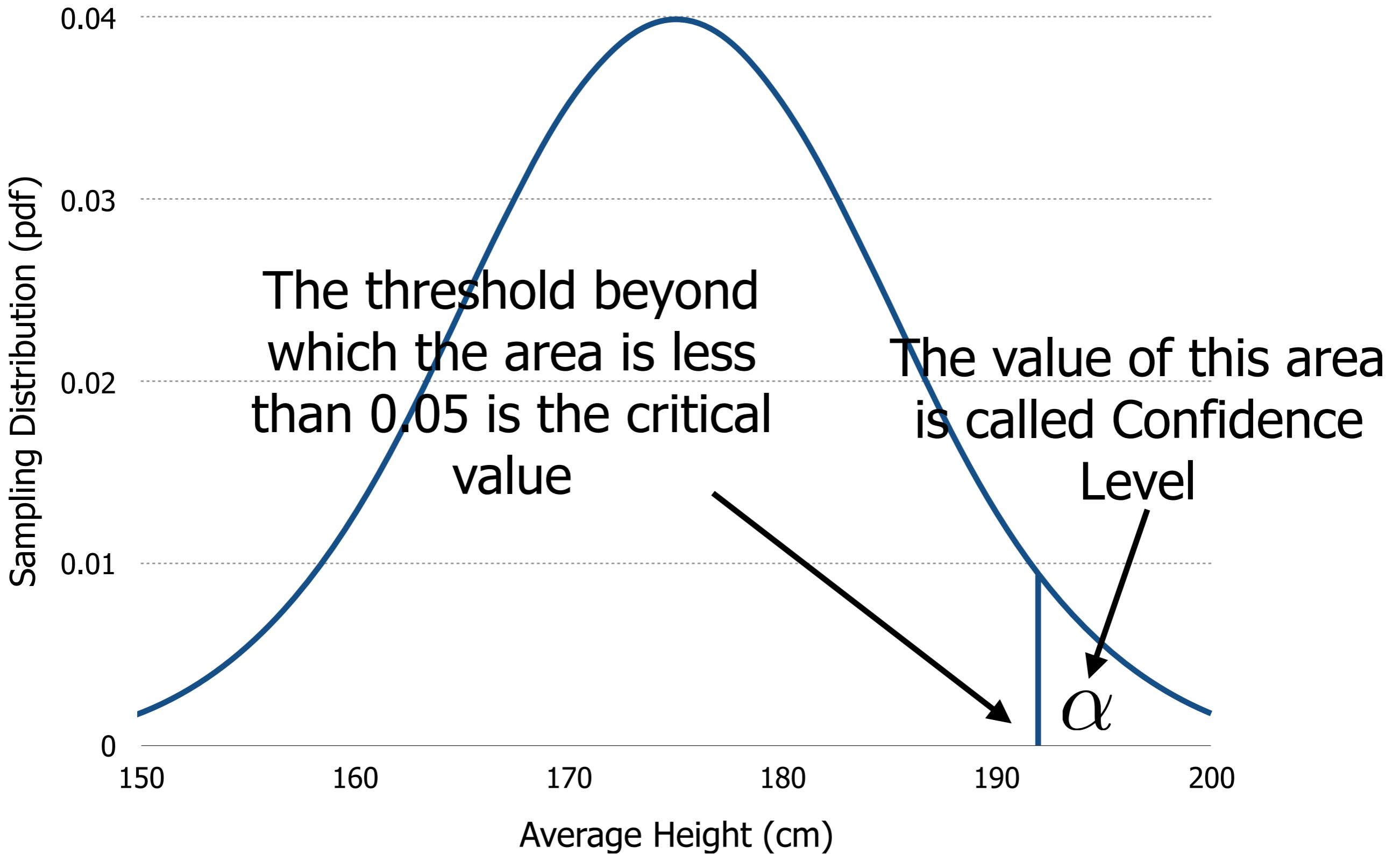
The Null Hypothesis (II)

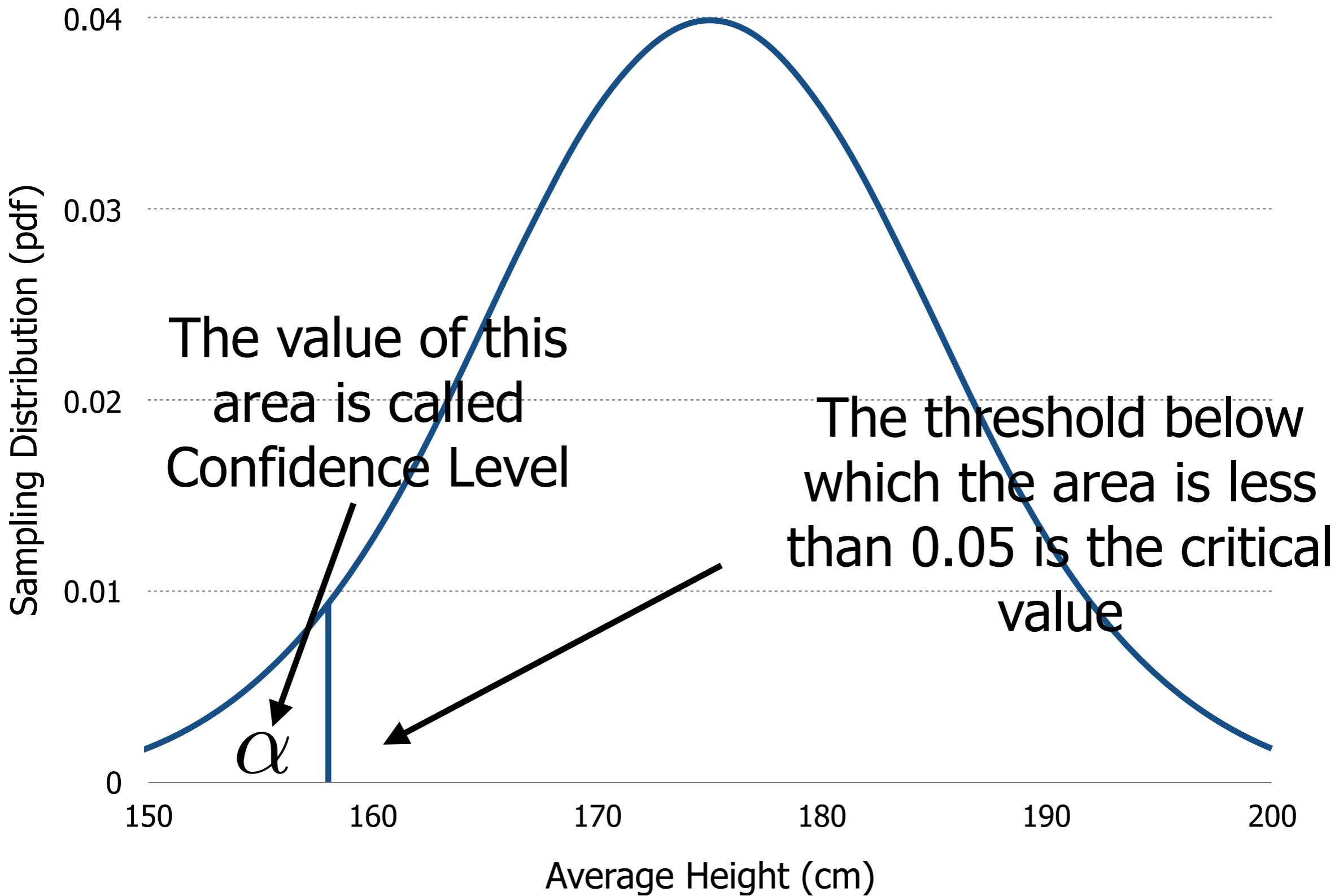
“The one thing on which all statisticians agree is that we can never claim to have ‘proved’ the null hypothesis.”

D.C.Howell, “Statistical Methods for Psychology”, Chapter 4,
Cengage Learning, 2009.

Outline

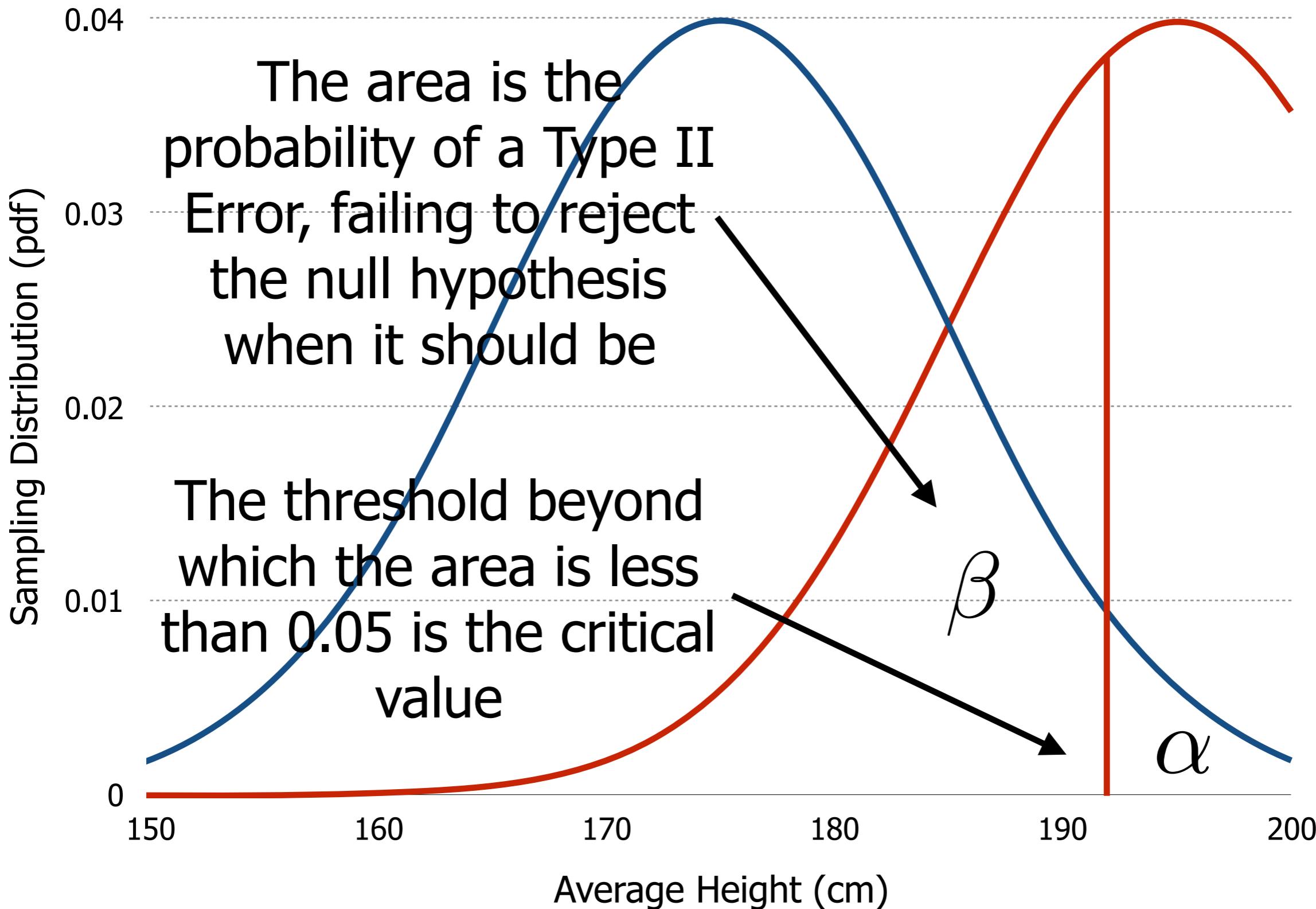
- Sampling Distributions
- Hypothesis Testing
- The Null Hypothesis
- Error Types
- Directionality





Type I Errors

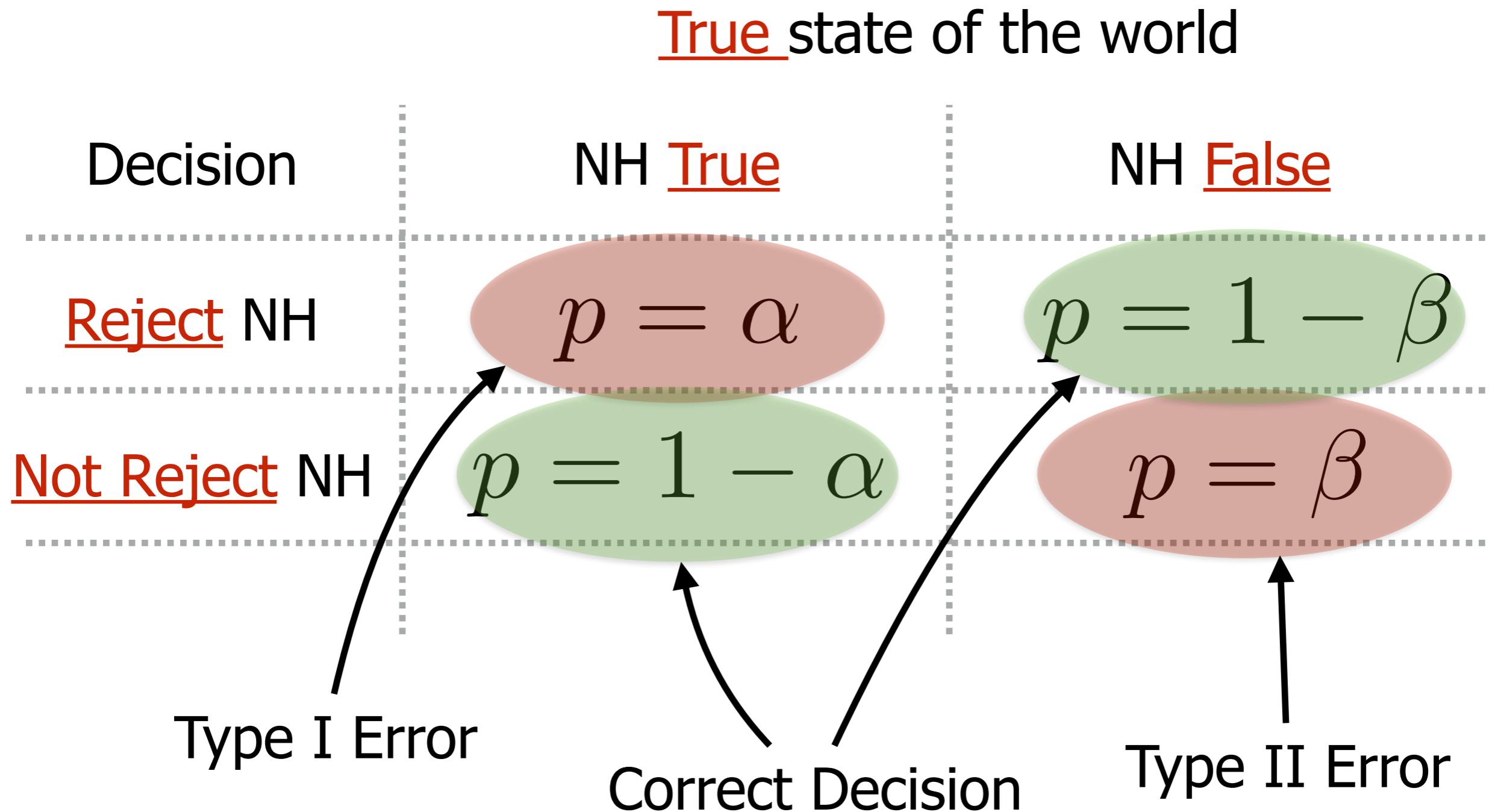
- The people below or beyond the critical values are “normal” (not necessarily primary school pupils or basketball players);
- The value $p=0.05$ is the probability of rejecting the null hypothesis when it is true, a Type I Error;
- It is a conventional confidence level widely accepted in experimental practice.



Type II Errors

- A Type II Error takes place when the null hypothesis should be rejected, but it does not;
- Reducing the probability of a Type I Error automatically increases the probability of a Type II Error;
- A good tradeoff requires one to know the distribution of a statistic in the population where the null hypothesis is false (e.g., the basketball players).

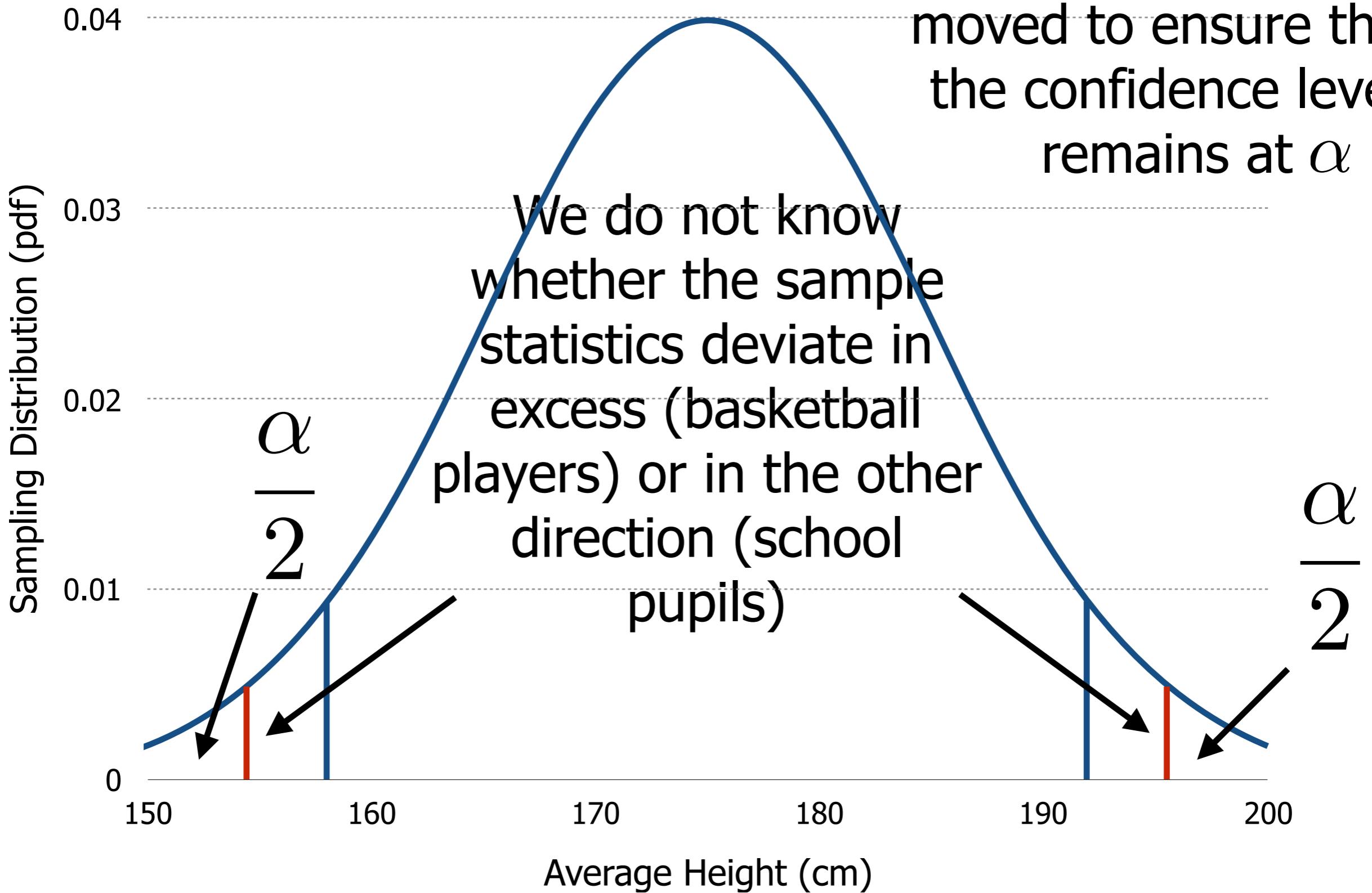
Decision Making Process



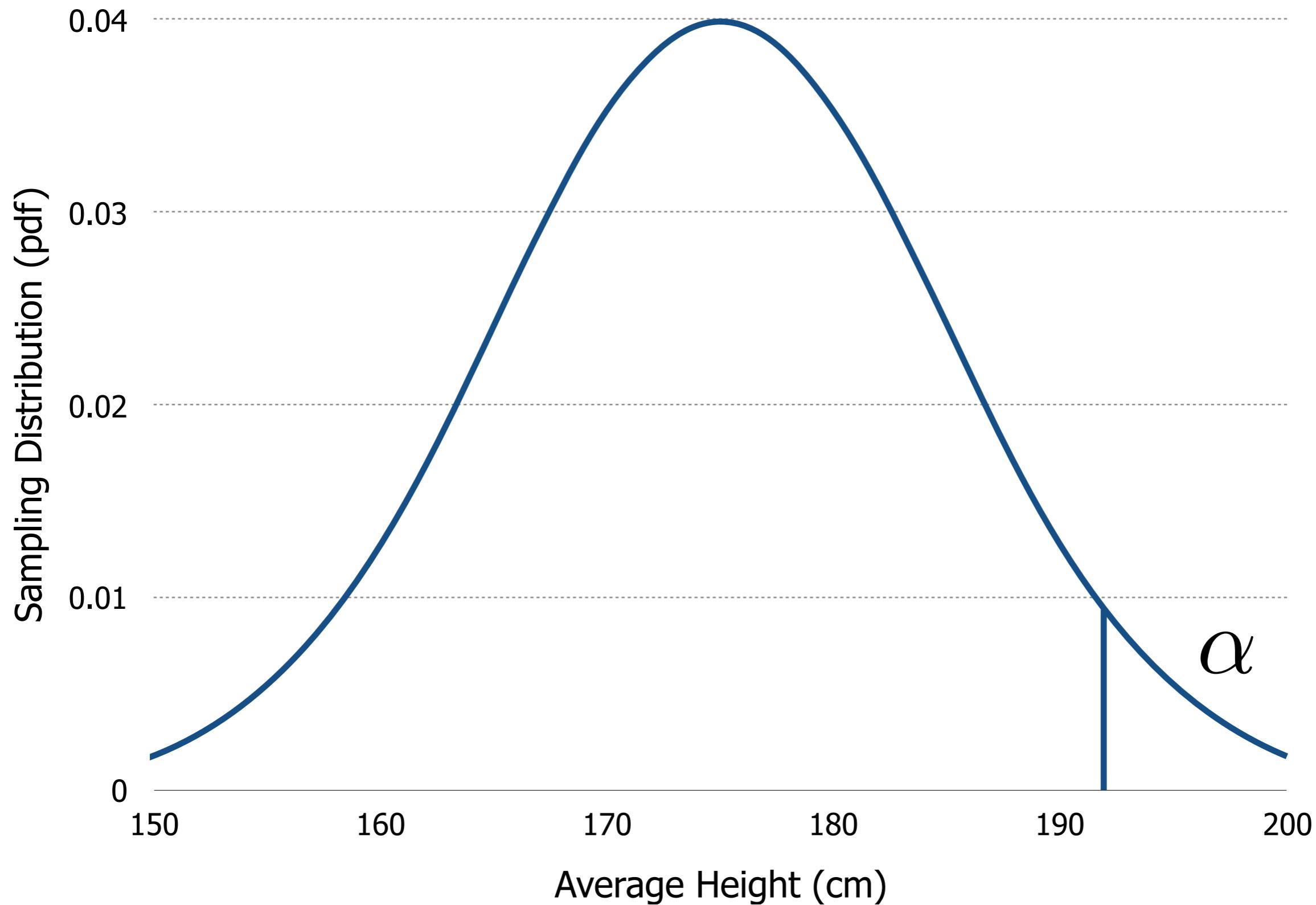
Outline

- Sampling Distributions
- Hypothesis Testing
- The Null Hypothesis
- Error Types
- Directionality

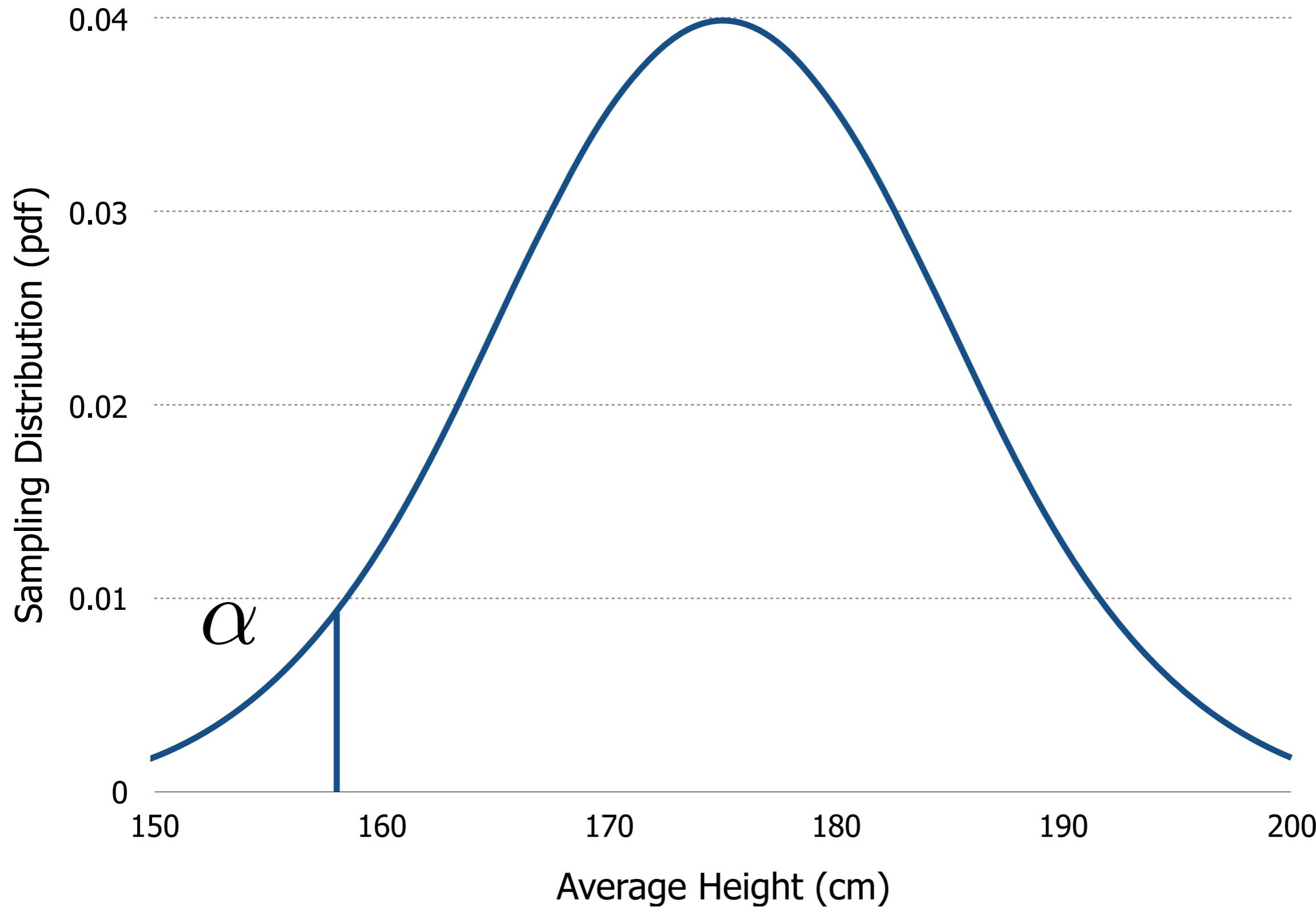
The critical values are moved to ensure that the confidence level remains at α



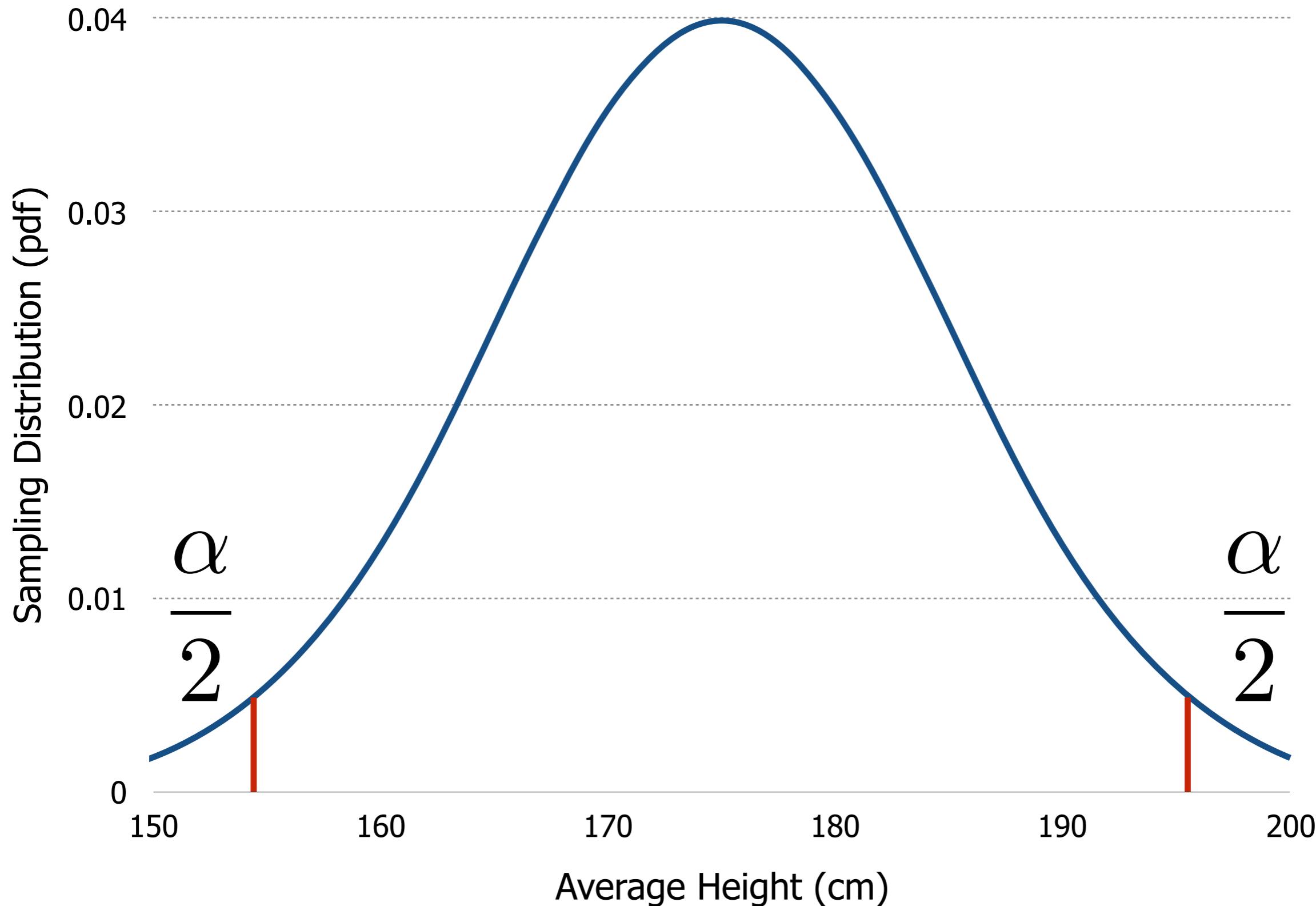
One-Tailed (Directional) Test



One-Tailed (Directional) Test



Two-Tailed (Nondirectional) Test



One vs Two-Tailed

- Two tailed tests ensure that deviations in both senses are taken into account;
- It rarely happens that a research hypothesis can be stated in directional terms with sufficient certainty;
- Both types of test are correct (they just tell a different story), but a two-tailed test is, on average, a safer choice.

Thank You!

The Chi Square

Computational Social Intelligence - Lecture 04

Prof. Alessandro Vinciarelli
School of Computing Science &
Institute of Neuroscience and Psychology

<http://www.dcs.gla.ac.uk/vincia>
Alessandro.Vinciarelli@glasgow.ac.uk



University
of Glasgow

EPSRC
Engineering and Physical Sciences
Research Council

FNSNF

This lecture is based on the following text
(available on Moodle):

- D.C.Howell, “Statistical Methods for Psychology”, Chapter 6, Sections 6.1, 6.3 and 6.4 (excluding subsection “Correcting for Continuity”), Cengage Learning, 2009.

The extra-material available in the pdf of the text does not need to be studied

Outline

- Very Quick Recap
- Goodness of Fit Test (One Way Classification)
- Two Way Classification
- Conclusions

Outline

- Very Quick Recap
- Goodness of Fit Test (One Way Classification)
- Two Way Classification
- Conclusions

Hypothesis Testing (Main)

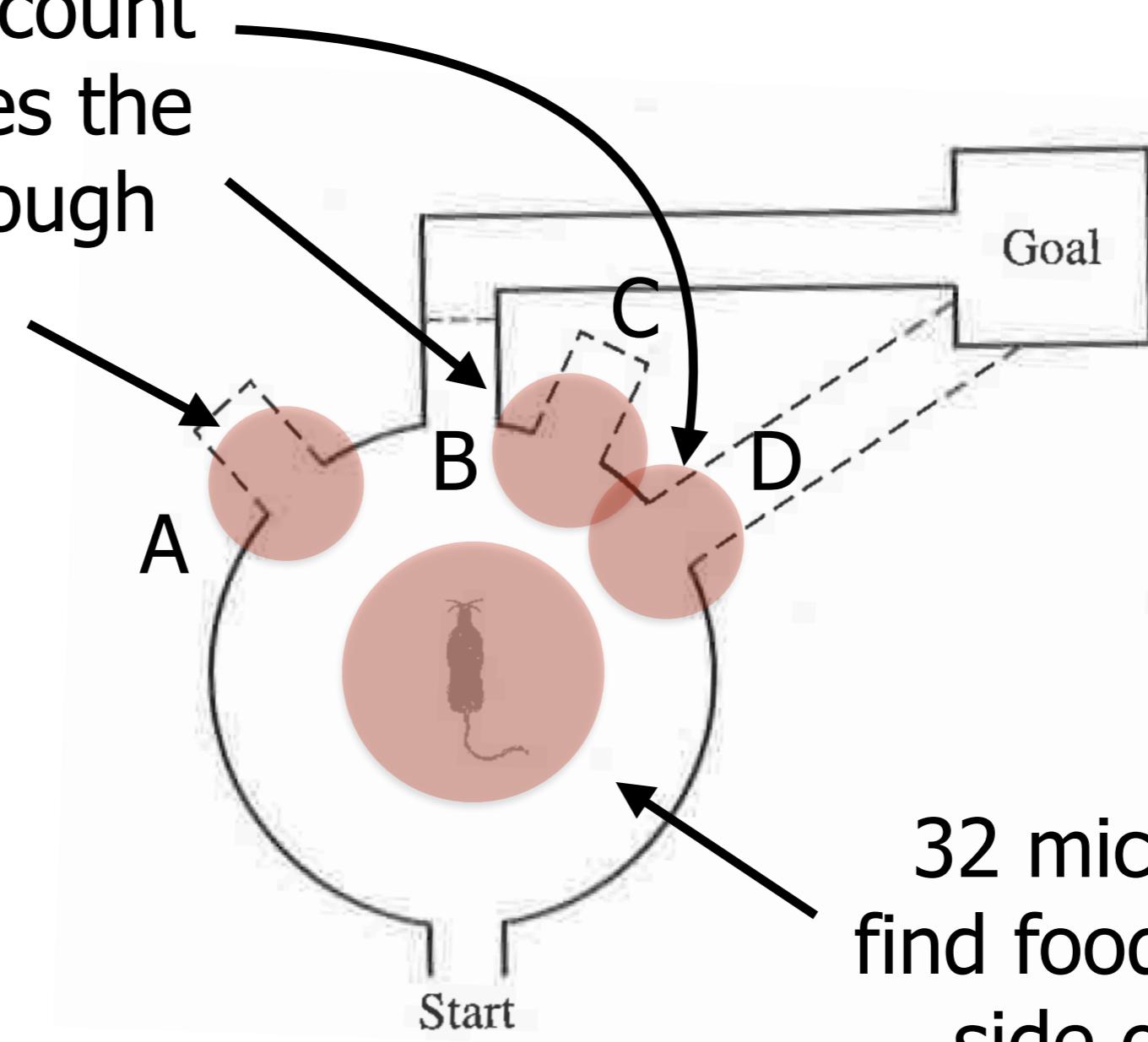
Ingredients

- Statistic: Any measurement that can be extracted from a data sample;
- Sampling Distribution: probability density function of the statistic when the Null Hypothesis is true;
- Confidence Level: an acceptable probability of doing a Type I Error (rejecting the Null Hypothesis when it should not), typically 5%.

Outline

- Very Quick Recap
- Goodness of Fit Test (One Way Classification)
- Two Way Classification
- Conclusions

When the other alleys
are open, the
experimenters count
how many times the
mice pass through
them



32 mice “learn” to
find food on the right
side of the alley

Tolman, Ritchie and Kalish, "Studies in spatial learning. II. Place Learning vs Response Learning", Journal of Experimental Psychology, 36(3):221, 1946.

Research Hypothesis

- Research Hpothesis: The mice learn that the food is on the right and tend to select alleys that go in such a direction;
- Null Hypothesis: The mice select randomly one of the alleys.

Results

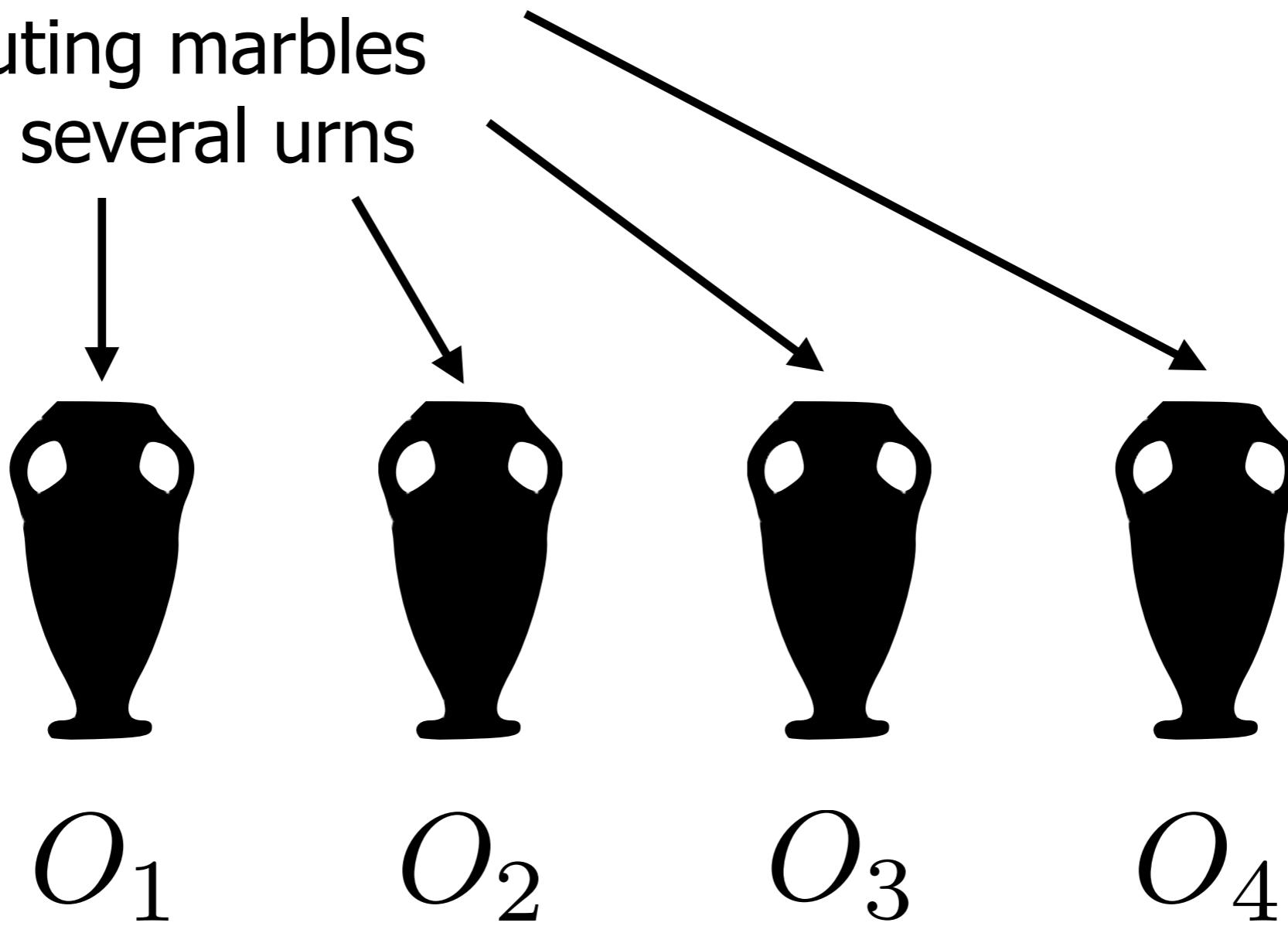
Alley Chosen	A	B	C	D
Observed (O)	4	5	8	15
Expected (E)	8	8	8	8

Tolman, Ritchie and Kalish, "Studies in spatial learning. II. Place Learning vs Response Learning", Journal of Experimental Psychology, 36(3):221, 1946.

Observations and Expectations

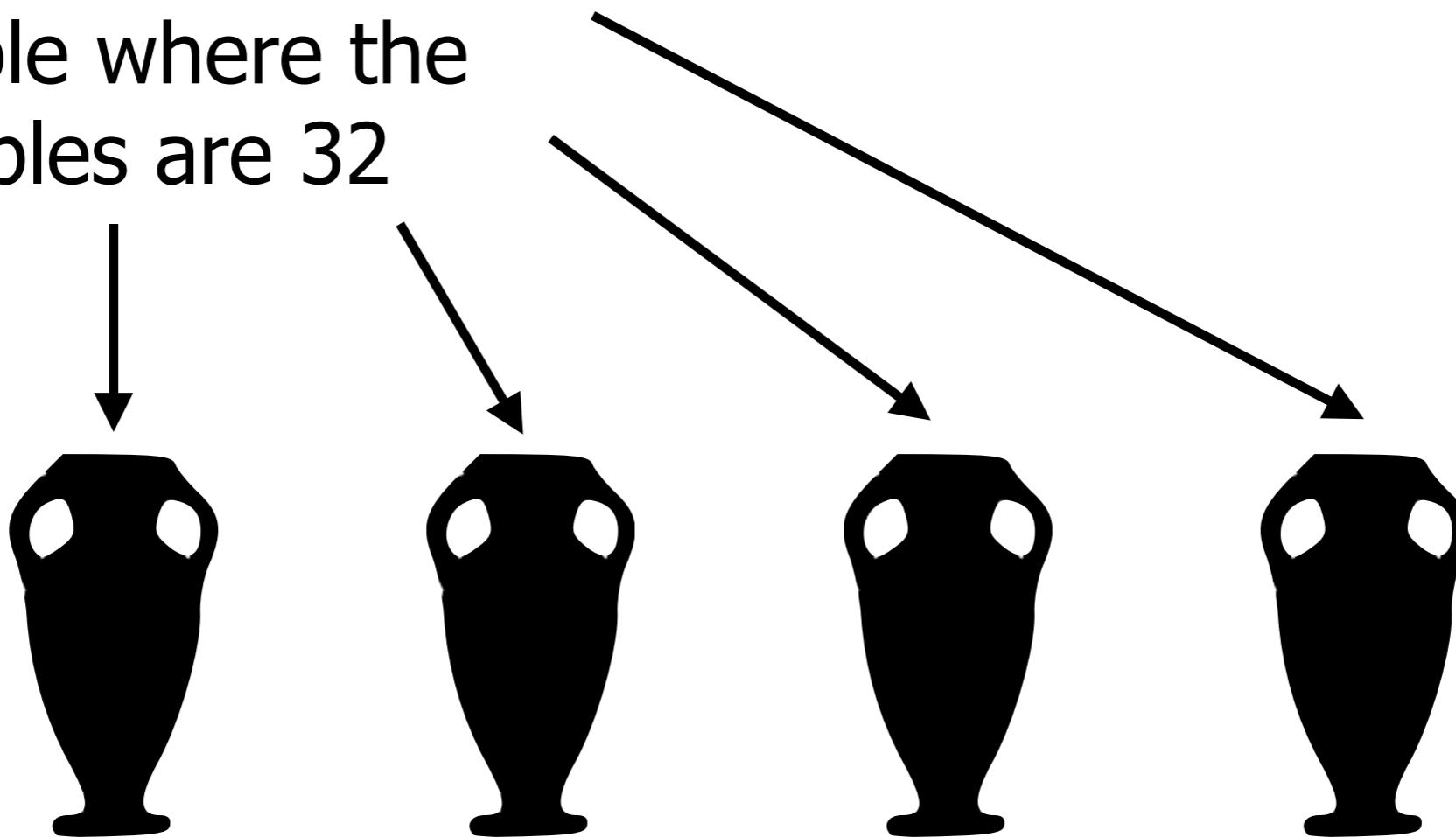
- Observed frequencies are those actually observed in the data (they do not stem from a decision of the experimenter);
- Expected frequencies are those that would be observed if the null hypothesis was true (they stem from the experimental design);
- It is necessary to find a suitable statistic and its sampling distribution.

Consider a process
distributing marbles
across several urns



The O values are the
numbers of marbles
observed in the urns

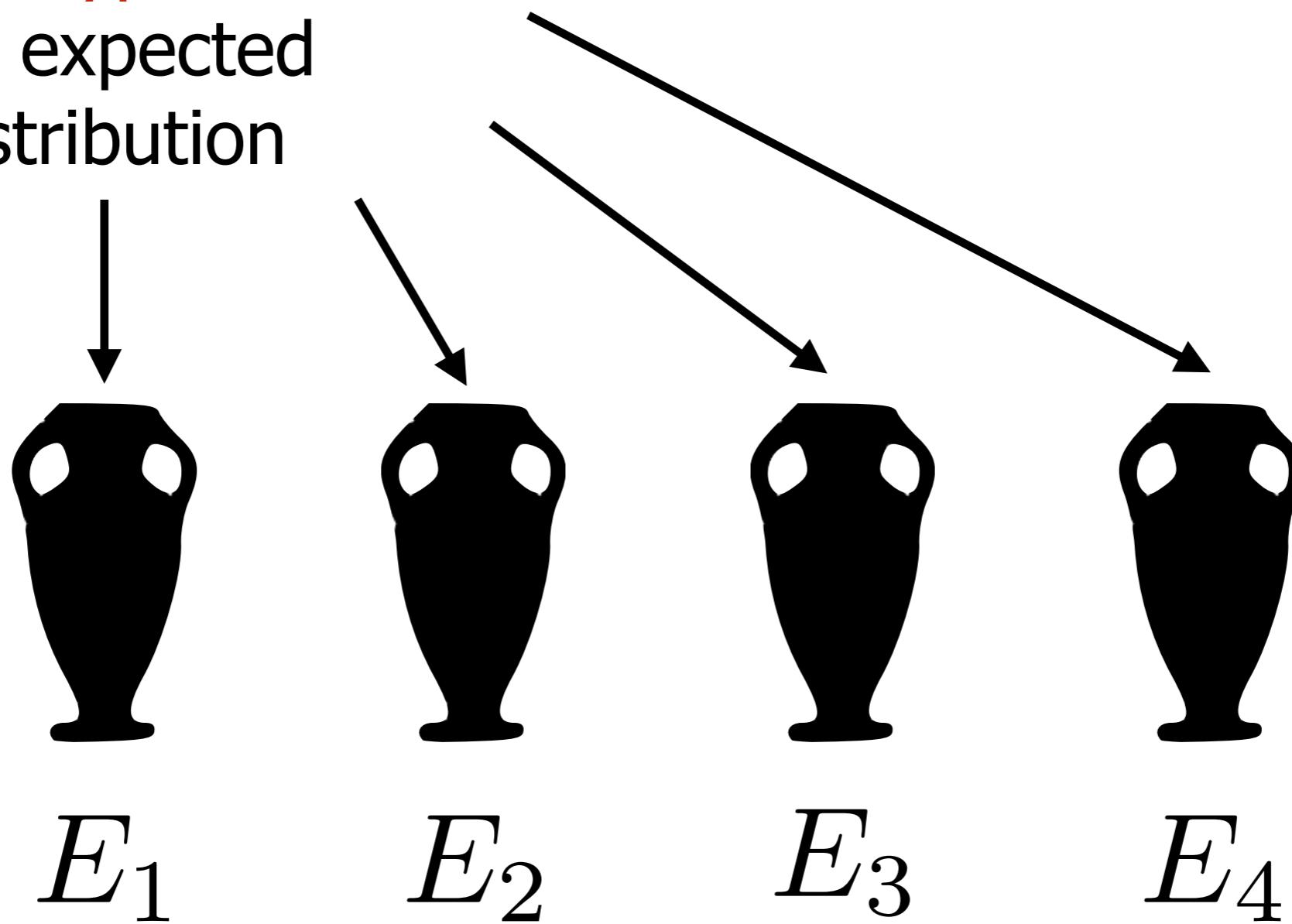
Consider a particular example where the marbles are 32



$$O_1 = 4 \quad O_2 = 5 \quad O_3 = 8 \quad O_4 = 15$$

The O values are the numbers of marbles **observed** in the urns

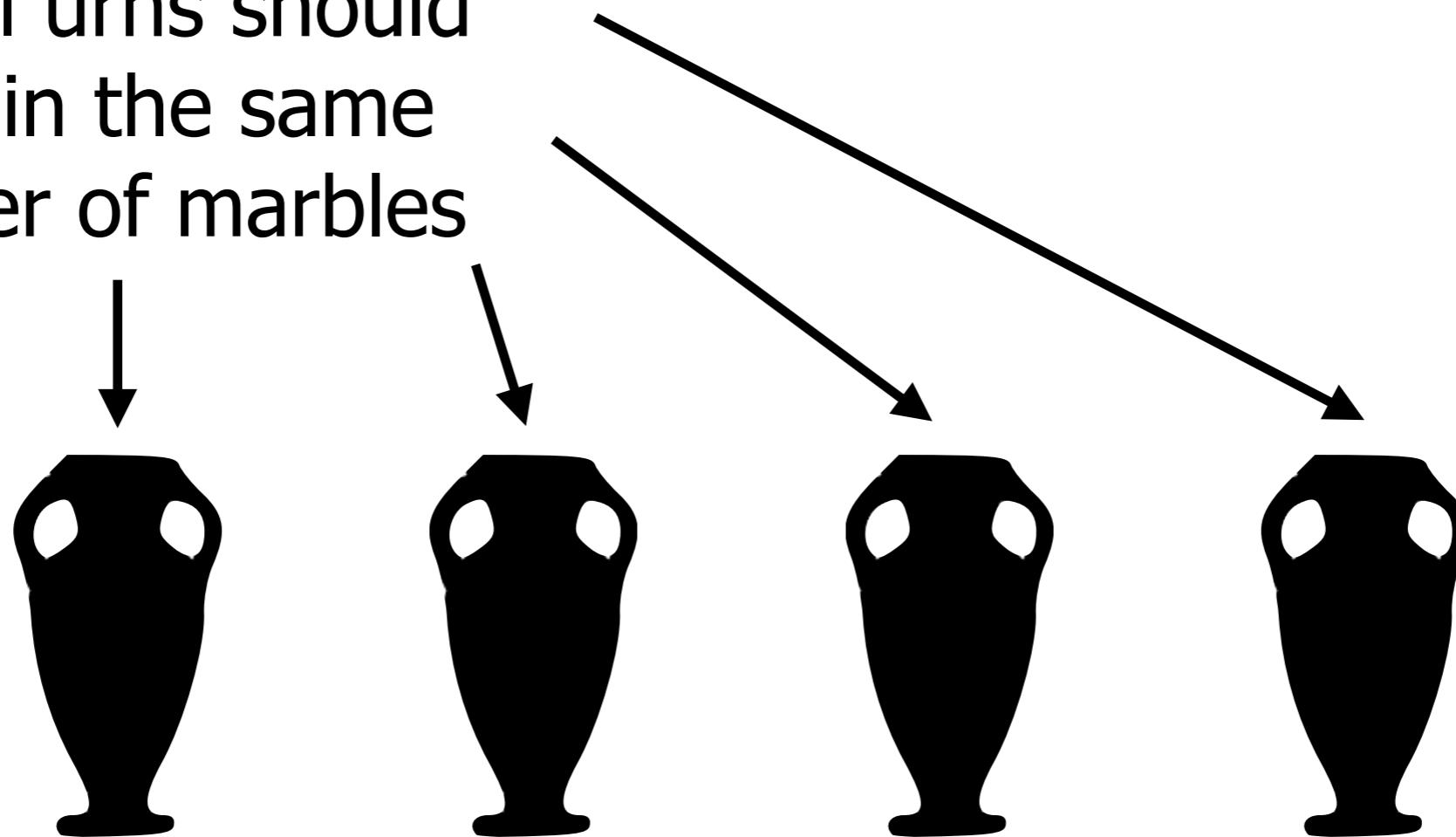
The null hypothesis is
the expected
distribution



The E values are the
numbers of marbles
expected in the urns

The question is
whether O and E
values are different

The null hypothesis is
that all urns should
contain the same
number of marbles



$$E_1 = E_2 = E_3 = E_4 = 8$$

The E values are all
the same to reflect
the null hypothesis

The E values are set
according to the null
hypothesis

The Chi Square

$$\chi^2 = \sum_{i=1}^N \frac{(O_i - E_i)^2}{E_i}$$

This variable tests whether there is a matching between the observations (O) and the expectations (E)

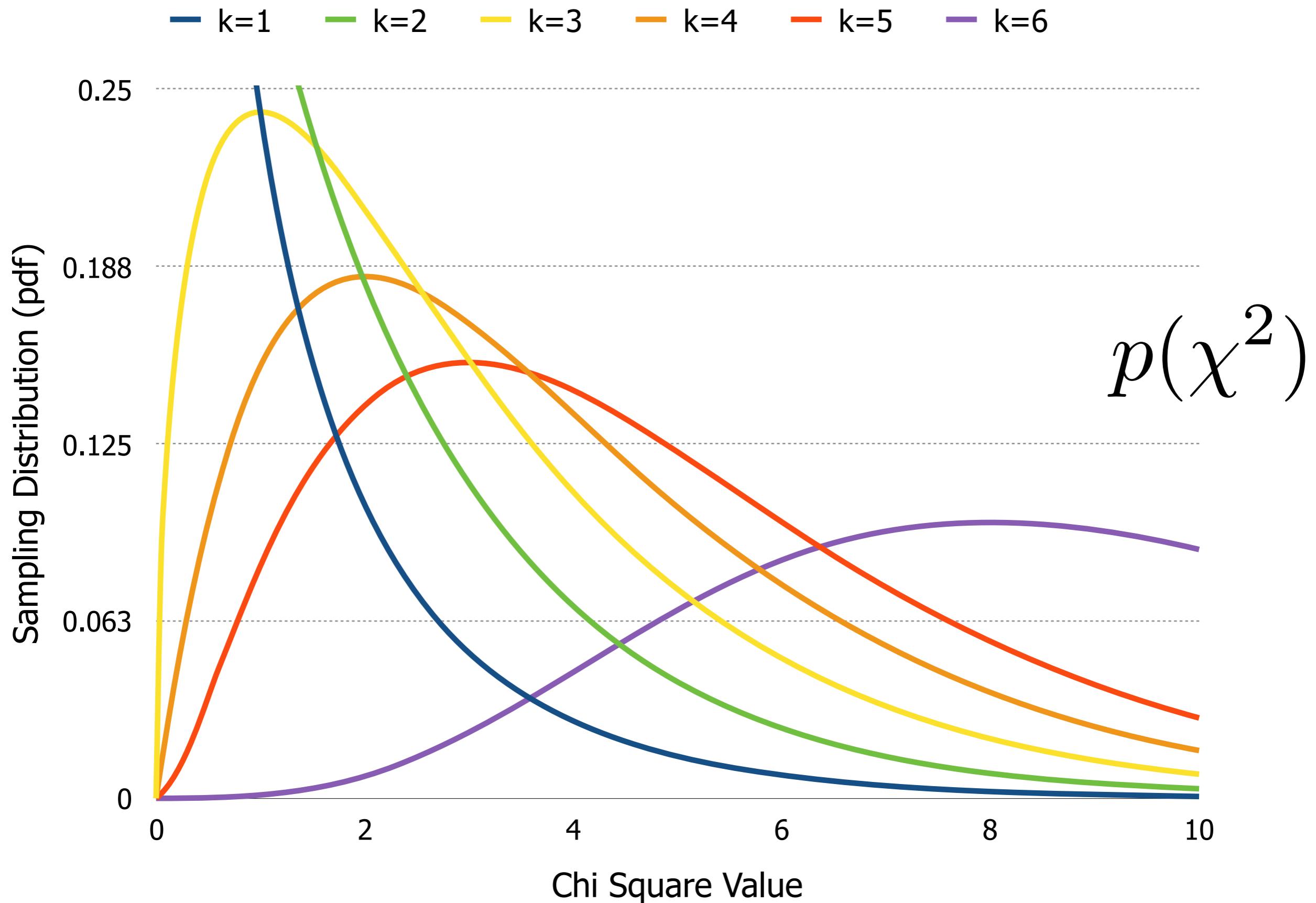
Sum over all values being compared

The probability density function is known when the null hypothesis is true

The parameter "k" corresponds to the degrees of freedom

$$p(\chi^2) = \frac{1}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)} (\chi^2)^{\frac{k}{2}-1} e^{-\frac{1}{2}\chi^2}$$

The value of "k" corresponds to the number of observations decreased by one



Degrees of Freedom

- The values of the Observations have to respect constraints;
- In the case of the One Way Classification, the constraint is the sum, no more than 32 mice (or marbles) can be observed;
- If there are N observations, the number of degrees of freedom is then N-1.

The O values are inserted in the expression of the Chi Square variable

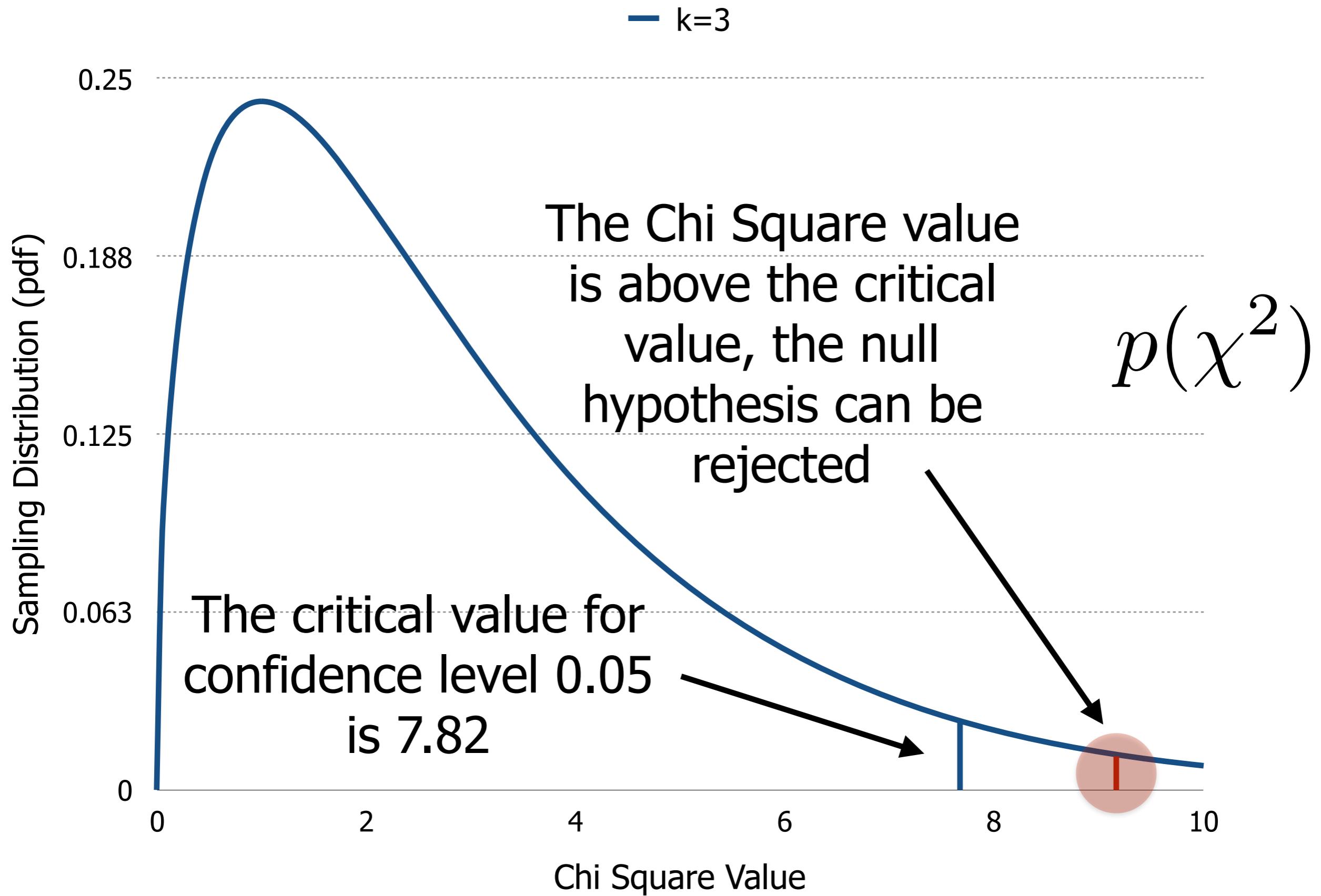
$$\chi^2 = \frac{(4 - 8)^2}{8} + \frac{(5 - 8)^2}{8} + \frac{(8 - 8)^2}{8} + \frac{(15 - 8)^2}{8}$$

$\chi^2 = 9.25$

The diagram illustrates the calculation of the Chi-Square statistic. It shows four terms in the formula, each representing the difference between an observed value (O) and an expected value (E), squared and divided by E . Red arrows point from the O values (4, 5, 8, 15) to their respective terms. Green arrows point from the E values (8) to the denominator of each term.

The E values are inserted in the expression of the Chi Square variable

The Chi Square is a random variable and its value depends on the O and E values, with the O being observed and the E reflecting the null hypothesis



Fake Example

- The outcome of the test depends on both Observations and Expectations;
- Imagine a (fake) experiment in which the expectations are different because the apparatus is different (e.g., there is food at the entrance of the alleys);
- The observations might remain the same, but the expectations change.

The O values are inserted in the expression of the Chi Square variable

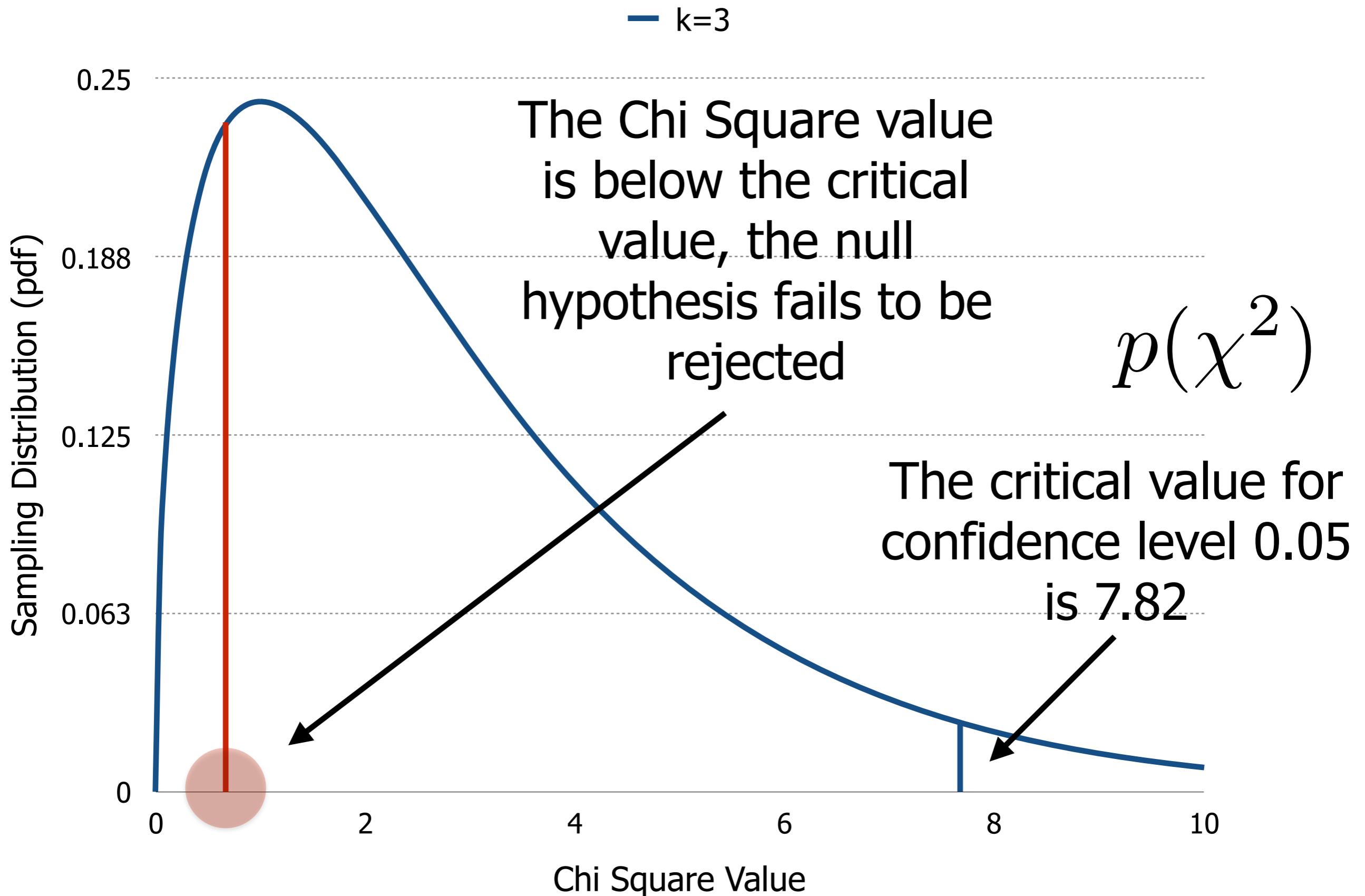
$$\chi^2 = \frac{(4 - 5)^2}{5} + \frac{(5 - 5)^2}{5} + \frac{(8 - 9)^2}{9} + \frac{(15 - 13)^2}{13}$$

$\chi^2 = 0.61$

The diagram illustrates the calculation of the Chi-Square statistic. It shows four terms in the formula, each representing the difference between an observed value (O) and an expected value (E), squared and divided by E . Red arrows point from the O values (4, 5, 8, 15) to their respective terms. Green arrows point from the E values (5, 5, 9, 13) to their respective terms.

The E values are inserted in the expression of the Chi Square variable

The Chi Square is a random variable and its value depends on the O and E values, with the O being observed and the E reflecting the null hypothesis



Recap

- The value of the Chi Square depends on the data (through the O values) and on the null hypothesis (through the E values);
- The O values cannot be changed because they correspond to the data observed in an experiments;
- The E values must be set according to the null hypothesis to be tested.

Outline

- Very Quick Recap
- Goodness of Fit Test (One Way Classification)
- Two Way Classification
- Conclusions

Research Hypothesis

- Research Hpothesis: An aggressor tends to be considered guilty more frequently when the victim appears to be less faulty;
- Null Hypothesis: An aggressor tends to be considered guilty irrespectively of how faulty the victim appears to be.

Contingency Tables (I)

Fault	Guilty	Not Guilty	Total
Low	153	24	177
High	105	76	181
Total	258	100	358

C_1 C_2 Marginals

R_1

R_2

Pugh, "Contributory fault and rape convictions: Loglinear models for blaming the victim", Social Psychology Quarterly, 46(3):233-242, 1983

Contingency Tables (II)

- The elements of the table are Observations (how many times two characteristics coexist in a sample);
- The table shows whether one characteristic (a variable) is contingent or associated to the other;
- The sums over the values of a row or a column are called marginals.

Expected value in
cell “ij” when the null
hypothesis is true

Total number of
“marbles” in row “i”

“N” is the total
number of marbles in
the table

Fraction of “marbles”
in column “j”

$$E_{ij} = R_i \frac{C_j}{N}$$

Expected value in
cell “ij” when the null
hypothesis is true

Total number of
“marbles” in column
“j”

$$E_{ij} = \frac{R_i}{N} C_j$$

“N” is the total
number of marbles in
the table

Fraction of “marbles”
in row “i”

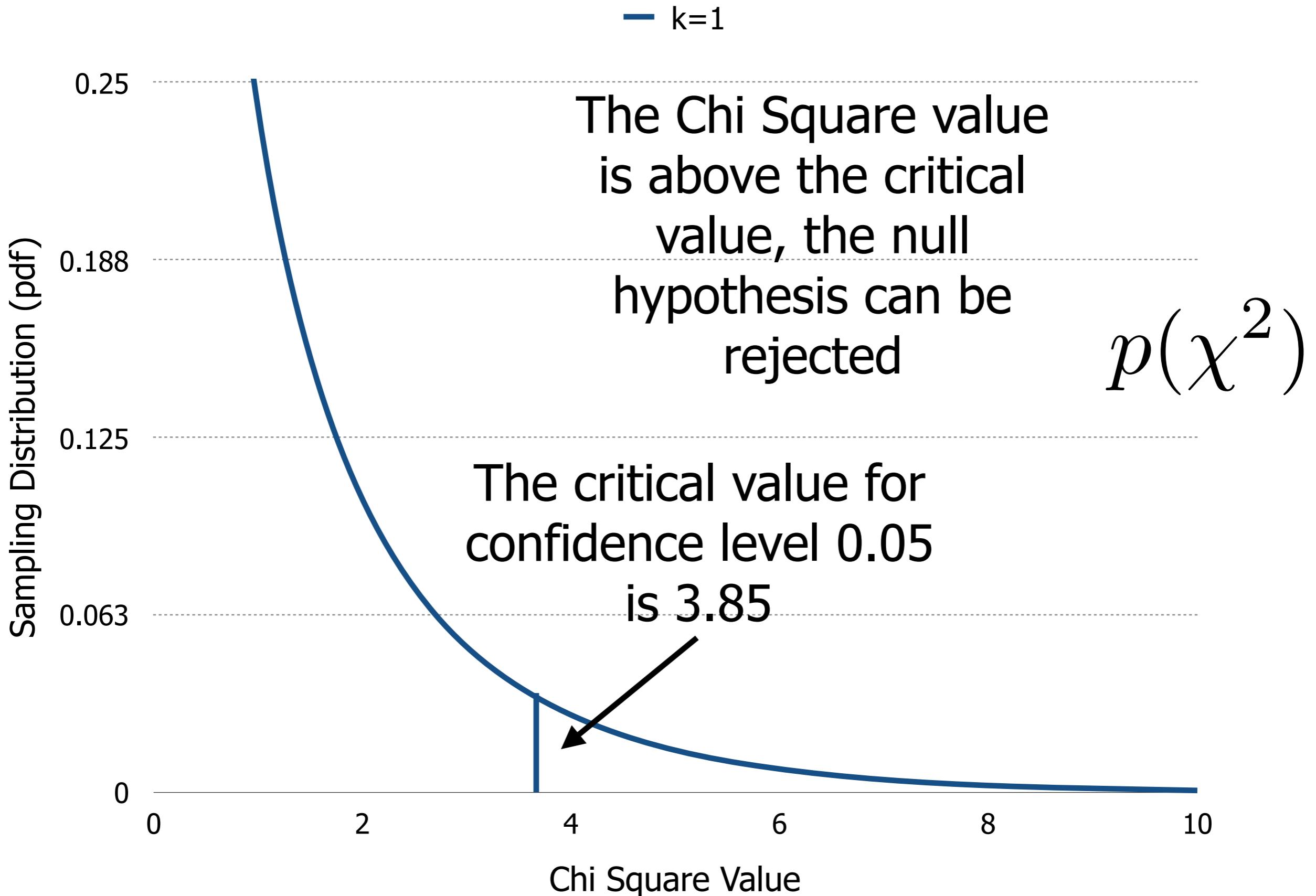
Number of
Rows

Number of
Columns

$$k = (R - 1)(C - 1) = 1$$

$$\chi^2 = \sum_{i=1}^R \sum_{j=1}^C \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 35.9$$

Sum over all elements
of the contingency
table



Outline

- Very Quick Recap
- Goodness of Fit Test (One Way Classification)
- Two Way Classification
- Conclusions

Conclusions

- The Chi Square test is useful when the observations take the form of counts (how many times an event of interest occurs);
- One way classification can show how well the observations fit the expectations;
- Two way classification can show how much two variables of interest are associated.

Thank You!

Observing Behaviour

Computational Social Intelligence - Lecture 05

Prof. Alessandro Vinciarelli
School of Computing Science &
Institute of Neuroscience and Psychology

<http://www.dcs.gla.ac.uk/vincia>
Alessandro.Vinciarelli@glasgow.ac.uk



University
of Glasgow

EPSRC
Engineering and Physical Sciences
Research Council

FNSNF

This lecture is based on the following text
(available on Moodle):

- Chapter 3 of “Measuring Behaviour” P.Martin & P.Bateson, Cambridge University Press (2007);
- Vinciarelli, Chatziliaoannou & Esposito, “When the Words are not Everything: The Use of Laughter, Fillers, Back-Channel, Silence and Overlapping Speech in Phone Calls”, Frontiers in ICT, 2:4, 2015.

Outline

- The 11 Steps of Behaviour Analysis
- An Example: Mobile Phone Conversations
- Conclusions

Outline

- The 11 Steps of Behaviour Analysis
- An Example: Mobile Phone Conversations
- Conclusions

1.Ask a question.

“Before any scientific problem is investigated, some sort of question will have been formulated [...] may initially be a broad one [...] Such a question is not a hypothesis.”

Martin & Bateson, “Measuring Behaviour”,
Cambridge University Press, 2007 (Chapter 3)

2. Make Preliminary Observations.

“A period of preliminary observation is generally invaluable in deciding what measurements to make and should be regarded as crucial part of any study.”

Martin & Bateson, “Measuring Behaviour”,
Cambridge University Press, 2007 (Chapter 3)

3. Identify the Behavioural Variables that Need to be Measured.

“The form of the research and the variables that are to be measured should then be chosen as to provide the best account of what you have observed.”

Martin & Bateson, “Measuring Behaviour”,
Cambridge University Press, 2007 (Chapter 3)

4. Choose Suitable Recording Methods.

“No observer can record behaviour without selecting some features [...] and ignoring others. The selection [...] reflects the questions you asked at the beginning of the study.”

Martin & Bateson, “Measuring Behaviour”,
Cambridge University Press, 2007 (Chapter 3)

5. Collect and Analyse the Data.

“[...] plan in advance how much data you will need to collect in order to obtain a clear conclusion [...] Use the appropriate statistical tools to for analysing the data.”

Martin & Bateson, “Measuring Behaviour”,
Cambridge University Press, 2007 (Chapter 3)

6. Formulate Precise Hypotheses.

"A clear hypothesis invites a direct test
[...] hypotheses may be tested by
observing natural variation in a
population as well as by performing
experiments."

Martin & Bateson, "Measuring Behaviour",
Cambridge University Press, 2007 (Chapter 3)

7. Make Predictions from the Hypotheses.

“A clear hypothesis should, by a process of straightforward reasoning, give rise to one or more specific predictions that can be tested empirically.”

Martin & Bateson, “Measuring Behaviour”,
Cambridge University Press, 2007 (Chapter 3)

8.Design the Tests.

“The variables that are to be measured should then be chosen so as to provide the best test of the different predictions made by competing hypotheses.”

Martin & Bateson, “Measuring Behaviour”,
Cambridge University Press, 2007 (Chapter 3)

9.Run Tests of your Hypotheses.

“Use the same measurement procedures throughout and try, if possible, to collect data ‘blind’ so that you **do not** unconsciously select data that fit your hypothesis.”

Martin & Bateson, “Measuring Behaviour”,
Cambridge University Press, 2007 (Chapter 3)

10. Analyse the Results of your Tests.

“Employ the appropriate statistical tools, both for presenting and exploring the data, and for testing the hypotheses.”

Martin & Bateson, “Measuring Behaviour”,
Cambridge University Press, 2007 (Chapter 3)

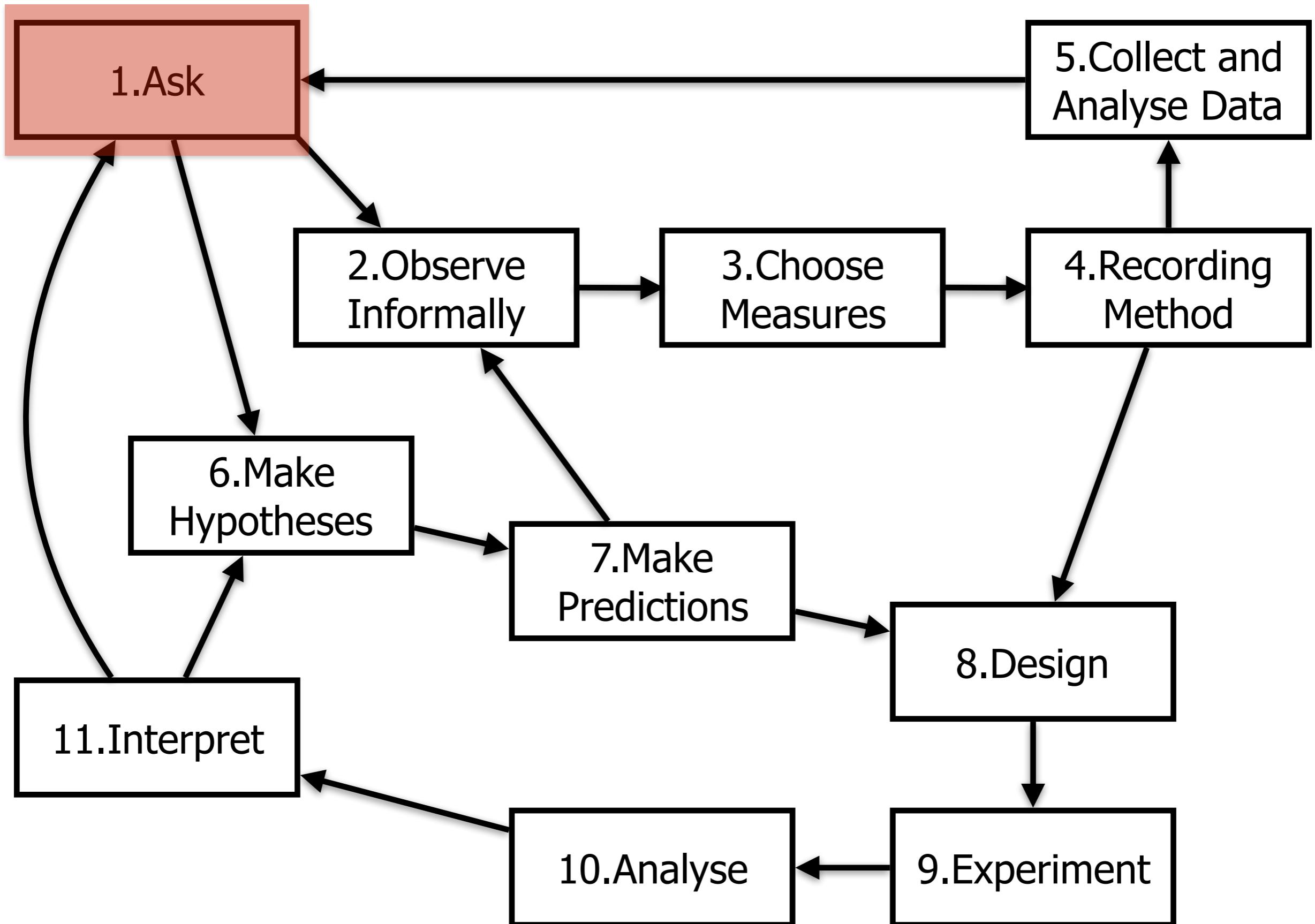
11. Consider Alternative Interpretations of the Evidence.

“Do not draw more conclusions than the data support, but do try to formulate a list of questions and ideas suggested by the data that could form the basis of future research.”

Martin & Bateson, “Measuring Behaviour”,
Cambridge University Press, 2007 (Chapter 3)

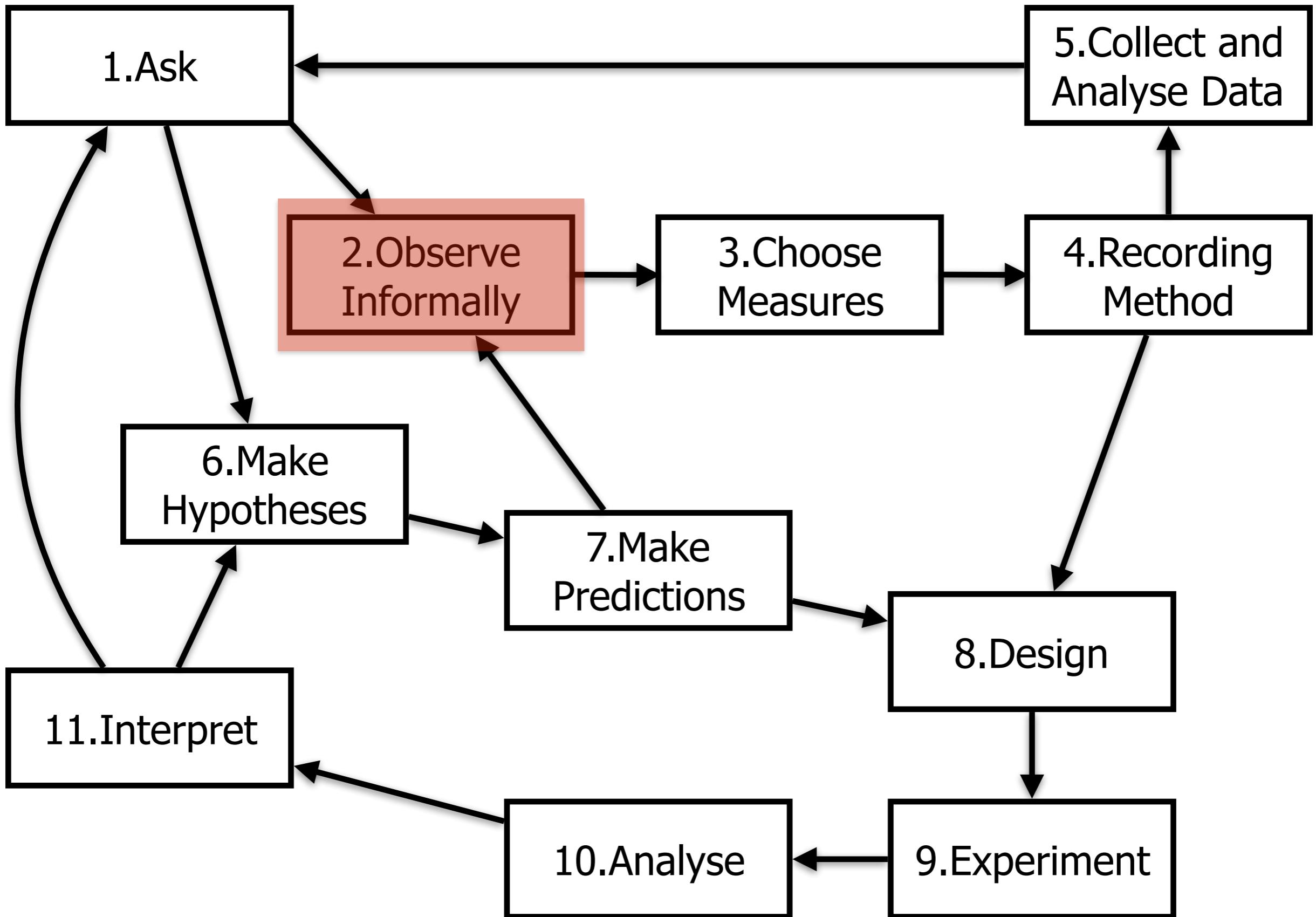
Outline

- The 11 Steps of Behaviour Analysis
- An Example: Mobile Phone Conversations
- Conclusions



1.Ask a question.

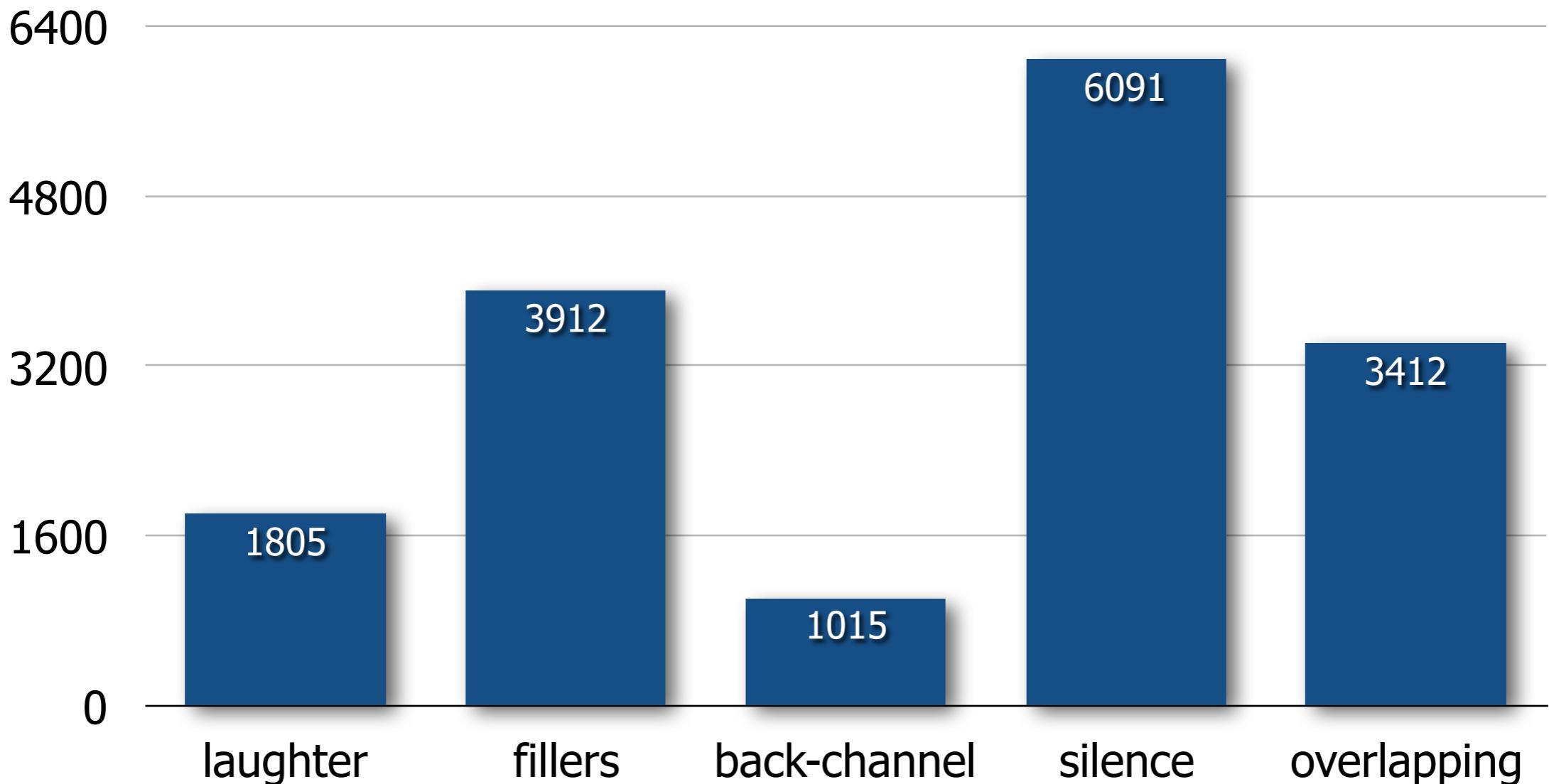
Is there a relationship between the use of
nonverbal communication and major
social dimensions (gender and role)?



2. Make Preliminary Observations

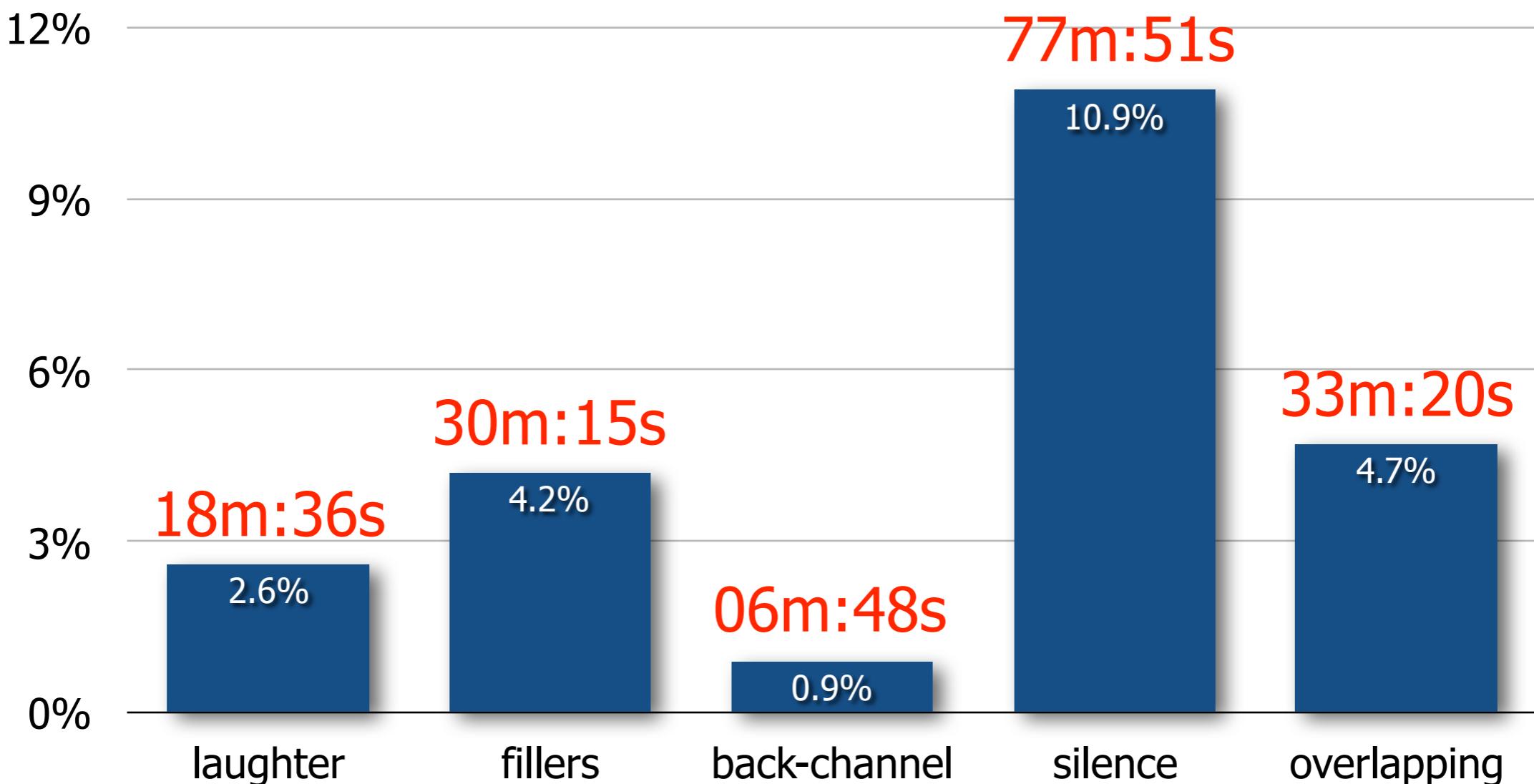
Number of Calls	60
Number of Subjects	120
Total Length	11h : 48m : 24s
Average Length	11m : 48s
Audio Sampling Frequency	44kHz
Gyrosopes Sampling	68Hz
Psychometric Questionnaires	2
Total Annotated Cues	16,235

2. Make Preliminary Observations.

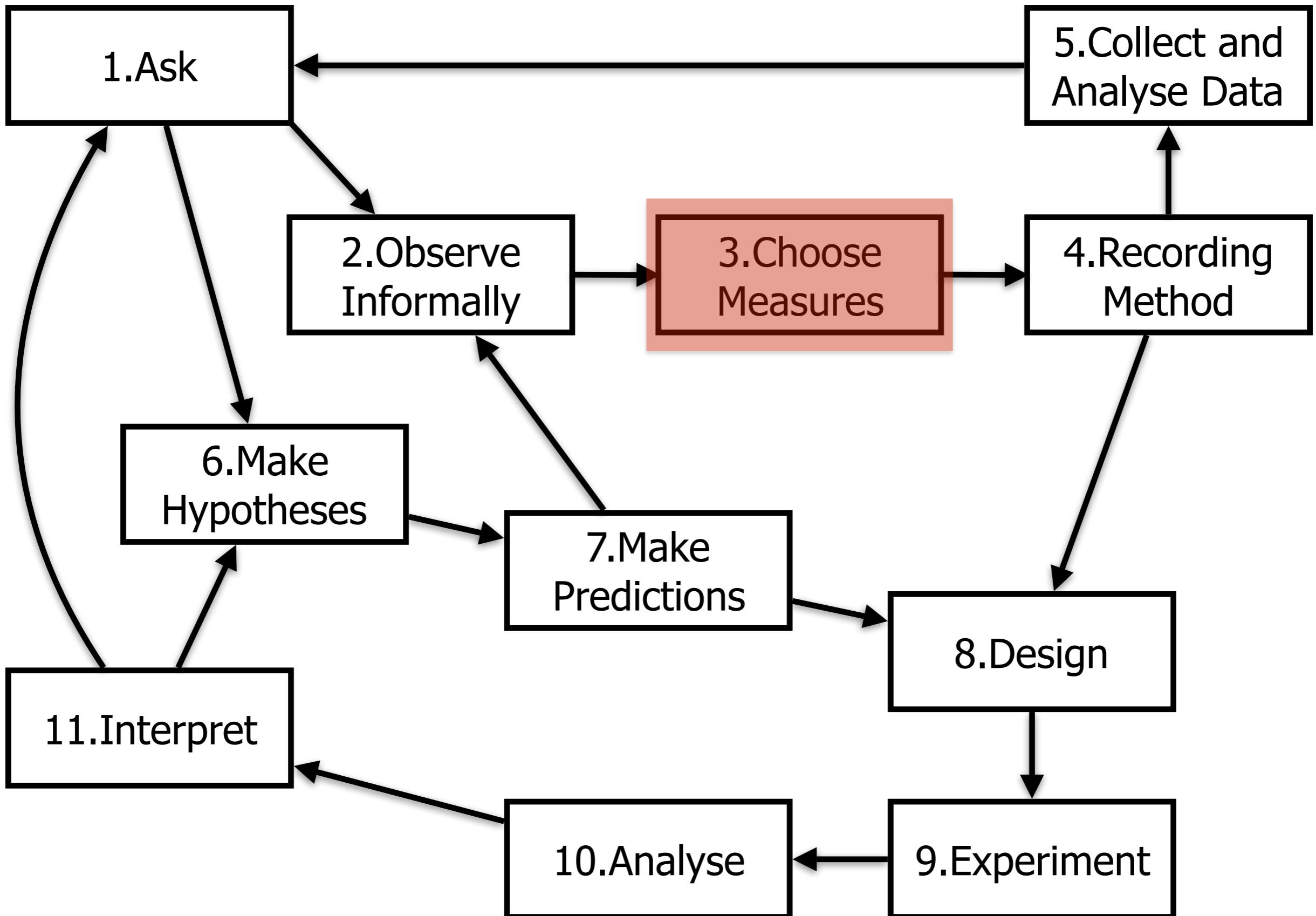


Vinciarelli, Chatzioannou & Esposito, "When Words are Not Everything: the use of laughter, fillers, back-channel, silence, and overlapping speech in phone calls", Frontiers in ICT, 2(4), 2015.

2. Make Preliminary Observations.

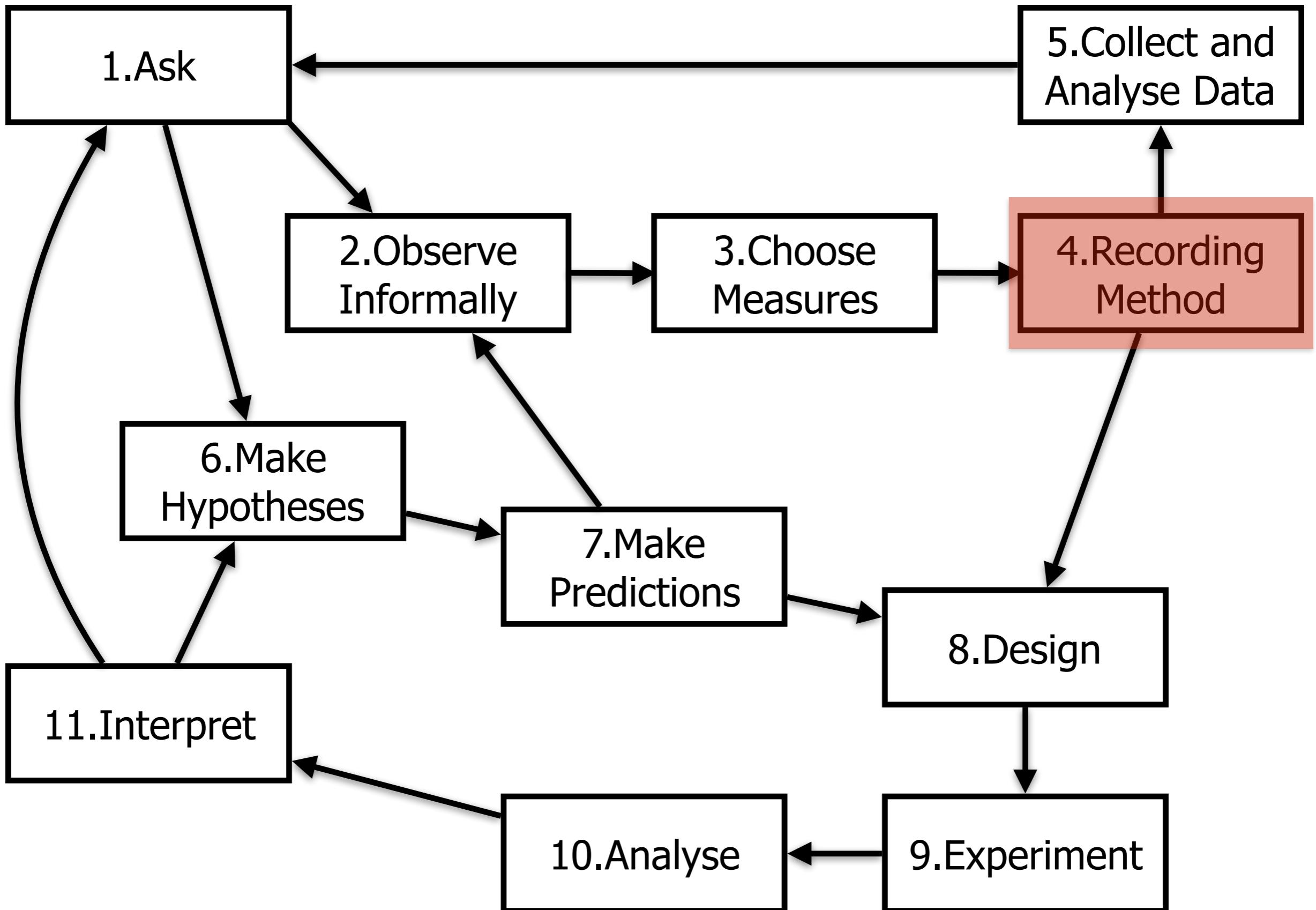


Vinciarelli, Chatzioannou & Esposito, "When Words are Not Everything: the use of laughter, fillers, back-channel, silence, and overlapping speech in phone calls", Frontiers in ICT, 2(4), 2015.



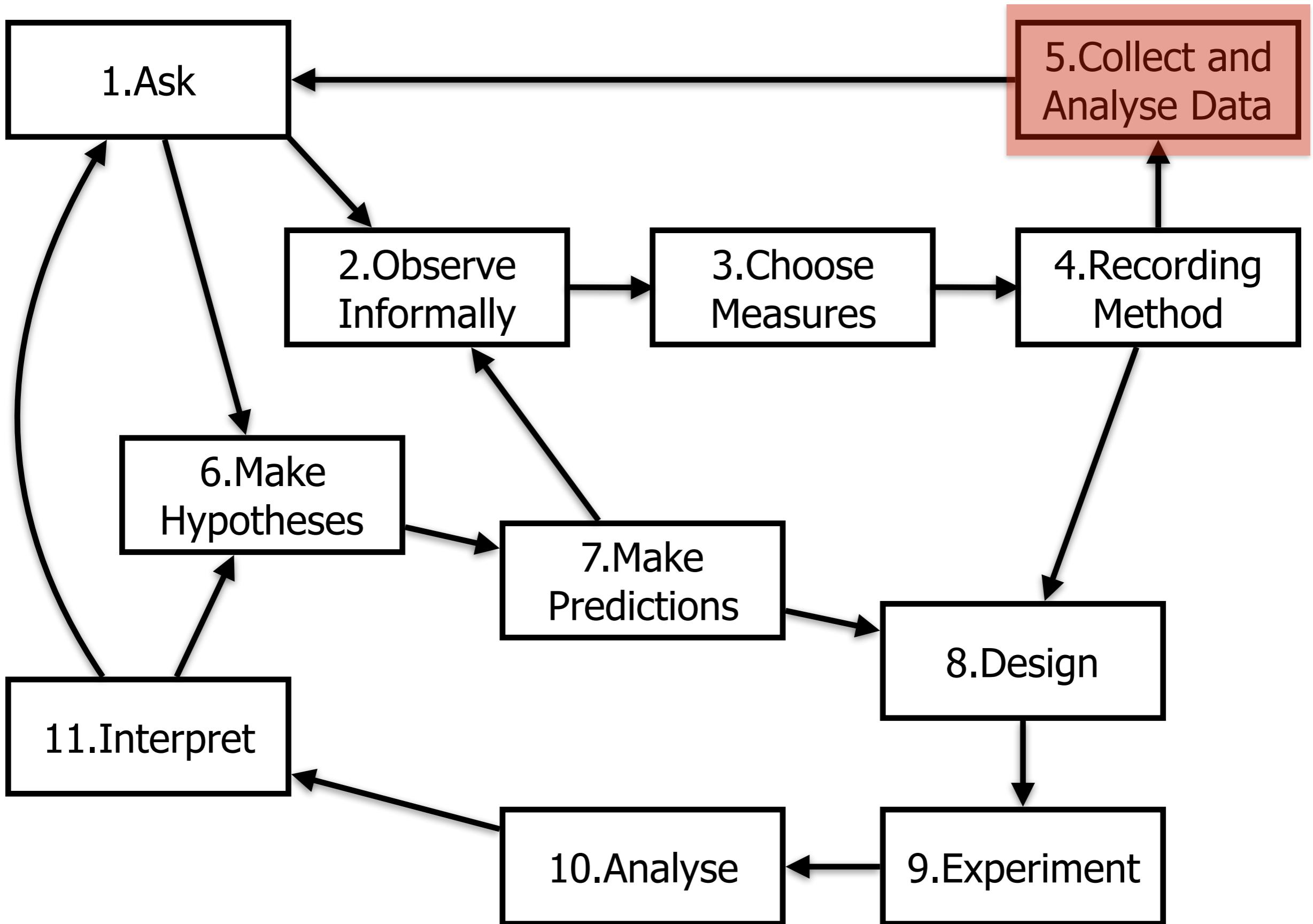
3. Identify the Behavioural Variables that Need to be Measured.

Number of occurrences for different types of subjects (female and male or callers and receivers) of silences, laughter, interruptions, fillers and back-channel

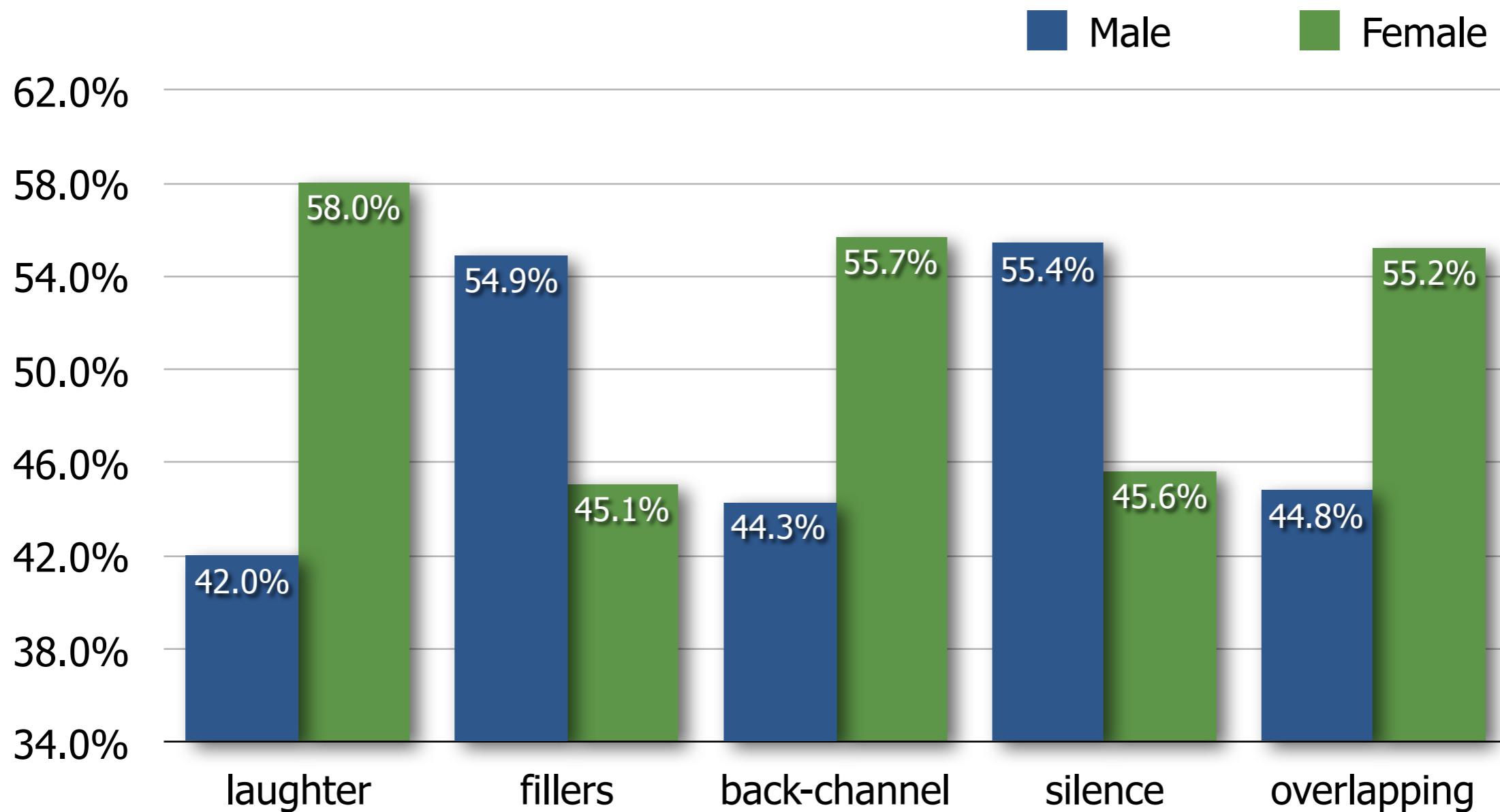


4. Choose Suitable Recording Methods.



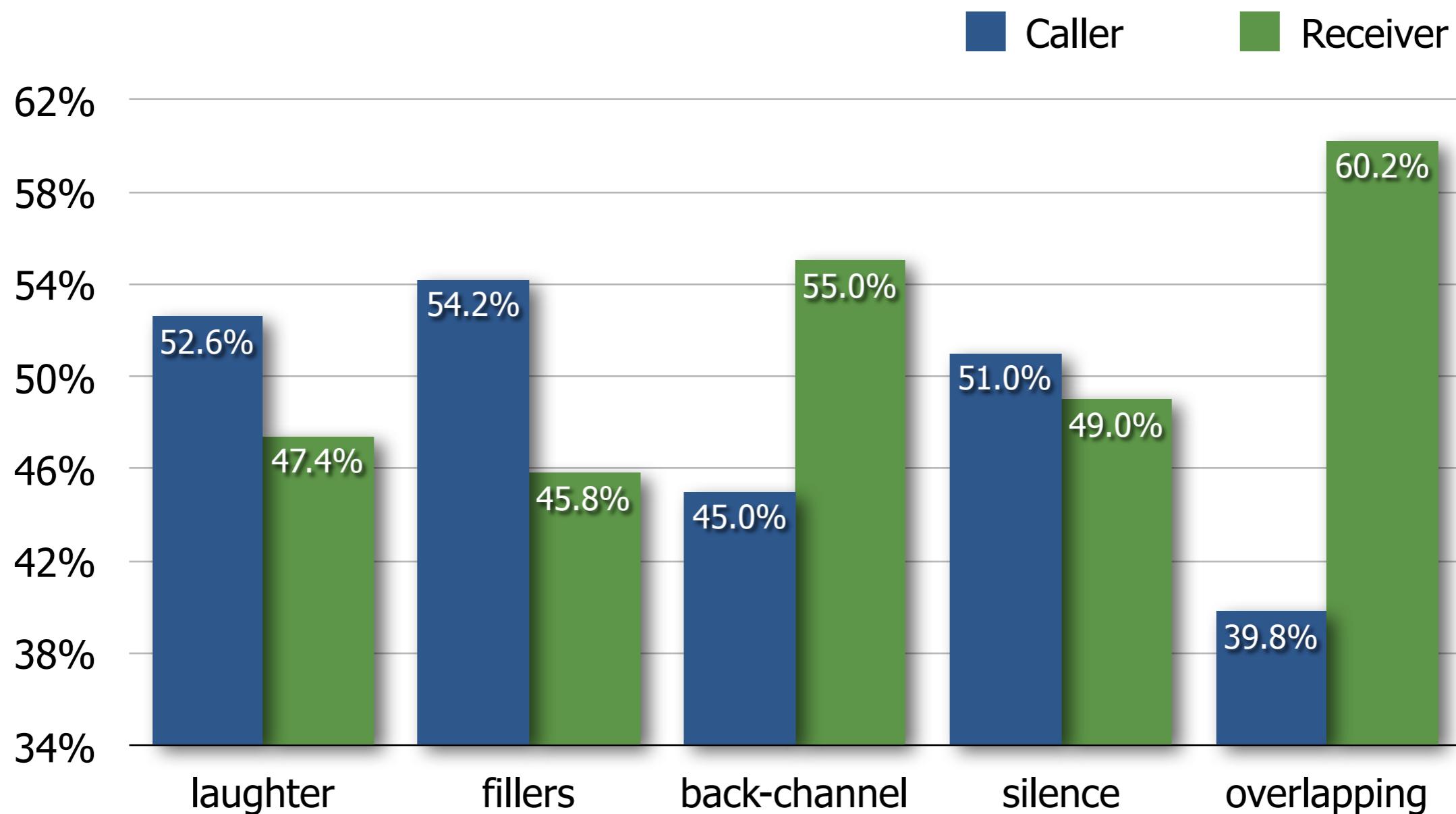


5. Collect and Analyse the Data.

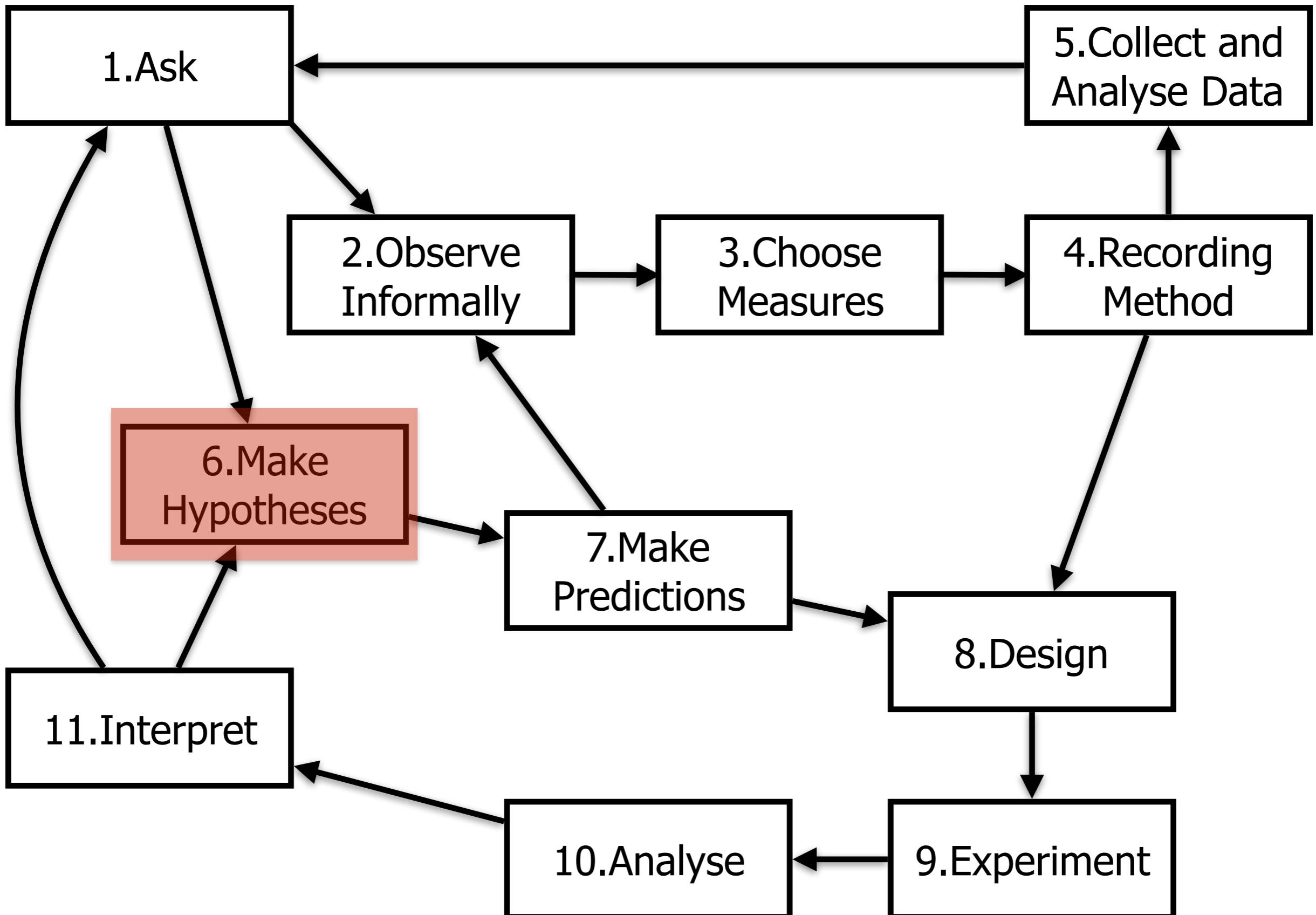


Vinciarelli, Chatzioannou & Esposito, "When Words are Not Everything: the use of laughter, fillers, back-channel, silence, and overlapping speech in phone calls", Frontiers in ICT, 2(4), 2015.

5. Collect and Analyse the Data.

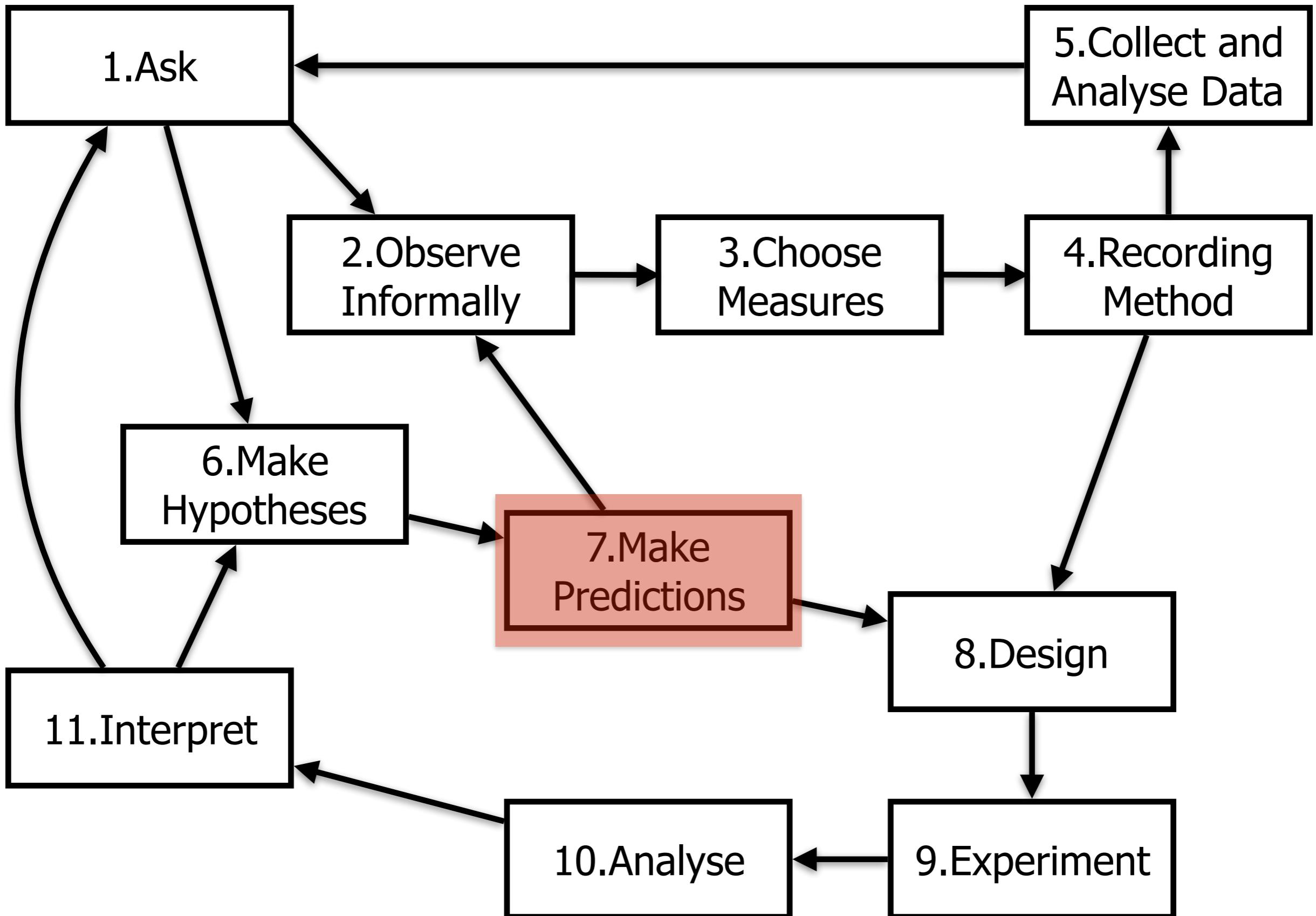


Vinciarelli, Chatzioannou & Esposito, "When Words are Not Everything: the use of laughter, fillers, back-channel, silence, and overlapping speech in phone calls", Frontiers in ICT, 2(4), 2015.



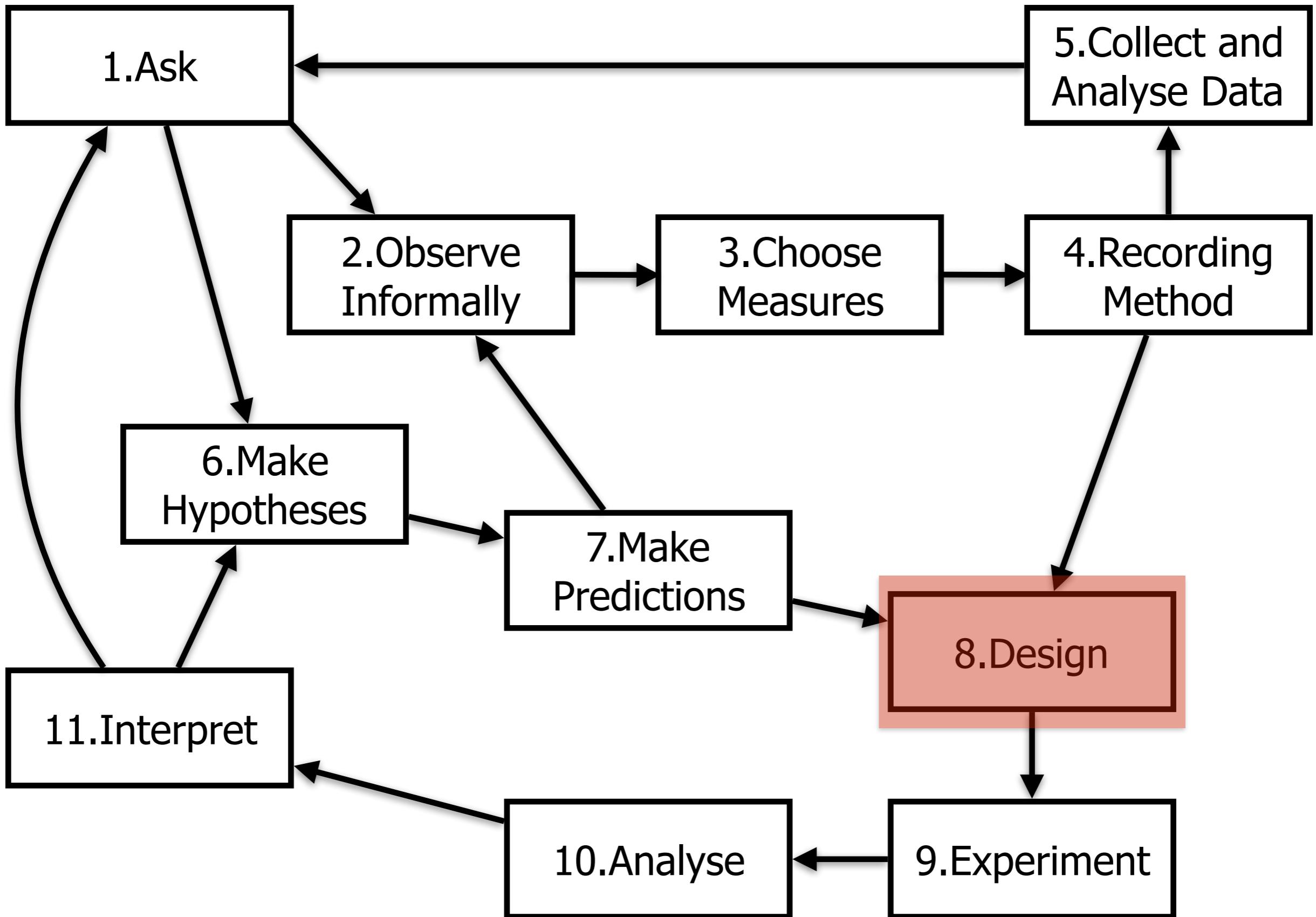
6. Formulate Precise Hypotheses.

There is a statistically significant relationship between social dimensions (gender and role) and number of occurrences of frequent nonverbal cues



7. Make Predictions from the Hypotheses.

The observed differences, at least for some cues, are statistically significant

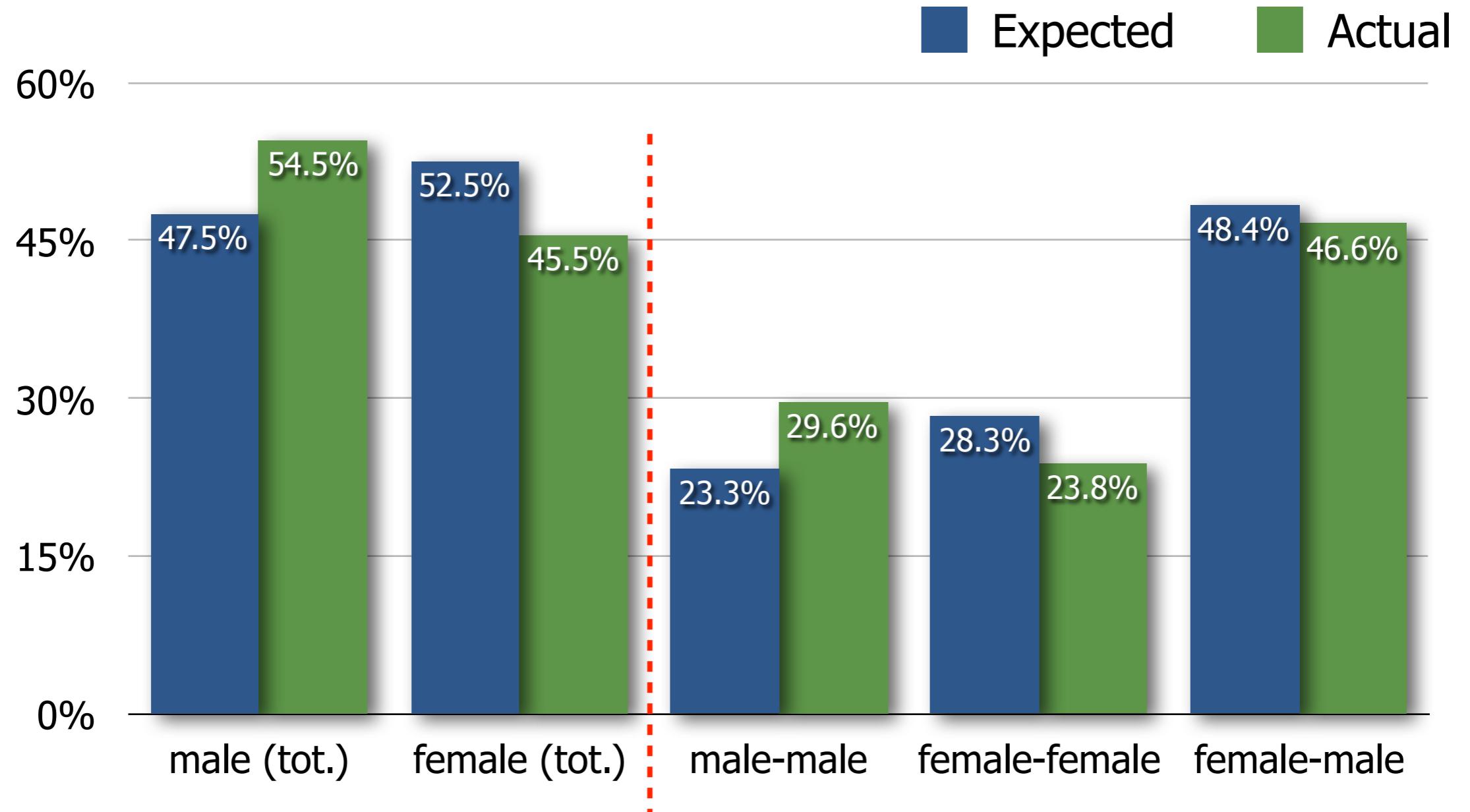


8.Design the Tests.

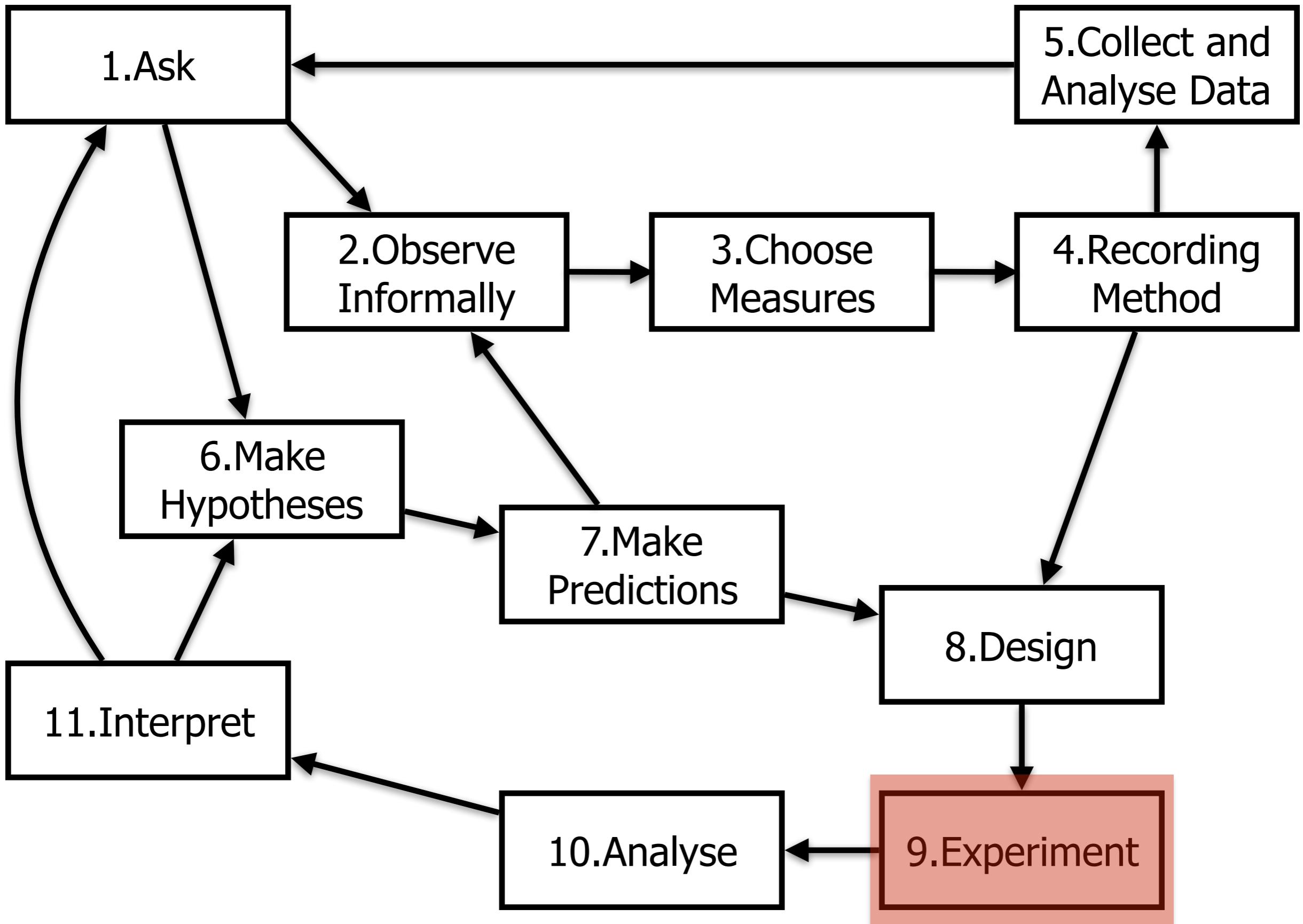
$$\chi^2 = \sum_{k=1}^C \frac{(O_k - E_k)^2}{E_k}$$

This variable tests whether there is a matching between the observations (O) and the expectations (E)

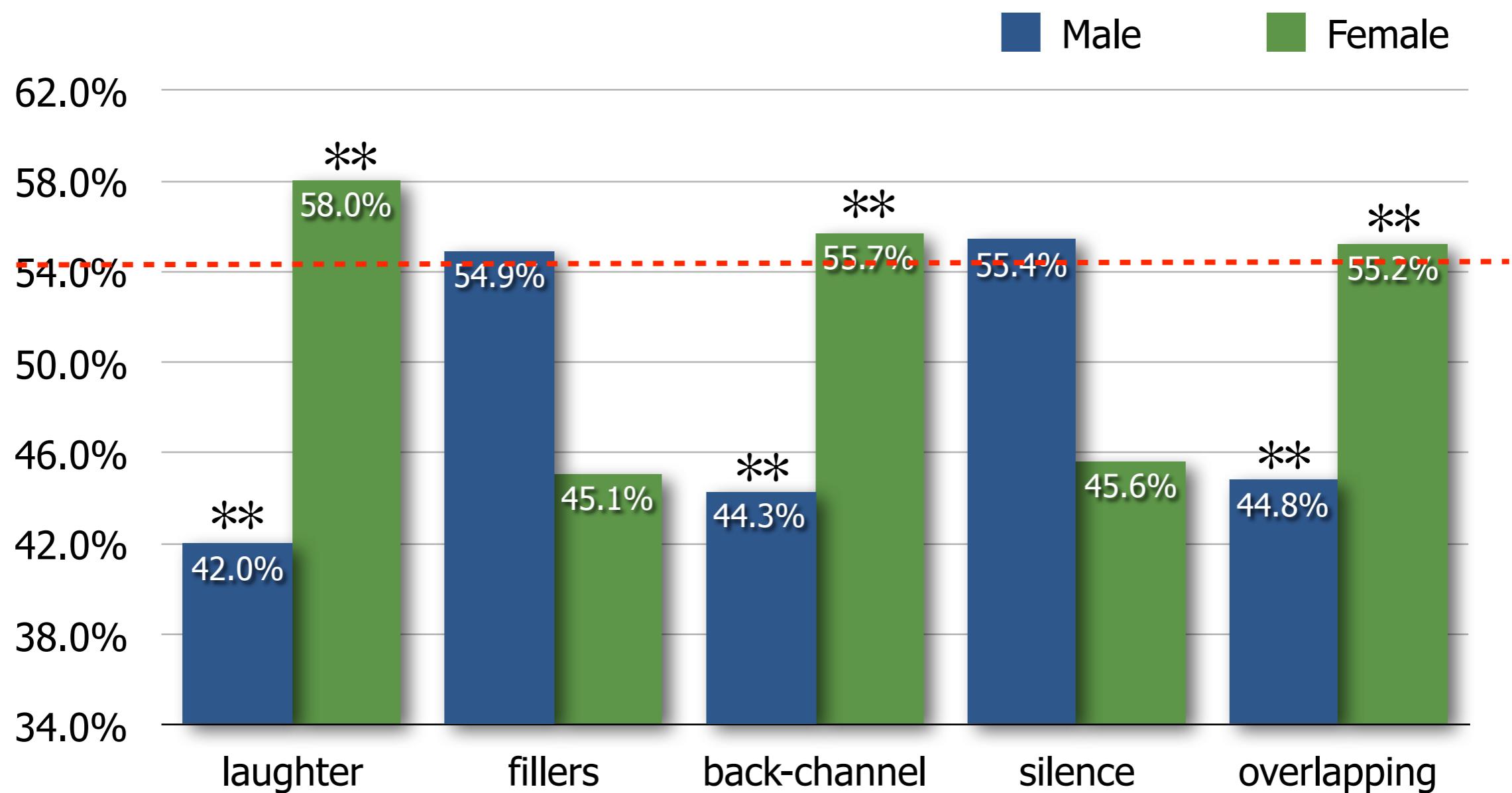
Sum over all categories being compared (female vs male or callers vs receivers)



Vinciarelli, Chatzioannou & Esposito, "When Words are Not Everything: the use of laughter, fillers, back-channel, silence, and overlapping speech in phone calls", Frontiers in ICT, 2(4), 2015.

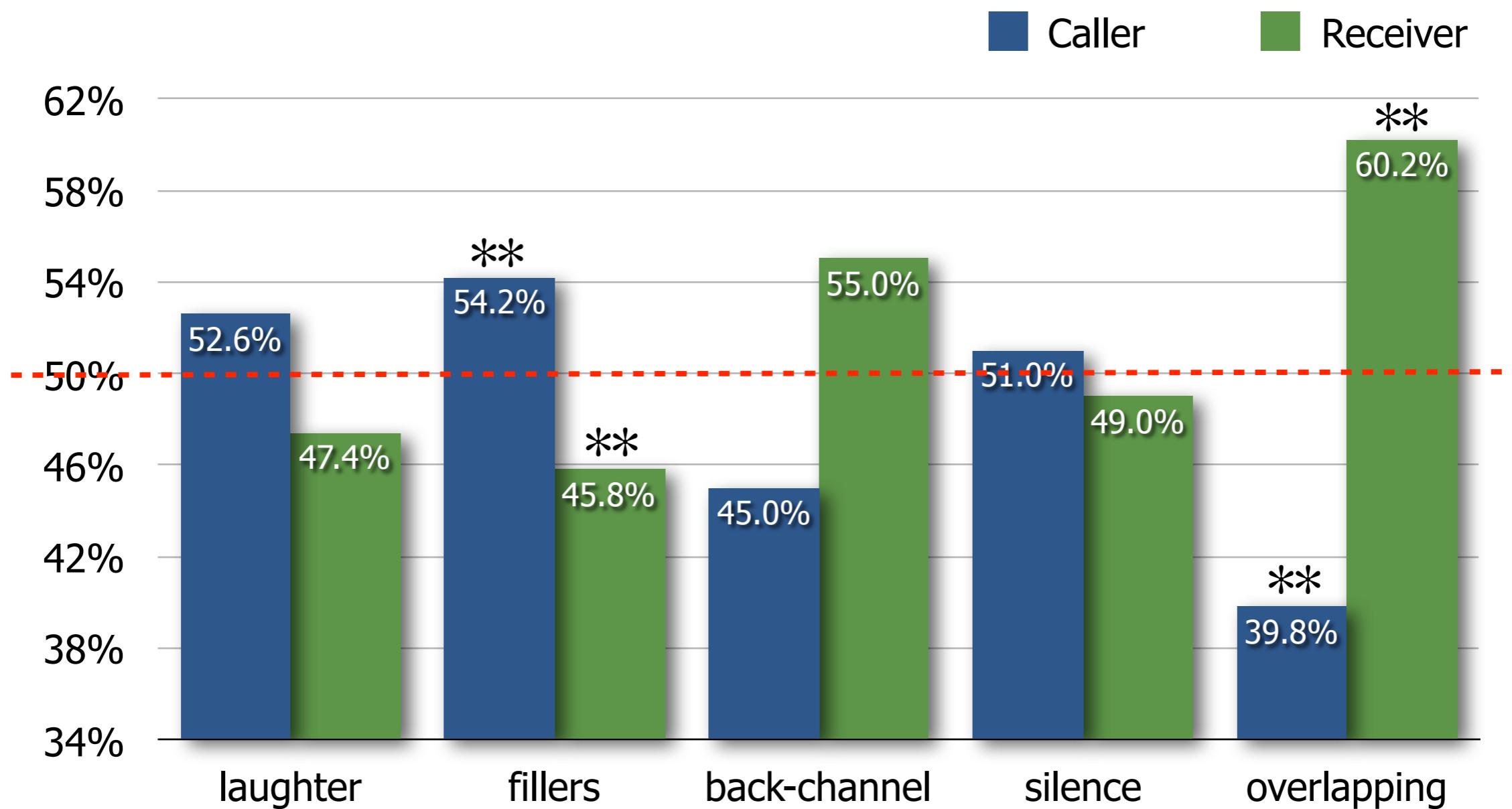


9.Run Tests of your Hypotheses.

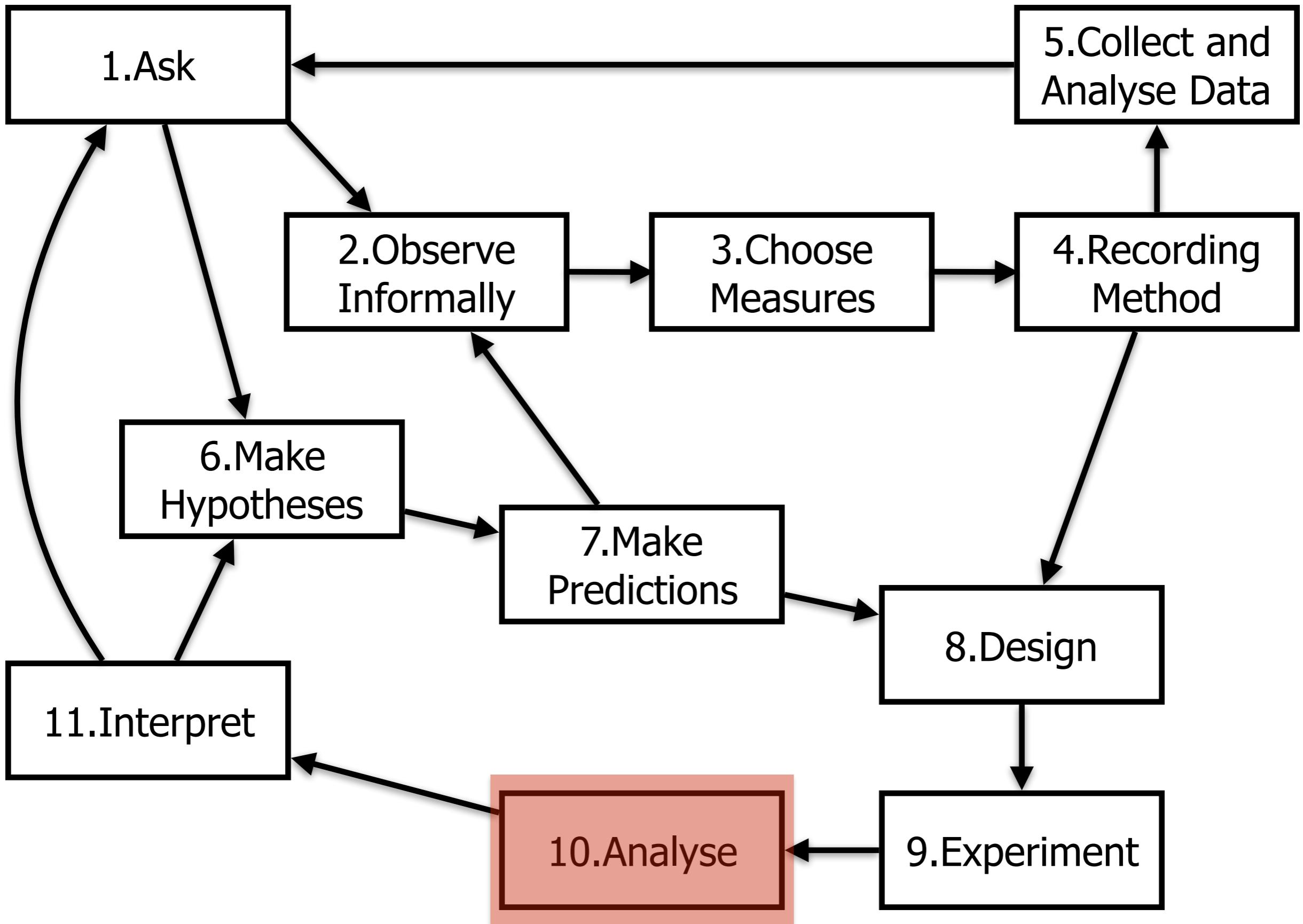


Vinciarelli, Chatzioannou & Esposito, "When Words are Not Everything: the use of laughter, fillers, back-channel, silence, and overlapping speech in phone calls", Frontiers in ICT, 2(4), 2015.

9.Run Tests of your Hypotheses.

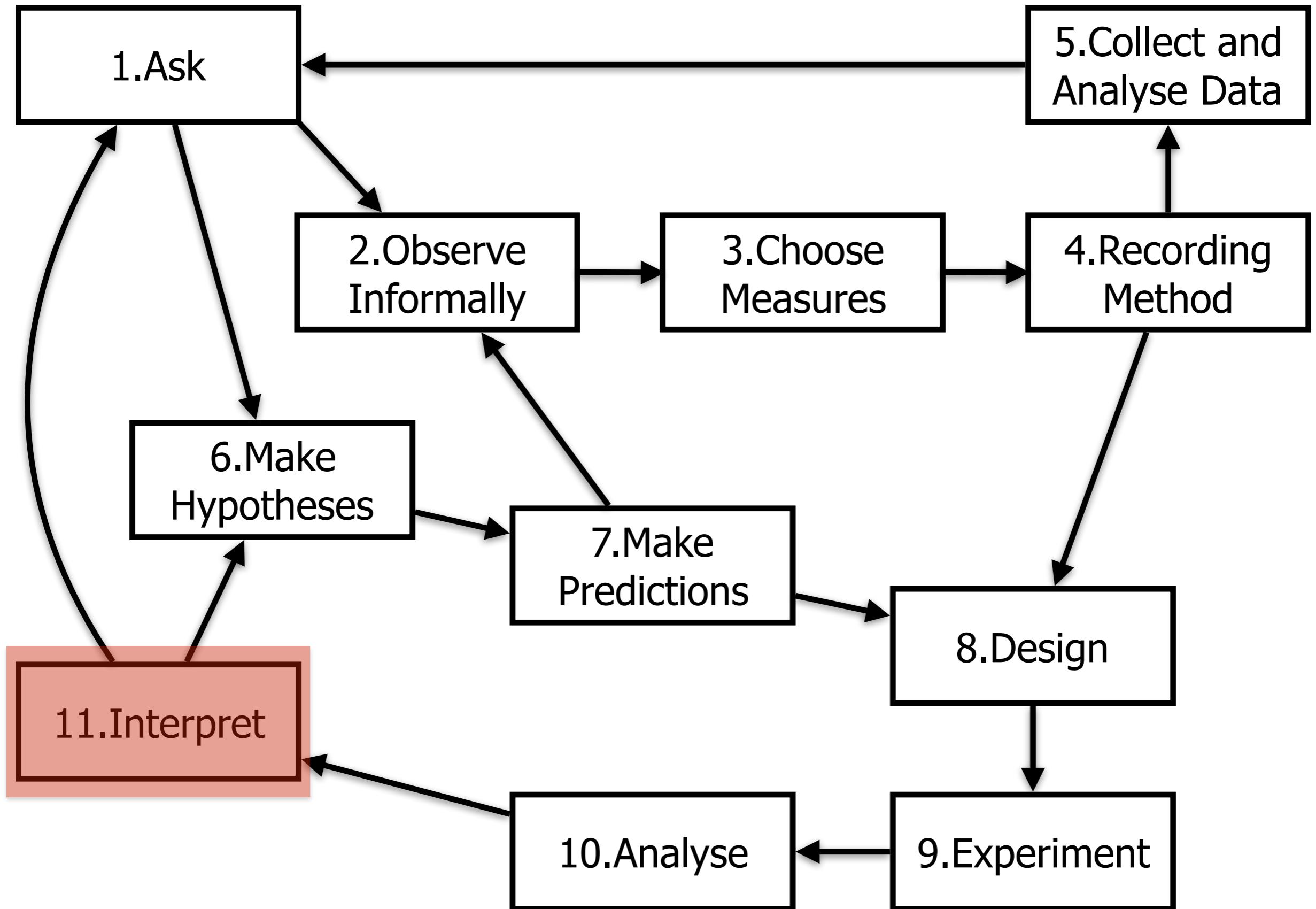


Vinciarelli, Chatzioannou & Esposito, "When Words are Not Everything: the use of laughter, fillers, back-channel, silence, and overlapping speech in phone calls", Frontiers in ICT, 2(4), 2015.



10.Analyse the Results of your Tests.

There is a relationship between the use of nonverbal communication and major social dimensions (both gender and role)



11. Consider Alternative Interpretations of the Evidence.

It is possible that the differences in observed behaviour reflect differences in perceived social status

Gender Differences.

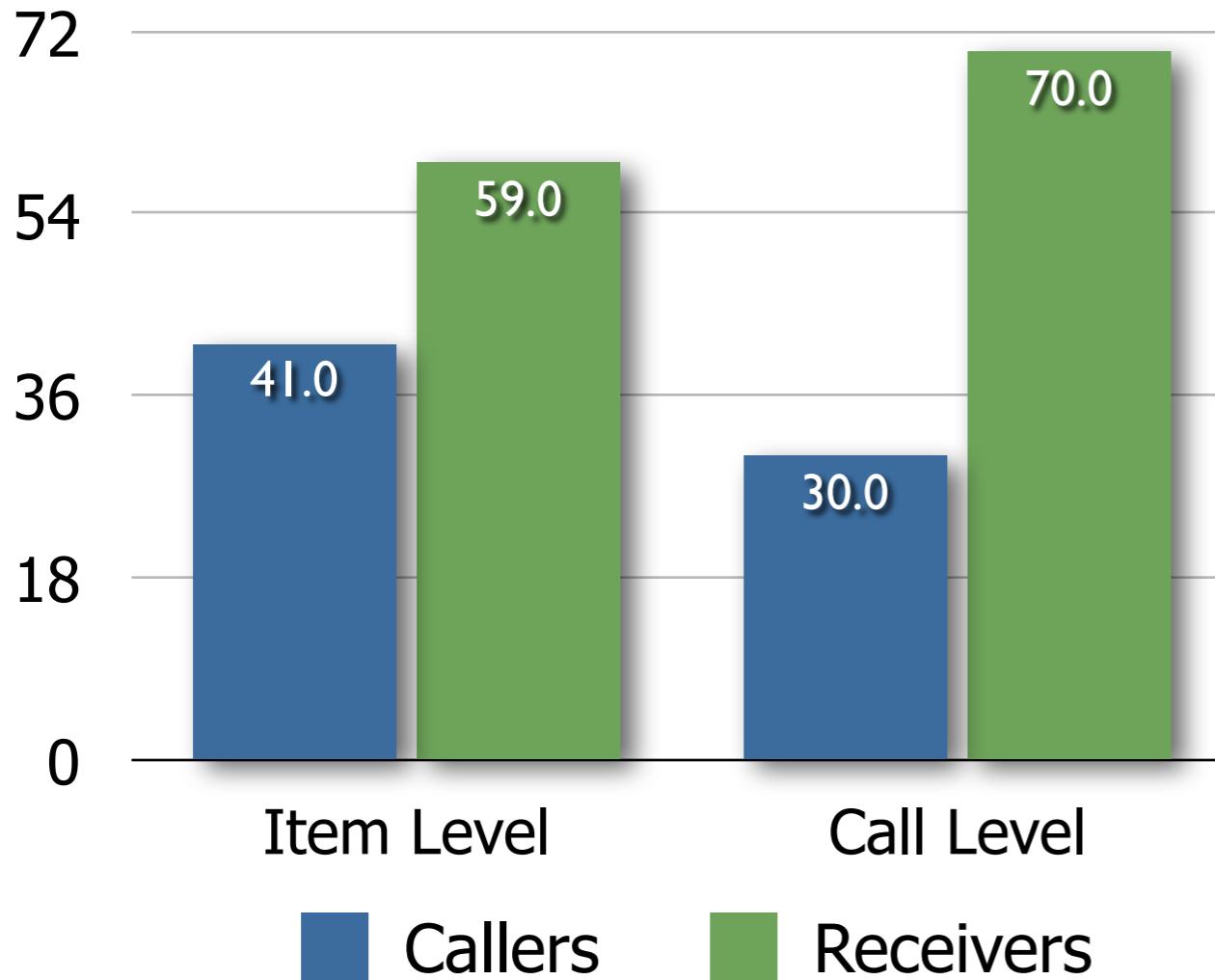
“[...] men and women are generally perceived as differing in status (importance, dominance, power, etc.) and also that they often feel themselves to differ in this way”

Leffler, Gillespie & Conaty, “The Effects of Status Differentiation on Nonverbal Behavior”, Social Psychology Quarterly, 45(3):153–161, 1982

Role Differences.

“[...] when communicating with a higher status person, the lower status person [...] has more filled and unfilled pauses than normal”

Richmond, McCroskey & Payne, “Nonverbal Behavior in Interpersonal Relations”, Prentice Hall, 1991



Receivers win against callers in 59% of the negotiations (70% of the times at the call level).

Calling or receiving makes the difference ($p < 0.005$).

Vinciarelli, Salamin & Polychroniou, "Negotiating Over Mobile Phones: Calling or Being Called can Make the Difference", Cognitive Computation, Vol. 6, no. 4, pp. 677-688, 2014.

Outline

- The 11 Steps of Behaviour Analysis
- An Example: Mobile Phone Conversations
- Conclusions

Conclusions

- Measuring behaviour requires one to carefully design the experiments in terms of measurable and quantitative observations
- Any conclusion about the data must be based on sound statistical approaches
- It is important to provide an interpretation of the observation, possibly based on the relevant literature (psychology, ethology, sociology, etc.)

Student's t

Computational Social Intelligence - Lecture 06

Prof. Alessandro Vinciarelli
School of Computing Science &
Institute of Neuroscience and Psychology

<http://www.dcs.gla.ac.uk/vincia>
Alessandro.Vinciarelli@glasgow.ac.uk



University
of Glasgow

EPSRC
Engineering and Physical Sciences
Research Council

FNSNF

This lecture is based on the following text
(available on Moodle):

- D.C.Howell, “Statistical Methods for Psychology”, Chapter 7, Sections 7.1, 7.2, 7.3 (until page 186 included), 7.5 (until page 203 included), Cengage Learning, 2009.

Outline

- The Gaussian Distribution
- From “z” to “t”
- One Sample Test
- Means Comparison

Outline

- The Gaussian Distribution
- From “z” to “t”
- One Sample Test
- Means Comparison



Carl Friedrich Gauss
1777-1855

The Gaussian (or Normal) pdf

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

“x” is a random variable

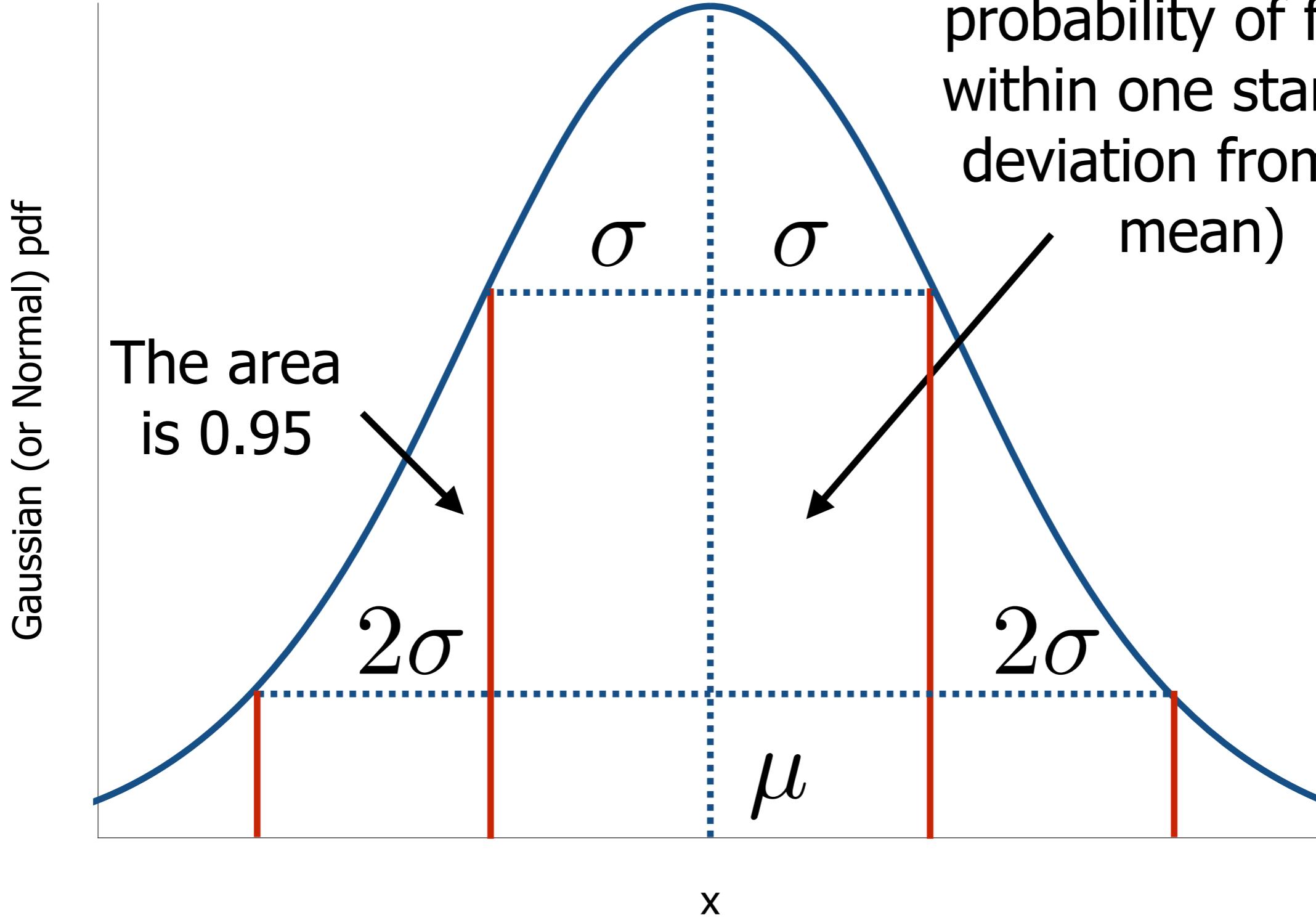
The mean

The variance

The standard deviation

The diagram illustrates the Gaussian probability density function (pdf) with three annotations pointing to specific components:

- An arrow points from the text "The mean" to the term $(x - \mu)^2$ in the exponent.
- An arrow points from the text "The variance" to the term $2\sigma^2$ in the denominator.
- An arrow points from the text "The standard deviation" to the term $\sqrt{2\pi}\sigma$ in the denominator.



“z” is a random variable known as the z-transform of x

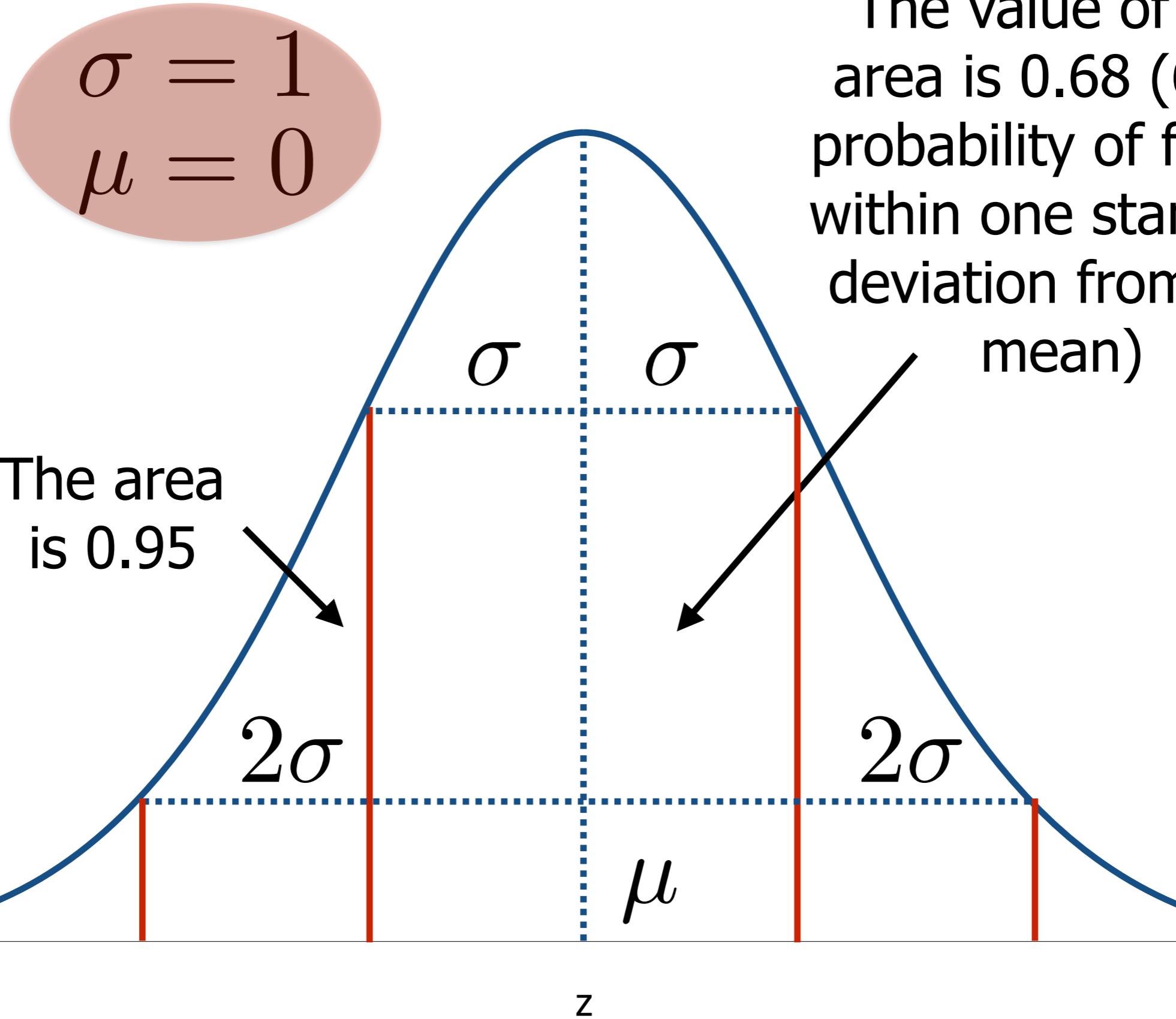
“x” is a random variable that follows a normal pdf

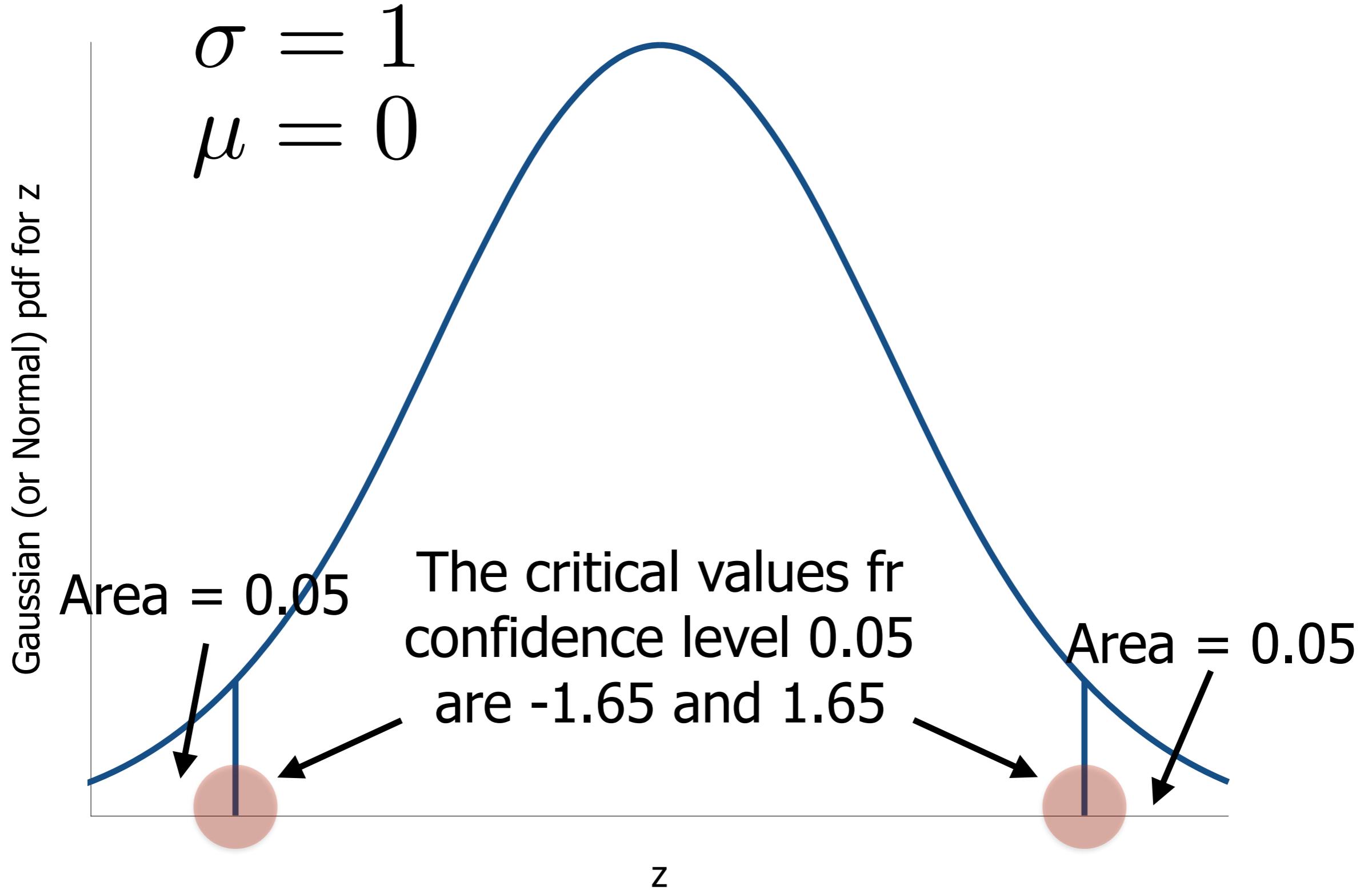
$$z = \frac{x - \mu}{\sigma}$$

The standard deviation of the normal distribution that variable “x” follows

The mean of the normal distribution that variable “x” follows

Gaussian (or Normal) pdf





(Toy) Research Hypothesis

- If the Gaussian is a sampling distribution, it can be used for hypothesis testing;
- Research Hypothesis: A person with temperature 38C has fever;
- Null Hypothesis: A person with temperature 38C does not have fever;
- Mean and variance for the people that have no fever are 37.0 and 0.01, respectively.

The variable value to be tested (temperature in the toy example)

$$z = \frac{x - \mu}{\sigma} = \frac{38.0 - 37.0}{0.1} = 10.0$$

The values of variable, mean and standard deviation are inserted

The value of z is above the critical value, the null hypothesis can be rejected

The equation can help to find the variable value (temperature) beyond which the null hypothesis can be rejected

$$z \geq 1.65 \Rightarrow \frac{x - \mu}{\sigma} \geq 1.65$$

$$x \geq \mu + 1.65 \cdot \sigma \Rightarrow x \geq 37.165$$

The (toy) temperature threshold beyond which the null hypothesis can be rejected

Outline

- The Gaussian Distribution
- From “z” to “t”
- One Sample Test
- Means Comparison

Central Limit Theorem

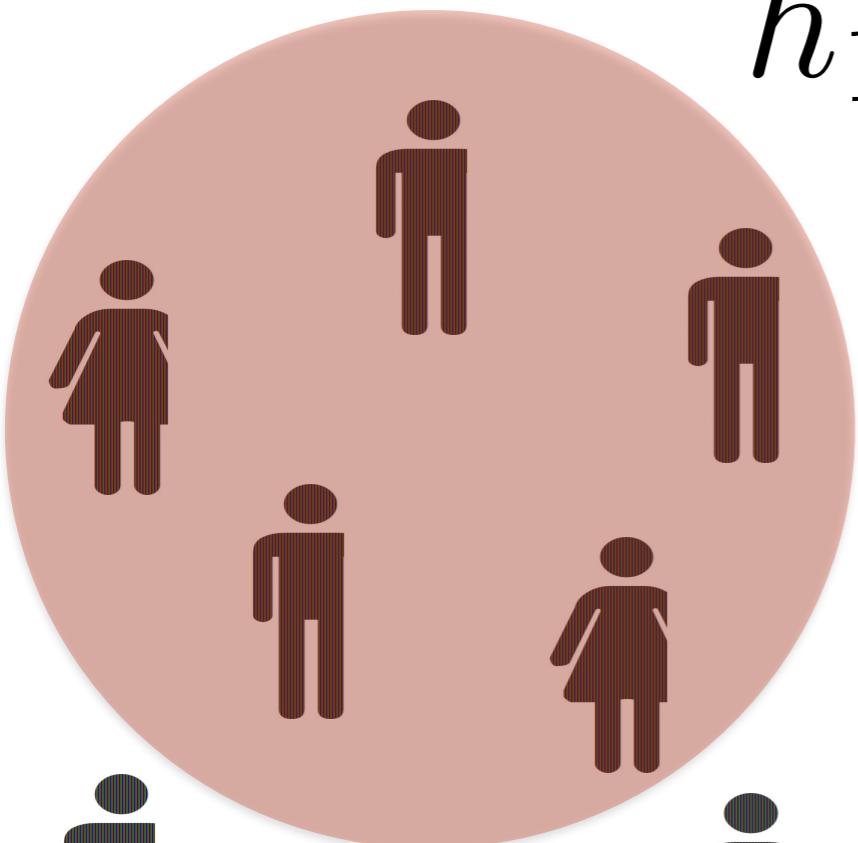
Given a population with mean μ and variance σ^2 , the sampling distribution of the mean will have mean $\mu_{\bar{x}} = \mu$ and variance $\sigma_{\bar{x}}^2 = \sigma^2/n$. The distribution will approach the normal distribution as the sample size n increases.

The height of
individual i in the
population

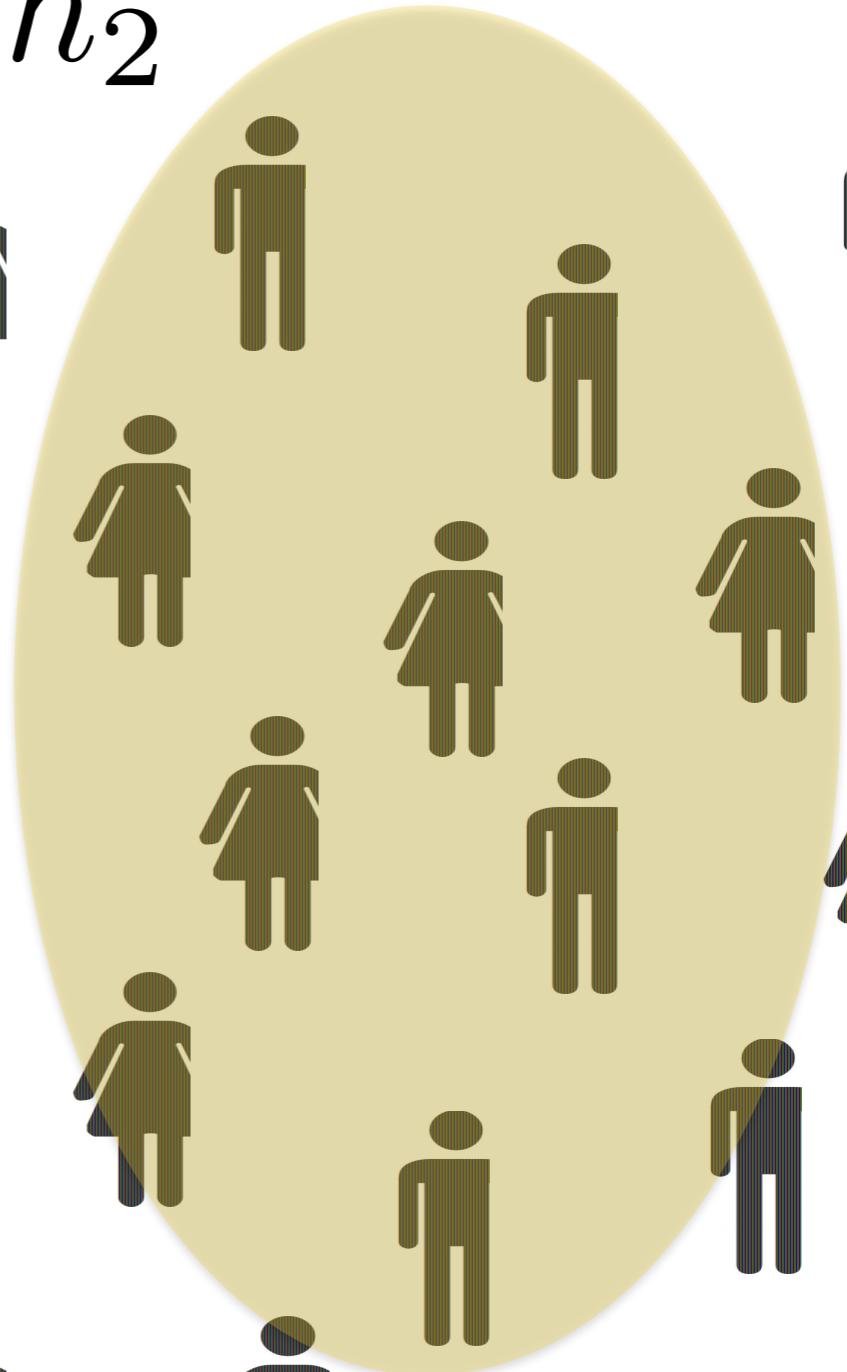
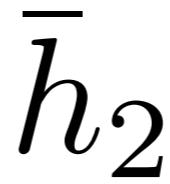
$$\bar{h} = \frac{1}{N} \sum_{i=1}^N h_i$$

The average height of
the individuals in the
population

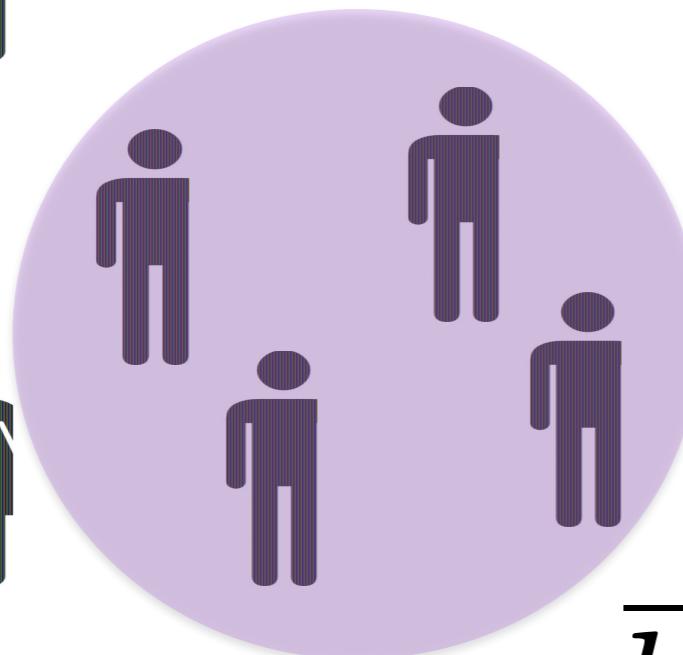


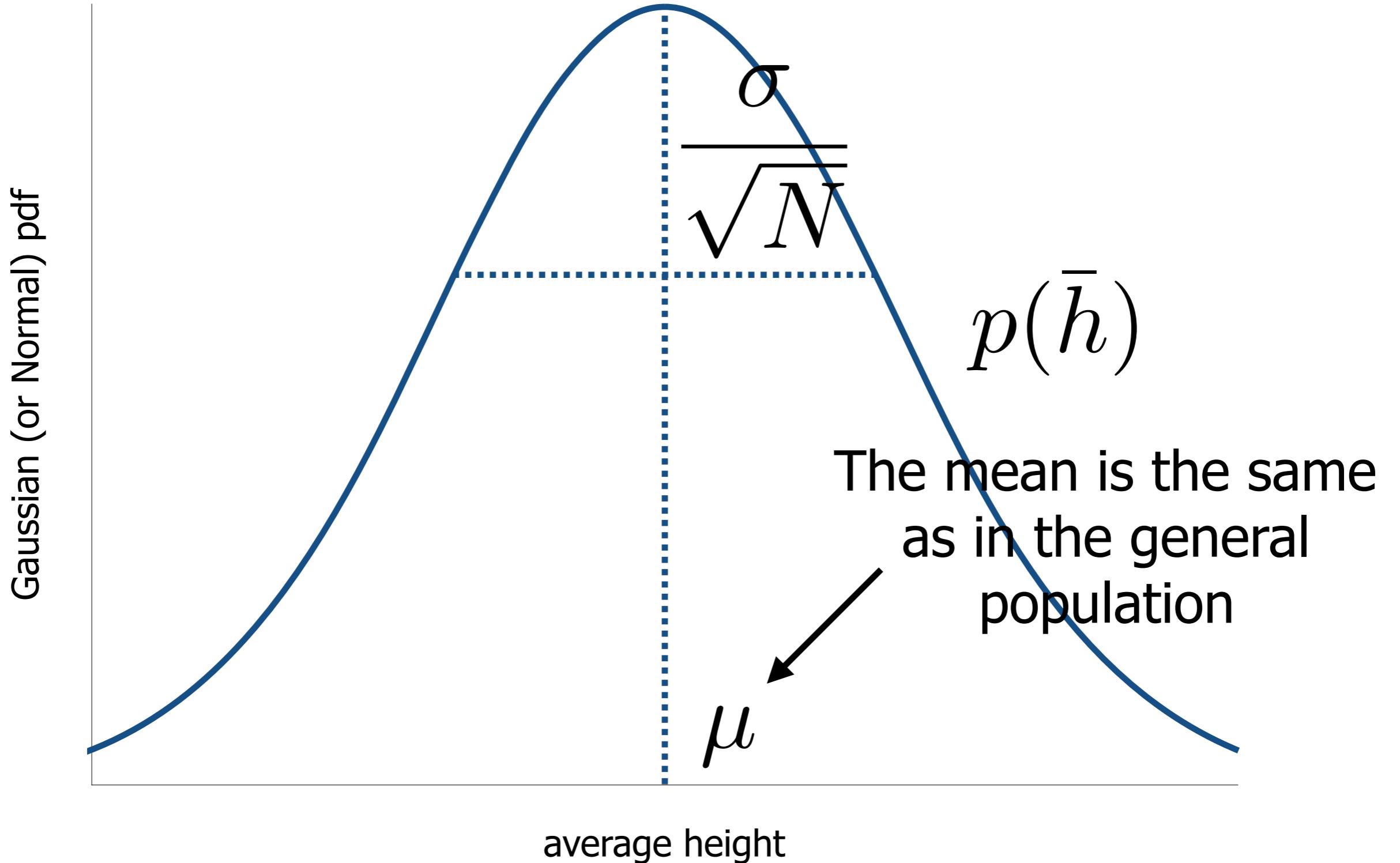


\bar{h}_1



\bar{h}_3





It is possible to apply
the z-transform to the
average age

It is still necessary to
know mean and
variance of the
general population

$$z = \frac{\bar{h} - \mu}{\frac{\sigma}{\sqrt{N}}}$$

For a given average
height, the number of
samples increases the
value of z

Recap

- The consequence of the Central Limit Theorem is that the sampling distribution of a sample average is a Gaussian distribution;
- It is possible to apply the z-transform and test whether the z value is beyond the critical threshold (1.65 for confidence level 0.05);
- It is still necessary to know mean and variance of the general population (rare in practice).

The height of
individual i in the
population

$$\bar{h} = \frac{1}{N} \sum_{i=1}^N h_i$$

The average height of
the individuals in the
population



The sample
variance

The N basketball
players

$$s^2 = \frac{1}{N-1} \sum_{i=1}^N (h_i - \bar{h})^2$$

We still need to know
the mean

The variance is replaced with the sample variance

$$z = \frac{\bar{h} - \mu}{\frac{s}{\sqrt{N}}} = \frac{\bar{h} - \mu}{\sqrt{\frac{s^2}{N}}} = \frac{\bar{h} - \mu}{\frac{s}{\sqrt{N}}}$$

The diagram illustrates the derivation of the z-score formula. It shows three equivalent forms of the formula:

$$z = \frac{\bar{h} - \mu}{\frac{s}{\sqrt{N}}} = \frac{\bar{h} - \mu}{\sqrt{\frac{s^2}{N}}} = \frac{\bar{h} - \mu}{\frac{s}{\sqrt{N}}}$$

Arrows indicate the transformation from the first form to the second, and from the second to the third. Circles highlight the N in the denominators of the second and third forms.

For a given average height, the number of samples increases the value of z

The Student's t
random variable

The sampling
distribution of t when
the null hypothesis is
true is known

$$t = \frac{\bar{h} - \mu}{\sqrt{\frac{s^2}{N}}}$$

The pdf of t (the sampling distribution) when the Null Hypothesis is true

$$p(t) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi} \Gamma\left(\frac{k}{2}\right)} \left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}}$$

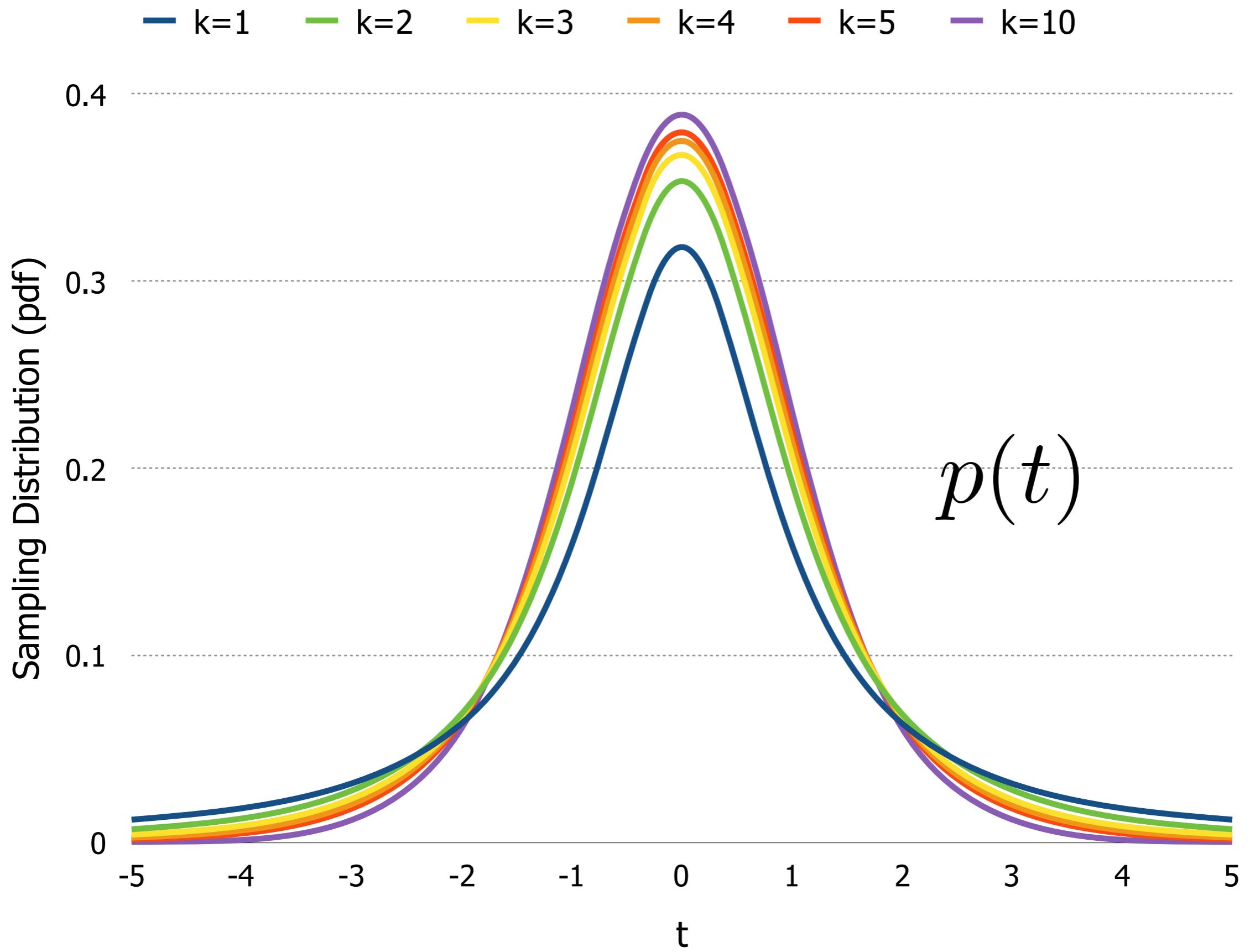
The parameter k is the number of degrees of freedom

The value of the sample variance is a constraint to be respected

$$s^2 = \frac{1}{N-1} \sum_{i=1}^N (h_i - \bar{h})^2$$

$$k = N - 1$$

Once the first $N-1$ values of height have been set, the last can be obtained from the sample variance





William Sealy Gosset
1876-1937

Recap

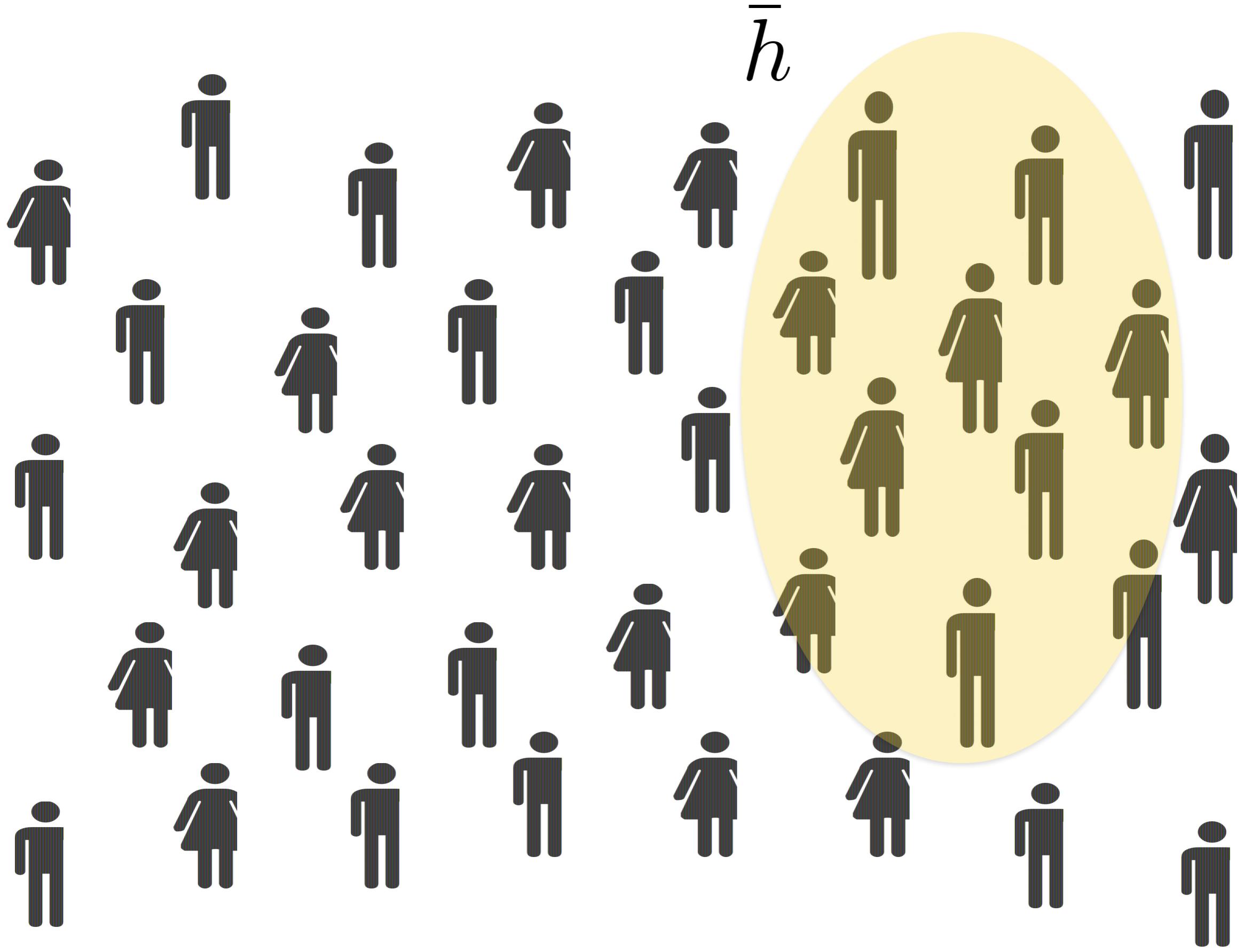
- The Student's t and its sampling distribution allow one to perform hypothesis testing when the mean of the population is known;
- When t (a function of the data) is beyond the critical value corresponding to a confidence level, it is possible to reject the NH;

Outline

- The Gaussian Distribution
- From “z” to “t”
- One Sample Test
- Means Comparison

Research Hypothesis

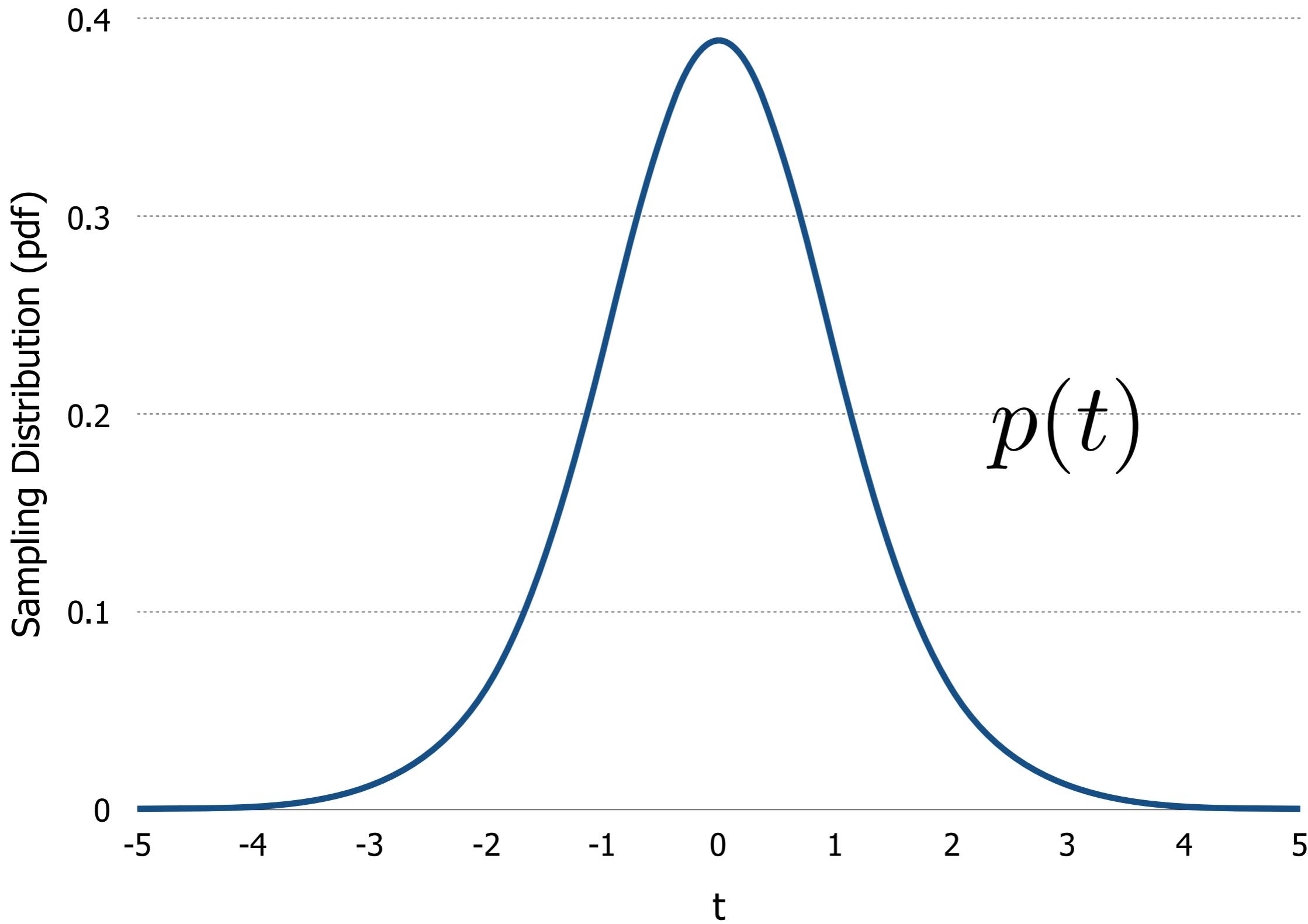
- Research Hypothesis: Basketball players are, on average, taller than the rest of the population;
- Null Hypothesis: Basketball players are, on average, as tall as the rest of the population.



The Student's t
random variable

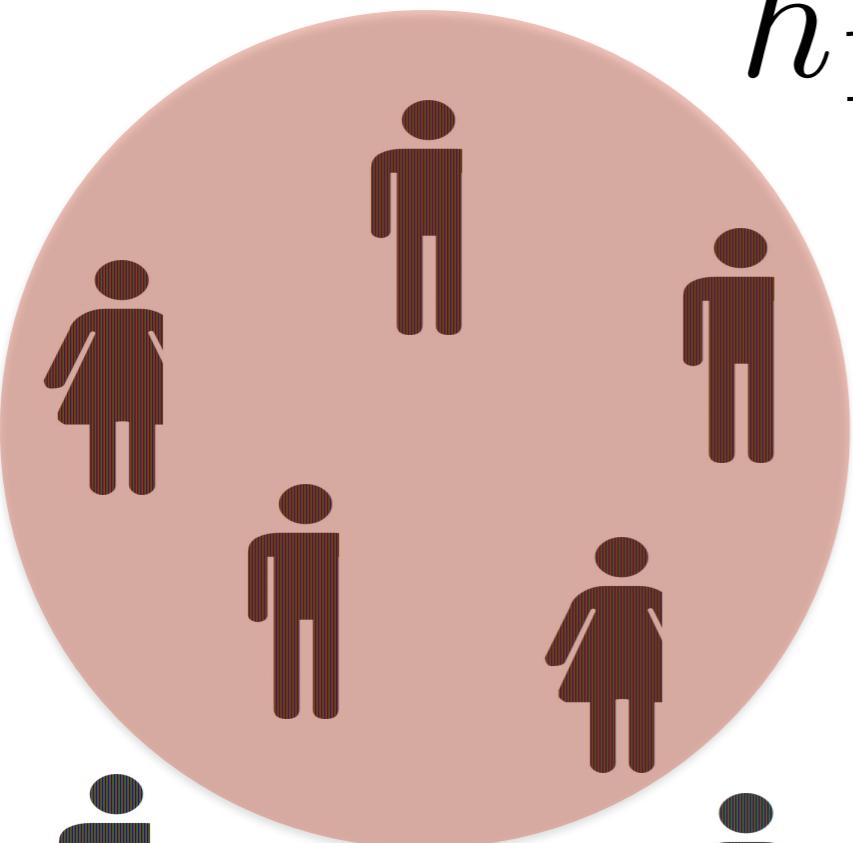
The sampling
distribution of t when
the null hypothesis is
true is known

$$t = \frac{\bar{h} - \mu}{\sqrt{\frac{s^2}{N}}}$$

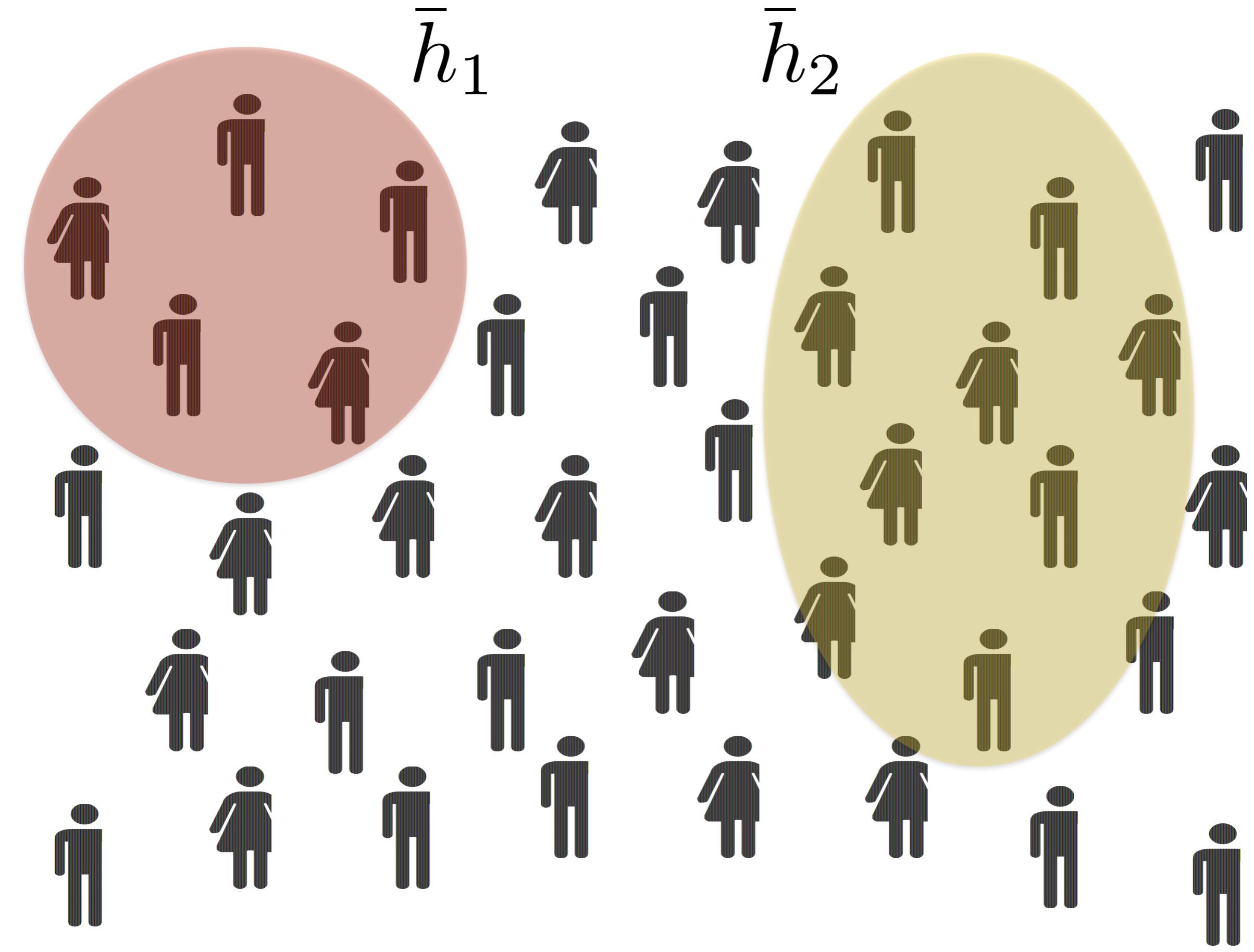
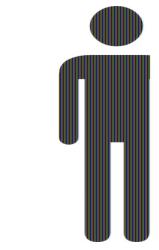


Outline

- The Gaussian Distribution
- From “z” to “t”
- One Sample Test
- Means Comparison



\bar{h}_1



\bar{h}_2

Variance Sum Law

“The variance of a sum or difference of two independent variables is equal to the sum of their variances.”

D.C.Howell, “Statistical Methods for Psychology”, Chapter 7,
Cengage Learning, 2009.

The random variable
is the difference
between the means

The mean of the
sampling distribution
(Central Limit
Theorem)

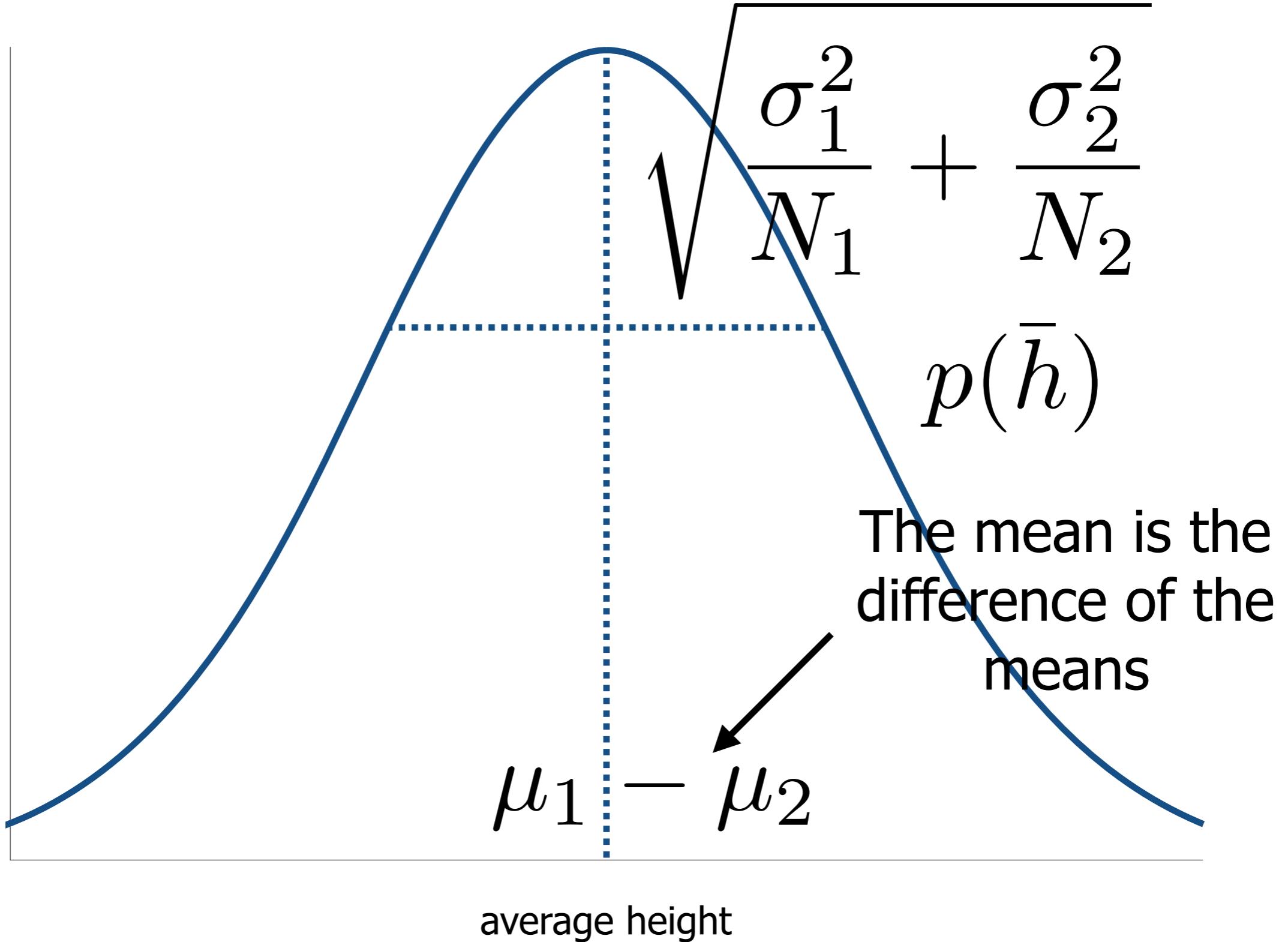
$$\bar{h} = \mu_1 - \mu_2 = \bar{h}_1 - \bar{h}_2$$

$$\sigma_{\bar{h}}^2 = \sigma_{\bar{h}_1}^2 + \sigma_{\bar{h}_2}^2 = \frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}$$

The variance of the
difference between
the means

The Variance Sum
Law

Gaussian (or Normal) pdf



Student's t

The Null Hypothesis is
that such a difference
is null

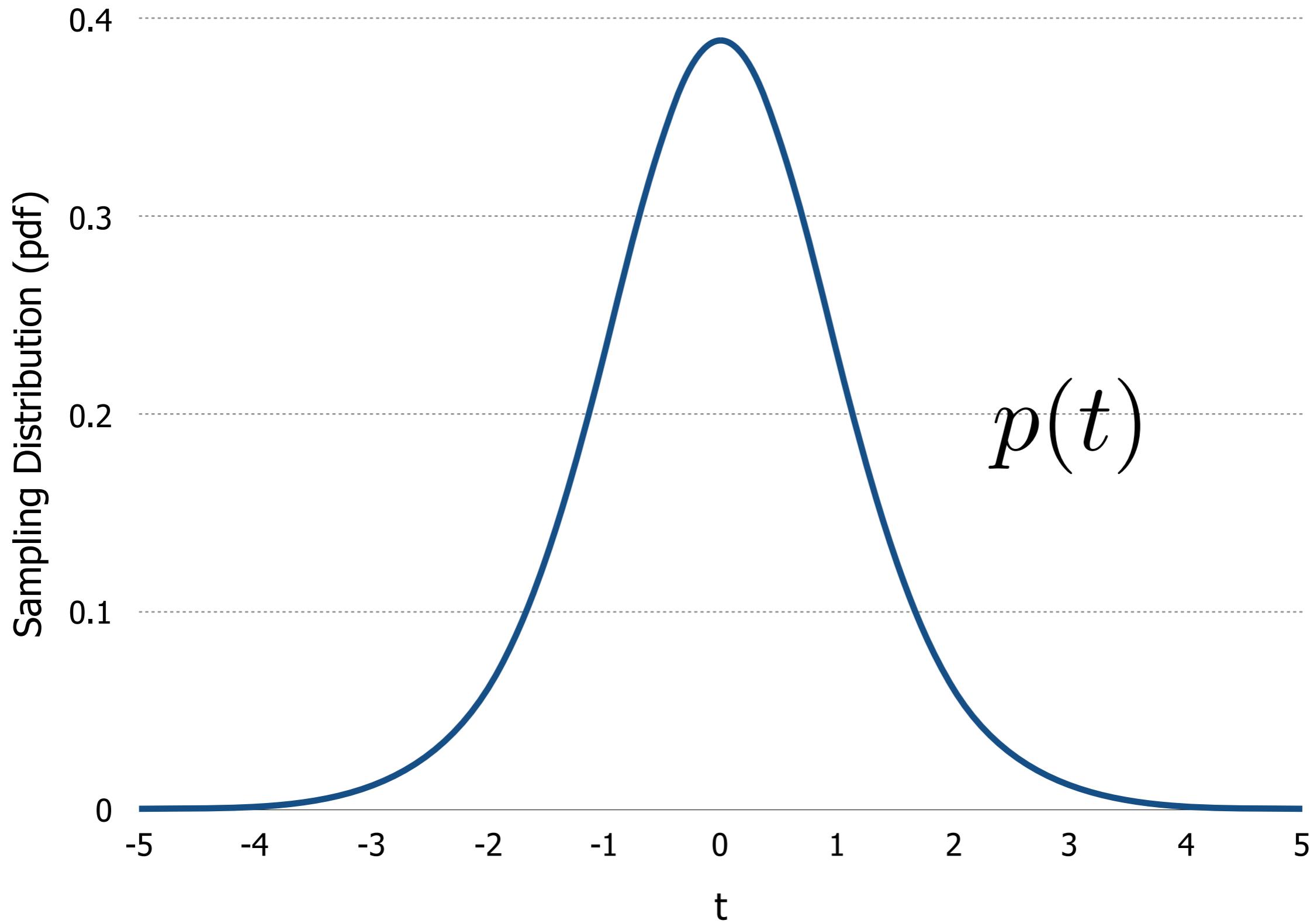
$$t = \frac{\bar{h}_1 - \bar{h}_2 - \mu_1 - \mu_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

Student's t

$$t = \frac{\bar{h}_1 - \bar{h}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

$$k = (N_1 - 1) + (N_2 - 1)$$

The degrees of freedom is the sum of the individual degrees of freedom



Research Hypothesis

- Research Hypothesis: The difference between the means is different from zero;
- Null Hypothesis: The difference between the means is null;
- The name “Null Hypothesis” comes from this case.

Recap

- The Student's t and its sampling distribution allow one to perform hypothesis testing when the mean of the population is known;
- When the statistic is the difference between the means extracted from different samples, then the mean of the population is null.

Thank You!

Psychometrics

Computational Social Intelligence - Lecture 07

Prof. Alessandro Vinciarelli
School of Computing Science &
Institute of Neuroscience and Psychology

<http://www.dcs.gla.ac.uk/vincia>
Alessandro.Vinciarelli@glasgow.ac.uk



University
of Glasgow

EPSRC
Engineering and Physical Sciences
Research Council

FNSNF

This lecture is based on the following text
(available on Moodle):

- T.R.Hinkin, "A Brief Tutorial on the Development of Measures for Use in Survey Questionnaires", Cornell University, 1998

Important

- You are not expected to study the article associated to this lecture (there will be no questions about at the exam);
- However, you are expected to know how to use a questionnaire and to acquire the related terminology;
- The slides provide you with all the information you need for the exam.

Outline

- Introduction
- The Example of Personality
- The Six Steps of Scale Development
- Conclusions

Outline

- Introduction
- The Example of Personality
- The Six Steps of Scale Development
- Conclusions

Psychometrics

1. “The branch of psychology concerned with the design and use of psychological tests.
2. The application of statistical and mathematical techniques to psychological testing”

Collins English Dictionary

Psychological Constructs

"A construct is a representation of something that does not exist as an observable dimension of behavior."

T.R.Hinkin, "A Brief Tutorial on the Development of Measures for Use in Survey Questionnaires", Cornell University, 1998.

Examples

- Personality;
- Emotions;
- Attitude;
- Intention;
- Interpersonal attraction;
- etc.

Outline

- Introduction
- **The Example of Personality**
- The Six Steps of Scale Development
- Conclusions

Personality Self-Assessment

Personality

“[Latent construct that accounts] for individuals’ characteristic patterns of thought, emotion, and behavior together with the psychological mechanisms - hidden or not - behind those patterns.”

D. Funder, “Personality,” Annual Reviews of Psychology, 52:197–221, 2001

The Big-Five

“The Big Five personality factors appear to provide a set of highly replicable dimensions that parsimoniously and comprehensively describe most phenotypic individual differences.”

Saucier, Goldberg, “The Language of Personality: Lexical Perspectives on the Five-Factor Model”, in “The Five-Factor Model of Personality”, Wiggins (ed.), 21-50, 1996

The Big-Five Traits

- **Extraversion:** Active, Assertive, Energetic, ...
- **Agreeableness:** Appreciative, Forgiving, Generous, Kind, Sympathetic, Trusting, ...
- **Conscientiousness:** Efficient, Organized, Planful, Reliable, Responsible, Thorough, ...
- **Neuroticism:** Anxious, Self-pitying, Tense, Touchy, Unstable, Worrying, ...
- **Openness:** Artistic, Curious, Imaginative, ...

The Big-Five Inventory 10

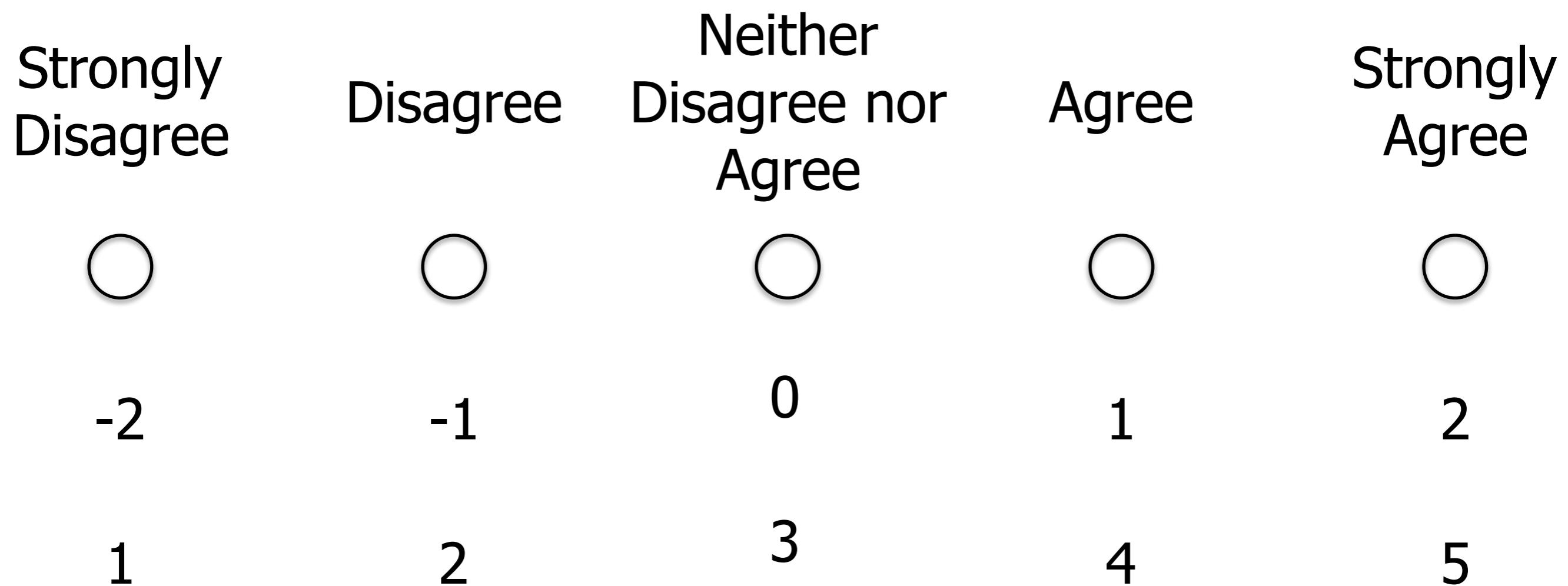
ID	Item	SD	D	NA	A	SA
1	I am reserved					
2	I am generally trusting					
3	I am lazy					
4	I am relaxed, I handle stress well					
5	I have few artistic interests					
6	I am outgoing, sociable					
7	I tend to find faults with others					
8	I do a thorough job					
9	I get nervous easily					
10	I have an active imagination					

Rammstedt and John, "Measuring Personality in One Minute or Less: A 10-item short version of the BFI", Journal of Research in Personality, 41(1): 203-212, 2007

The Items

- An item is a statement or a question about an observable aspect of behaviour;
- Every item is associated to a Likert Scale expected to quantify how correct the statement is;
- The items are expected to be relevant to the construct that the questionnaire aims at measuring.

Likert Scales (I)



Likert Scales (II)

"Likert (1932) developed the scales to be composed of 5 equal appearing intervals with a neutral midpoint [...] Coefficient alpha reliability [...] has been shown to increase up to the use of five points, but then it levels off [...]."

T.R.Hinkin, "A Brief Tutorial on the Development of Measures for Use in Survey Questionnaires", Cornell University, 1998.

Outline

- Introduction
- The Example of Personality
- **The Six Steps of Scale Development**
- Conclusions

Scale Development

“Scale development clearly involves a bit of art as well as a lot of science.”

T.R.Hinkin, “A Brief Tutorial on the Development of Measures for Use in Survey Questionnaires”, Cornell University, 1998.

1.Item Generation

“The key to successful item generation is the development of a well articulated theoretical foundation that would indicate the content domain for the new measure.”

T.R.Hinkin, “A Brief Tutorial on the Development of Measures for Use in Survey Questionnaires”, Cornell University, 1998.

Two Approaches

- The items must be relevant to the construct being addressed;
- Deductive: The items are deduced from the theory underlying the construct under exam (more common and reliable);
- Inductive: The items are designed to define the construct (less common).

2. Questionnaire Administration

"The items should now be presented to a sample representative of the actual population of interest [to confirm] expectations regarding the psychometric properties of the new measure."

T.R.Hinkin, "A Brief Tutorial on the Development of Measures for Use in Survey Questionnaires", Cornell University, 1998.

Test

- The sample should be selected according to the construct being targeted (e.g., people without children should not fill questionnaires about parenting);
- The sample should be large enough to allow statistical analysis of the results;
- The sample should be large enough to avoid individual biases.

3. Initial Item Reduction

“[...] it is recommended that factor analysis is used to further refine the new scales [...] This creates a more parsimonious representation of the original set of observations [...]”

T.R.Hinkin, “A Brief Tutorial on the Development of Measures for Use in Survey Questionnaires”, Cornell University, 1998.

Main Criteria

- The items that do not show sufficient variance should be removed;
- The items that do not correlate sufficiently with the others should be removed;
- Factor analysis is the approach most commonly adopted.

4. Confirmatory Factor Analysis

“[...] confirmatory factory analysis should be just that - a confirmation that the prior analyses have been conducted thoroughly and appropriately [...]”

T.R.Hinkin, “A Brief Tutorial on the Development of Measures for Use in Survey Questionnaires”, Cornell University, 1998.

Main Criteria

- After the initial reduction of the items, the statistical properties of the items should be improved or, at least, confirmed;
- It is important to administer the test to a sample different from the one involved in the previous steps.

5. Convergent/Discriminant Validity

"[...] examining the extent to which the scales correlate with other measures designed to assess similar constructs (convergent validity) and to which they do not correlate with dissimilar measures (discriminant validity)."

T.R.Hinkin, "A Brief Tutorial on the Development of Measures for Use in Survey Questionnaires", Cornell University, 1998.

Sanity Check

- The outcome of the questionnaire should be aligned with other, established questionnaires targeting the main construct;
- The outcome of the questionnaire should not correlate with the outcome of other questionnaires targeting other constructs.

6. Replication

“It would now be necessary to collect another set of data from an appropriate sample and repeat the scale-testing process with the new scales.”

T.R.Hinkin, “A Brief Tutorial on the Development of Measures for Use in Survey Questionnaires”, Cornell University, 1998.

Test

- The outcome of a questionnaire depends both on the test and on the people that fill it;
- Analysing the outcome of the questionnaire over multiple samples ensures that the dependence on those filling it is attenuated;
- If a questionnaire is effective, it will be adopted more likely in practice (empirical a-posteriori confirmation).

Outline

- Introduction
- The Example of Personality
- The Six Steps of Scale Development
- Conclusions

Conclusions

- A questionnaire is the result of an empirical process driven by rigorous scientific criteria;
- The statistical properties of the questionnaire's outcomes are the main evaluation criteria;
- The effectiveness of a questionnaire in addressing professional and scientific problems makes it its adoption more or less likely.

Thank You!

Relationships

Computational Social Intelligence - Lecture 08

Prof. Alessandro Vinciarelli
School of Computing Science &
Institute of Neuroscience and Psychology

<http://www.dcs.gla.ac.uk/vincia>
Alessandro.Vinciarelli@glasgow.ac.uk



University
of Glasgow

EPSRC
Engineering and Physical Sciences
Research Council

FNSNF

This lecture is based on the following text
(available on Moodle):

- D.C.Howell, “Statistical Methods for Psychology”, Chapter 9, pp. 244-257 (included), pp. 273-274 (included), Cengage Learning, 2009.

Outline

- Introduction
- Regression
- Correlation
- Conclusions

Outline

- Introduction
- Regression
- Correlation
- Conclusions

Relationships

“When we are concerned with relationships [...] the experimenter is interested in showing that the dependent variable is some function of the independent variable.”

D.C.Howell, “Statistical Methods for Psychology”, Chapter 9,
Cengage Learning, 2009.

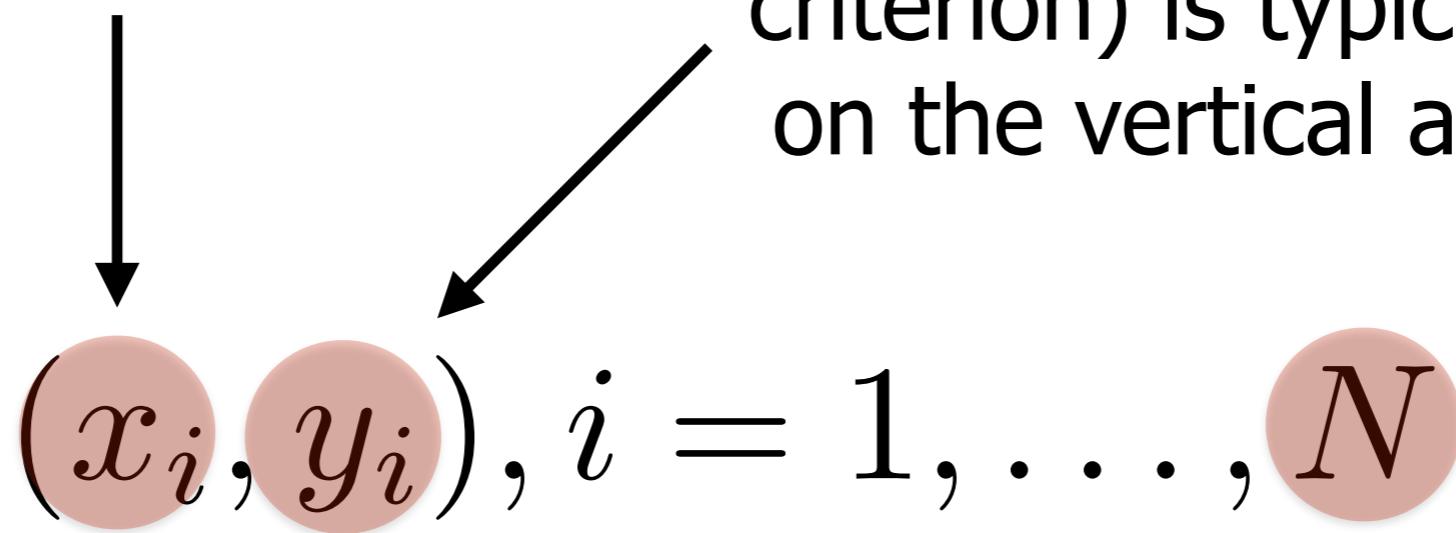
Scatterplot

“One of the most useful techniques for gaining insight into [a] relationship is a scatterplot (also called a scatter diagram or scattergram).”

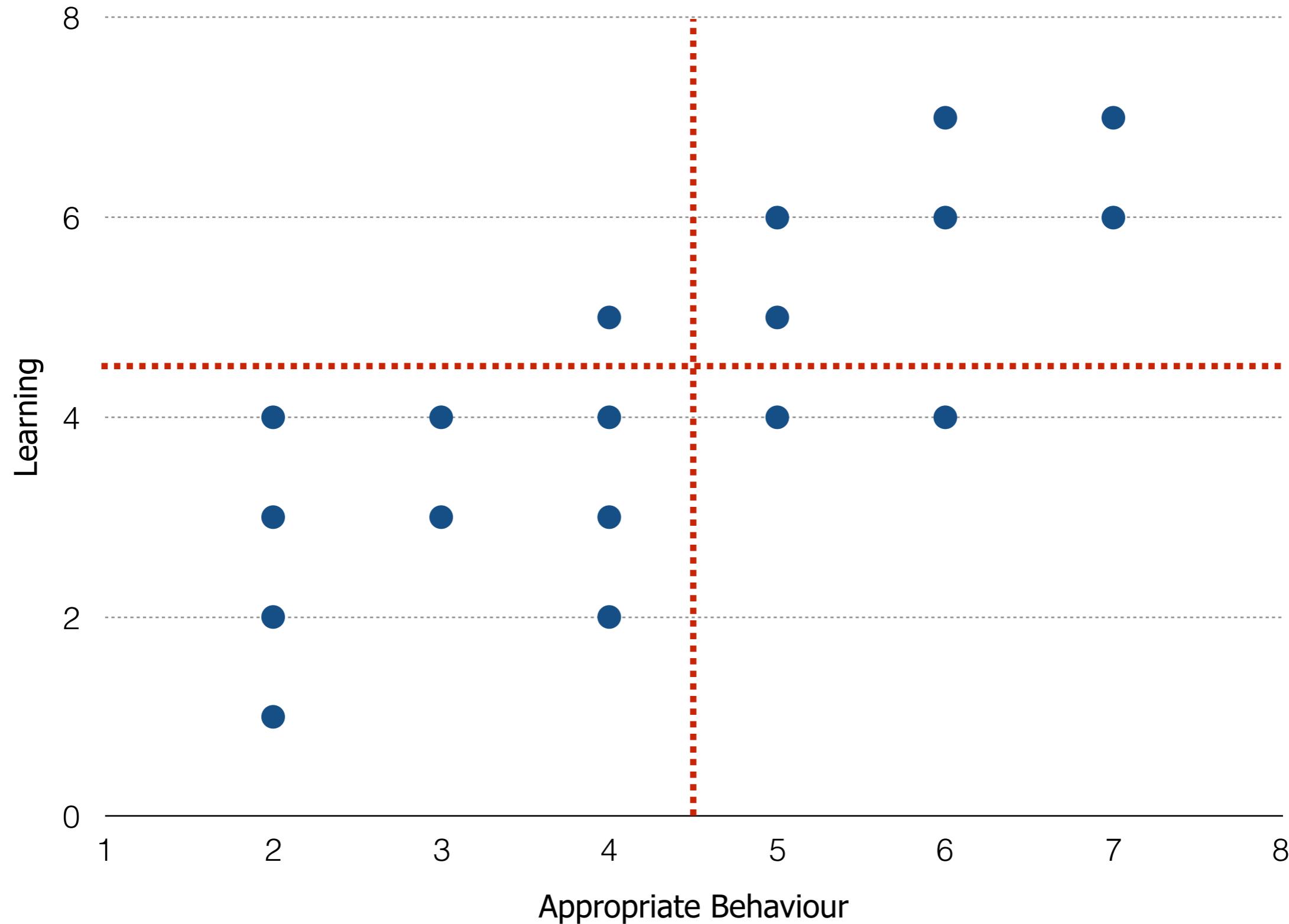
D.C.Howell, “Statistical Methods for Psychology”, Chapter 9,
Cengage Learning, 2009.

The independent variable (the predictor) is typically on the horizontal axis

The dependent variable (the criterion) is typically on the vertical axis



The number of pairs at disposition

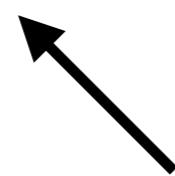


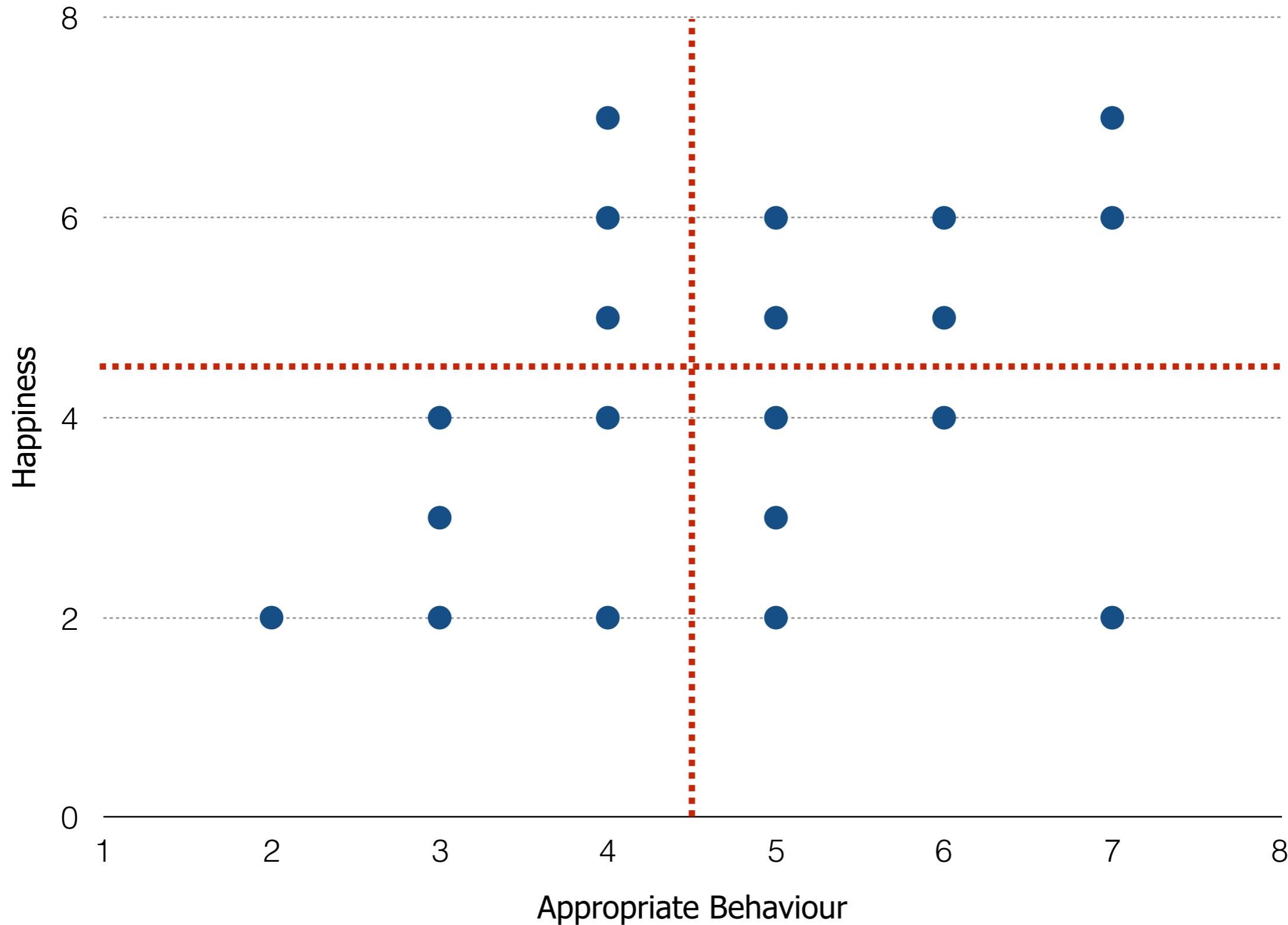
Appropriate
Behaviour



	Above Mean	Below Mean
Above Mean	7	1
Below Mean	2	10

Learning



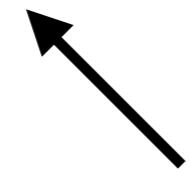


Happiness



	Above Mean	Below Mean
Above Mean	6	3
Below Mean	5	6

Appropriate
Behaviour



The covariance of X
and Y

$$\sigma_{xy} = \frac{\sum_{i=1}^N (x_i - \bar{x})(y - \bar{y})}{N - 1}$$

The average
value of X

The average
value of Y

The total number of
pairs (x,y)

Covariance of
Appropriate
Behaviour and
Learning

$$\sigma_{xy} = 2.34$$

$$\sigma_{xy} = 0.96$$

Covariance of
Appropriate
Behaviour and
Happiness

Covariance

- The expression of the covariance is symmetric with respect to X and Y, its value does not change by switching predictor and criterion;
- It captures the relationship between variables as a systematic tendency to be on the same (or opposite) side of the mean;
- The value of the covariance is difficult to interpret.

Recap

- Some random variables are expected to change according to one another;
- Conventionally, one of the two variables acts as predictor while the other acts as criterion;
- Switching between predictor and criterion does not change the results, the relationship has no direction.

Outline

- Introduction
- Regression
- Correlation
- Conclusions

The estimated value
of y when using a
linear plot

The predictor (the
independent variable)

$$\hat{y} = a + bx$$

The intercept The slope

The diagram illustrates the components of a linear regression equation. The equation is $\hat{y} = a + bx$. Three circles highlight the terms a , b , and x . Arrows point from the text 'The estimated value of y when using a linear plot' to the circle containing a . Another arrow points from 'The predictor (the independent variable)' to the circle containing x . A third arrow points from 'The intercept' to the circle containing a .

The prediction error

The estimated value
of y when using a
linear plot

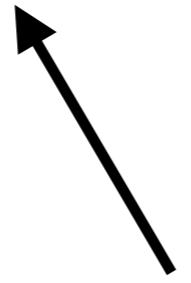
$$E = \sum_{i=1}^N (y_i - \hat{y}_i)^2 = \sum_{i=1}^N (y_i - a - bx_i)^2$$

What is the “best” value for
slope and intercept?

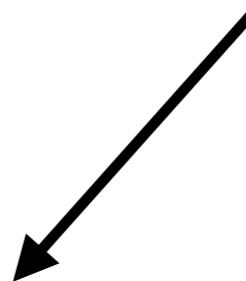
The derivative of the Error with respect to the intercept



$$\frac{\partial E}{\partial a} = 0$$



The derivative of the Error with respect to the slope



$$\frac{\partial E}{\partial b} = 0$$



The results of the equations are the values of intercept and slope that correspond to the minimum Error

$$b = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2} = \frac{\text{The covariance of X and Y}}{\text{The variance of X}} = \frac{\sigma_{xy}}{\sigma_x^2}$$

The slope

The covariance of X and Y

The variance of X

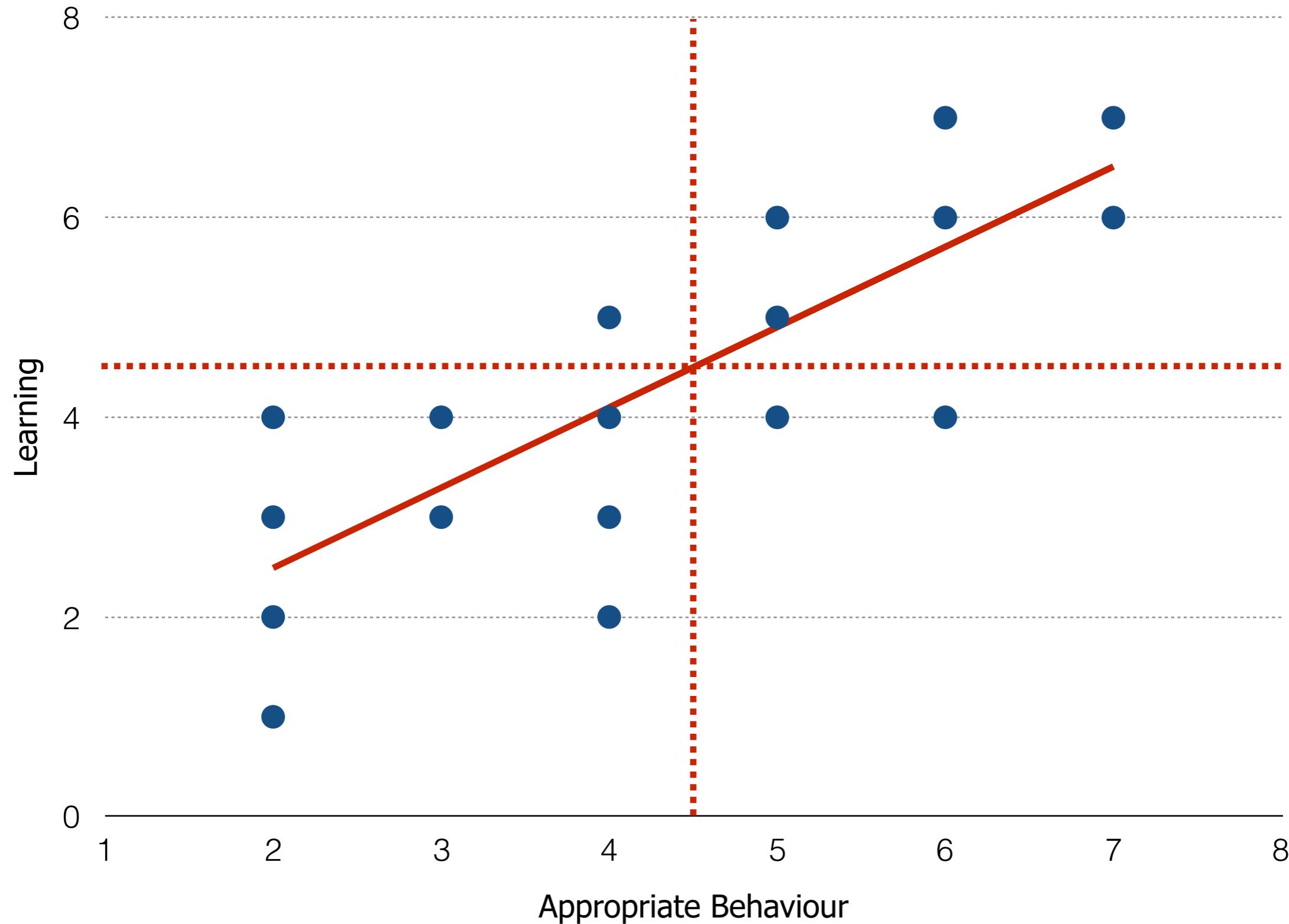
The averages
of X and Y

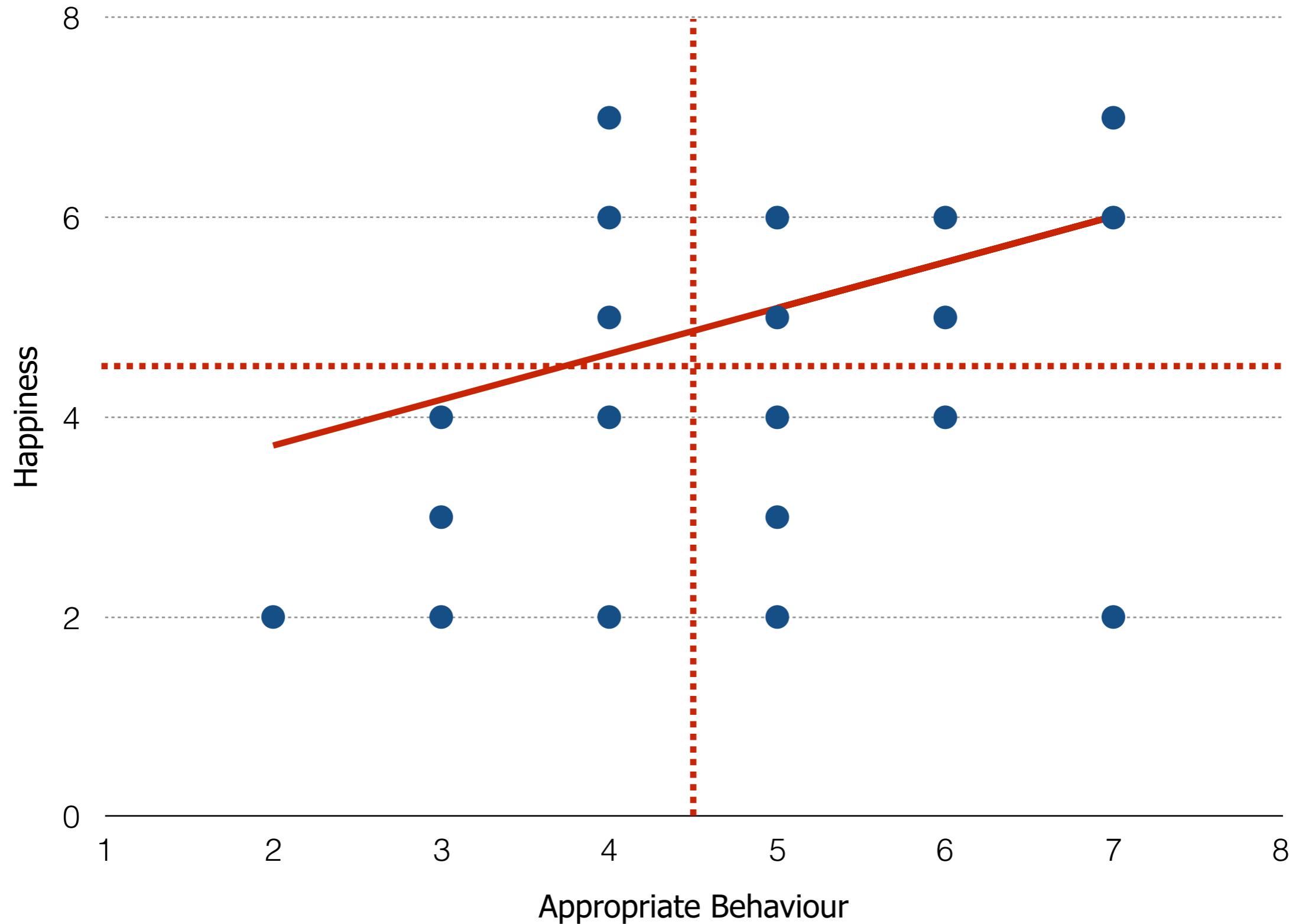
The intercept

$$a = \frac{1}{N} \sum_{i=1}^N x_i - b \frac{1}{N} \sum_{i=1}^N y_i$$

The slope

The total number of
pairs (x,y)





Average Prediction
Error (Appropriate
Behaviour vs
Learning)

$$\frac{E}{N} = 1.07$$

$$\frac{E}{N} = 3.00$$

Average Prediction
Error (Appropriate
Behaviour vs
Happiness)

Recap

- The regression allows one to express the criterion as a function of the predictor;
- The use of a line is a constraint that is not necessarily respected by the data, it is an approximation;
- It is difficult to quantify how well the data fits the relationship.

Outline

- Introduction
- Regression
- Correlation
- Conclusions

Correlation

“The degree to which the points cluster around the regression line (in other words, the degree to which the actual values of Y agree with the predicted values) is related to the correlation between X and Y.”

D.C.Howell, “Statistical Methods for Psychology”, Chapter 9,
Cengage Learning, 2009.

The Pearson
correlation between X
and Y

The covariance of X
and Y

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

The standard
deviation of X

The standard
deviation of Y

The covariance
of X and Y (multiplied
by N-1)

The Pearson
correlation coefficient
between X and Y

$$\frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2 \sum_{i=1}^N (y_i - \bar{y})^2}}$$

The variance of X
(multiplied by N-1)

The variance of Y
(multiplied by N-1)

The correlation can
be adjusted when N
(the number of pairs)
is small

The non-adjusted
value of the
correlation

$$r_{adj} = \sqrt{1 - \frac{(1 - r^2)(N - 1)}{N - 2}}$$

The total number of
pairs (x,y)

The Student's t variable can show whether r is statistically significant

The total number of pairs (x,y)

$$t = \frac{r \sqrt{N - 1}}{\sqrt{1 - r^2}}$$

The value of the correlation

Research Hypothesis

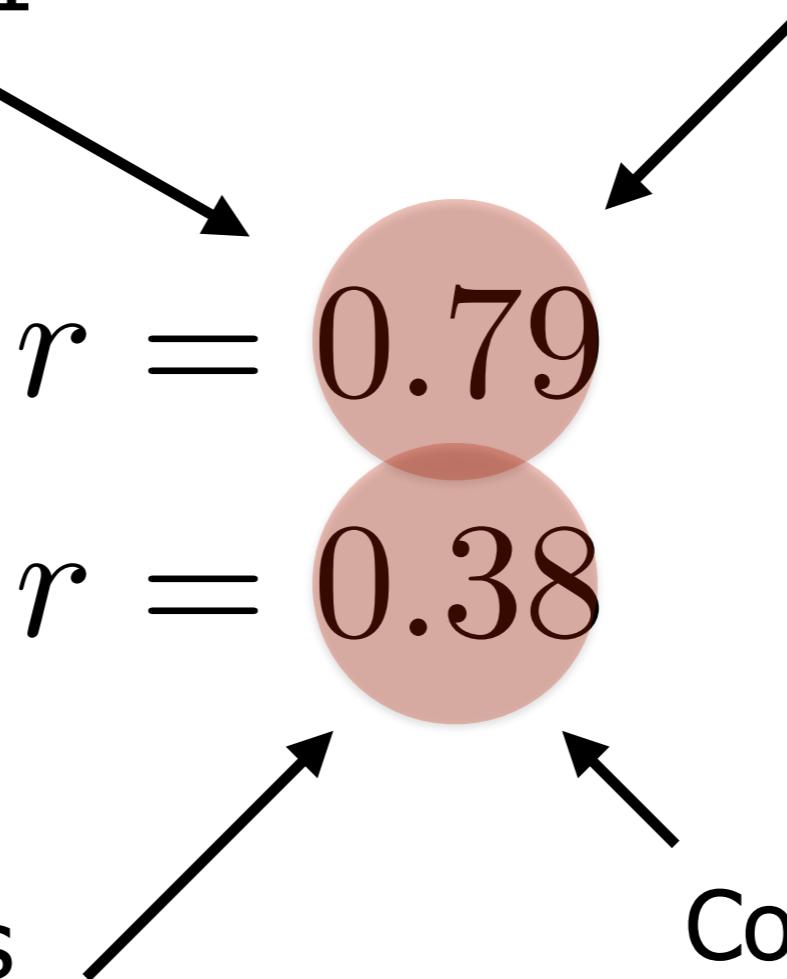
- Research Hypothesis: The correlation is higher (or lower for the negative values) than what is expected by chance;
- Null Hypothesis: The correlation is lower (or higher for the negative values) than what is expected by chance.

The Null Hypothesis
can be rejected with
confidence level 0.01
(two-tailed)

Correlation between
Appropriate
Behaviour and
Learning

The Null Hypothesis
fails to be rejected

Correlation between
Appropriate
Behaviour and
Happiness



Recap

- The correlation measures the fraction of common variance with respect to the variance of the individual variables;
- The values of the correlation are comparable and easy to interpret;
- It is possible to test whether the correlation is statistically significant, i.e., higher than what is expected by chance.

Outline

- Introduction
- Regression
- Correlation
- Conclusions

Conclusions

- It is possible to measure the relationship between two variables, i.e., their tendency to change according to each other;
- The relationship has no direction (it is not possible to say whether one variable influences the other or vice versa);
- However, the analysis of the relationships can provide insight about the phenomena under exam.

Thank You!

Judgment's Studies

Computational Social Intelligence - Lecture 09

Prof. Alessandro Vinciarelli
School of Computing Science &
Institute of Neuroscience and Psychology

<http://www.dcs.gla.ac.uk/vincia>
Alessandro.Vinciarelli@glasgow.ac.uk



University
of Glasgow

EPSRC
Engineering and Physical Sciences
Research Council

FNSNF

This lecture is based on the following text
(available on Moodle):

- R.R.Rosenthal, "Conducting Judgment Studies: Some Methodological Issues", in "The New Handbook of Methods in Nonverbal Behavior Research", J.A.Harrigan, R.Rosenthal and K.R.Scherer (eds.), pp 199-211 (included), pp 213-214 (Cronbach's Alpha), 2008.

Outline

- Introduction
- Reliability
- Conclusions

Outline

- Introduction
 - Reliability
 - Conclusions

The Big-Five Inventory 10

ID	Item	SD	D	NA	A	SA
1	I am reserved					
2	I am generally trusting					
3	I am lazy					
4	I am relaxed, I handle stress well					
5	I have few artistic interests					
6	I am outgoing, sociable					
7	I tend to find faults with others					
8	I do a thorough job					
9	I get nervous easily					
10	I have an active imagination					

Rammstedt and John, "Measuring Personality in One Minute or Less: A 10-item short version of the BFI", Journal of Research in Personality, 41(1): 203-212, 2007

The Big-Five Inventory 10

ID	Item	SD	D	NA	A	SA
1	This person is reserved					
2	... is generally trusting					
3	... is person lazy					
4	... is relaxed, handles stress well					
5	... has few artistic interests					
6	... is outgoing, sociable					
7	... tends to find faults with others					
8	... does a thorough job					
9	... gets nervous easily					
10	... has an active imagination					

Rammstedt and John, "Measuring Personality in One Minute or Less: A 10-item short version of the BFI", Journal of Research in Personality, 41(1): 203-212, 2007

Judgment Studies

"The term 'judgment studies' refers most generally to those studies in which behaviors, persons, objects or concepts are evaluated by one or more judges, raters, coders, or categorizers, referred to collectively as 'judges'."

R.R.Rosenthal, "Conducting Judgment Studies: Some Methodological Issues",
in "The New Handbook of Methods in Nonverbal Behavior Research",
J.A.Harrigan, R.Rosenthal and K.R.Scherer (eds.), 2008.

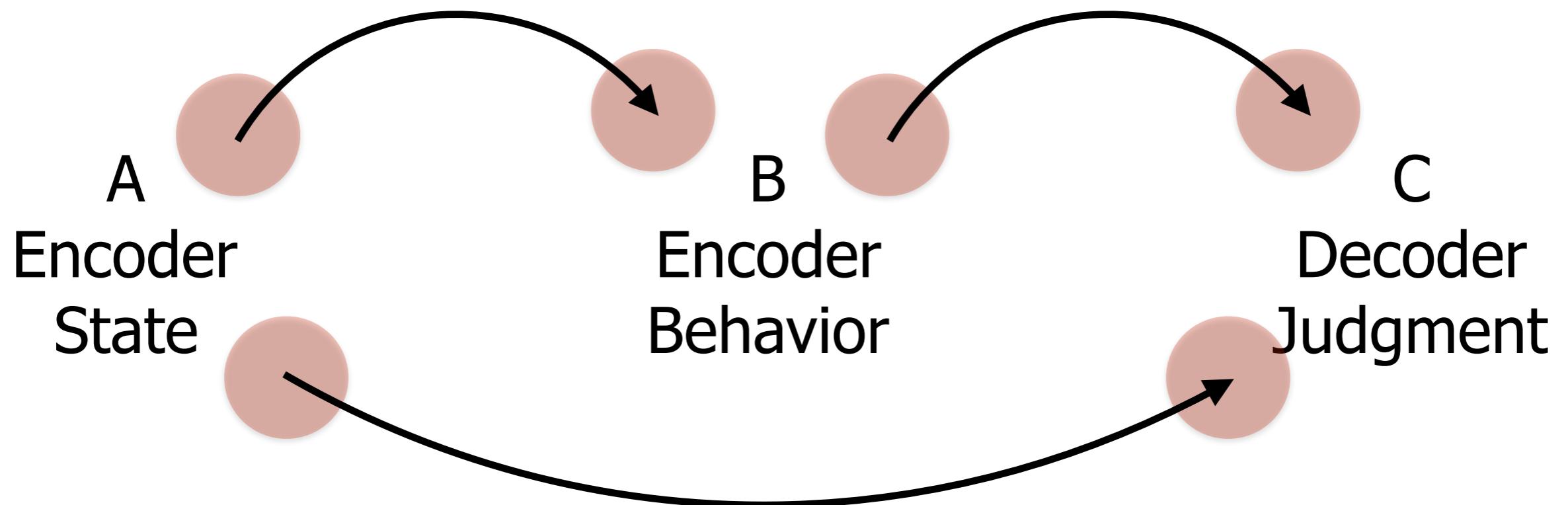
Types of Judgment Studies

Dimensions	Examples
Type of Variable	Dependent vs Independent
Measurement Units	Physical vs Psychological
Reliability	Lower vs Higher
Social Meaning	Lower vs Higher

R.R.Rosenthal, "Conducting Judgment Studies: Some Methodological Issues",
in "The New Handbook of Methods in Nonverbal Behavior Research",
J.A.Harrigan, R.Rosenthal and K.R.Scherer (eds.), 2008.

How does the
encoder manifest her/
his state through her/
his behaviour?

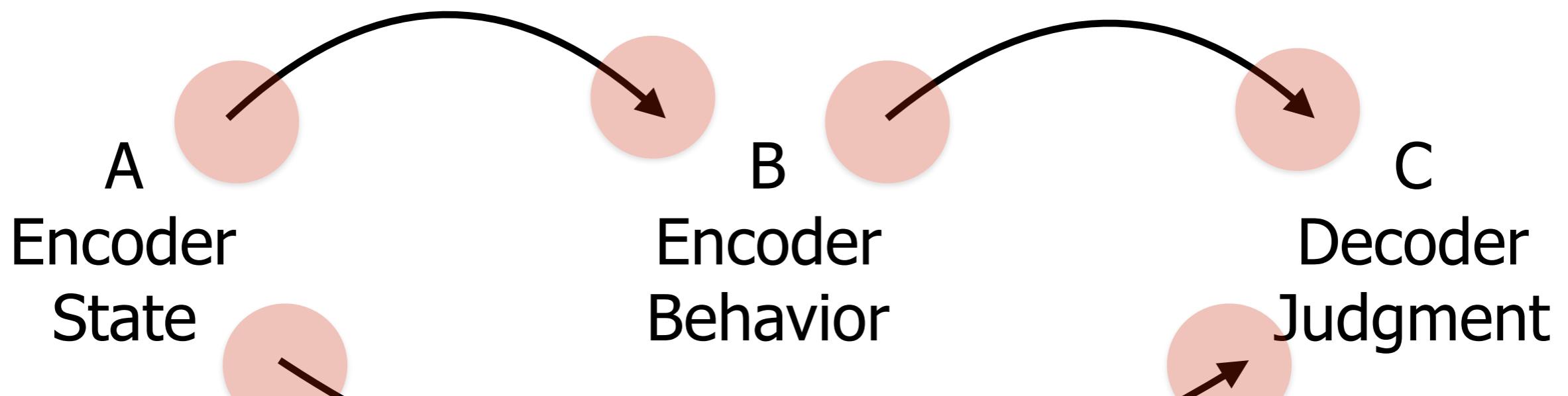
How do decoders
interpret the
behaviour of the
encoder



What is the state the
decoder attributes to
the encoder?

The State is the independent variable,
the behaviour the dependent one

The behaviour is the independent variable,
the judgment the dependent one



The State is the independent variable,
the judgment is the dependent one

Main Issues in Judgment Studies

- Reliability: How reliable are the judgments?
How many judges are necessary to obtain reliable judgments?
- Selection: How to select the judges?
- Composition: How to combine different judgments to form composite variables?

Outline

- Introduction
- Reliability
- Conclusions

Reliability

- The consensus among multiple judges suggests that there is consistency between observations and judgments;
- The reliability can be thought of as the measure of the consensus among multiple judges;
- In principle, the higher the consensus, the higher the reliability.

Percentage of times
two judges agree

Number of times two
judges agree

$$R = \left(\frac{A}{A + D} \right) 100$$

Number of times two
judges disagree

Decision of Judge A

	Frown	No-Frown
Frown	98	1
No-Frown	1	0

Decision of Judge C

	Frown	No-Frown
Frown	49	1
No-Frown	1	49

Decision of Judge B

Decision of Judge D

$$R = 98\%$$

Decision of Judge A
(1 for Frown and -1
for No-Frown)

$$\frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2 \sum_{i=1}^N (y_i - \bar{y})^2}}$$

Average decision of Judge A

Average decision of Judge B

The diagram illustrates the formula for covariance. At the top, two boxes define 'Decision of Judge A' (1 for Frown and -1 for No-Frown) for both the numerator and denominator. Below the formula, two circles represent the average decisions: a red circle labeled \bar{x} for Judge A and a red circle labeled \bar{y} for Judge B. Arrows point from the text labels to these average circles. The formula itself is a fraction. The numerator is the sum from $i=1$ to N of the product of the deviation of x_i from \bar{x} and the deviation of y_i from \bar{y} . The denominator is the square root of the sum from $i=1$ to N of the square of the deviation of x_i from \bar{x} , multiplied by the sum from $i=1$ to N of the square of the deviation of y_i from \bar{y} .

Decision of Judge A
(1 for Frown and -1
for No-Frown)

Decision of Judge C
(1 for Frown and -1
for No-Frown)

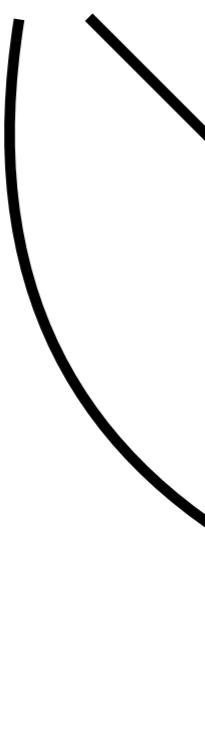
$$\frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2 \sum_{i=1}^N (y_i - \bar{y})^2}}$$

Average decision of
Judge C

Decision of Judge D
(1 for Frown and -1
for No-Frown)

Average decision of
Judge D

The level of
agreement is the
same but the
correlation is different



A diagram consisting of a vertical line with two curved arrows pointing downwards to two circular nodes. The top node contains the text $r_{AB} = -0.01$ and the bottom node contains the text $r_{CD} = 0.96$. Both nodes have a light brown background.

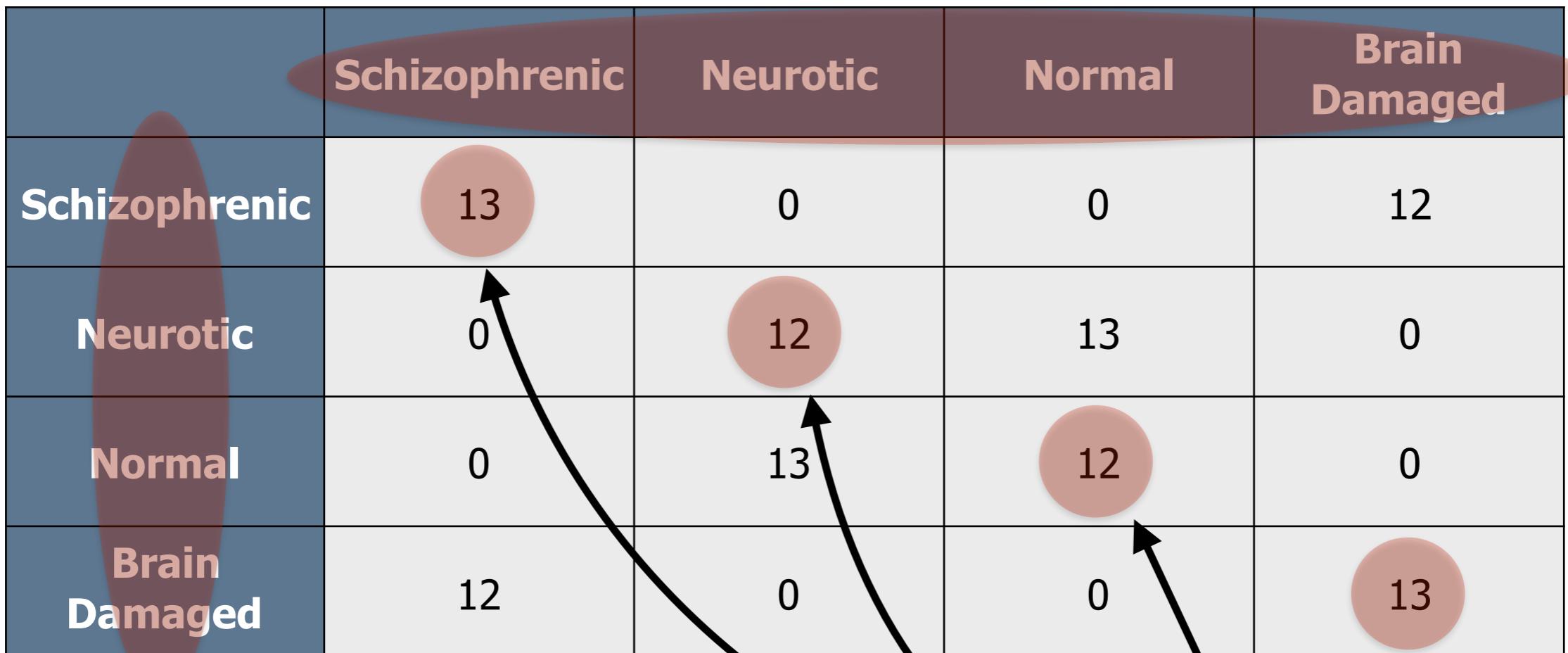
$$r_{AB} = -0.01$$
$$r_{CD} = 0.96$$

The correlation takes
into account the
variance in the
judgment

Limits

- The percentage of Agreement can be high simply because there is no variance in the judgments;
- If there is no variance, it is not possible to say what happens when there are different judgments;
- The same value of R can correspond to different values of correlation.

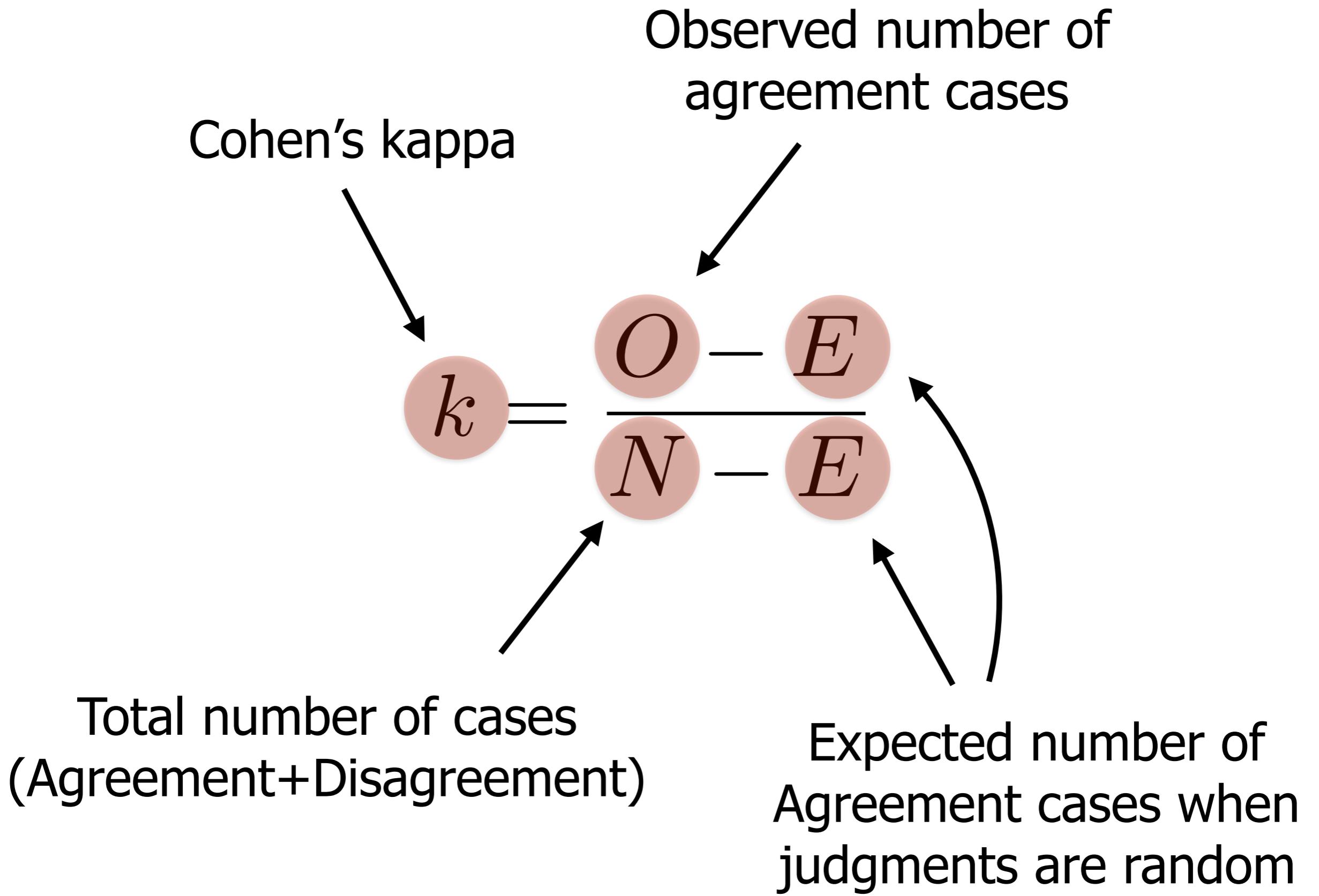
The decision of Judge A



	Schizophrenic	Neurotic	Normal	Brain Damaged
Schizophrenic	13	0	0	12
Neurotic	0	12	13	0
Normal	0	13	12	0
Brain Damaged	12	0	0	13

The decision of
Judge B

Agreement

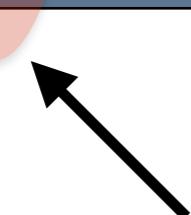


The decision of Judge A



Schizophrenic	Schizophrenic	Neurotic	Normal	Brain Damaged
Schizophrenic	13	0	0	12
Neurotic	0	12	13	0
Normal	0	13	12	0
Brain Damaged	12	0	0	13

The decision of
Judge B



Agreement

Expected number of
agreement cases in
cell “ii”

Marginal of Column
and Row “i”

$$E_{ii} = \frac{R_i}{N} \cdot \frac{C_i}{N} \cdot N = \frac{R_i C_i}{N}$$

Probability of falling
in Column “i” and
Row “i”

Total number of cases

$$k = \frac{O - E}{N - E} = \frac{50 - 25}{100 - 25} = 0.\bar{3}$$

The decision of Judge A

	Schizophrenic	Neurotic	Normal	Brain Damaged
Schizophrenic	13	0	0	12
Neurotic	0	12	13	0
Normal	0	13	12	0
Brain Damaged	12	0	0	13

The decision of
Judge B

$$k=0.04$$

Limits

- The value of k compares the observed agreement and the agreement expected when the judgments are random;
- When there are more than two categories (2×2 tables), it is not clear whether the kappa value applies equally to all of them.

The average correlation between judges

The correlation between judges "i" and "j"

$$r = \frac{2 \sum_{i=1}^N \sum_{j=i+1}^N r_{ij}}{N(N-1)}$$

The number of judges

The Effective (or
Spearman Brown)
Reliability

The average
correlation between
judges

$$R_{SB} = \frac{Nr}{1 + (N - 1) \cdot r}$$

The number of judges

The diagram illustrates the Spearman-Brown formula. It features a large circle containing the formula $R_{SB} = \frac{Nr}{1 + (N - 1) \cdot r}$. A curved arrow points from the left towards the 'Nr' term. Another curved arrow points from the right towards the 'r' term. Labels with arrows point to each part of the formula: 'The Effective (or Spearman Brown) Reliability' points to R_{SB} ; 'The average correlation between judges' points to r ; and 'The number of judges' points to Nr .

Limits

- The effective reliability provides an indication of a how much associated are the judgments of two random judges;
- It is an average value that does not say whether all judges are equally correlated with one another.

The Three Judges

Encoders	A	B	C	Total
1	5	6	7	18
2	3	6	4	13
3	3	4	6	13
4	2	2	3	7
5	1	4	4	9

$$S^2_{tot}$$

$$S^2_B$$

The variance of the scores for one judge

The variance of the total for each encoder

The Cronbach's alpha

$$\alpha = \frac{N}{N - 1} \left(\frac{S^2_{tot}}{S^2_{tot} - \sum_j S^2_j} \right)$$

The variance of the total for each encoder

The number of judges

Sum over all judges

$$-\sum_j S^2_j$$

$$S^2_{tot}$$

The variance of the scores for one judge

Limits

- It avoids the calculation of multiple correlations when the number of judges is high;
- It tends to give the same values as the other reliability measures considered in this lecture (it is affected by the same limitations).

Outline

- Introduction
- Reliability
- Conclusions

Conclusions

- Judgment studies allow one to answer questions on how inner states are expressed and perceived;
- Reliability measures are expected to quantify the extent to which multiple judgments agree with one another;
- It is not sufficient that multiple judges agree, they must have high mutual correlation.

Thank You!

Synthetic Impressions (I)

Computational Social Intelligence - Lecture 10

Prof. Alessandro Vinciarelli
School of Computing Science &
Institute of Neuroscience and Psychology

<http://www.dcs.gla.ac.uk/vincia>
Alessandro.Vinciarelli@glasgow.ac.uk



University
of Glasgow

EPSRC
Engineering and Physical Sciences
Research Council

FNSNF

This lecture is based on the following texts
(available on Moodle):

- Deshmukh, Craenen, Vinciarelli & Foster,
"Shaping Robot Gestures to Shape Users' Perception: The Effect of Amplitude and Speed on Godspeed Ratings", Proceedings of the International Conference on Human-Agent Interaction, 2018;

This lecture is based on the following texts
(available on Moodle):

- Craenen, Deshmukh, Foster & Vinciarelli,
“Shaping Gestures to Shape Personalities:
Interplay Between Gesture Parameters,
Attributed Personality Traits and Godspeed
Scores”, Proceedings of the IEEE International
Symposium on Robot and Human Interactive
Communication, 2018.

Outline

- Synthetic Impressions
- Gestures and Godspeed Scores
- Gestures and Personality
- Conclusions

Outline

- Synthetic Impressions
- Gestures and Godspeed Scores
- Gestures and Personality
- Conclusions

The Godspeed Questionnaire

“[...] standardised measurement tools for human robot interaction (HRI) [...] to compare the results from different studies [...] measurements of five key concepts in HRI.”

Bartneck et al., “Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots”, International Journal of Social Robotics, 1(1):71-81, 2009.

1. Anthropomorphism

“Anthropomorphism refers to the attribution of a human form, human characteristics, or human behaviour to nonhuman things such as robots, computers, and animals.”

Bartneck et al., “Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots”, International Journal of Social Robotics, 1(1):71-81, 2009.

2. Animacy

"The classic perception of life, which is often referred to as animacy, is based on [...] 'moving of one's own accord'"

Bartneck et al., "Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots", International Journal of Social Robotics, 1(1):71-81, 2009.

3.Likability

“[...] positive impressions [are] to some degree dependent on the visual and vocal behavior [...] and that positive first impressions (e.g., likeability) [...] often lead to more positive evaluations [...]”

Bartneck et al., “Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots”, International Journal of Social Robotics, 1(1):71-81, 2009.

4. Perceived Intelligence

“[...] perceived intelligence of a robot will depend on its competence. To monitor the progress being made in robotic intelligence it is important to have a good measurement tool.”

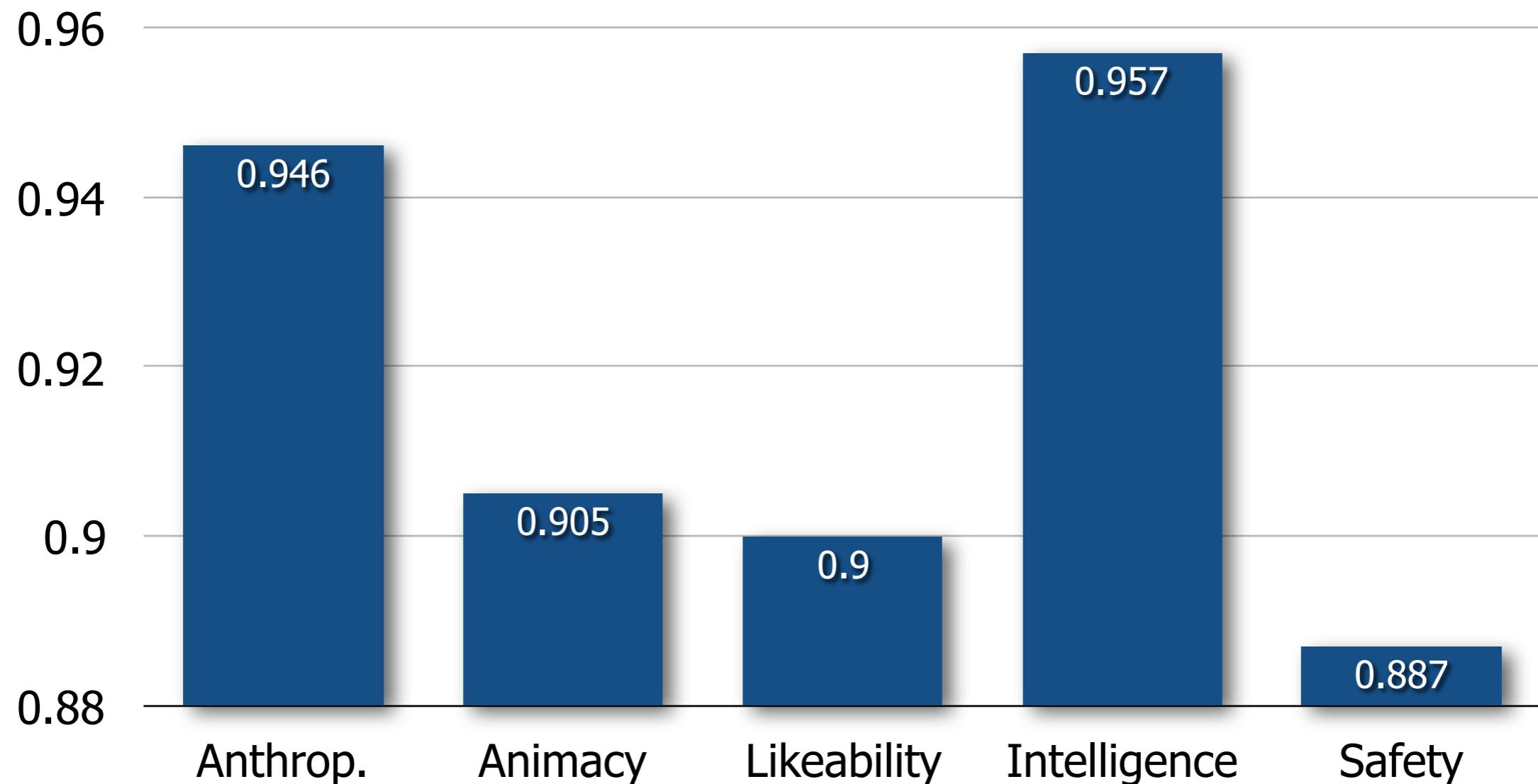
Bartneck et al., “Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots”, International Journal of Social Robotics, 1(1):71-81, 2009.

5. Perceived Safety

“Perceived safety describes the user’s perception of the level of danger when interacting with a robot, and the user’s level of comfort during the interaction.”

Bartneck et al., “Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots”, International Journal of Social Robotics, 1(1):71-81, 2009.

Reliability



Deshmukh, Craenen, Vinciarelli & Foster, "Shaping Robot Gestures to Shape Users' Perception: The Effect of Amplitude and Speed on Godspeed Ratings", Proc. of the International Conference on Human-Agent Interaction, 2018

Outline

- Synthetic Impressions
- Gestures and Godspeed Scores
- Gestures and Personality
- Conclusions

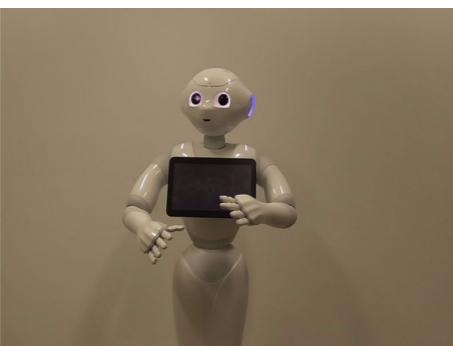
Why Gestures?

“[Gestures] often are used to communicate when distance or noise renders vocal communication impossible [...] expressing concepts that also are expressed verbally.”

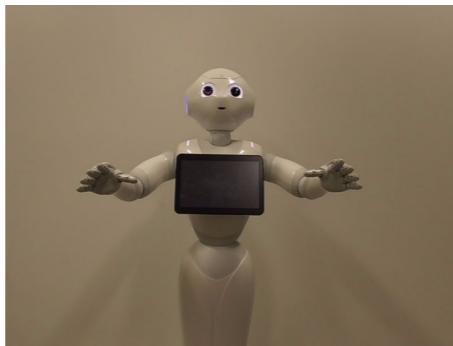
Krauss et al., “Lexical Gestures and Lexical Access: a process model”, in “Language and Gesture”, McNeill (ed.), Cambridge University Press, 2000

The Gestural Stimuli

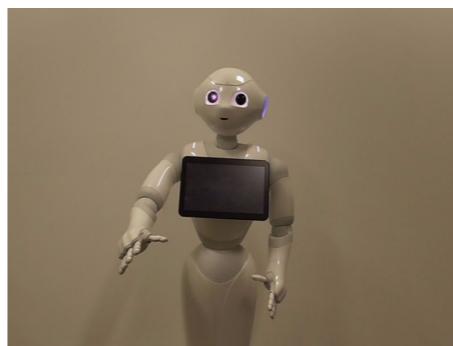
Disengage



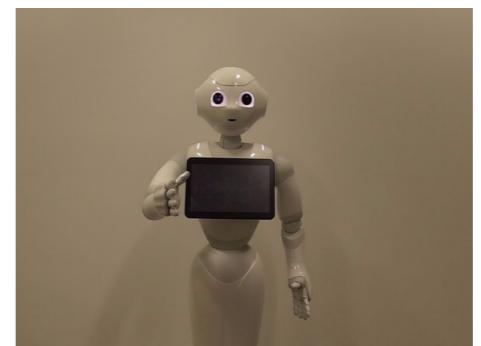
Engage



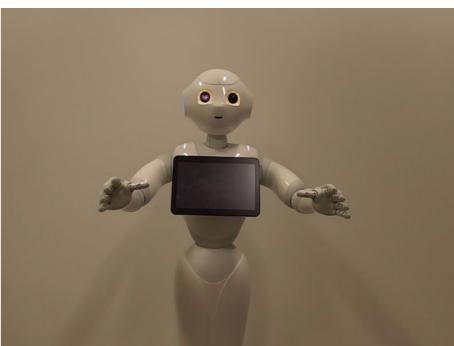
Pointing



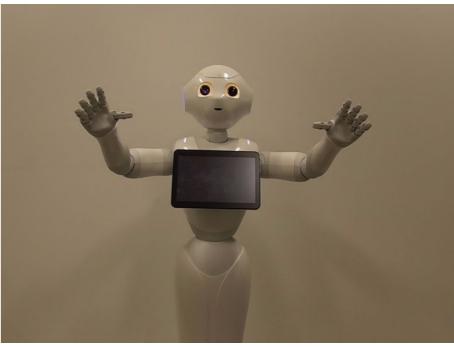
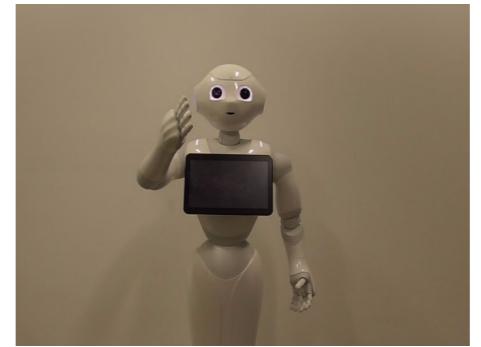
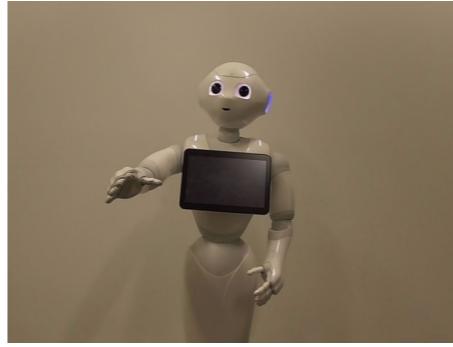
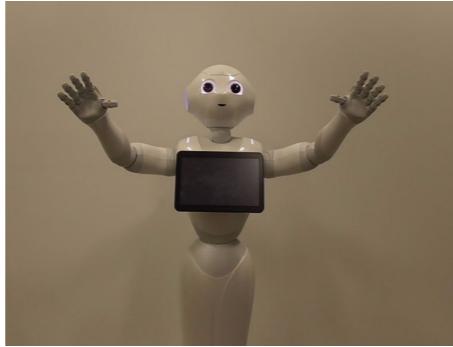
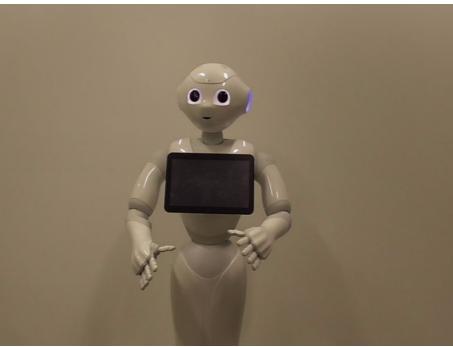
Head
Touching



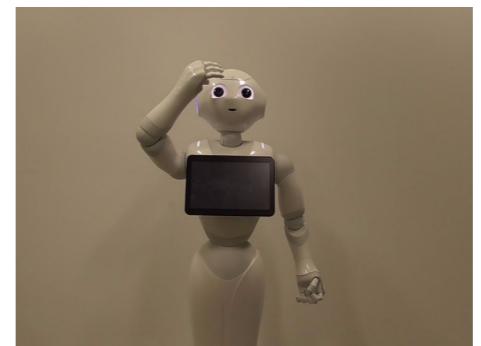
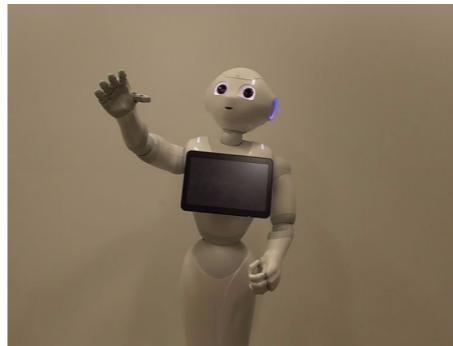
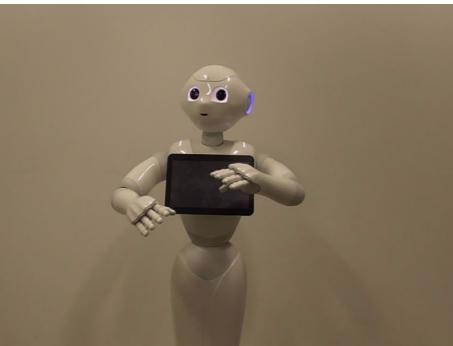
Cheering



$\alpha = 0.50$



$\alpha = 0.75$

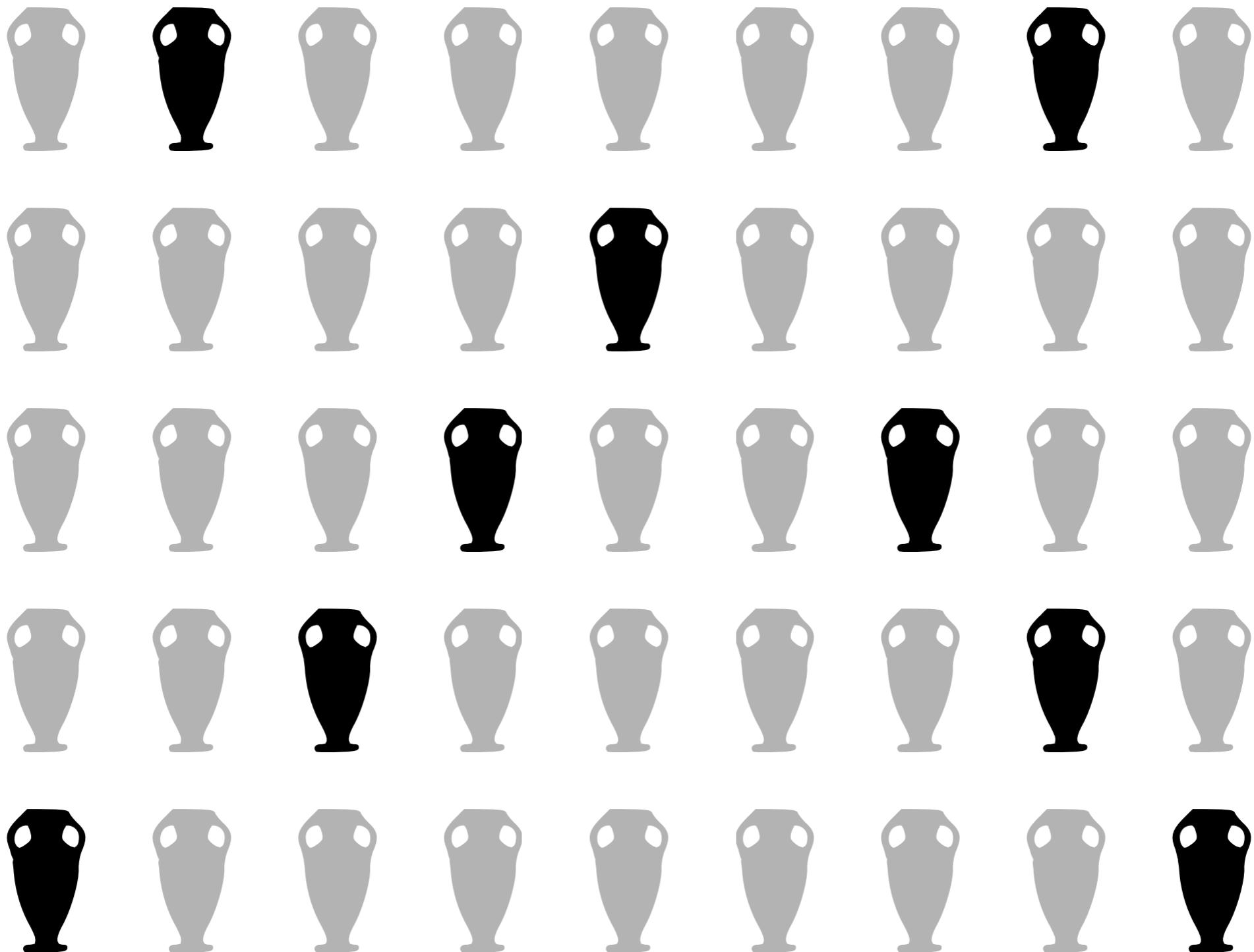


$\alpha = 1.00$

The Setting

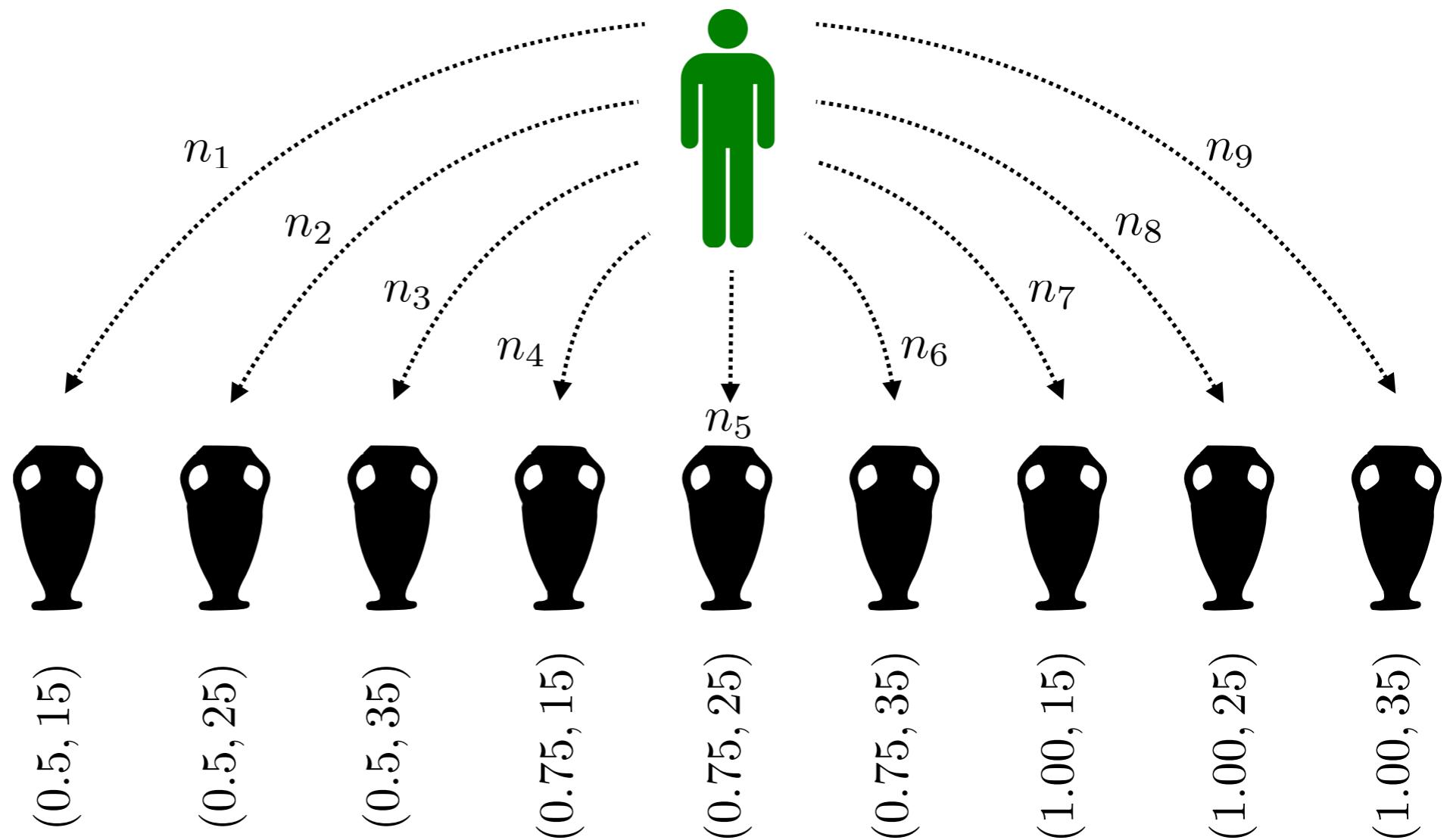


30 observers have filled Godspeed questionnaire and Big-Five Inventory 10 (self and attributed) while rating different interpretations for all 45 stimuli.



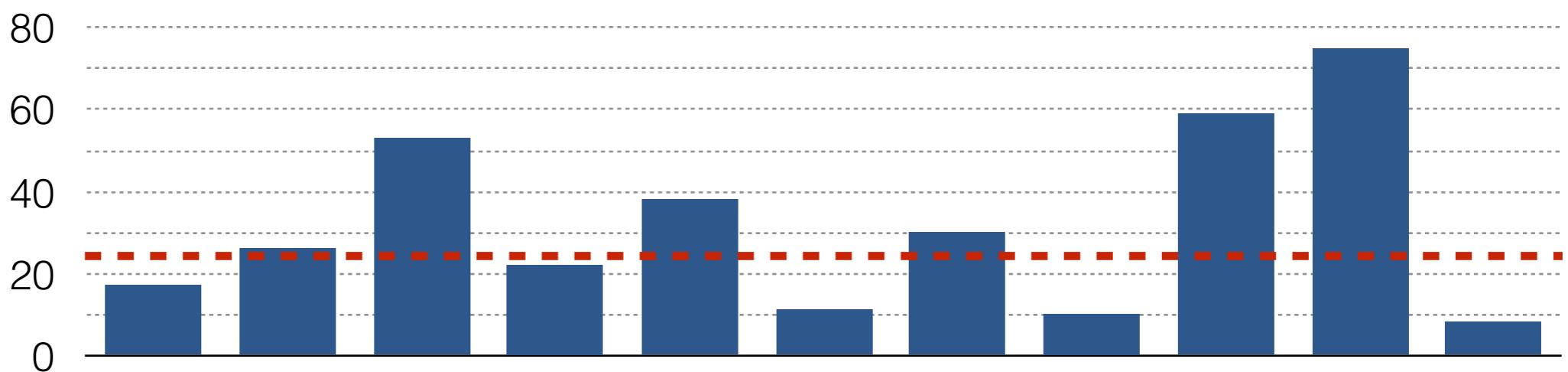
The stimuli are like urns where the observers can throw a number of votes (the black urns are to the 9 variants of a core gesture)

One observer



The 9 versions of
a core gesture
for one dimension

After all 30
observers
have rated



A matrix for a specific stimulus (speed and amplitude)

$$S^{(\alpha, \lambda)}$$

An element is the score of observer "i" for GS dimension "k"

$$\{s_{ik}^{(\alpha, \lambda)}\}$$

$$c_j^{(\alpha, \lambda)}$$

Total number of points along GS dimension "j" for one stimulus

$$c_j^{(\alpha, \lambda)} = \sum_{i=1}^N s_{ij}^{(\alpha, \lambda)}$$

Sum over the elements of column "j" of the matrix

The total number of points along GS dimension “j”

$$T_j = \sum \sum c_j^{(\alpha, \lambda)}$$

A large black arrow points downwards from the explanatory text above to the mathematical equation below. Below the equation, two smaller black arrows point upwards from two circular nodes containing the symbols α and λ respectively, towards their corresponding variables in the term $c_j^{(\alpha, \lambda)}$.

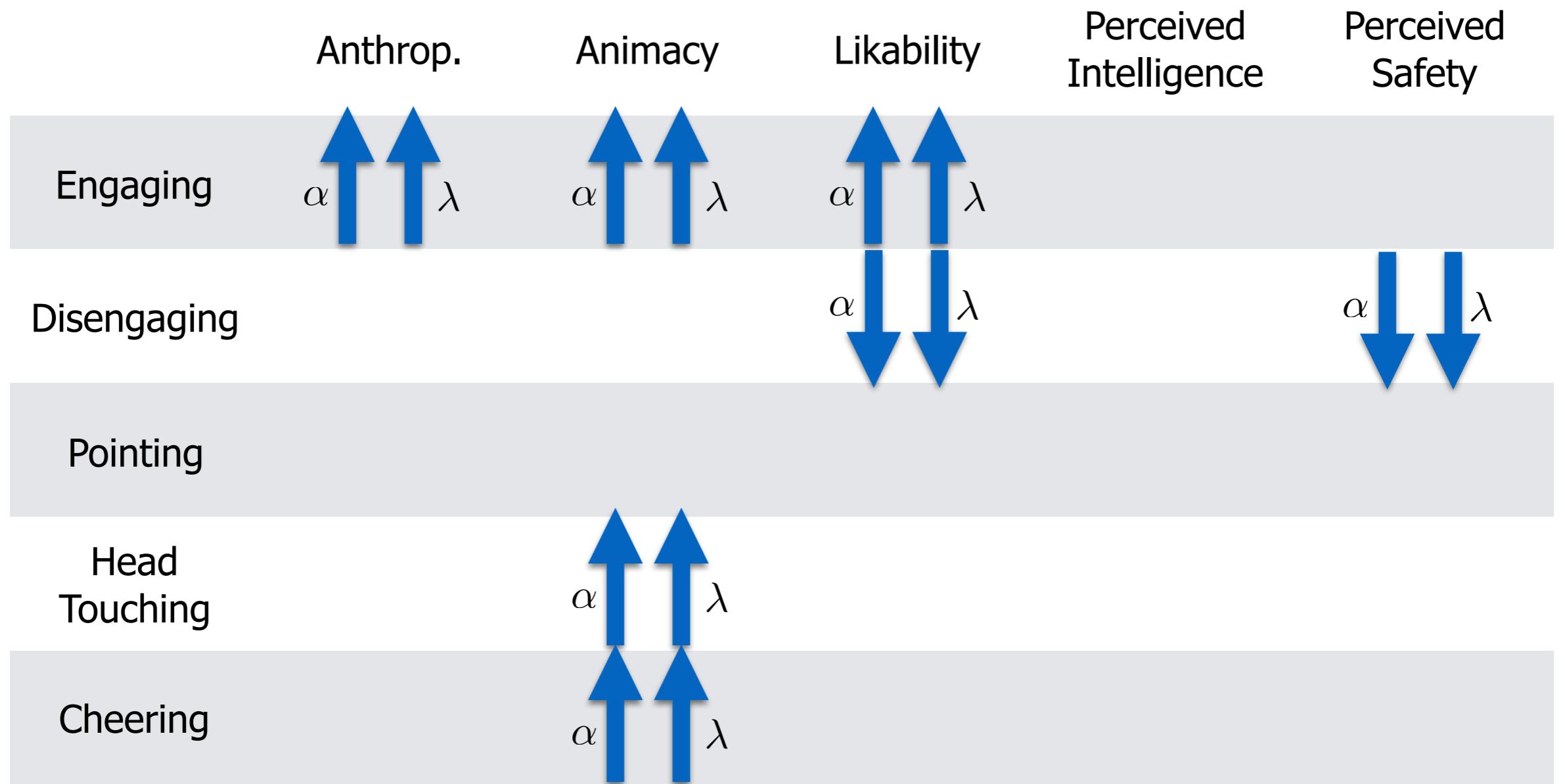
Sum over all values of amplitude and speed

$$\chi^2 = \sum_{\alpha} \sum_{\lambda} \frac{(c_j^{(\alpha, \lambda)} - E)^2}{E}$$

$$E = \frac{1}{9} T_j$$

Chi Square variable for testing whether the distribution of the points across the 9 variants of the same core gesture is uniform

Average over all variants of the same core gesture



Significant effects after False Discovery Rate Correction (arrows pointing upwards account for positive relationships and vice versa)

Recap

- There is a relationship between the “shape” of a gesture (amplitude and speed) and the perception of the users (Godspeed scores);
- Animacy and Likability are the dimensions along which there is more interaction;
- The core gesture “Pointing” does not show any interaction between shape and perception.

Outline

- Synthetic Impressions
- Gestures and Godspeed Scores
- **Gestures and Personality**
- Conclusions

The Big Five Traits

“The Big Five Personality Factors appear to provide a set of highly replicable dimensions that parsimoniously and comprehensively describe most phenotypic individual differences”

Saucier, Goldberg, “The Language of Personality: Lexical Perspectives on the Five-Factor Model”, in “The Five-Factor Model of Personality”, Wiggins (ed.), 21-50, 1996

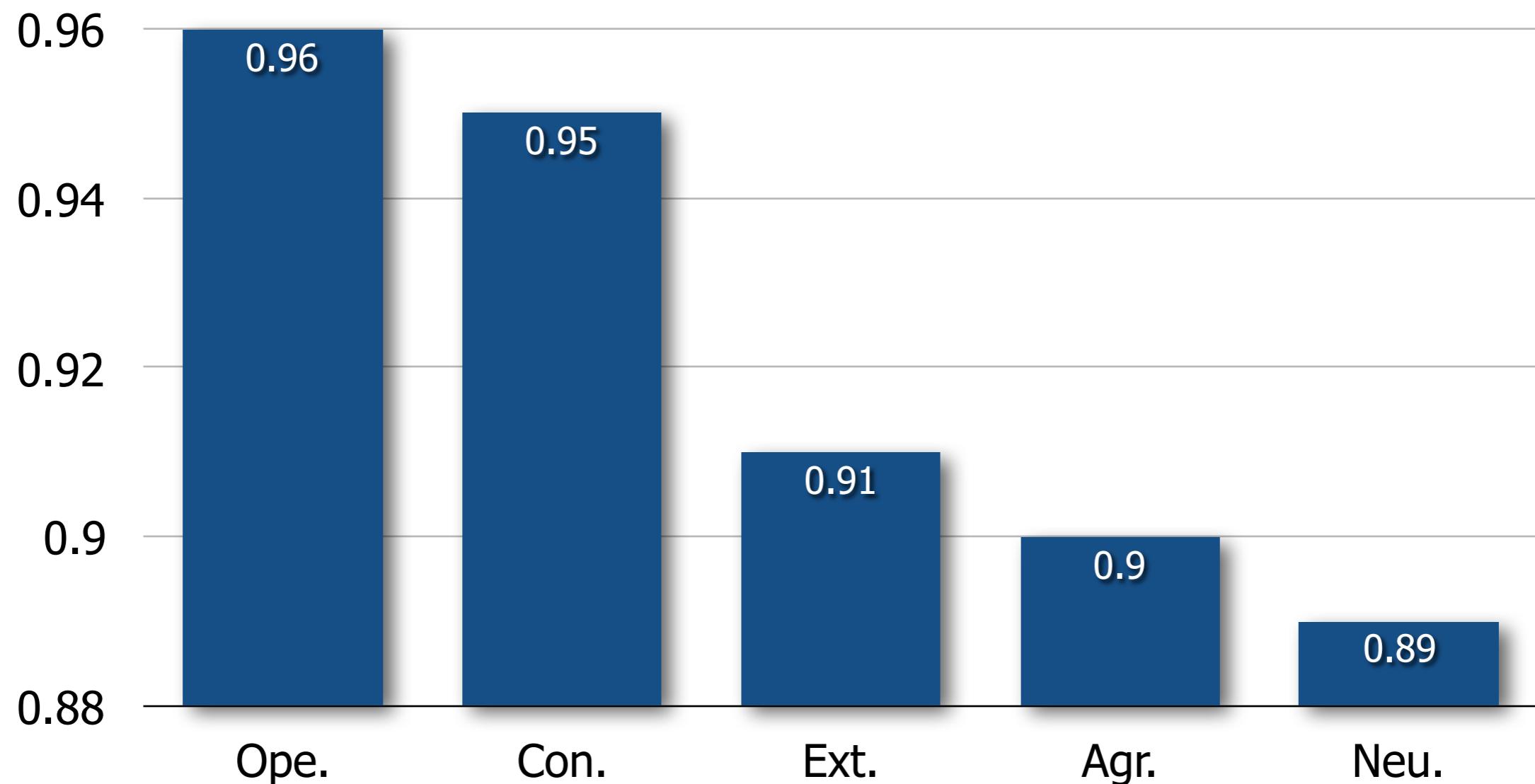
- **Extraversion**: Active, Assertive, Energetic, Outgoing;
- **Agreeableness**: Appreciative, Forgiving, Generous, Kind, Sympathetic, Trusting;
- **Conscientiousness**: Efficient, Organised, Planful, Reliable, Responsible, Thorough;
- **Neuroticism**: Anxious, Self-pitying, Tense, Touchy, Unstable, Worrying;
- **Openness**: Artistic, Curious, Imaginative, Insightful;

Saucier, Goldberg, "The Language of Personality: Lexical Perspectives on the Five-Factor Model", in "The Five-Factor Model of Personality", Wiggins (ed.), 21-50, 1996

This robot is reserved	E	-
This robot is generally trusting	A	+
This robot tends to be lazy	C	-
This robot is relaxed, handles stress well	N	-
This robot has few artistic interests	O	-
This robot is outgoing, sociable	E	+
This robot tends to find faults with others	N	+
This robot does a thorough job	C	+
This robot gets nervous easily	A	-
This robot has an active imagination	O	+

Rammstedt and John, "Measuring Personality in One Minute or Less: A 10-item short version of the BFI", Journal of Research in Personality, 41(1):203-212, 2007

Reliability



Craenen, Deshmukh, Foster & Vinciarelli, "Shaping Gestures to Shape Personalities: Interplay Between Gesture Parameters, Attributed Personality Traits and Godspeed Scores", Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication, 2018.

A matrix for a specific stimulus (speed and amplitude)

An element is the score of observer "i" for B5 trait "k"

$$A^{(\alpha, \lambda)} = \{a_{ik}^{(\alpha, \lambda)}\}$$

$$t_j^{(\alpha, \lambda)} = \sum_{i=1}^N a_{ij}^{(\alpha, \lambda)}$$

Total number of points along B5 trait "j" for one stimulus

Sum over the elements of column "j" of the matrix

The total number of points along B5 trait “j”

$$T_j = \sum \sum t_j^{(\alpha, \lambda)}$$

A large black arrow points down to the equation $T_j = \sum \sum t_j^{(\alpha, \lambda)}$. Below the equation, two smaller arrows point from circles containing the symbols α and λ to the corresponding variables in the equation.

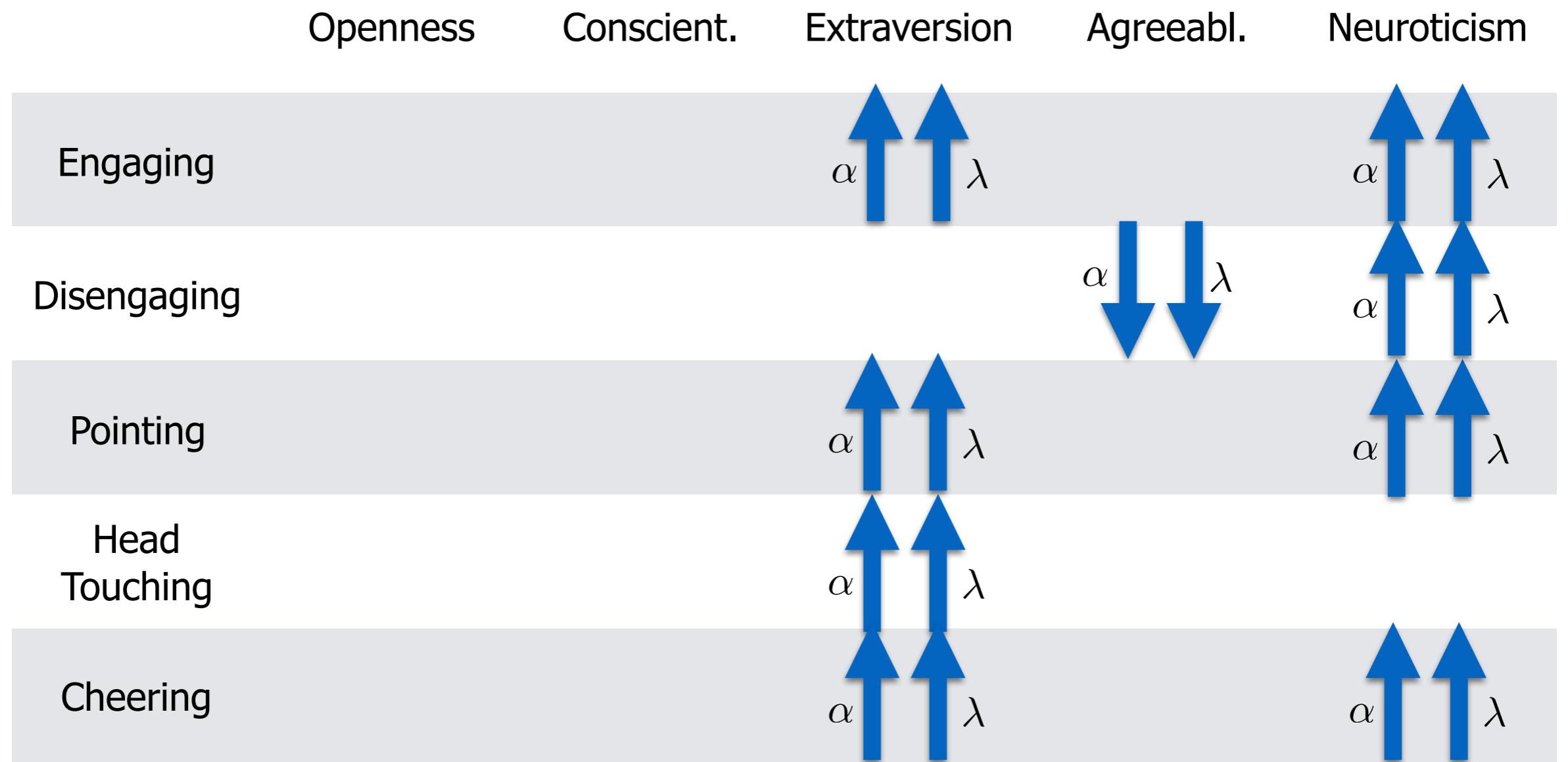
Sum over all values of amplitude and speed

$$\chi^2 = \sum_{\alpha} \sum_{\lambda} \frac{(t_j^{(\alpha, \lambda)} - E)^2}{E}$$

$$E = \frac{1}{9} T_j$$

Chi Square variable for testing whether the distribution of the points across the 9 variants of the same core gesture is uniform

Average over all variants of the same core gesture



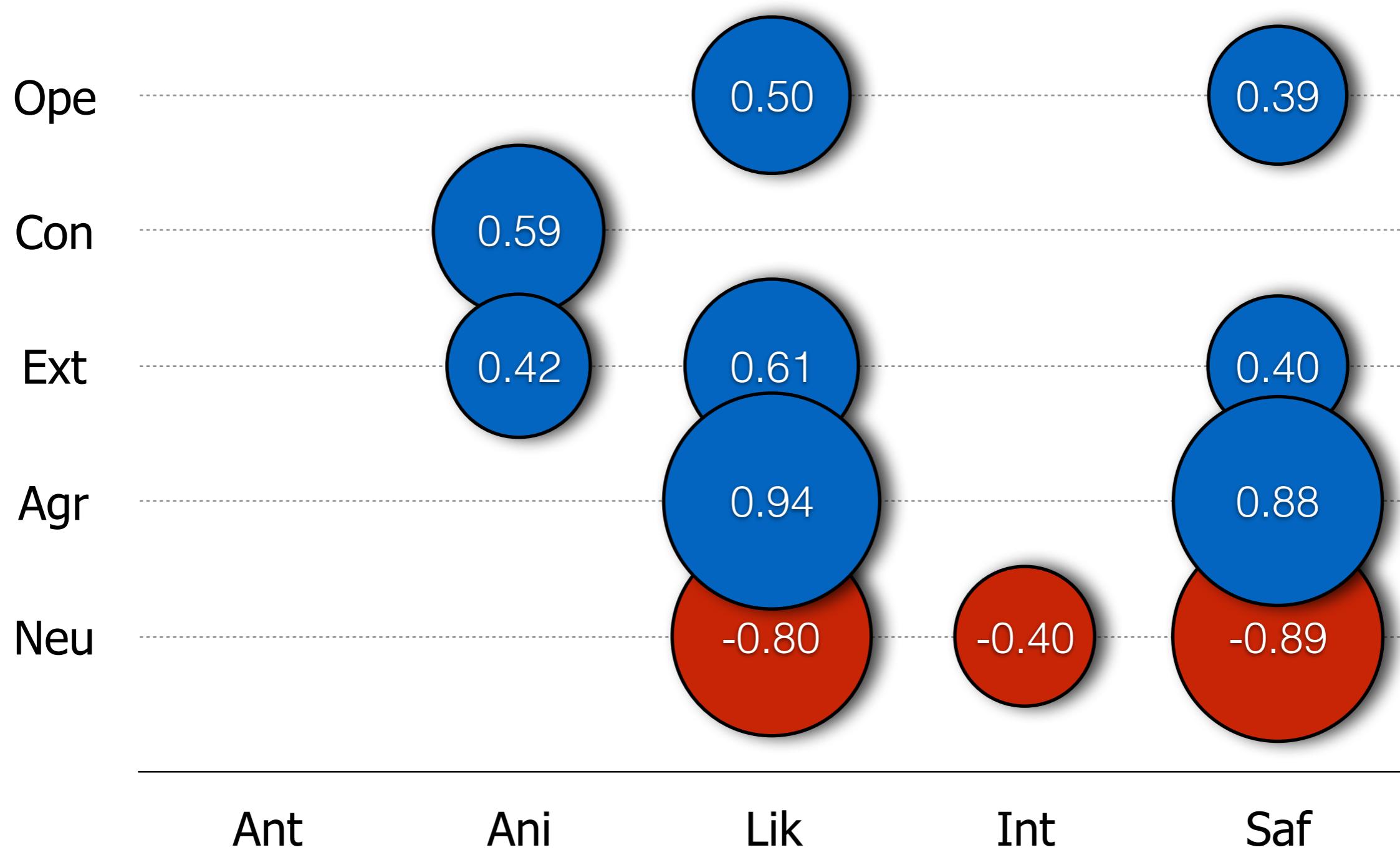
Significant effects after False Discovery Rate Correction (arrows pointing upwards account for positive relationships and vice versa)

The Spearman Correlation Coefficient

Difference between rank of trait and rank of GS score for the same stimulus

$$r = 1 - \frac{6 \sum_{k=1}^M d(t_k, g_k)}{M(M^2 - 1)}$$

The Spearman Correlation Coefficient is more robust to outliers than the most common Pearson Correlation



Relationship between Godspeed scores and Big Five traits (effects observed after application of the False Discovery Rate Correction)

Recap

- There is a relationship between the “shape” of a gesture (amplitude and speed) and the Big Five traits attributed by the users;
- There is a significant interplay between Godspeed Scores and Big-Five Traits;
- It is possible to change the perception of the users by changing the personality impressions that the robots convey.

Outline

- Synthetic Impressions
- Gestures and Godspeed Scores
- Gestures and Personality
- Conclusions

Conclusions

- However simple, gestures give rise to a wide spectrum of synthetic impressions;
- Overall, the impressions appear to follow principles and laws observed in human-human interactions;
- The next step is the collection of data in real-world settings.

Thank You!

Special thanks to:

- Bart Craenen
- Amol Deshmukh
- Mary Ellen Foster

Synthetic Impressions (II)

Computational Social Intelligence - Lecture 11

Prof. Alessandro Vinciarelli
School of Computing Science &
Institute of Neuroscience and Psychology

<http://www.dcs.gla.ac.uk/vincia>
Alessandro.Vinciarelli@glasgow.ac.uk



University
of Glasgow

EPSRC
Engineering and Physical Sciences
Research Council

FNSNF

This lecture is based on the following texts
(available on Moodle):

- Craenen, Deshmukh, Foster & Vinciarelli, "Do We Really Like Robots that Match our Personality? The Case of Big-Five Traits, Godspeed Scores and Robotic Gestures", Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication, 2018.

This lecture is based on the following texts
(available on Moodle):

- Deshmukh, Craenen, Foster & Vinciarelli, "The More I Understand it, the Less I Like it: The Relationship Between Understandability and Godspeed Scores for Robotic Gestures", Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication, 2018.

Outline

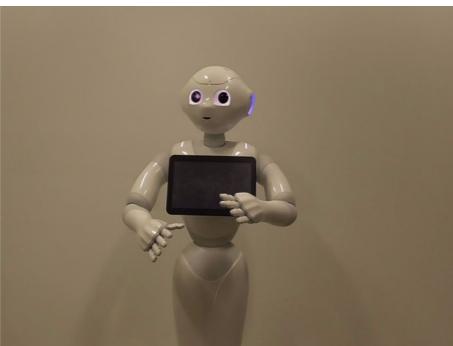
- Synthetic Impressions
- Gestures and the Attraction Paradigm
- Gestures and Understandability
- Conclusions

Outline

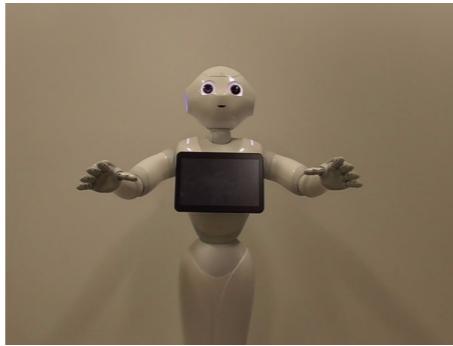
- Synthetic Impressions
- Gestures and the Attraction Paradigm
- Gestures and Understandability
- Conclusions

The Gestural Stimuli

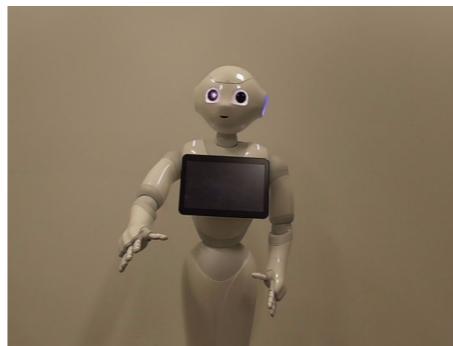
Disengage



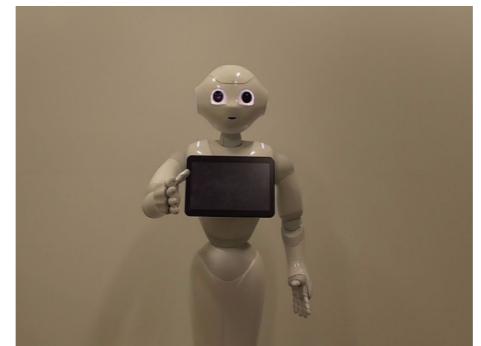
Engage



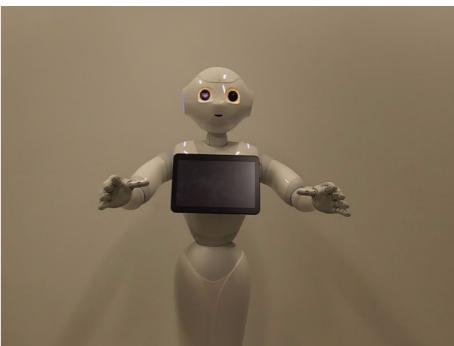
Pointing



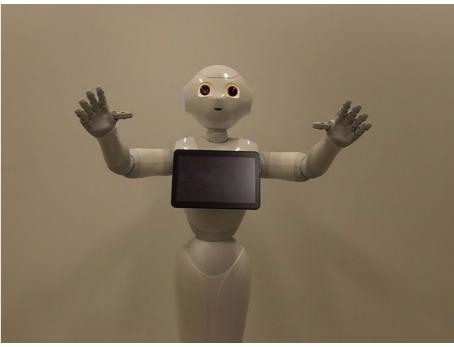
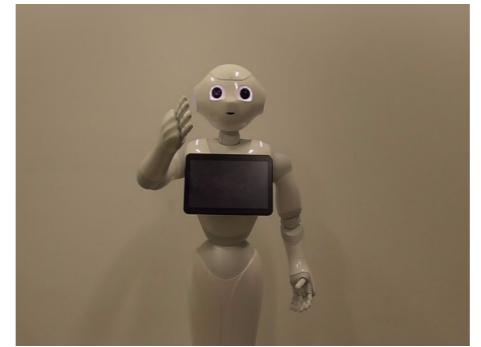
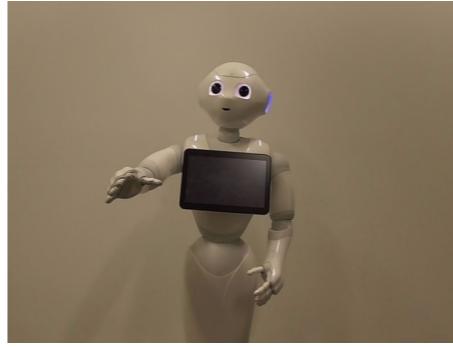
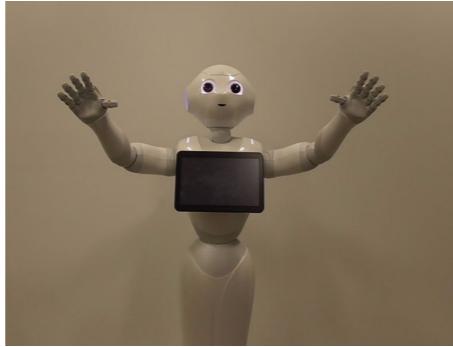
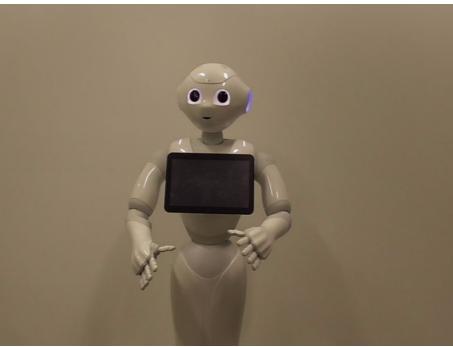
Head
Touching



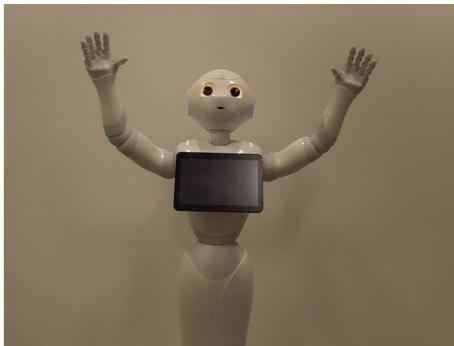
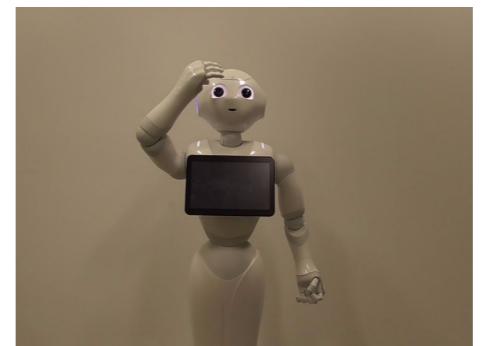
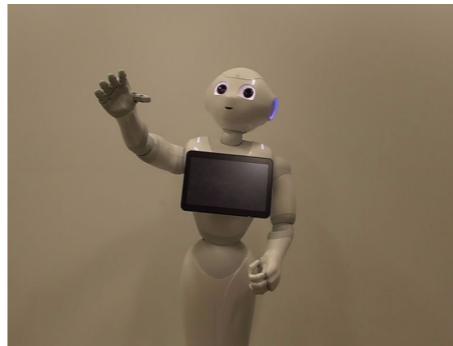
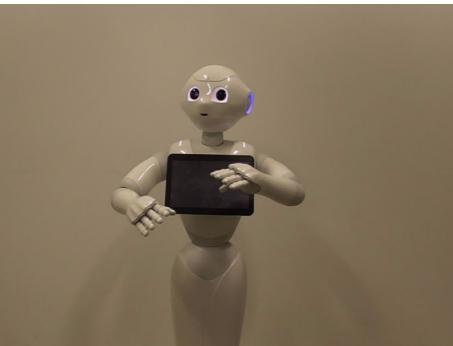
Cheering



$\alpha = 0.50$



$\alpha = 1.00$



The Setting



30 observers have filled Godspeed questionnaire and Big-Five Inventory 10 (self and attributed) while rating different interpretations for all 45 stimuli.

Outline

- Synthetic Impressions
- Gestures and the Attraction Paradigm
- Gestures and Understandability
- Conclusions

The Attraction Paradigm (I)

“[...] evaluations of strangers appear to be affected by degree of similarity, and statistical analysis confirms this impression.”

Byrne, “An Overview (and Underview) of Research and Theory Within the Attraction Paradigm”, Journal of Social and Personal Relationships, 14(3):417-431, 1997.

The Attraction Paradigm (II)

“[...] perceived similarity predicted attraction in no-interaction, short-interaction, and existing relationship studies.”

Montoya, Horton & Kirchner, “Is Actual Similarity Necessary for Attraction? A Meta-Analysis of Actual and Perceived Similarity”, Journal of Social and Personal Relationships, 25(6):899-922, 2008.

Distance between
self-assessed and
attributed traits for
stimulus "k"

"T" is the total
number of traits

$$d_k = \left[\sum_{j=1}^T (t_j^{(s)} - t_{jk}^{(a)})^2 \right]^{\frac{1}{2}}$$

The expression
corresponds to one of
the observers

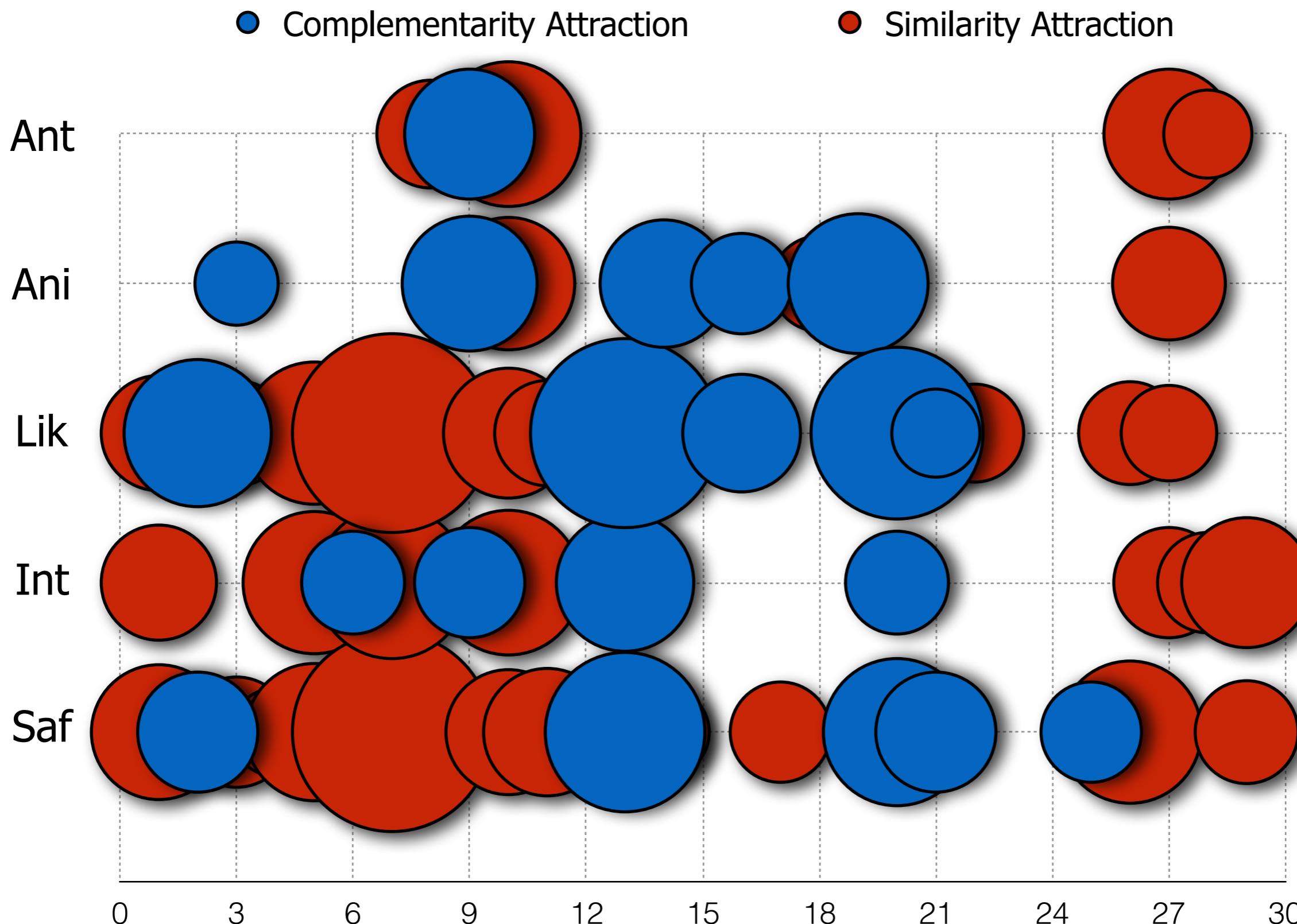
Trait "j" attributed to
stimulus "k"

The Spearman Correlation Coefficient

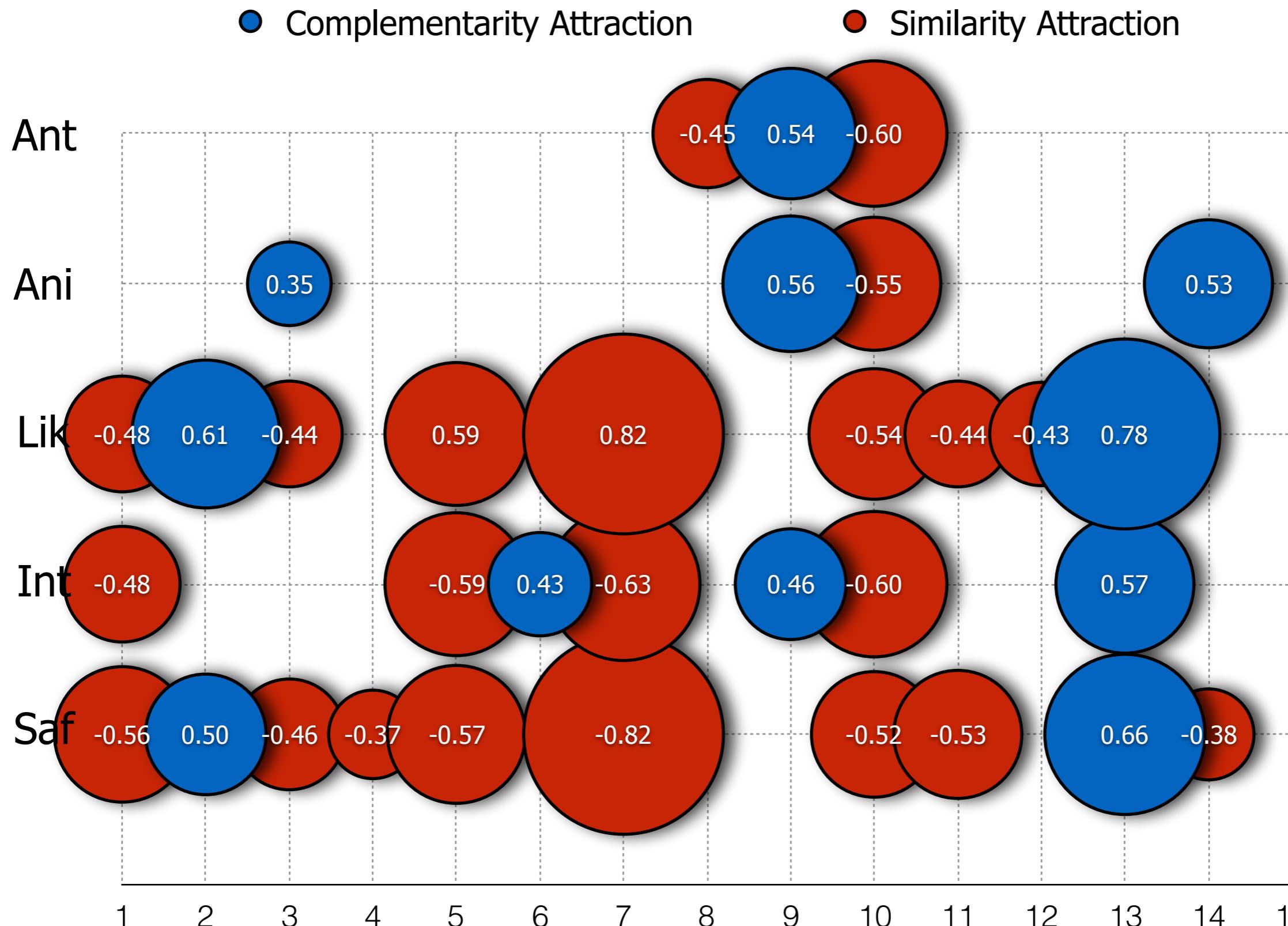
Difference between rank of distance and rank of GS score for the same stimulus

$$r = 1 - \frac{6 \sum_{k=1}^M d(d_k, g_k)}{M(M^2 - 1)}$$

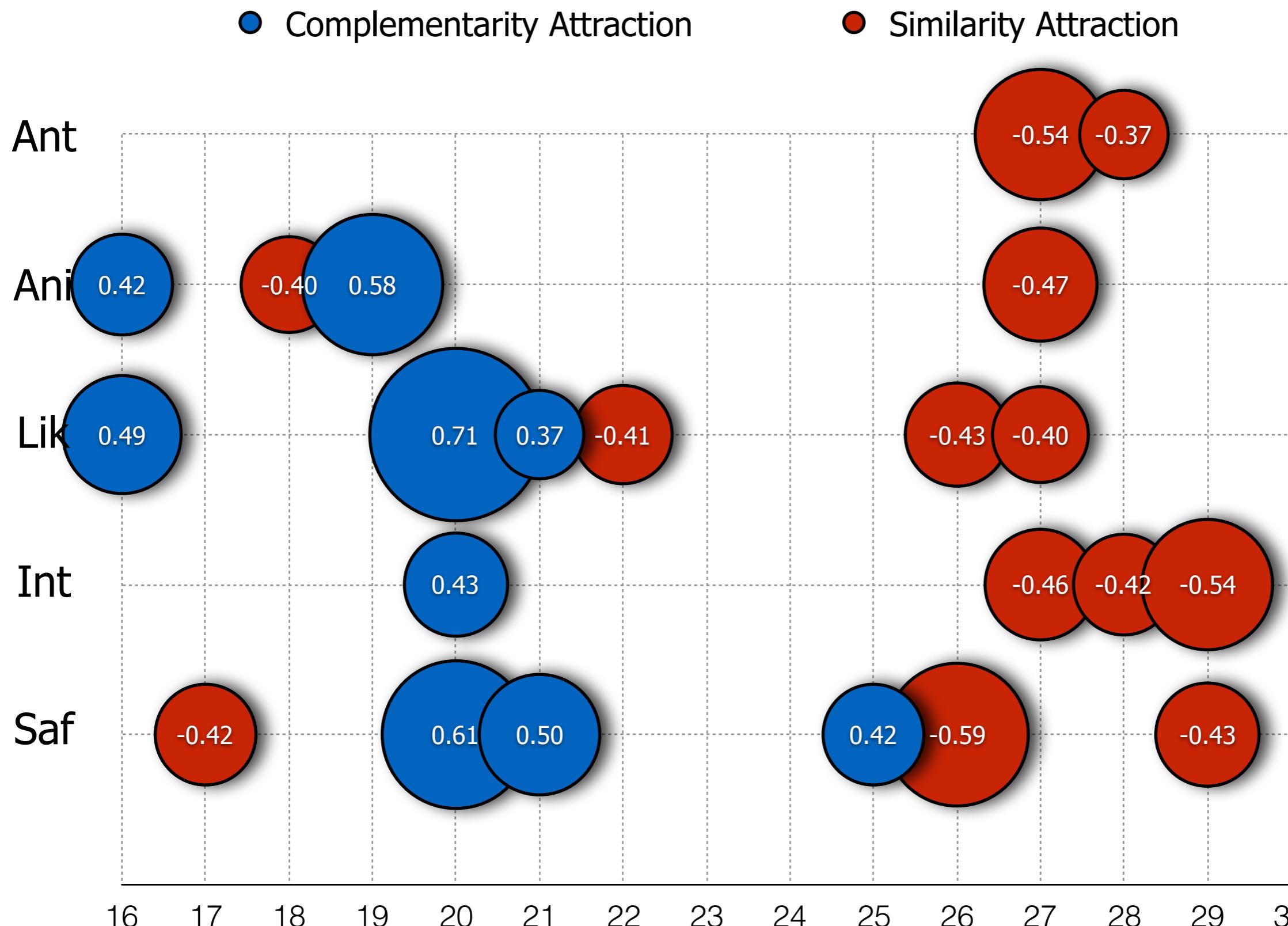
The Spearman Correlation Coefficient is more robust to outliers than the most common Pearson Correlation



Relationship between Godspeed scores and Big Five traits (effects observed after application of the False Discovery Rate Correction)



Relationship between Godspeed scores and Big Five traits (effects observed after application of the False Discovery Rate Correction)



Relationship between Godspeed scores and Big Five traits (effects observed after application of the False Discovery Rate Correction)

Recap

- The attraction paradigm appears to apply, to a large extent, to Human-Robot Interaction;
- Out of 30 observers, 16 show similarity-attraction, 9 show complementarity-attraction, and 2 show mixed effects;
- The attraction paradigm can be exploited effectively only if it is possible to understand which of the effects is taking place.

Outline

- Synthetic Impressions
- Gestures and the Attraction Paradigm
- **Gestures and Understandability**
- Conclusions

Emblems

“[gestures] steadily linked to a meaning, so that the two make a signal-meaning pair [...] like it happens, for instance, with the lexical items of a verbal lexicon”

Poggi, “Mind, hands, face and body. A goal and belief view of multimodal communication”, Weidler 2007

Interpretation

The observers have been asked to rate 10 possible interpretations of every gesture:

Getting Distracted; Aggressing; Flirting; Pointing; Complaining; Cheering; Reflecting; Teasing; Rejecting; and Welcoming.

A matrix for a specific stimulus (speed and amplitude)

An element is the score of observer "i" for interpretation "k"

$$M^{(\alpha, \lambda)} = \{m_{ik}^{(\alpha, \lambda)}\}$$

$$u_j^{(\alpha, \lambda)} = \sum_{i=1}^N m_{ij}^{(\alpha, \lambda)}$$

Total number of points for interpretation "j" for one stimulus

Sum over the elements of column "j" of the matrix

Probability of one interpretation being voted

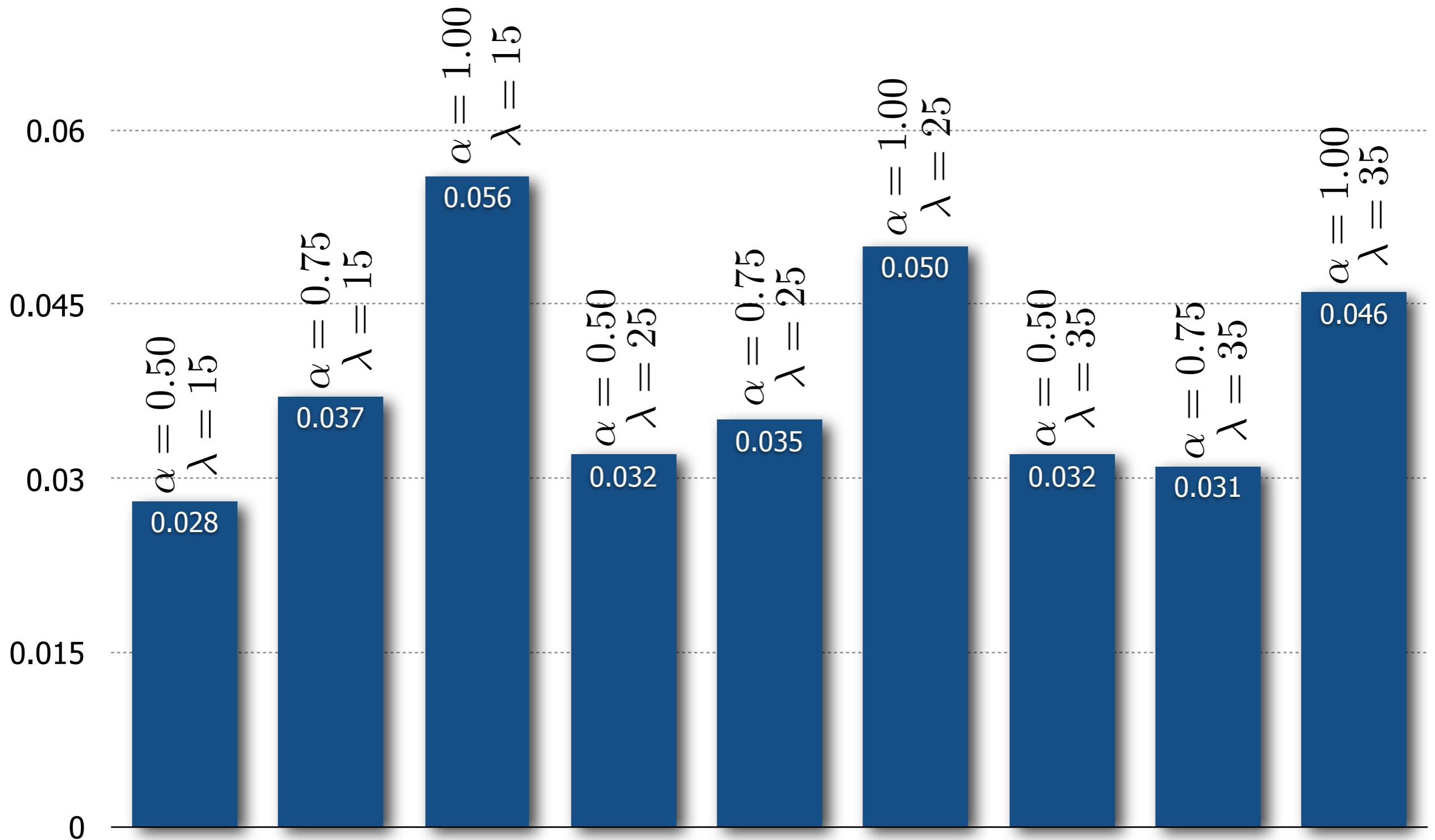
$$p_k = \frac{u_k}{\sum_{i=1}^N \sum_{j=1}^T m_{ij}}$$

Sum over all elements of matrix "M"

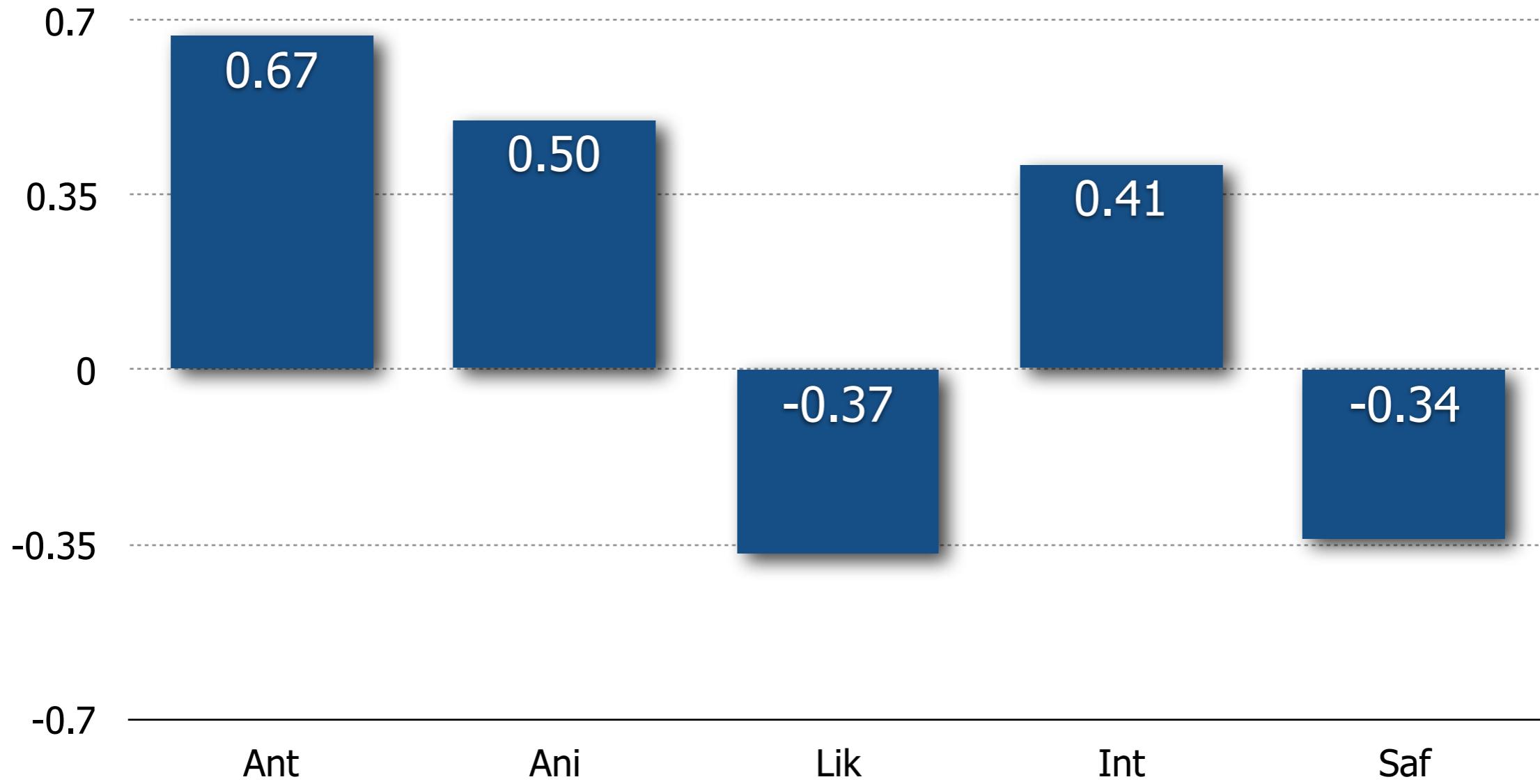
$$U = 1 - \frac{- \sum_{j=1}^T p_j \log p_j}{\log T}$$

Understandability
(high when interpretation only attracts many votes)

Entropy (measuring how uniform the distribution is)



The Understandability is higher when there is no dampening.



Correlations between Understandability and Godspeed Scores (all values are statistically significant)

Recap

- There is an association between changes in amplitude and changes in understandability;
- There is a statistically significant correlation between understandability and Godspeed scores;
- The interplay appears to reproduce the incompatibility between social and task skills.

Outline

- Synthetic Impressions
- Gestures and Godspeed Scores
- Gestures and Personality
- Gestures and the Attraction Paradigm
- Gestures and Understandability
- Conclusions

Conclusions

- However simple, gestures give rise to a wide spectrum of synthetic impressions;
- Overall, the impressions appear to follow principles and laws observed in human-human interactions;
- The next step is the collection of data in real-world settings.

Thank You!

Special thanks to:

- Bart Craenen
- Amol Deshmukh
- Mary Ellen Foster

Bayesian Decision Theory

Computational Social Intelligence - Lecture 12

Prof. Alessandro Vinciarelli
School of Computing Science &
Institute of Neuroscience and Psychology

<http://www.dcs.gla.ac.uk/vincia>
Alessandro.Vinciarelli@glasgow.ac.uk



University
of Glasgow

EPSRC
Engineering and Physical Sciences
Research Council

FNSNF

Texts (see Moodle)

This lecture is based on the following text
(available on Moodle):

- Chapter 5 of F.Camastra and A.Vinciarelli,
“Machine Learning for Audio, Image and Video
Processing”, Springer Verlag, 2008.

Outline

- Introduction
- Prior Rule
- Posterior Rule
- Conclusions

Outline

- Introduction
- Prior Rule
- Posterior Rule
- Conclusions

Thomas Bayes (1701-1761)



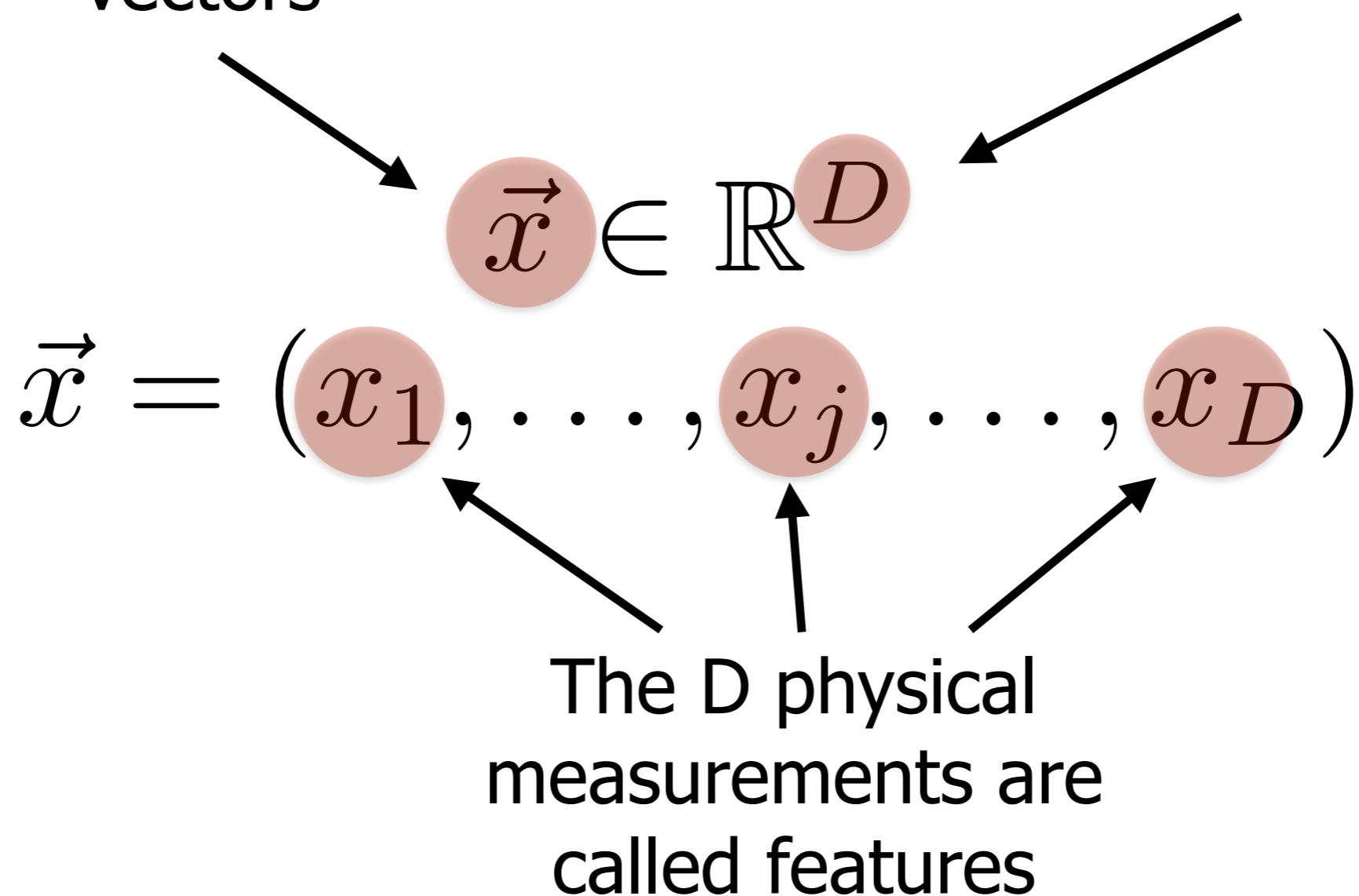
Hence the name Bayesian Decision Theory

Bayesian Decision Theory

- Bayesian Decision Theory is a statistical approach for the formalisation of common sense;
- It is one of the main approaches that Artificial Intelligence technologies adopt to “make decisions”;
- Bayesian Decision Theory has nothing to do with the way humans make decisions.

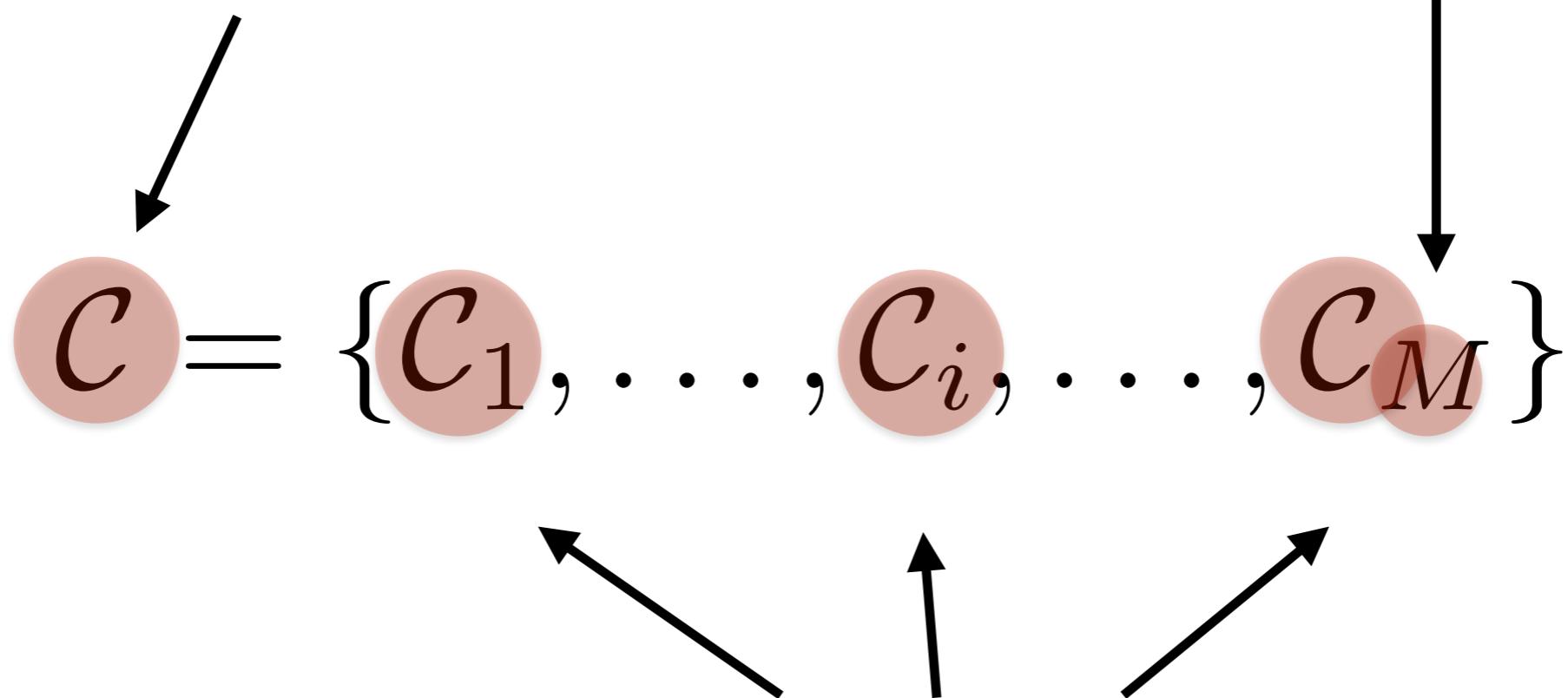
The Feature Vectors

The data is represented with D-dimensional vectors



The Decisions

The set C includes M possible decisions



The M decision are called classes

AI Decisions

In AI, decision means to map a feature vector into a decision

$$\vec{x} \rightarrow C_j$$
$$C_j \in \mathcal{C}$$

The decision must belong to the set of the M possible decisions

Toy Example (Logic)

The vector represents a student

$$\vec{x} = (AC, Ex)$$
$$C = \{passed, failed\}$$

The decisions is easy because there is a rule (e.g., both AC and Ex above C3)

AC and Ex are Assessed Exercise and Exam scores, respectively

The machine must decide whether it is passed or failed

Toy Example (AI)

The vector represents a patient

$$\vec{x} = (T, P)$$
$$C = \{ill, healthy\}$$

T and P are temperature and blood pressure, respectively

The difficulty is that there is uncertainty, it is not possible to solve the problem with a rule

The machine must decide whether a person is ill or healthy

AI and Decisions

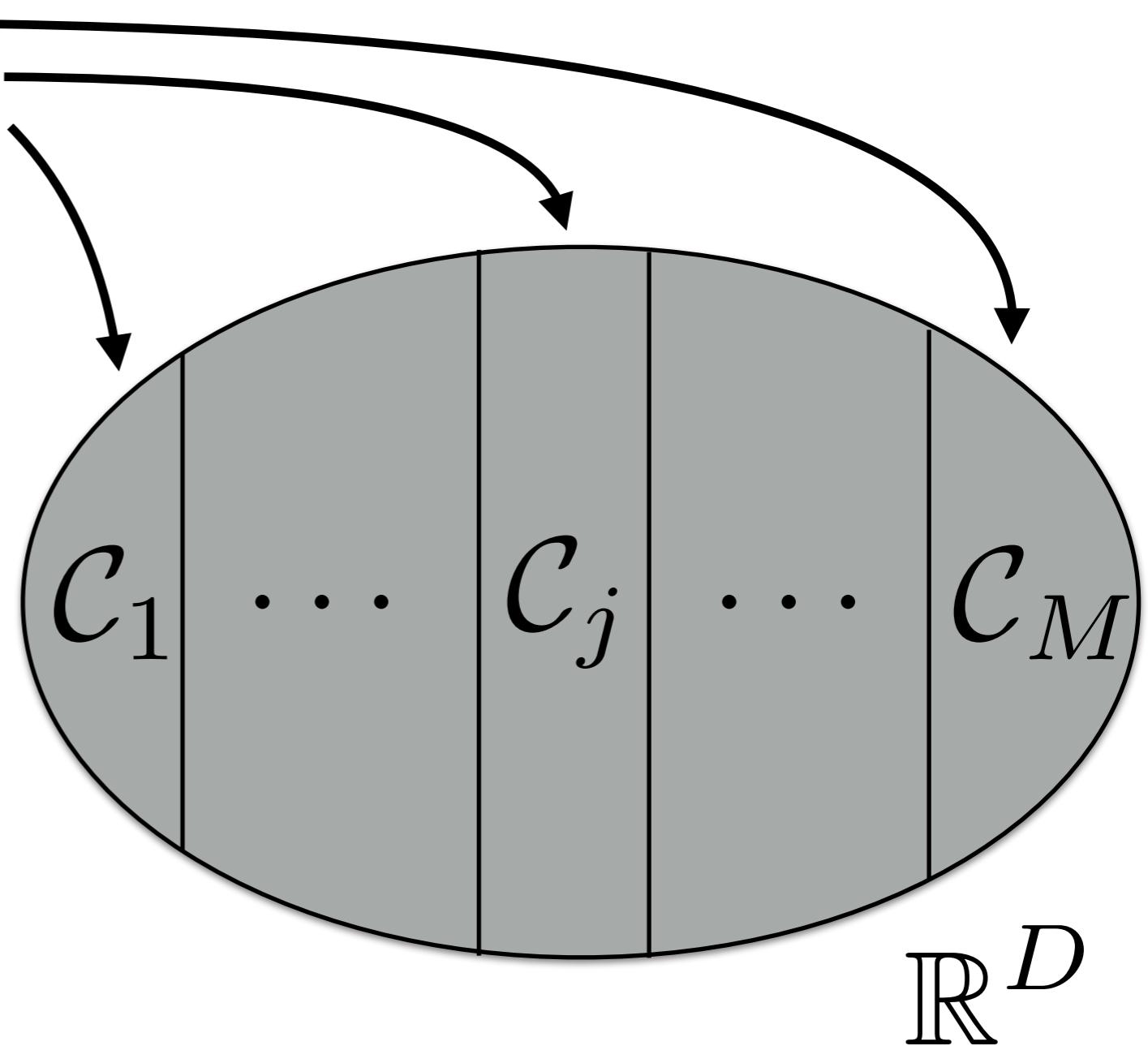
- When all the necessary information is available and there is a rule, logic is sufficient and no decision theory is necessary (typical of IT);
- When the information is partial and there is uncertainty, logic is not sufficient and Bayesian Decision Theory is necessary (typical of AI);
- The decision process is often referred to as “classification” (the decisions can be thought of as classes or categories).

Outline

- Introduction
- Prior Rule
- Posterior Rule
- Conclusions

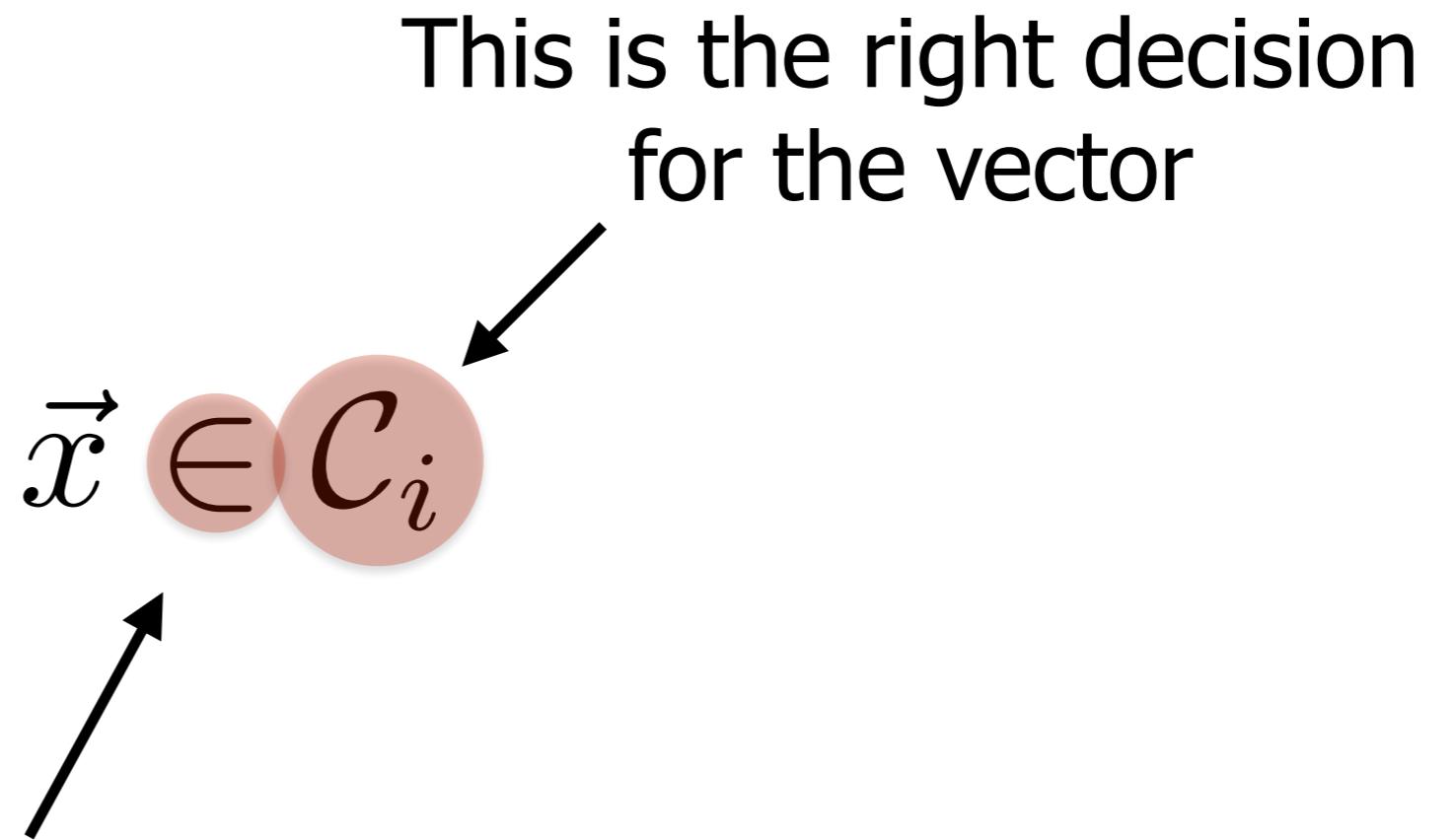
From Decisions to Classes (I)

Every decision corresponds to a subset of the space, a class



The classes are disjoint

From Decisions to Classes (II)



It is equivalent to
saying that the vector
belongs to class i

A-Priori Probability

A-priori probability of
class i

$$p(\vec{x} \in C_i) = p(C_i)$$
$$\sum_{k=1}^M p(C_k) = 1$$

The classes correspond
to mutually exclusive
events

Prior Decision Rule

The class is the one
with the highest a-priori
probability

$$C^* = \arg \max_{C_k \in C} p(C_k)$$

The decision approach
takes into account all
possible decisions

Toy Example (I)

There are two mutually exclusive classes

$$\mathcal{C} = \{\mathcal{C}_1 = \textit{woman}, \mathcal{C}_2 = \textit{man}\}$$

$p(\mathcal{C}_1) = 0.15$

$p(\mathcal{C}_2) = 0.85$

The class with the highest a-priori probability is “man”

Toy Example (II)

The problem is to decide what is the outcome after rolling the dice



$$C = \left\{ \begin{array}{l} C_1 = 1 \\ C_2 = 2 \\ C_3 = 3 \\ C_4 = 4 \\ C_5 = 5 \\ C_6 = 6 \end{array} \right\}$$

Each class corresponds to one of the possible outcomes

Toy Example (II)

The a-priori probability
is the same for all
classes

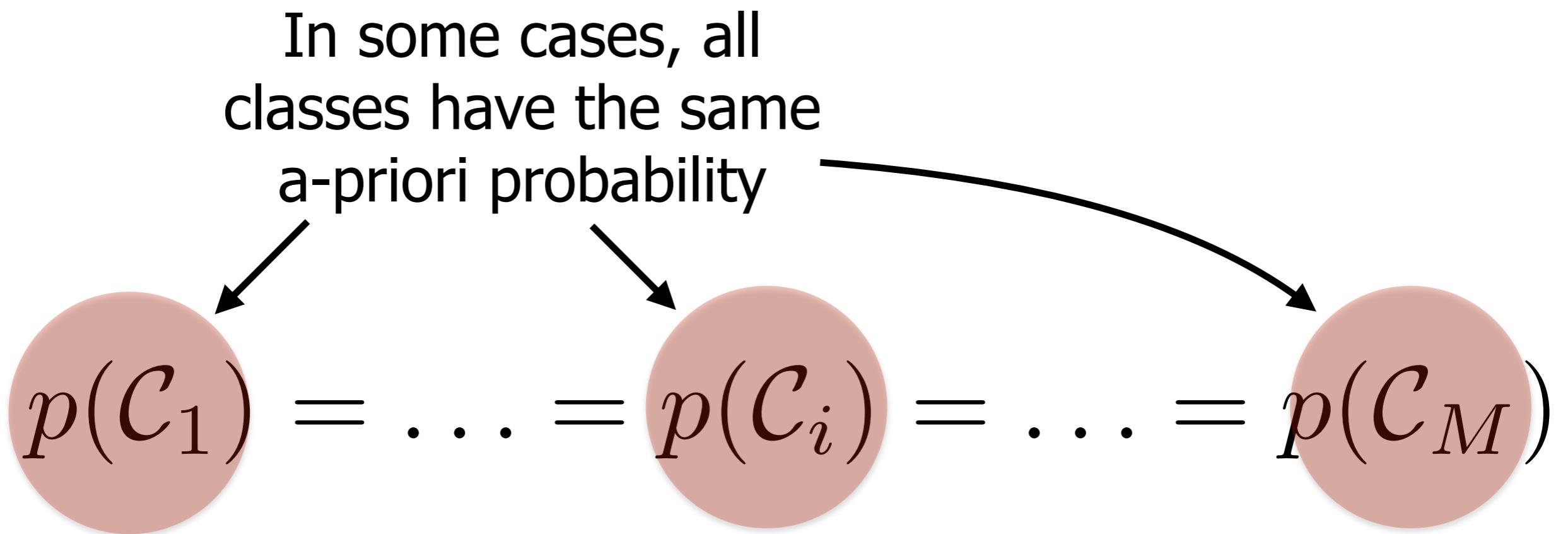
$$p(C_i) = \frac{1}{6}$$

$i = 1, 2, \dots, 6$

$$\arg \max_{C_i \in \mathcal{C}} p(C_i) = ?$$

It is not possible to
identify the highest a-
priori probability

Uninformative Prior



The prior rule is like rolling a dice and the prior is uninformative

Recap

- The prior rule can be used when the only available information is the a-priori probability of the classes;
- The common sense suggests to make the decision corresponding to the class with the highest a-priori probability;
- The prior rule does not take into account the features.

Outline

- Introduction
- Prior Rule
- Posterior Rule
- Conclusions

Taking the Features into Account

The joint probability of class and feature vector

$$p(C_i, \vec{x}) = p(C_i | \vec{x})p(\vec{x}) = p(\vec{x} | C_i)p(C_i)$$

The Product Law allows one to write the joint probability in two ways

The Bayes Theorem

A-posteriori probability
(or posterior) of class i

$$p(C_i | \vec{x}) = \frac{p(\vec{x} | C_i)p(C_i)}{p(\vec{x})}$$

The evidence

Likelihood of class i with
respect to the feature
vector

$$p(\vec{x})$$

The a-priori probability
of class i

The Evidence

Evidence

$$p(\vec{x}) = \sum_{i=1}^M p(\vec{x}|\mathcal{C}_i)p(\mathcal{C}_i)$$

A combination of
product and addition
laws

The Bayes Theorem

The sum over the posteriors is 1

$$\sum_{i=1}^M p(C_i | \vec{x}) = \frac{\sum_{i=1}^M p(\vec{x} | C_i) p(C_i)}{\sum_{k=1}^M p(\vec{x} | C_k) p(C_k)} = 1$$

The evidence is a normalisation constant

Outline

- Introduction
- Bayesian Decision Theory
- Conclusions

Error Probability

Probability of error if
the right decision is j

$$p(\text{err} | \vec{x}) = \sum_{i \neq j} p(C_i | \vec{x})$$

The sum over all
posteriors except the
posterior of j

Error Probability

The class that corresponds to the highest posterior

$$C^* = \arg \max_{C_k \in C} p(C_k | \vec{x})$$

The value of the posterior is checked for all possible classes

The posterior takes the features into account

Posterior Rule

The expression of the priors according to the Bayes Theorem

$$C^* = \arg \max_{C_k \in C} \frac{p(\vec{x}|C_k)p(C_k)}{p(\vec{x})} =$$

$$= \arg \max_{C_k \in C} p(\vec{x}|C_k)p(C_k)$$

The evidence is the same for all classes and it can be eliminated

Toy Example

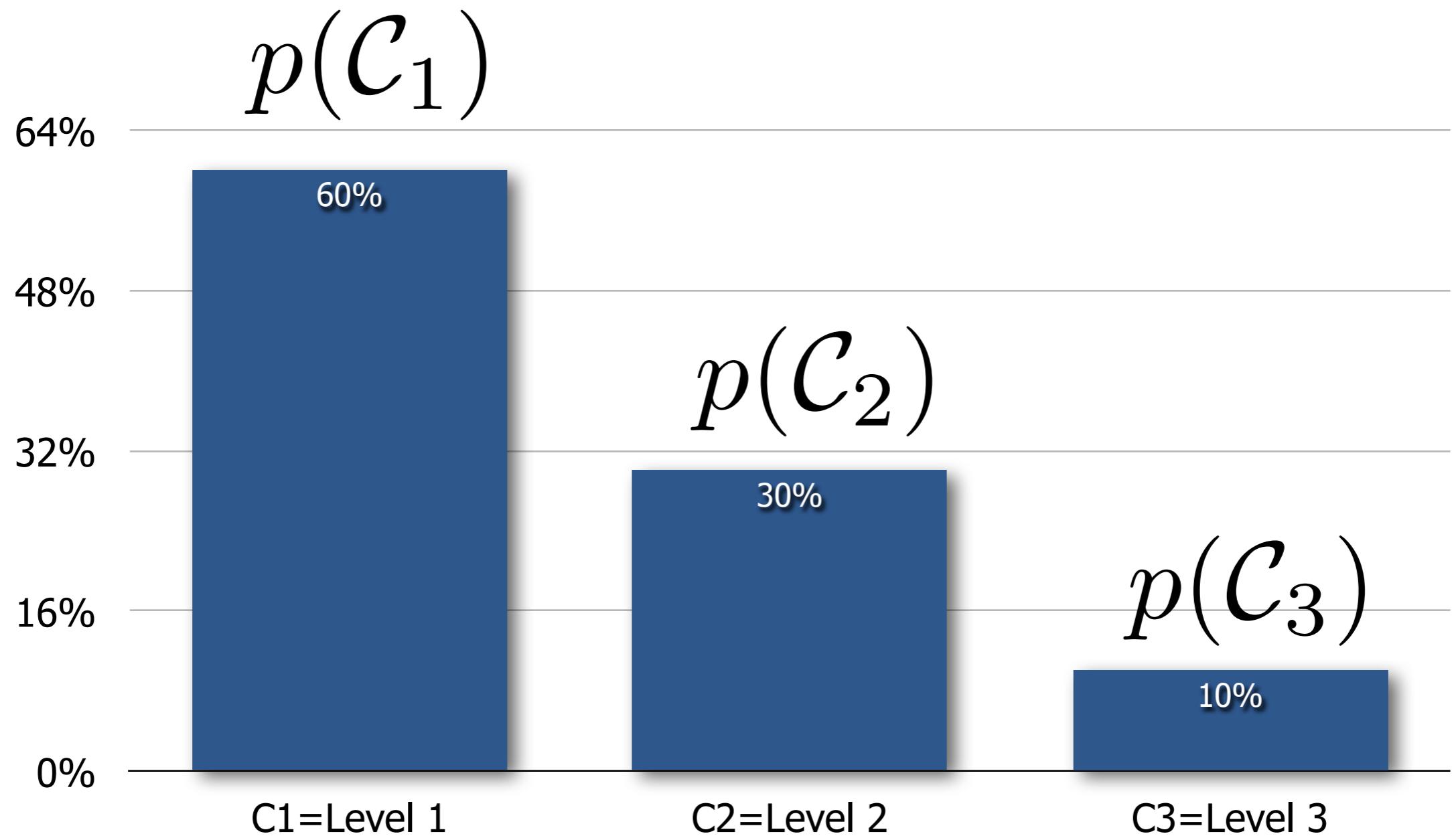
The feature vector
represents a student

The only feature is the
age of the student

$$\vec{x} = (\text{age})$$
$$C = \{L1, L2, L3\}$$

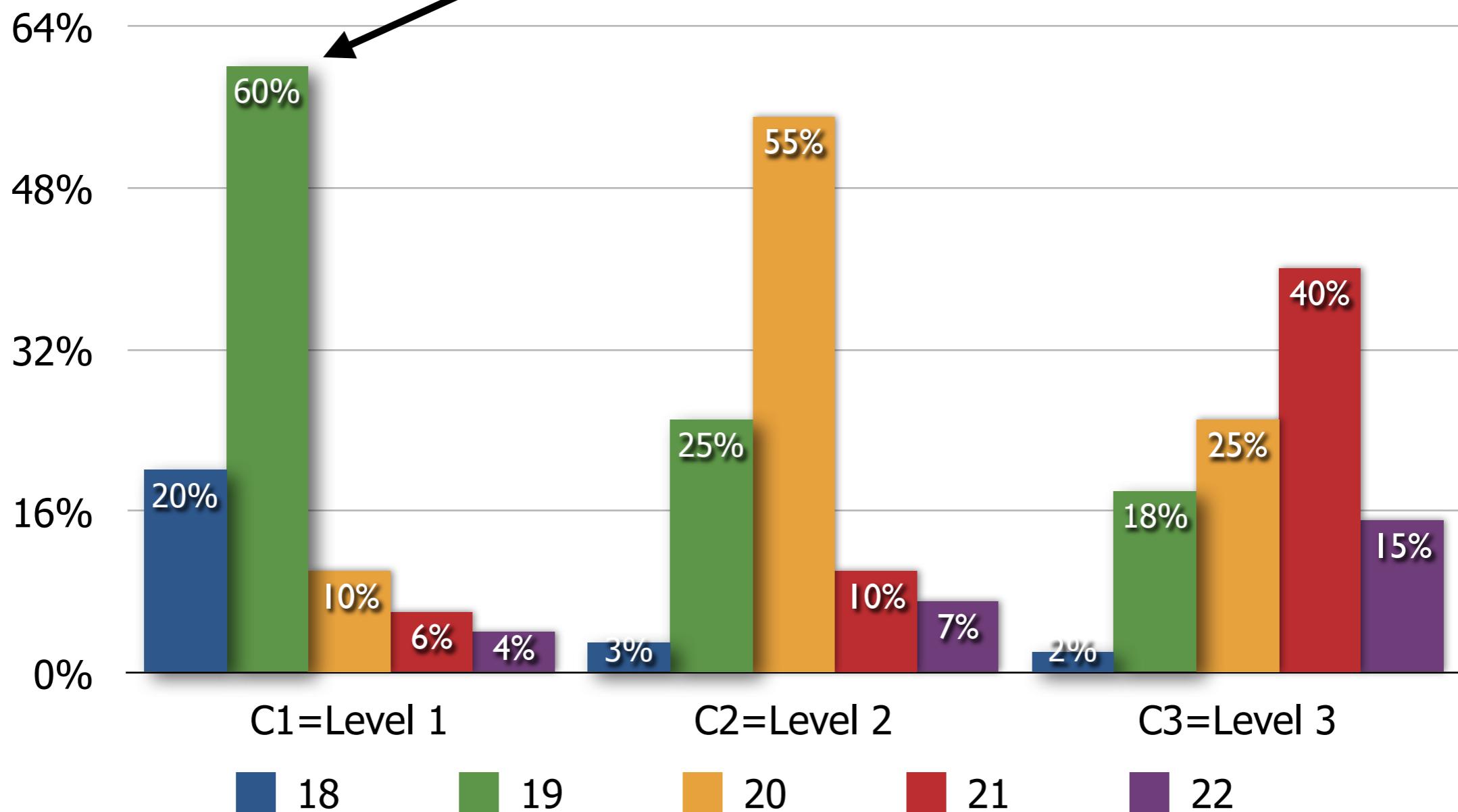
The classes correspond
to the Levels 1, 2 and 3

Priors



Likelihoods

$$p(x = 19 | \mathcal{C}_1)$$



Decision

The age is known, but
the level is not

$$\vec{x} = (21)$$

$$p(\vec{x}|\mathcal{C}_1)p(\mathcal{C}_1) = 0.036$$

$$p(\vec{x}|\mathcal{C}_2)p(\mathcal{C}_2) = 0.030$$

$$p(\vec{x}|\mathcal{C}_3)p(\mathcal{C}_3) = 0.040$$

$$\mathcal{C}_3 = \arg \max_{\mathcal{C}_i \in \mathcal{C}} p(\vec{x}|\mathcal{C}_i)p(\mathcal{C}_i)$$

Class C3 corresponds to
the highest posterior

Recap

- Unlike the prior rule, the posterior rule takes into account the features;
- The goal of the posterior rule is to minimise the error probability (in this respect it formalises common sense);
- The main problem left open is how to estimate the a-priori and a-posteriori probabilities involved in the problem.

Outline

- Introduction
- Prior Rule
- Posterior Rule
- Conclusions

Conclusions

- In the Bayesian Decision Theory, making a decision means to map a feature vector into one of M predefined decisions;
- The set of the feature vectors for which the right decision is the same can be thought of as a class (a subset of the feature space);
- The decision process is often referred to as classification.

Discriminant Functions

Computational Social Intelligence - Lecture 13

Prof. Alessandro Vinciarelli
School of Computing Science &
Institute of Neuroscience and Psychology

<http://www.dcs.gla.ac.uk/vincia>
Alessandro.Vinciarelli@glasgow.ac.uk



University
of Glasgow

EPSRC
Engineering and Physical Sciences
Research Council

FNSNF

Texts (see Moodle)

This lecture is based on the following text
(available on Moodle):

- Chapter 5 of F.Camastra and A.Vinciarelli,
“Machine Learning for Audio, Image and Video
Processing”, Springer Verlag, 2008.

Outline

- Quick Recap
- Conditional Risk and Bayes Decision Rule
- Binary Classification and Likelihood Ratio
- Zero-One Loss and Posterior Rule
- Gaussian Discriminant Functions
- Conclusions

Outline

- Quick Recap
- Conditional Risk and Bayes Decision Rule
- Binary Classification and Likelihood Ratio
- Zero-One Loss and Posterior Rule
- Gaussian Discriminant Functions
- Conclusions

The Bayes Theorem

A-posteriori probability
(or posterior) of class i

$$p(C_i | \vec{x}) = \frac{p(\vec{x} | C_i)p(C_i)}{p(\vec{x})}$$

The evidence

Likelihood of class i with
respect to the feature
vector

$$p(\vec{x})$$

The a-priori probability
of class i

Error Probability

Probability of error if
the right decision is j

$$p(\text{err} | \vec{x}) = \sum_{i \neq j} p(C_i | \vec{x})$$

The sum over all
posteriors except the
posterior of j

Error Probability

The class that corresponds to the highest posterior

$$C^* = \arg \max_{C_k \in C} p(C_k | \vec{x})$$

The value of the posterior is checked for all possible classes

The posterior takes the features into account

Posterior Rule

The expression of the priors according to the Bayes Theorem

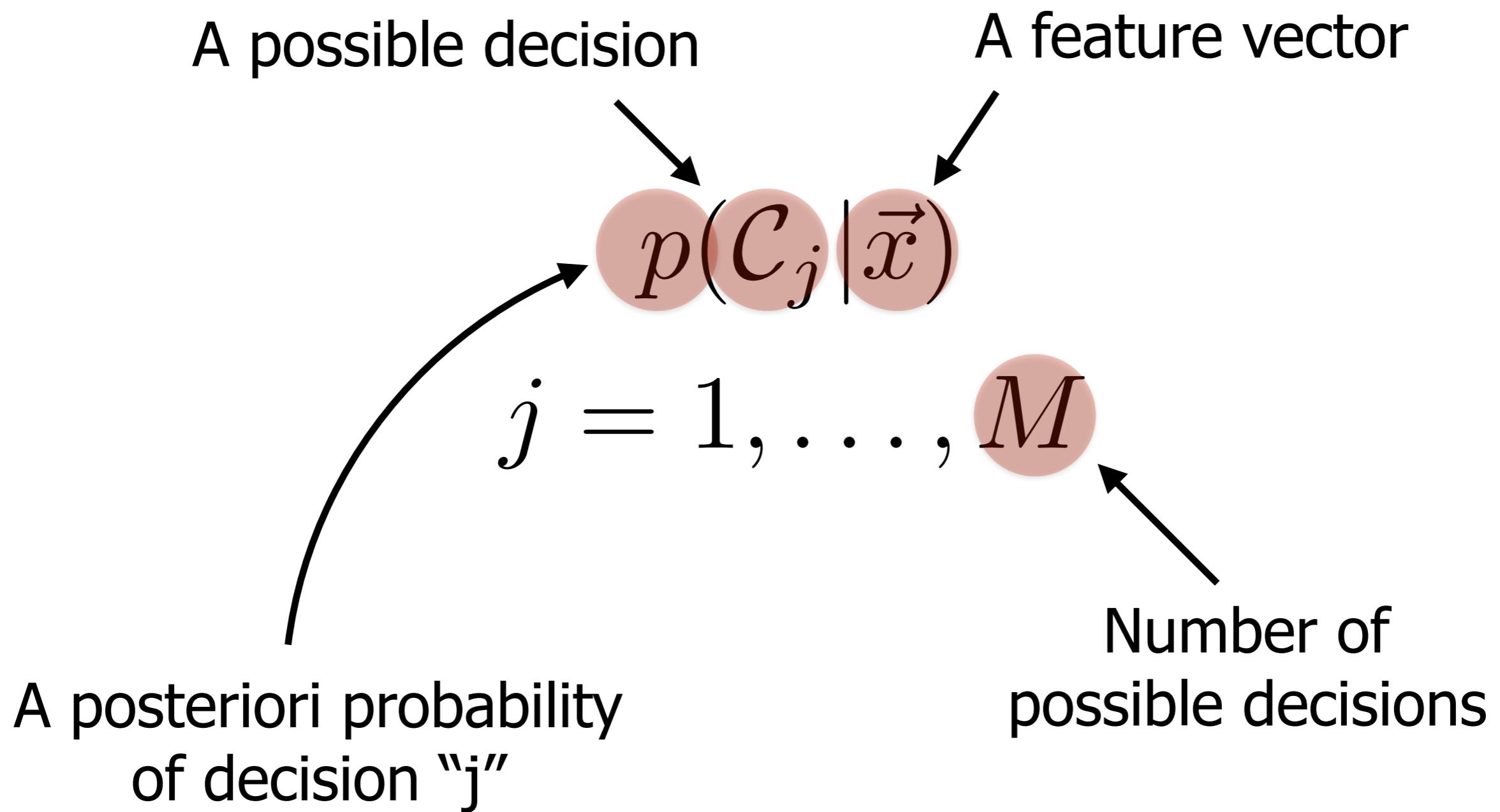
$$C^* = \arg \max_{C_k \in C} \frac{p(\vec{x}|C_k)p(C_k)}{p(\vec{x})} =$$

$$= \arg \max_{C_k \in C} p(\vec{x}|C_k)p(C_k)$$

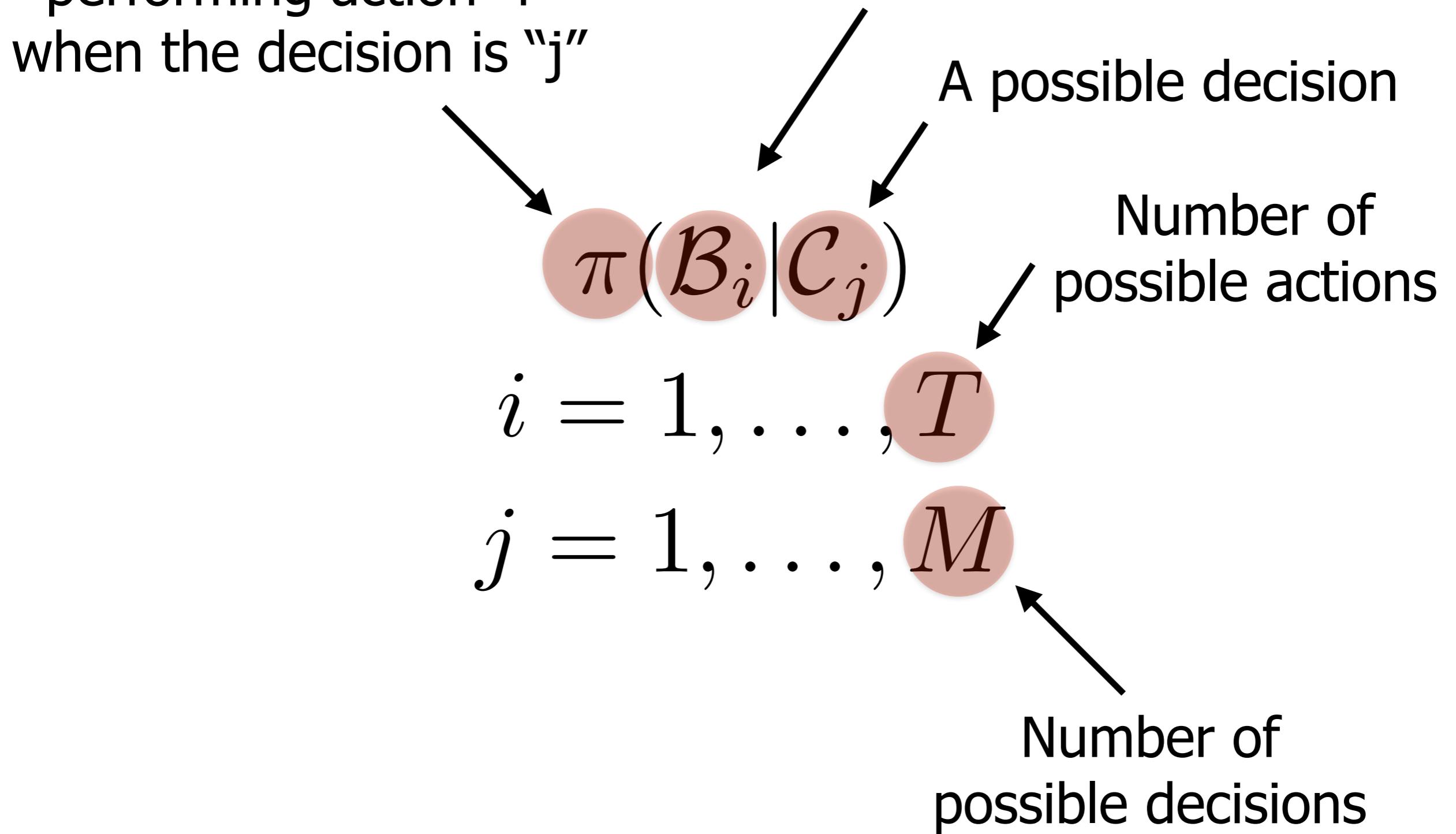
The evidence is the same for all classes and it can be eliminated

Outline

- Quick Recap
- Conditional Risk and Bayes Decision Rule
- Binary Classification and Likelihood Ratio
- Zero-One Loss or Posterior Rule
- Gaussian Discriminant Functions
- Conclusions



The loss resulting from performing action “i” when the decision is “j”



Expected loss (or
conditional risk)
associated to the action

$$\mathcal{R}(\mathcal{B}_i | \vec{x}) = \sum_{j=1}^M \pi(\mathcal{B}_i | C_j) p(C_j | \vec{x})$$

Sum over all possible
classes the vector can
be assigned to

A possible action

The posterior of the
class acts as a weight in
the sum

Bayes Decision Rule

The index of the action
that minimises the
conditional risk

$$\hat{k} = \arg \min_{k=1, \dots, T} \mathcal{R}(\mathcal{B}_k | \vec{x}) =$$

$$\mathcal{R}(\mathcal{B}_{\hat{k}} | \vec{x})$$

The Bayes Risk

Recap

- The conditional risk is the weighted sum of the losses associated to an action;
- The weights are the posteriors of the classes (the conditional risk is an expectation);
- The Bayes Decision Rule targets the decision that minimises the conditional risk.

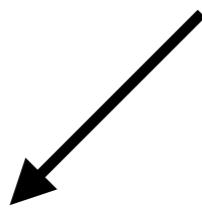
Outline

- Quick Recap
- Conditional Risk and Bayes Decision Rule
- **Binary Classification and Likelihood Ratio**
- Zero-One Loss and Posterior Rule
- Gaussian Discriminant Functions
- Conclusions

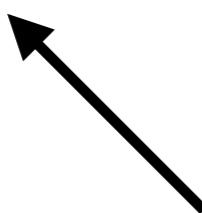
Binary Classification

- In a binary classification there are two classes and two actions;
- Each of the two actions is right for one class, but wrong for the other;
- For example, the classes are “healthy” and “ill” while the actions are “keep in the hospital” and “do not keep in the hospital”.

Expected loss (or
conditional risk)
associated to action 1



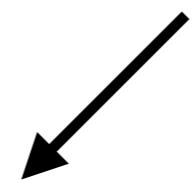
$$\mathcal{R}(\mathcal{B}_1 | \vec{x}) = \pi(\mathcal{B}_1 | \mathcal{C}_1)p(\mathcal{C}_1 | \vec{x}) + \pi(\mathcal{B}_1 | \mathcal{C}_2)p(\mathcal{C}_2 | \vec{x})$$



Expected loss (or
conditional risk)
associated to action 2

$$\mathcal{R}(\mathcal{B}_2 | \vec{x}) = \pi(\mathcal{B}_2 | \mathcal{C}_1)p(\mathcal{C}_1 | \vec{x}) + \pi(\mathcal{B}_2 | \mathcal{C}_2)p(\mathcal{C}_2 | \vec{x})$$

When the difference is
positive, action 1
minimises the expected
loss



$$\begin{aligned} & \mathcal{R}(\mathcal{B}_2 | \vec{x}) - \mathcal{R}(\mathcal{B}_1 | \vec{x}) = \\ &= [\pi(\mathcal{B}_2 | \mathcal{C}_1) - \pi(\mathcal{B}_1 | \mathcal{C}_1)] p(\mathcal{C}_1 | \vec{x}) + \\ &+ [\pi(\mathcal{B}_2 | \mathcal{C}_2) - \pi(\mathcal{B}_1 | \mathcal{C}_2)] p(\mathcal{C}_2 | \vec{x}) \end{aligned}$$

The Likelihood Ratio

$$\frac{p(\vec{x}|\mathcal{C}_1)}{p(\vec{x}|\mathcal{C}_2)} > \frac{\pi(\mathcal{B}_1|\mathcal{C}_2) - \pi(\mathcal{B}_2|\mathcal{C}_2)}{\pi(\mathcal{B}_2|\mathcal{C}_1) - \pi(\mathcal{B}_1|\mathcal{C}_1)} \frac{p(\mathcal{C}_2)}{p(\mathcal{C}_1)}$$

When the Likelihood Ratio satisfies the equation above, action 1 minimises the expected loss

Example

- The authentication of a face image for unlocking a device is an example of binary classification;
- Decision 1 is to accept the face image (the action is unlock), Decision 2 is to reject the face (the action is to keep the device locked);
- The faces are converted into feature vectors through image processing techniques.

Likelihood of accepting

$$\frac{p(\vec{x}|C_a)}{p(\vec{x}|C_r)}$$

Likelihood of rejecting

Loss resulting from unlocking when the right decision is reject

Loss resulting from locking when the right decision is reject

Loss resulting from locking when the right decision is accept

Loss resulting from unlocking when the right decision is accept

Likelihood of accepting

$$\frac{p(\vec{x}|C_a)}{p(\vec{x}|C_r)}$$

Likelihood of rejecting

Loss resulting from unlocking when the right decision is reject

Loss resulting from locking when the right decision is accept

Loss resulting from locking when the right decision is reject

Loss resulting from unlocking when the right decision is accept

Recap

- A binary decision problem can be addressed through the likelihood ratio;
- The losses and the priors determine the likelihood ratio threshold above which one of the two decisions is made;

Outline

- Quick Recap
- Conditional Risk and Bayes Decision Rule
- Binary Classification and Likelihood Ratio
- Zero-One Loss and Posterior Rule
- Gaussian Discriminant Functions
- Conclusions

The loss resulting from performing action “i” when the decision is “j”

$$\pi(\mathcal{B}_i | \mathcal{C}_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$$

There is no loss resulting from performing action “i” when the decision is “i”

The loss is the same for every mismatch between decision and action

The expected loss
(conditional risk) when
performing action “i”

The sum includes only
the cases in which “j”
and “i” are different

$$\begin{aligned}\mathcal{R}(\mathcal{B}_i | \vec{x}) &= \sum_{j \neq i} \pi(\mathcal{B}_i | \mathcal{C}_j) p(\mathcal{C}_j | \vec{x}) = \\ &= 1 - p(\vec{\mathcal{C}}_i | \vec{x})\end{aligned}$$

The highest posterior
minimises the
conditional risk

Recap

- In most cases, the number of actions is the same as the number of decisions;
- In general, for every class one decision is right while all the others are wrong;
- In such cases, the minimisation of the conditional risk is equivalent to the application of the posterior rule.

Outline

- Quick Recap
- Conditional Risk and Bayes Decision Rule
- Binary Classification and Likelihood Ratio
- Zero-One Loss and Posterior Rule
- **Gaussian Discriminant Functions**
- Conclusions

The index of the function gamma that corresponds to the maximum

$$\hat{k} = \arg \max_{k=1, \dots, M} \gamma_k(\vec{x})$$

$$\mathcal{G} = \{\gamma_1(\vec{x}), \dots, \gamma_M(\vec{x})\}$$

The discriminant functions

All functions in set G

The discriminant functions are as many as the classes

The opposite of the expected loss can be used as discriminant function

This relationship holds for a zero-one loss function

$$\gamma_k(\vec{x}) = -\mathcal{R}(\mathcal{B}_k | \vec{x}) = p(C_k | \vec{x}) - 1$$

$$\gamma_k(\vec{x}) \simeq p(C_k | \vec{x})$$

The discriminant functions are compared to one another, the additive constant can be dropped

The posterior of decision "k"

This relationship holds
thanks to the
Bayes Theorem

$$\gamma_k(\vec{x}) = \frac{p(\vec{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\vec{x})}$$

The logarithm is
monotonic, the result of
the comparison does
not change

$$\log \gamma_k(\vec{x}) = \log \frac{p(\vec{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\vec{x})}$$

$$\log \gamma_k(\vec{x}) \simeq \log p(\vec{x}|\mathcal{C}_k) + \log p(\mathcal{C}_k)$$

The evidence is the
same for all discriminant
functions and can be
dropped

The decision
corresponds to the
maximum value of the
logarithm

$$\hat{k} = \arg \max_{k \in [1, M]} \log p(\vec{x} | C_k) + \log p(C_k)$$

All possible values of k
are tested

The Likelihood

The probability of a vector is the joint probability of its components

$$p(\vec{x}|\mathcal{C}_k) = p(x_1, x_2, \dots, x_D | \mathcal{C}_k)$$

The dimensionality
of the vector

The probability of a vector is the joint probability of its components

Likelihood of an individual component

$$p(x_1, x_2, \dots, x_D | C_k) = \prod_{i=1}^D p(x_i | C_k)$$

It is true if the components are statistically independent given the class

The Naive Bayes classifier

The decision
corresponds to the
maximum value of the
logarithm

$$\hat{k} = \arg \max_{k \in [1, M]} \log p(\vec{x} | C_k) + \log p(C_k)$$

All possible values of k
are tested

$$\log p(\vec{x}|\mathcal{C}_k) = \sum_{i=1}^D \log p(x_i|\mathcal{C}_k)$$

It is true if the components are statistically independent given the class

Probability of observing
the value of feature “i”
when the class is “k”

$$p(x_i | C_k) = \frac{1}{\sqrt{2\pi}\sigma_{ik}} \exp \left[-\frac{(x_i - \mu_{ik})^2}{2\sigma_{ik}^2} \right]$$

Diagram illustrating the components of the Gaussian probability formula:

- An arrow points from the text "Probability of observing the value of feature “i” when the class is “k”" to the term $p(x_i | C_k)$.
- An arrow points from the text "Average of feature “i” in class “k”" to the term μ_{ik} .
- An arrow points from the text "Standard deviation of feature “i” in class “k”" to the term σ_{ik} .

Sum over the standard deviations of the individual features

True when features statistically independent given the class

$$\log p(\vec{x} | C_k) = -\sum_{i=1}^D \left[\log \sqrt{2\pi} \sigma_{ik} + \frac{(x_i - \mu_{ik})^2}{2\sigma_{ik}^2} \right]$$

Euclidean distance between feature vector and class average

Outline

- Quick Recap
- Conditional Risk and Bayes Decision Rule
- Binary Classification and Likelihood Ratio
- Zero-One Loss and Posterior Rule
- Gaussian Discriminant Functions
- Conclusions

Conclusions

- Discriminant functions allow one to implement the Bayes Decision Rule by checking their value for the feature vectors;
- The features can be assumed statistically independent given the class (the Naive Bayes Classifier);
- The discriminant functions can be thought of distances from the class averages.

Training and Test

Computational Social Intelligence - Lecture 14

Prof. Alessandro Vinciarelli
School of Computing Science &
Institute of Neuroscience and Psychology

<http://www.dcs.gla.ac.uk/vincia>
Alessandro.Vinciarelli@glasgow.ac.uk



University
of Glasgow

EPSRC
Engineering and Physical Sciences
Research Council

FNSNF

Texts (see Moodle)

This lecture is based on the following text
(available on Moodle):

- Chapter 5 of F.Camastra and A.Vinciarelli,
“Machine Learning for Audio, Image and Video
Processing”, Springer Verlag, 2008.

Outline

- Training for the likelihood
- Training for the priors
- K-fold and performance measurement
- Conclusions

Outline

- Training for the likelihood
- Training for the priors
- K-fold and performance measurement
- Conclusions

Training

- Training is the mathematical process through which the parameters of statistical distributions are adapted to training data;
- The process is often referred to as “learning from data”;
- In general, the training takes place by selecting the parameter values that are optimal according to a given criterion.

Sum over the standard deviations of the individual features

True when features statistically independent given the class

$$\log p(\vec{x} | C_k) = -\sum_{i=1}^D \left[\log \sqrt{2\pi} \sigma_{ik} + \frac{(x_i - \mu_{ik})^2}{2\sigma_{ik}^2} \right]$$

Euclidean distance between feature vector and class average

The training set

Every element of the
training set is a pair of a
feature vector and class

$$\mathcal{X} = \{(\vec{x}_1, c_1), \dots, (\vec{x}_N, c_N)\}$$

The class of training
feature vector "i"

$$c_i \in \mathcal{C} = \{c_1, \dots, c_M\}$$

The number of possible
decisions (classes)

The subset of the training set that includes the feature vectors belonging to class “k”

$$\mathcal{X}^{(k)} = \{\vec{x}_1, \dots, \vec{x}_K\}$$

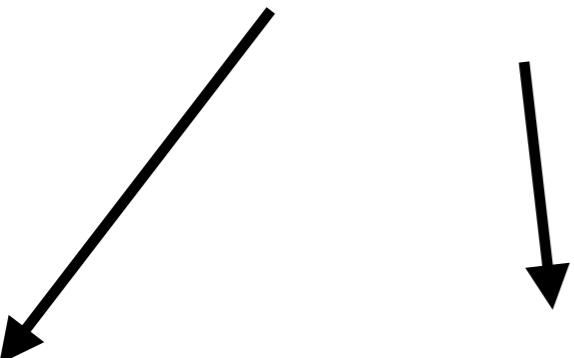
Feature vectors that belong to the training set and belong to class “k”

The likelihood of the training set

The product of the likelihoods of the individual feature vectors

$$p(\mathcal{X}^{(k)} | \mathcal{C}_k) = \prod_{i=1}^K p(\vec{x}_i | \mathcal{C}_k)$$

The log-likelihood of the
training set


$$\mathcal{L} = \log p(\mathcal{X}^{(k)} | \mathcal{C}_k) = \sum_{i=1}^K \log p(\vec{x}_i | \mathcal{C}_k)$$

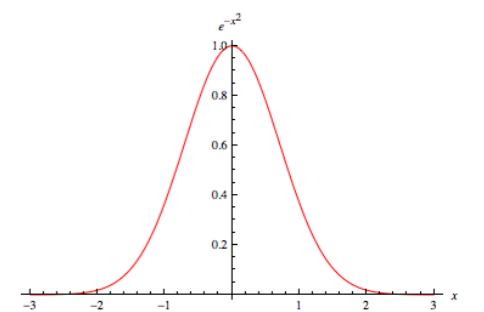
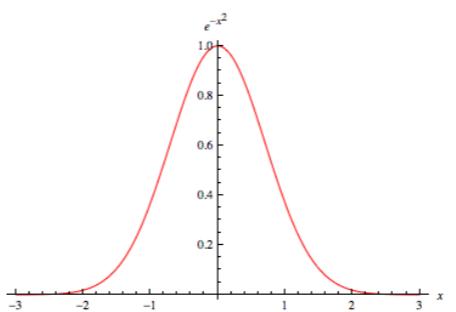
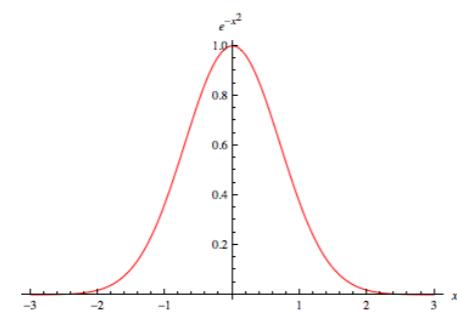
Dimensionality of the feature vectors

$$\mathcal{L} = -\sum_{i=1}^K \sum_{j=1}^D \left[\log \sqrt{2\pi} \sigma_{kj} + \frac{(x_{ij} - \mu_{kj})^2}{2\sigma_{kj}^2} \right]$$

Diagram illustrating the components of the loss function \mathcal{L} :

- Component "j" of vector "i"**: Points to the term $(x_{ij} - \mu_{kj})^2$.
- Sample mean of component "j" in the feature vectors belonging to class "k"**: Points to the term μ_{kj} .
- Standard deviation of component "j" in the feature vectors belonging to class "k"**: Points to the term σ_{kj} .
- Number of feature vectors belonging to class "k"**: Points to the term K .
- Dimensionality of the feature vectors**: Points to the term D .

$$\mathcal{X}^{(k)} = \begin{pmatrix} x_{11} & \dots & x_{1l} & \dots & x_{1D} \\ x_{21} & \dots & x_{2l} & \dots & x_{2D} \\ \dots & \dots & \dots & \dots & \dots \\ x_{K1} & \dots & x_{Kl} & \dots & x_{KD} \end{pmatrix}$$



$$p(x_1 | \mathcal{C}_k)$$

$$p(x_l | \mathcal{C}_k)$$

$$p(x_D | \mathcal{C}_k)$$

Setting the derivative to zero allows to find the value of the parameters that maximise the log likelihood

Sample mean of component “l” in the feature vectors belonging to class “k”

$$\frac{\partial \mathcal{L}}{\partial \sigma_{kl}} = \sum_{i=1}^K \left[\frac{(x_{il} - \mu_{kl})^2}{2\sigma_{kl}^3} - \frac{1}{\sigma_{kl}} \right] = 0$$

Standard deviation of component “j” in the feature vectors belonging to class “k”

Setting the derivative to zero allows to find the value of the parameters that maximise the log likelihood

$$\frac{\partial \mathcal{L}}{\partial \mu_{kl}} = 2 \sum_{i=1}^K \frac{(x_{il} - \mu_{kl})}{2\sigma_{kl}^2} = 0$$

The maximum likelihood estimate of the parameter

The sample mean of component “l” in the feature vectors belonging to class “k”

$$\mu_{kl} = \frac{1}{K} \sum_{i=1}^K x_{il}$$

Number of feature vectors belonging to class “k” in the training set

The maximum likelihood estimate of the parameter

The sample variance of component “l” in the feature vectors belonging to class “k”

$$\sigma_{kl}^2 = \frac{1}{K} \sum_{i=1}^K (x_{il} - \mu_{kl})^2$$

Number of feature vectors belonging to class “k” in the training set

$$\begin{pmatrix} x_{11} & \dots & x_{1l} & \dots & x_{1D} \\ x_{21} & \dots & x_{2l} & \dots & x_{2D} \\ \dots & \dots & \dots & \dots & \dots \\ x_{K1} & \dots & x_{Kl} & \dots & x_{KD} \end{pmatrix}$$

$$p(x_l | C_k) = \frac{1}{\sqrt{2\pi}\sigma_{kl}} e^{-\frac{(x_{kl} - \mu_{kl})^2}{2\sigma_{kl}^2}}$$

Recap

- For every component of the feature vectors and for every class there is a different Gaussian distribution;
- Means and standard deviations of the Gaussians have been set to maximise the likelihood over the training data;

Outline

- Training for the likelihood
- **Training for the priors**
- K-fold and performance measurement
- Conclusions

Number of samples
belonging to class “j”

$$n(\mathcal{C}_1) = n_1, \dots, n(\mathcal{C}_M) = n_M$$

$$\sum_{j=1}^M n_j = N$$

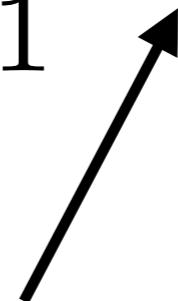
Sum over all numbers of samples belonging to the classes

Total number of samples

Probability of observing
the number of training
samples per class



$$p(n_1, \dots, n_M) = \prod_{j=1}^M p(C_j)^{n_j}$$



Prior of class “j”

$$\mathcal{L} = \sum_{j=1}^M n_j \log p(\mathcal{C}_j) + \lambda \left(1 - \sum_{j=1}^M p(\mathcal{C}_j) \right)$$

Log likelihood

Lagrange multiplier

The content of the parenthesis is null

Setting the derivative to zero allows to find the value of the parameters that maximise the log likelihood

$$\frac{\partial \mathcal{L}}{\partial p(C_l)} = \frac{n_l}{p(C_l)} - \lambda = 0$$

$$n_l = \lambda p(C_l)$$

The value of the Lagrange multiplier must still be found

$$N = \sum_{l=1}^M n_l = \lambda \sum_{l=1}^M p(\mathcal{C}_l) = \lambda$$

The value of the
Lagrange multiplier

$$n_l = \lambda p(C_l)$$

$$p(C_l) = \frac{n_l}{N}$$

The maximum likelihood
estimate of the prior of
class "l"

Recap

- The maximum likelihood estimate of the priors is the percentage of samples belonging to the classes in the training set;
- A possible alternative is to use a-priori knowledge about the problem under exam (e.g., it is known that men and women are 50% of the population);

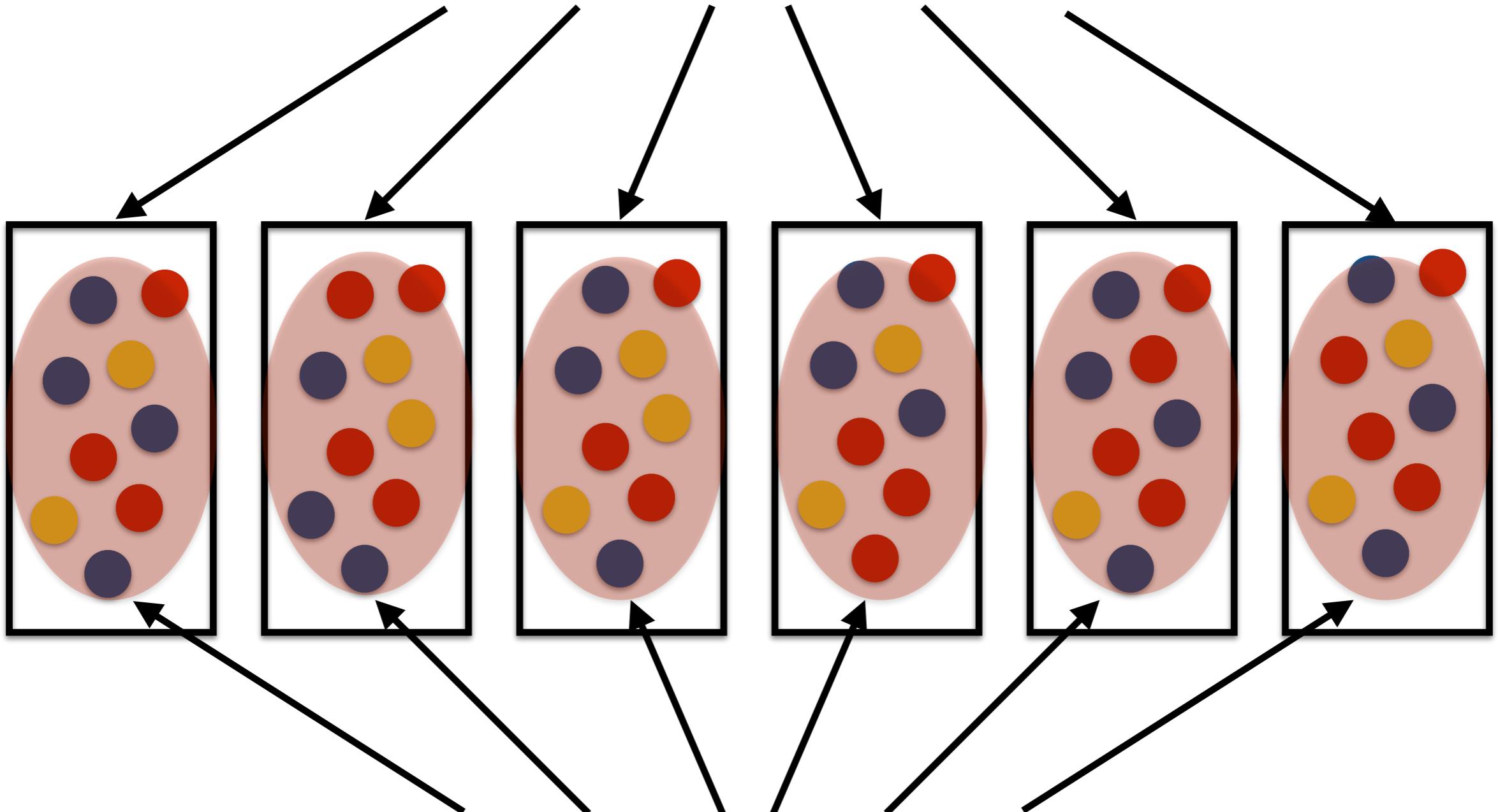
Outline

- Training for the likelihood
- Training for the priors
- K-fold and performance measurement
- Conclusions

Training and Test Set

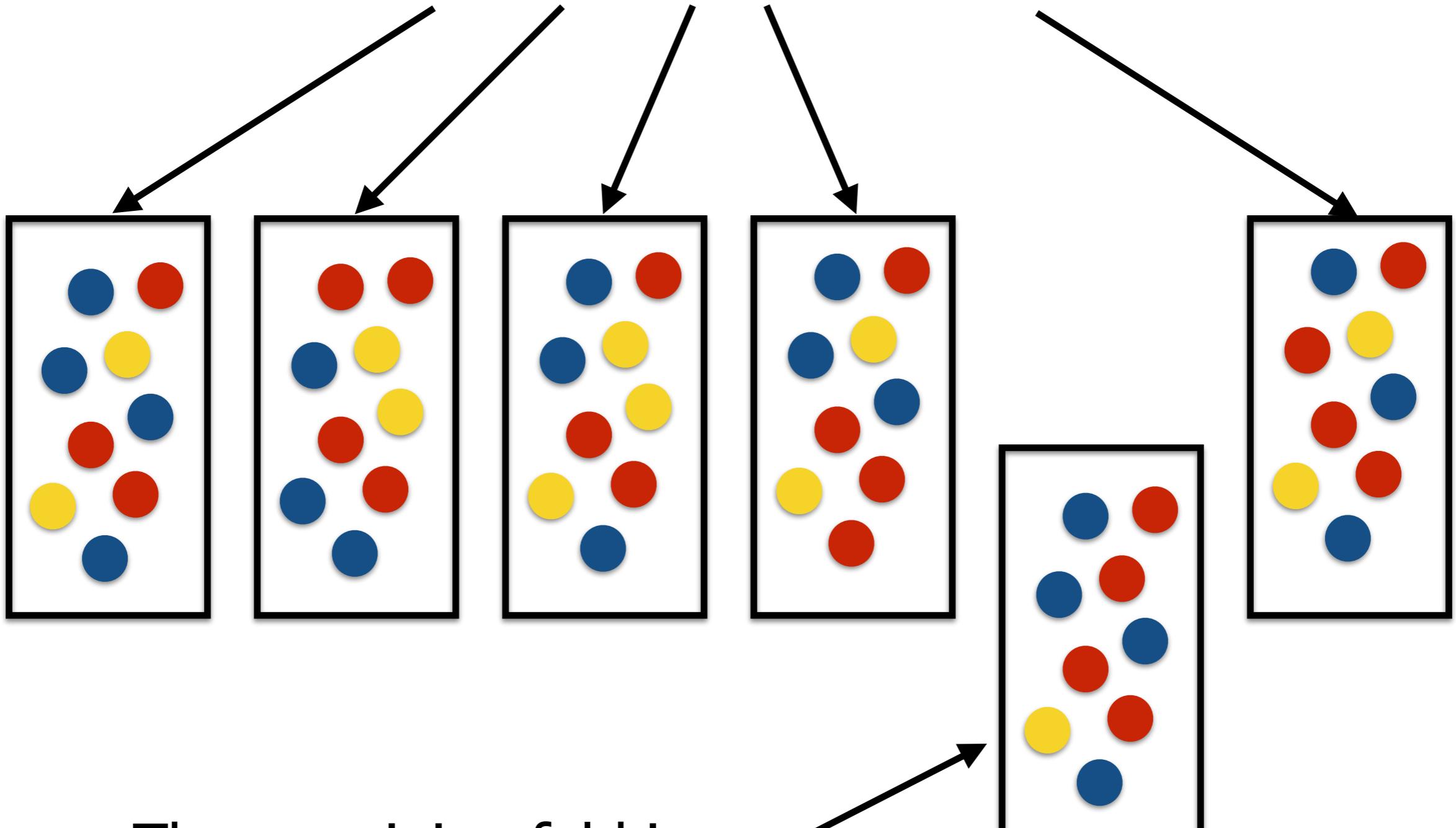
- There must be a separation between the data used for training and the data used for testing;
- The reason is that an approach must generalise, i.e., it must be capable to make the right decision for previously unseen data;
- The k-fold approach allows one to test over the whole dataset at disposition while keeping separated training and test set.

The folds are subsets of
the data at disposition



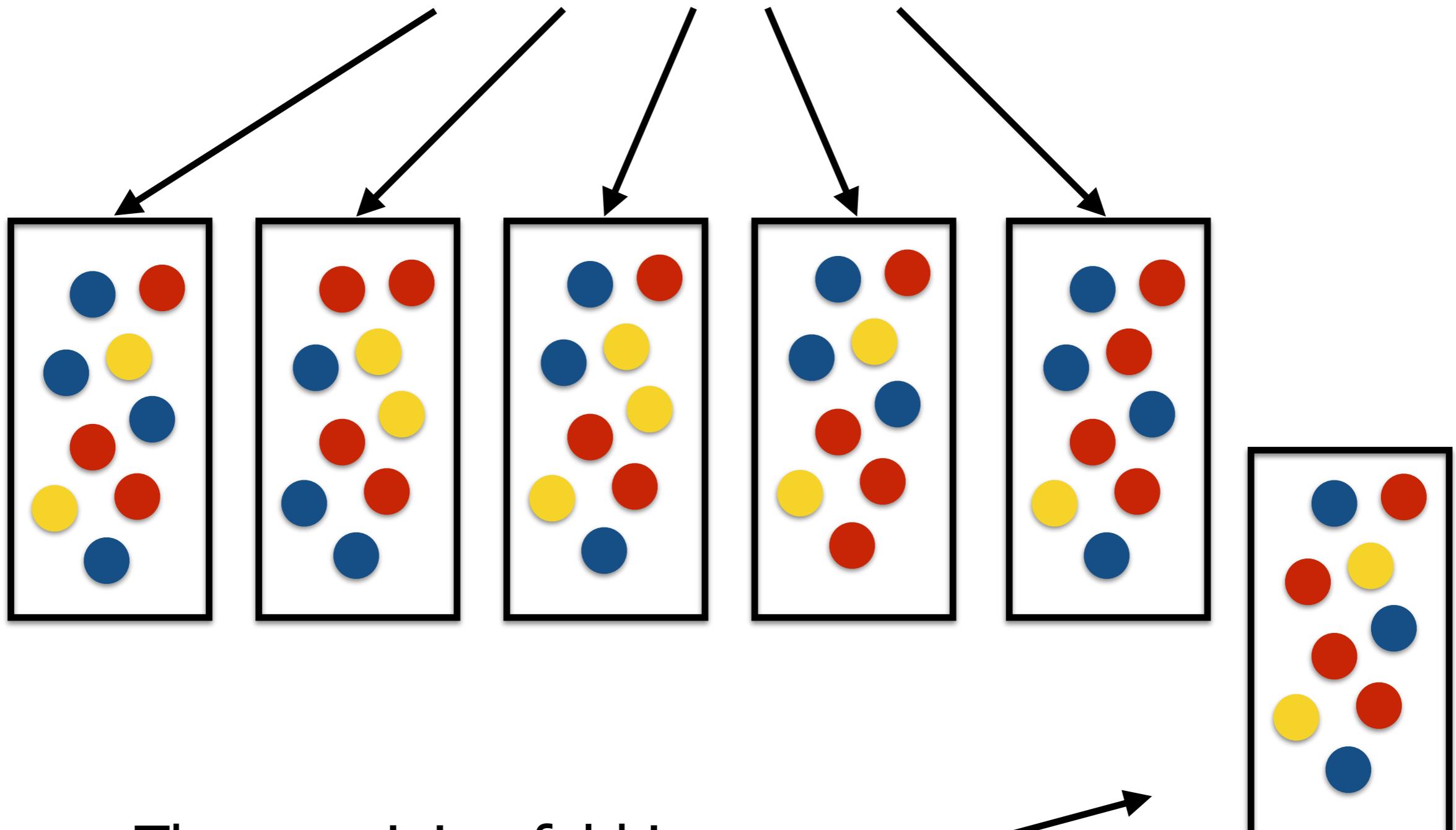
Number of samples and
class distribution are
roughly the same

K-1 folds are used for
training



The remaining fold is
used for testing

K-1 folds are used for
training



The remaining fold is
used for testing

K-Fold

- The dataset at disposition is split into K disjoint subsets (the folds) of roughly the same size;
- The class distribution is roughly the same in all folds (it is sufficient to select the folds through a random process);
- Every fold is used once as a test set and K-1 times as a part of the training set.

Performance measured
in terms of average loss

Action that minimises
the Bayes Risk

$$\alpha = \frac{1}{N} \sum_{i=1}^N \pi(\mathcal{B}_i^* | \vec{x}_i)$$

Loss associated to the
action that minimises
the Bayes Risk

Fraction of times the approach makes the wrong decision (error rate)

Action that minimises the Bayes Risk

$$\alpha = \frac{1}{N} \sum_{i=1}^N \pi(\mathcal{B}_i^* | \vec{x}_i)$$

In the Zero-One loss case, it is 1 when the action (the decision is wrong) and zero otherwise

Outline

- Training for the likelihood
- Training for the priors
- K-fold and performance measurement
- Conclusions

Conclusions

- The parameters of the discriminant functions are set by optimising a criterion;
- The maximisation of the likelihood is one of the training approaches most commonly adopted;
- Once the models are trained, it is possible to test them over unseen data and measure their performance.

Facial Expressions

Computational Social Intelligence - Lecture 15

Prof. Alessandro Vinciarelli
School of Computing Science &
Institute of Neuroscience and Psychology

<http://www.dcs.gla.ac.uk/vincia>
Alessandro.Vinciarelli@glasgow.ac.uk



University
of Glasgow

EPSRC
Engineering and Physical Sciences
Research Council

FNSNF

This lecture is based on the following text
(available on Moodle):

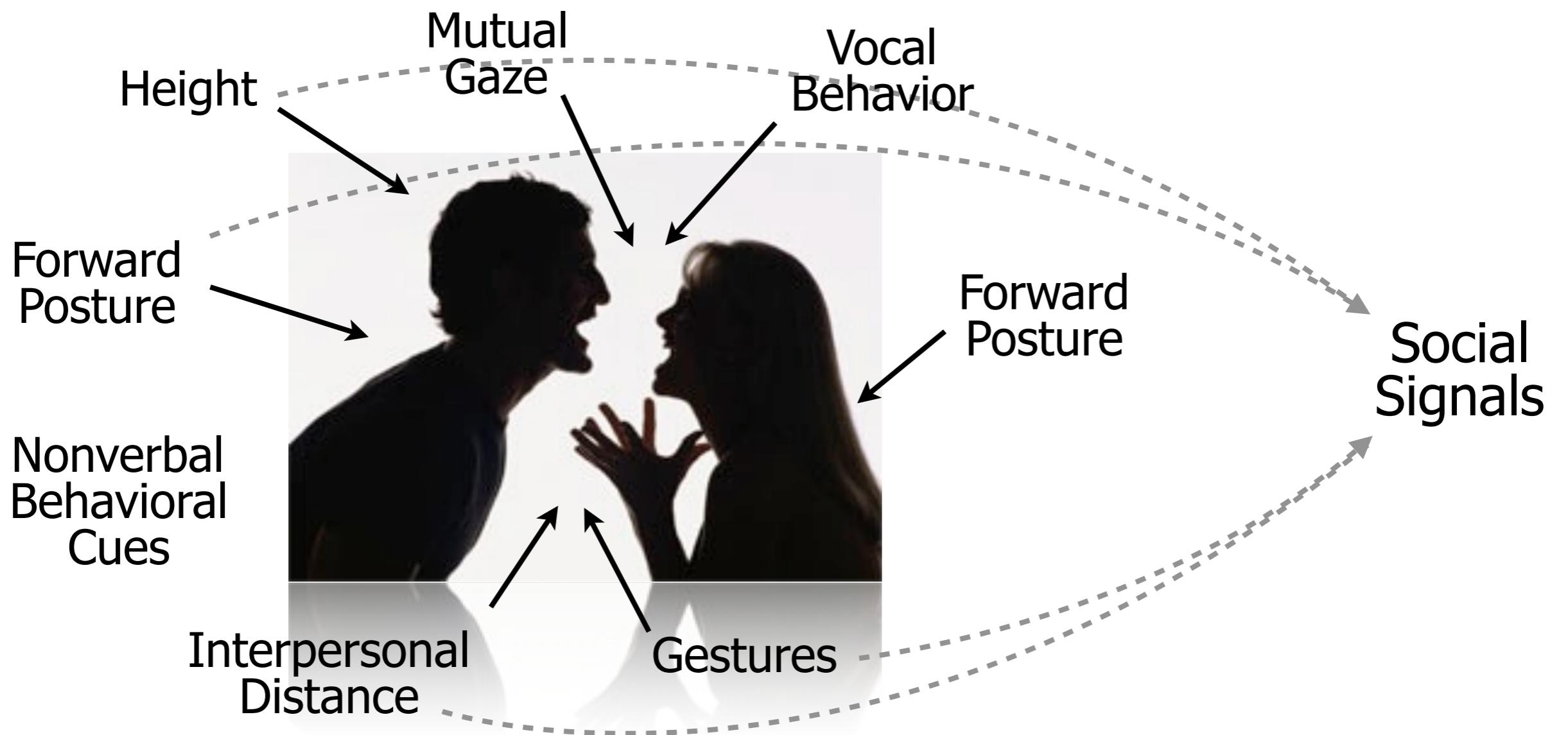
- Ekman and Friesen, “Measuring Facial Movements”, Environmental Psychology and Nonverbal Behavior, 1(1):56-75, 1976.
- Baltrušaitis, Robinson, and Morency, “Openface: an open source facial behavior analysis toolkit”, IEEE Winter Conference on Applications of Computer Vision, 2016.

Outline

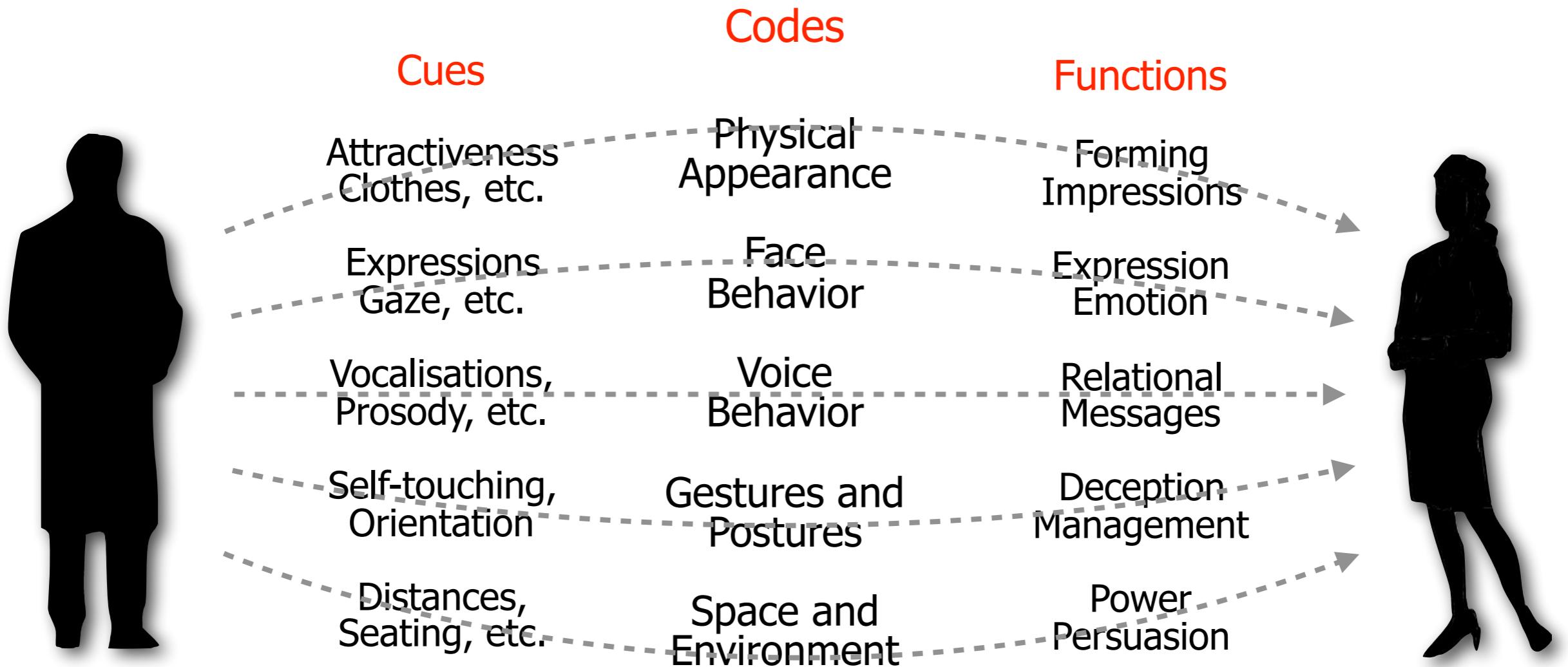
- Nonverbal Communication
- Facial Expressions
- Action Units
- Facial Expression Analysis
- Conclusions

Outline

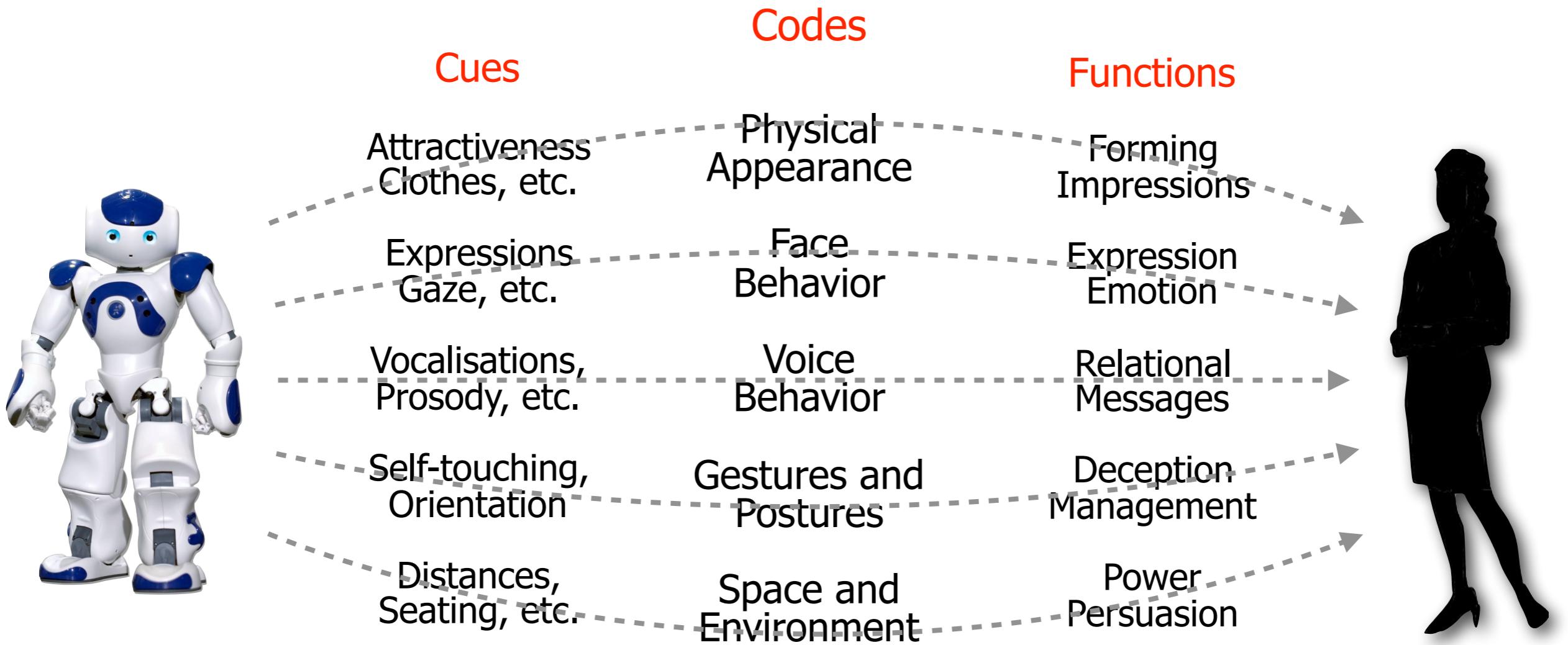
- Nonverbal Communication
- Facial Expressions
- Action Units
- Facial Expression Analysis
- Conclusions



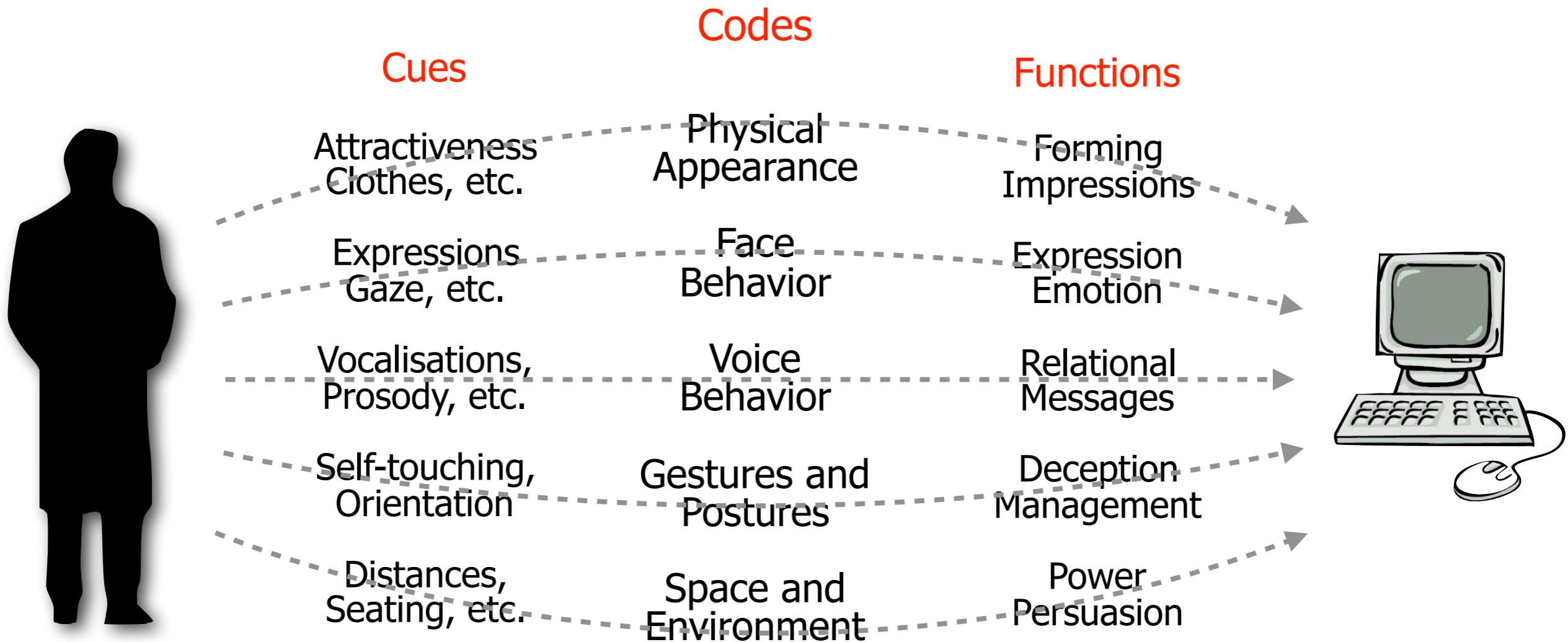
Vinciarelli, Pantic and Bourlard, "Social Signal Processing: Survey of an Emerging Domain", Journal of Image and Vision Computing, 27(12):1743-1759, 2009



Richmond and McCroskey, "Nonverbal Behaviors in Interpersonal Relations",
Allyn and Bacon, 1995



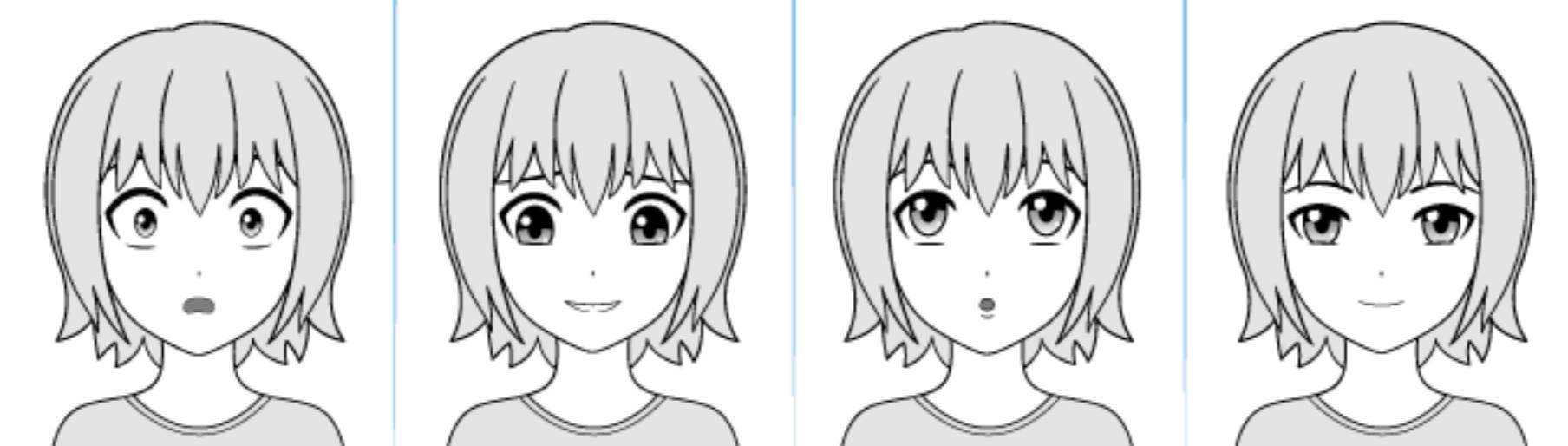
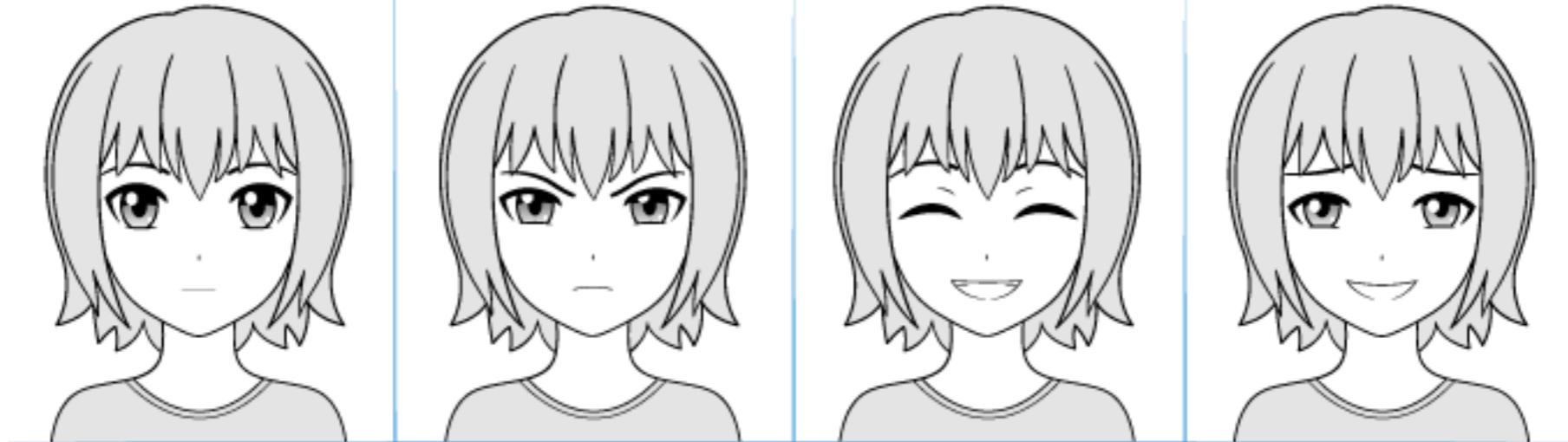
Richmond and McCroskey, "Nonverbal Behaviors in Interpersonal Relations",
Allyn and Bacon, 1995



Richmond and McCroskey, "Nonverbal Behaviors in Interpersonal Relations",
Allyn and Bacon, 1995

Outline

- Nonverbal Communication
- **Facial Expressions**
- Action Units
- Facial Expression Analysis
- Conclusions



<https://www.pinterest.co.uk/pin/214906213452277482/>

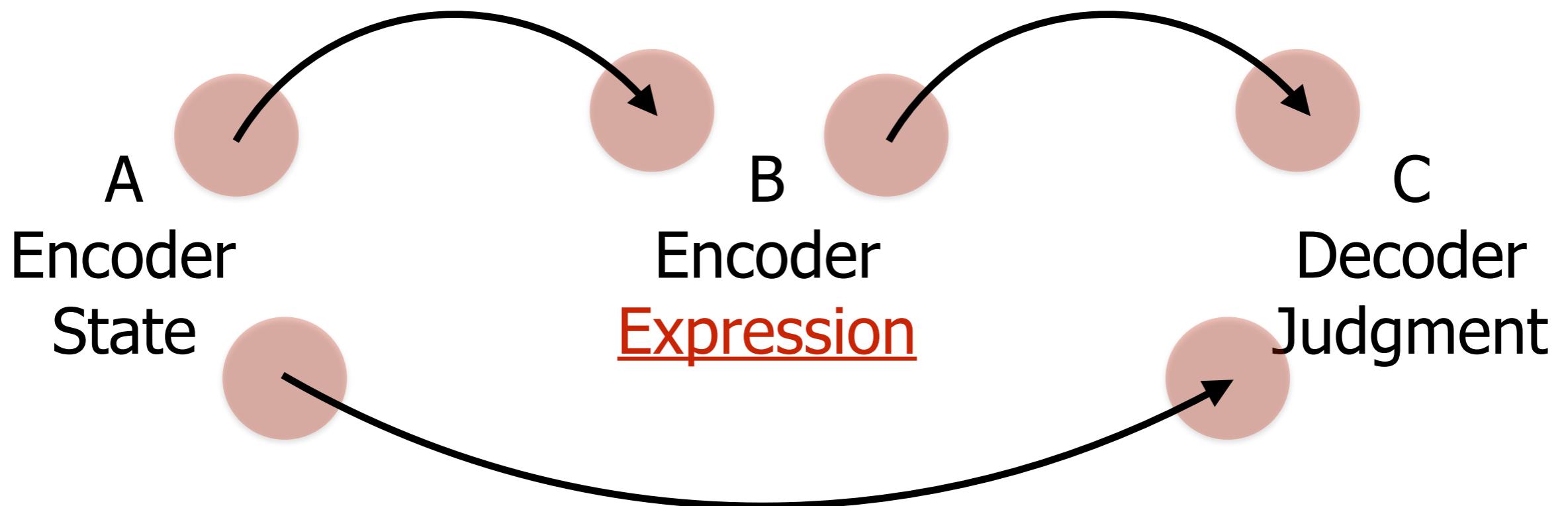
Facial Expressions

“Which movements signal emotion? [...] Do the same facial movements occur in the same social contexts in different cultures? Are certain facial actions inhibited in certain social settings? Which facial movements punctuate conversation, etc.?”

Ekman and Friesen, “Measuring Facial Movements”, Environmental Psychology and Nonverbal Behavior, 1(1):56-75, 1976

How does the encoder manifest her/his state through her/his expressions?

How do decoders interpret the expressions of the encoder



What is the state the decoder attributes to the encoder?

Facial Action Code (I)

"Our primary goal in developing the Facial Action Code (FAC) was to develop a comprehensive system which could distinguish all possible visually distinguishable facial movements."

Ekman and Friesen, "Measuring Facial Movements", Environmental Psychology and Nonverbal Behavior, 1(1):56-75, 1976

Facial Action Code (II)

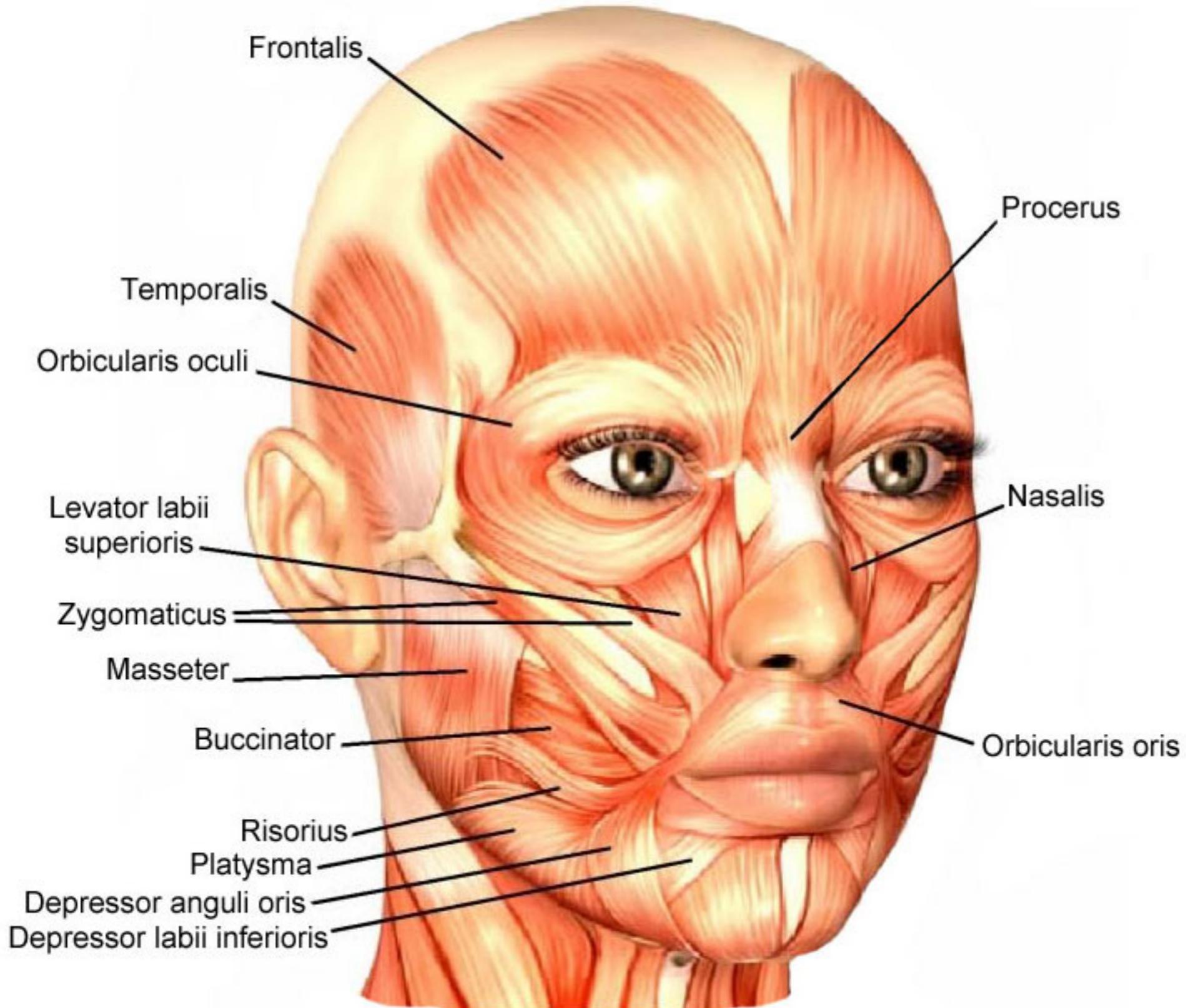
“[...] it deals with what is clearly visible in the face, ignoring invisible changes [...] based on our interest in what could have social consequences [...] could be applied to any record of behaviour [...]”

Ekman and Friesen, “Measuring Facial Movements”, Environmental Psychology and Nonverbal Behavior, 1(1):56-75, 1976

Facial Action Code (III)

“We chose to derive FAC from an analysis of the anatomical basis of facial movement. Since every facial movement is the result of muscular action, a comprehensive system could be obtained by discovering how each muscle of the face acts to change visible appearance.”

Ekman and Friesen, “Measuring Facial Movements”, Environmental Psychology and Nonverbal Behavior, 1(1):56-75, 1976

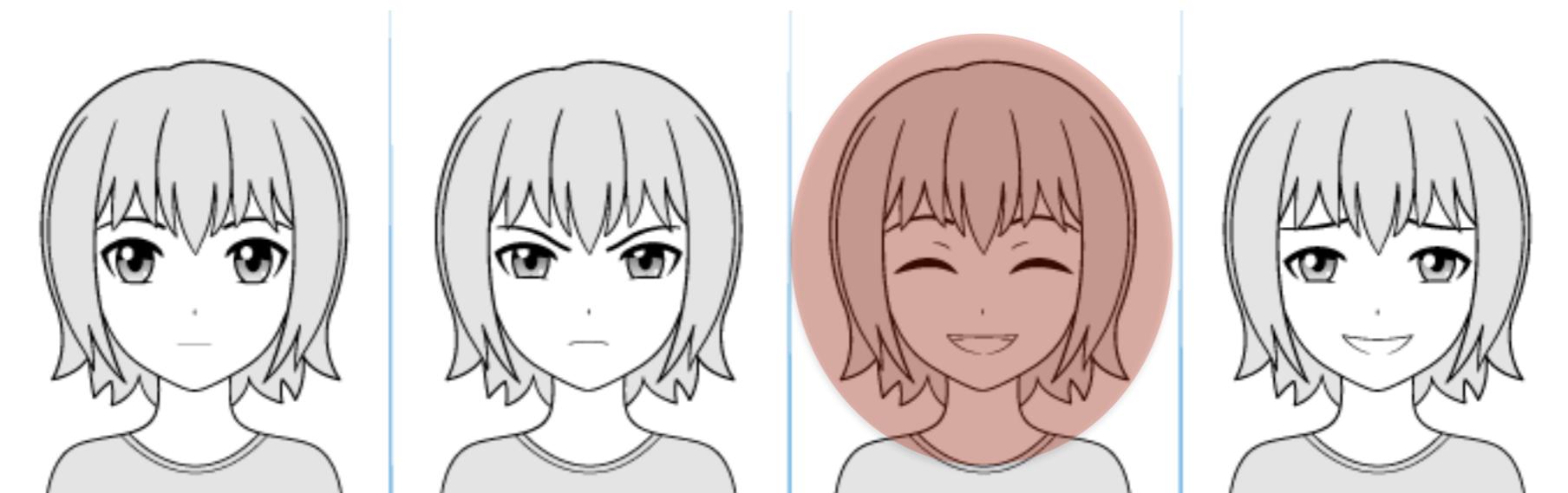


Facial Action Code (III)

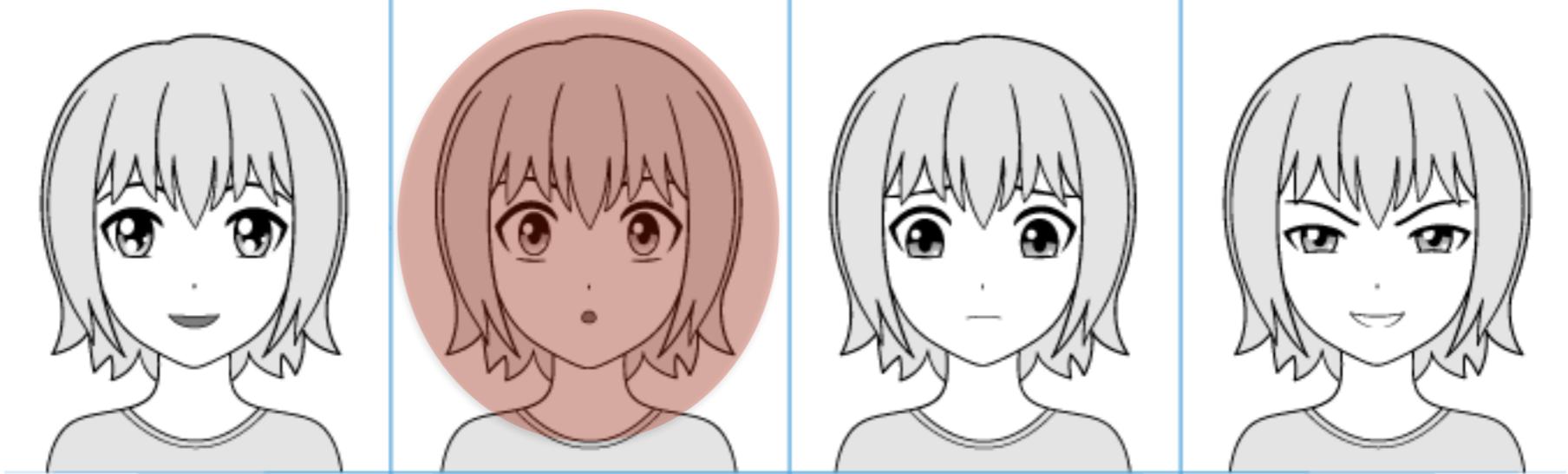
“[...] separate inference from description. We are interested in determining which facial behavior is playful, or puzzled, or sad, but such inferences about underlying state, antecedent, or consequent actions should rest upon evidence.”

Ekman and Friesen, “Measuring Facial Movements”, Environmental Psychology and Nonverbal Behavior, 1(1):56-75, 1976

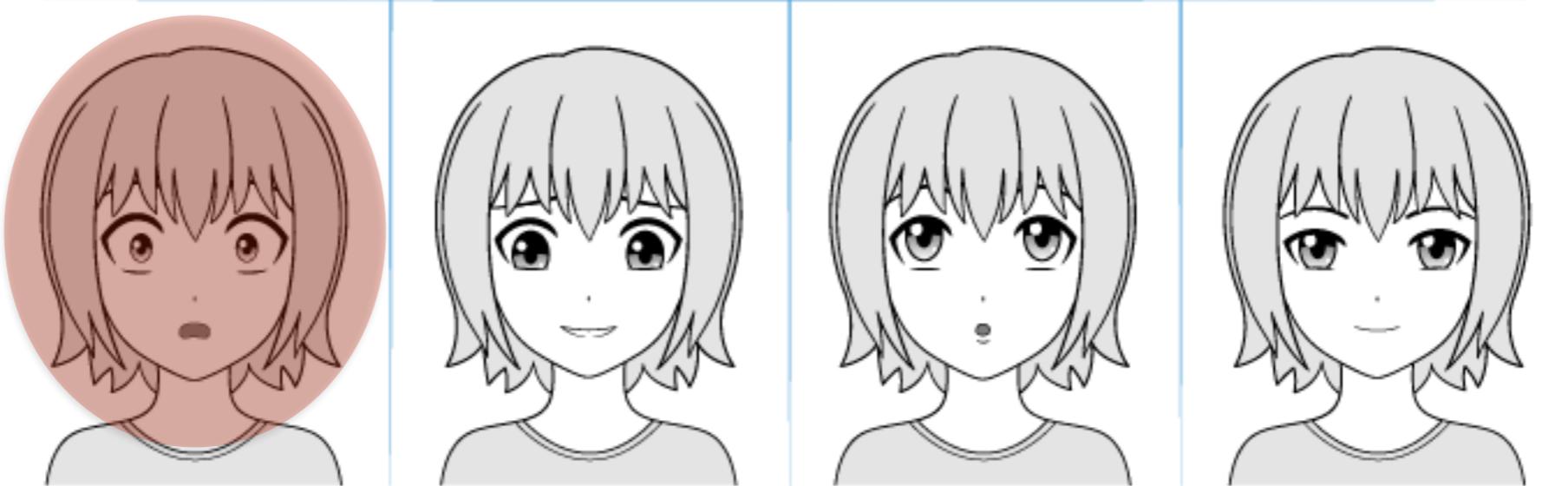
Happiness



Surprise



Fear



<https://www.pinterest.co.uk/pin/214906213452277482/>

Facial Action Code (III)

"[...] most of the appearance changes were additive. The characteristic appearance of each of the two AUs was clearly recognizable and virtually unchanged. There were a few AU combinations which were not additive."

Ekman and Friesen, "Measuring Facial Movements", Environmental Psychology and Nonverbal Behavior, 1(1):56-75, 1976

The Likelihood of a muscle activation pattern given a certain expression

Every component accounts for the activation of a muscle

$$p(\vec{x}|\mathcal{C}_k) = p(x_1, x_2, \dots, x_D | \mathcal{C}_k)$$

The diagram illustrates the decomposition of a vector \vec{x} into its components x_1, x_2, \dots, x_D . A large oval labeled $p(\vec{x}|\mathcal{C}_k)$ contains the equation. Below the equation, a diagonal arrow points from the left towards the first term $p(x_1|\mathcal{C}_k)$. From each term x_i in the sequence, a vertical arrow points downwards to a corresponding oval labeled $x_i|\mathcal{C}_k$.

The class can corresponds to a inner state (e.g., happy or sad)

The number of Action Units

The probability of a vector is the joint probability of its components

Likelihood of an individual component

$$p(x_1, x_2, \dots, x_D | C_k) = \prod_{i=1}^D p(x_i | C_k)$$

It is true if the components are statistically independent given the class

The Naive Bayes classifier

Outline

- Nonverbal Communication
- Facial Expressions
- Action Units
- Facial Expression Analysis
- Conclusions

Inner Brow Raiser (AU1)

Outer Brow Raiser (AU2)

Brow Lowerer (AU4)

Upper Lid Raiser (AU5)

Cheek Raiser (AU6)

Lid Tightener (AU7)

Nose Wrinkler (AU9)

Upper Lid Raiser (AU10)

Nasolabial Fold Deepener (AU11)

Lip Corner Puller (AU12)

Cheek Puffer (AU13)

Dimpler (AU14)

Lip Corner Depressor (AU15)

Lower Lip Depressor (AU16)

Chin Raiser (AU17)

Lip Puckerer (AU18)

Lip Stretcher (AU20)

Lip Funneler (AU22)

Lip Tightner (AU23)

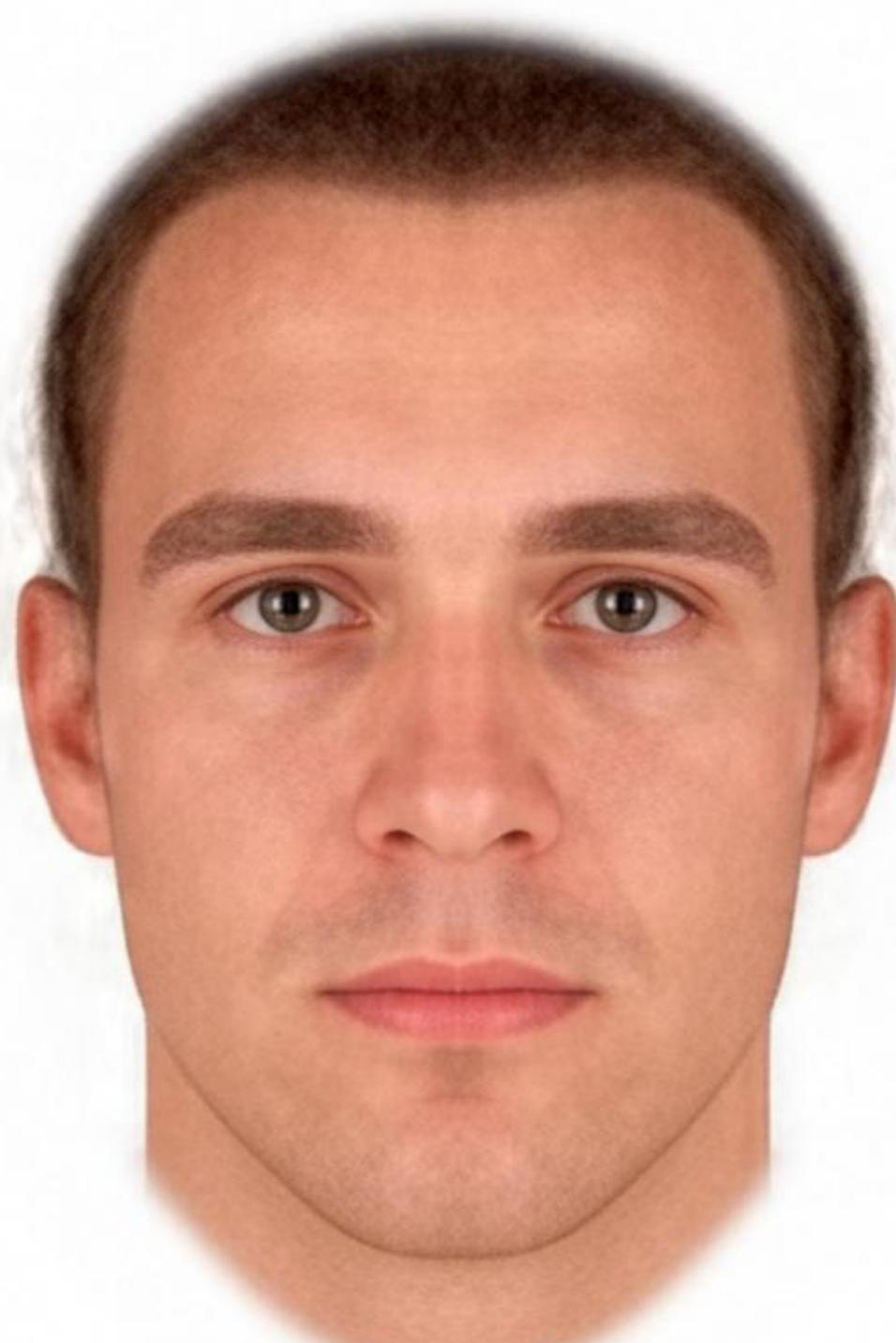
Lip Pressor (AU24)

Lips Part (AU25)

Jaw Drop (AU26)

Mouth Stretch (AU27)

Lip Suck (AU28)



Tongue Out (AU19)

Neck Tightener (AU21)

Jaw Thrust (AU29)

Jaw Sideways (AU30)

Jaw Clencher (AU31)

Lip Bite (AU32)

Cheek Blow (AU33)

Cheek Puff (AU34)

Cheek Suck (AU35)

Tongue Bulge (AU36)

Lip Wipe (AU37)

Nostril Dilator (AU38)

Nostril Compressor (AU39)

Lid Droop (AU41)

Slit (AU42)

Eyes Closed (AU43)

Squint (AU44)

Blink (AU45)

Wink (AU46)

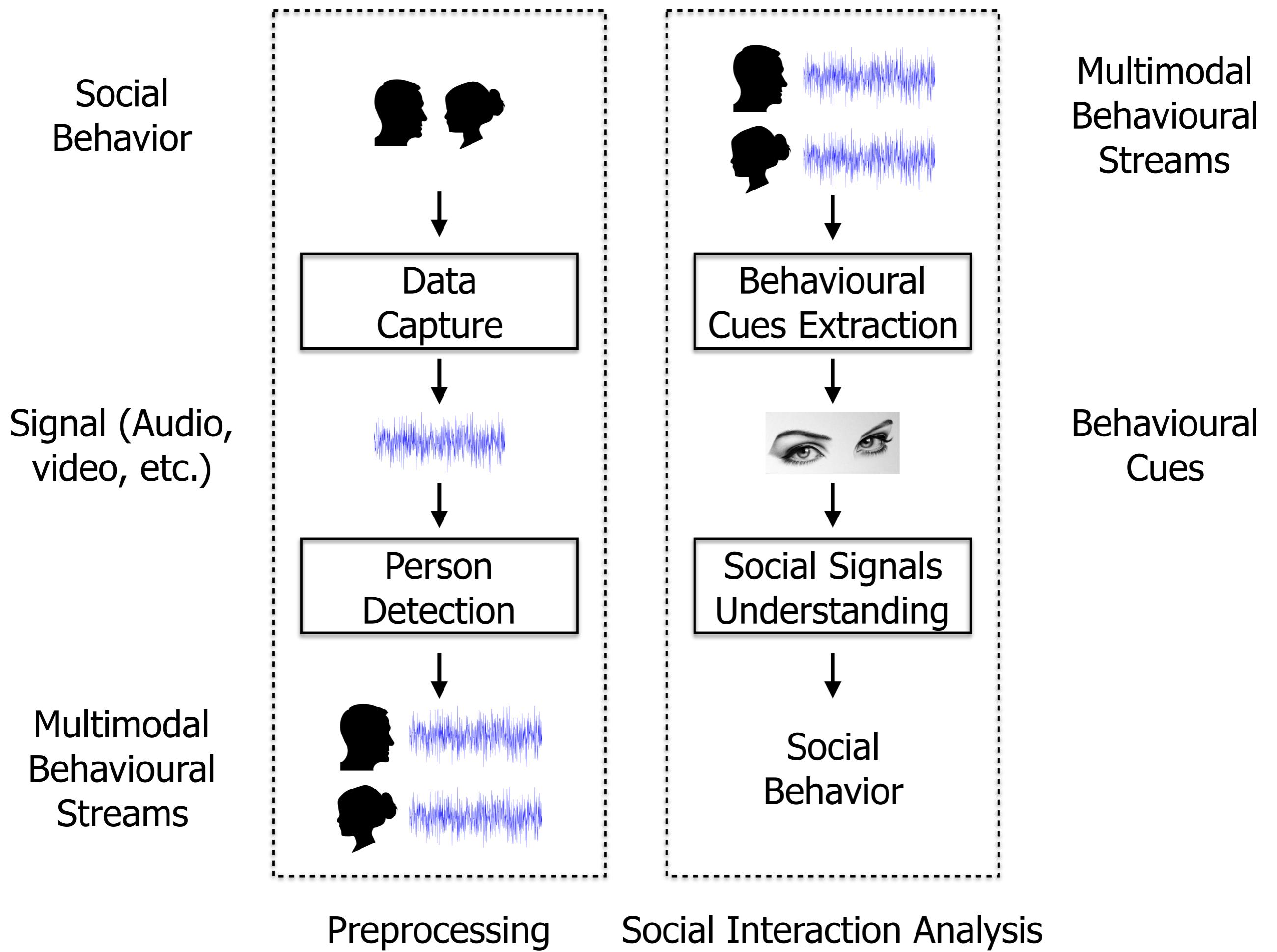
- <https://imotions.com/blog/facial-action-coding-system/>

Recap

- An Action Unit (AU) corresponds to the activation of one or more muscles (it is an “atom” of facial movement);
- A facial expression is a combination of some of the Action Units identified by Ekman and Friesen;
- Different facial expressions convey different social and psychological messages.

Outline

- Nonverbal Communication
- Facial Expressions
- Action Units
- **Facial Expression Analysis**
- Conclusions



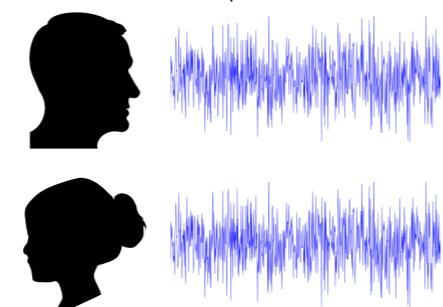
Facial Expressions

Video

Behavioural Streams

Video Recording

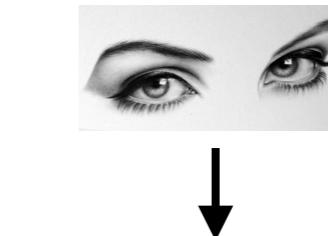
Face Detection



Landmarks and AUs Extraction

Facial Expression Understanding

Facial Expressions



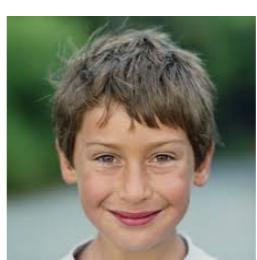
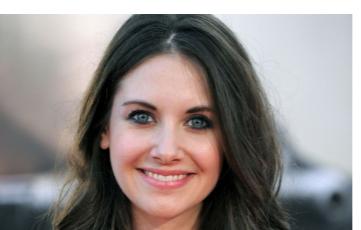
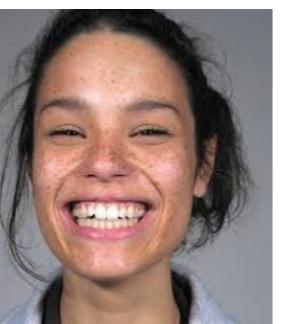
Preprocessing

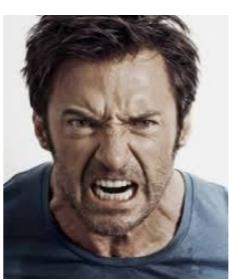
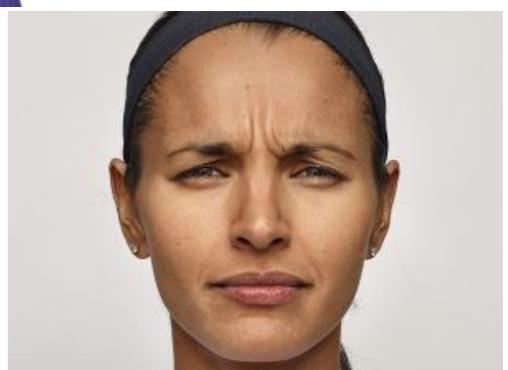
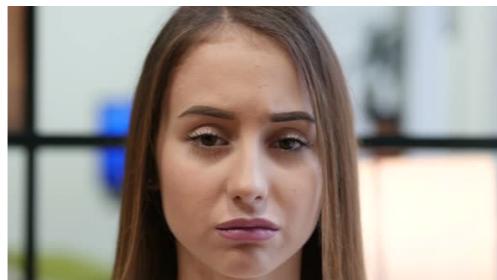
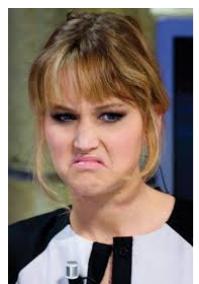
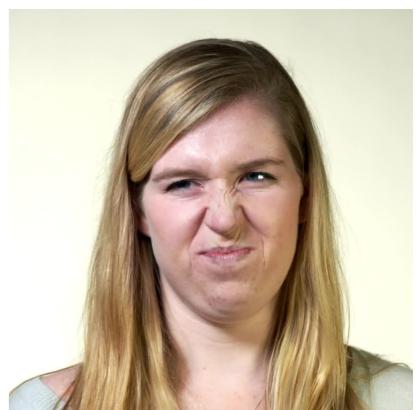
Social Interaction Analysis

Video Behavioural Streams

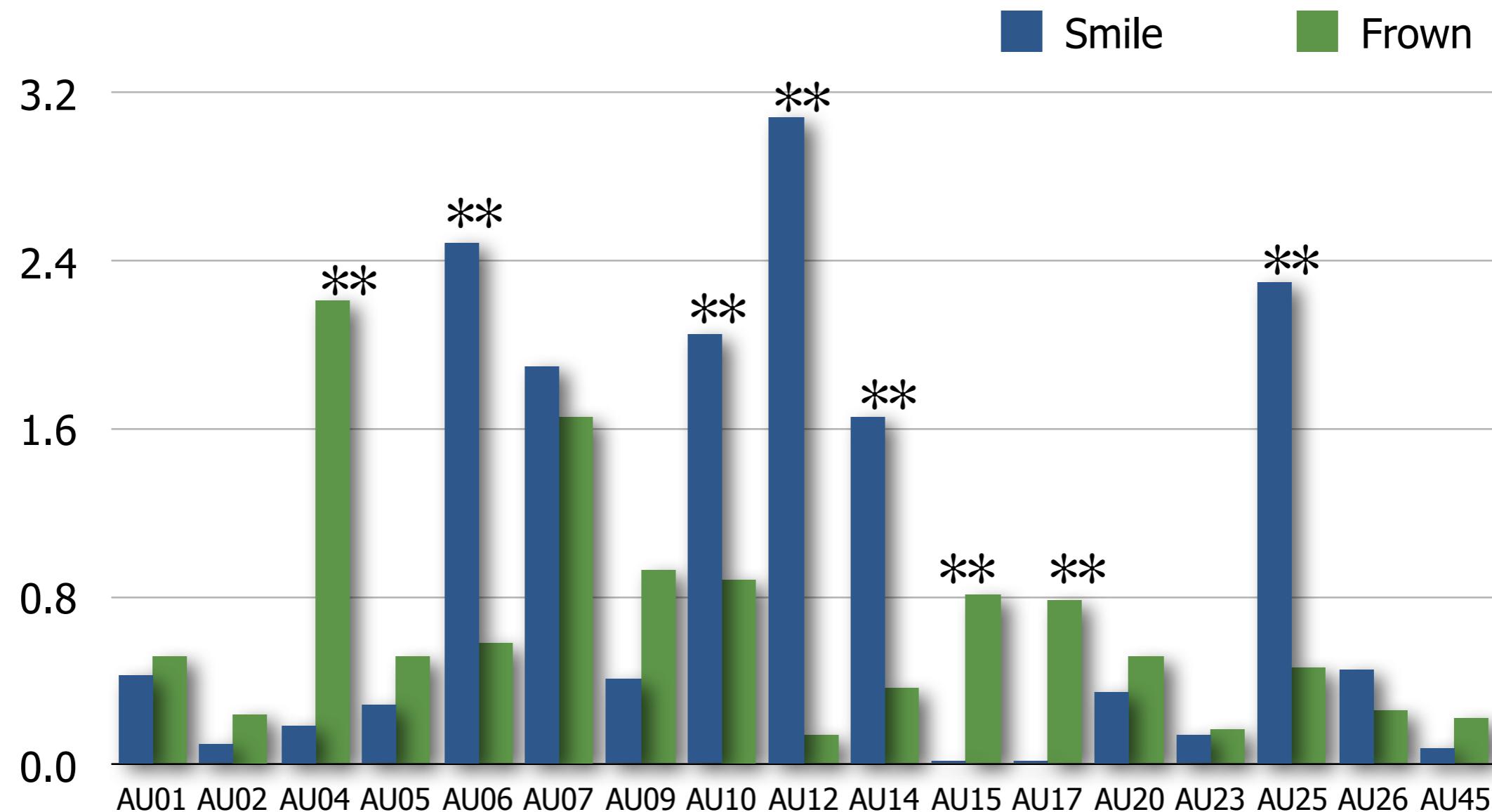
Behavioural Cues

<https://www.youtube.com/watch?v=QDQff5d1NJM>





9.Run Tests of your Hypotheses.



Outline

- Nonverbal Communication
- Facial Expressions
- Action Units
- Facial Expression Analysis
- Conclusions

Conclusions

- The analysis of facial expression is based on **Action Units**, movements associated with changes in facial appearance;
- There is a gap between the Action Units being displayed and the interpretation of an expression;
- Machines can automatically detect (and possibly interpret) Action Units.

Basic Signal Processing (I)

Computational Social Intelligence - Lecture 16

Prof. Alessandro Vinciarelli
School of Computing Science &
Institute of Neuroscience and Psychology

<http://www.dcs.gla.ac.uk/vincia>
Alessandro.Vinciarelli@glasgow.ac.uk



University
of Glasgow

EPSRC
Engineering and Physical Sciences
Research Council

FNSNF

Texts (see Moodle)

This lecture is based on the following text
(available on Moodle):

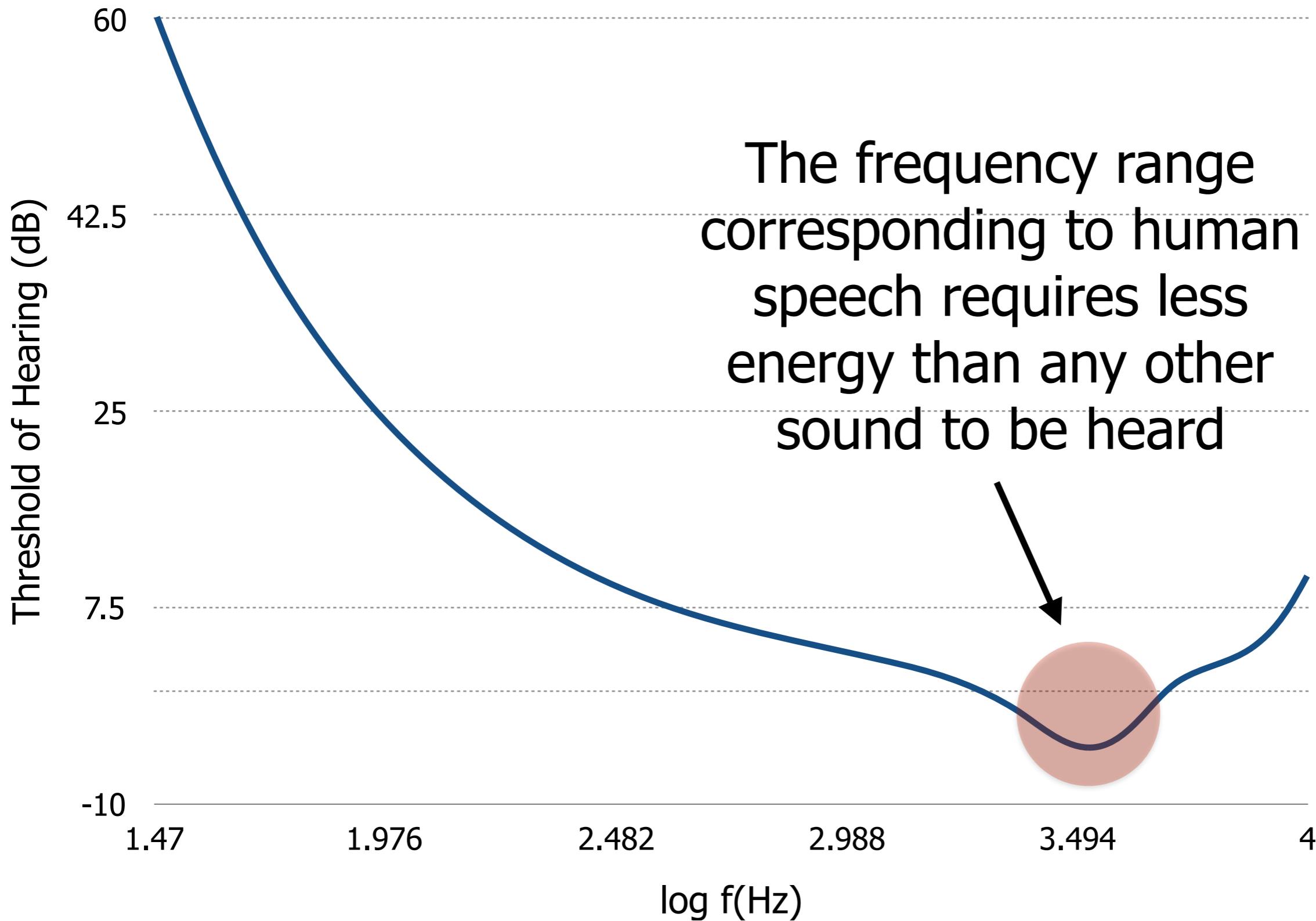
- F.Camastra and A.Vinciarelli, "Machine Learning for Audio, Image and Video Processing", Springer Verlag, Chapter 2, pp. 38-46, 2008.

Outline

- Introduction
- Time Domain Processing
- Short-Term Analysis
- Conclusions

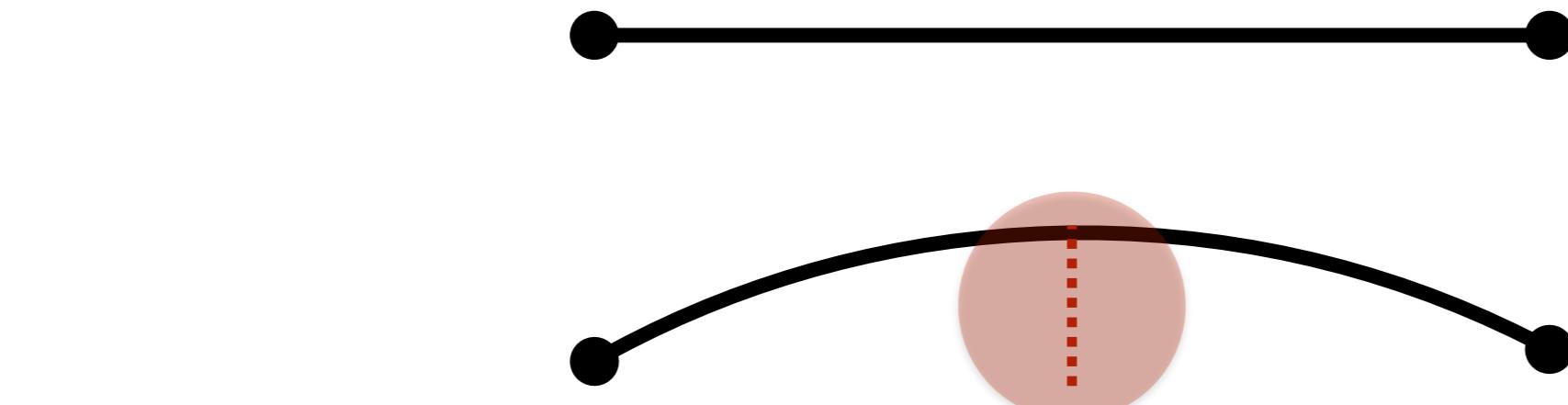
Outline

- Introduction
- Time Domain Processing
- Short-Term Analysis
- Conclusions

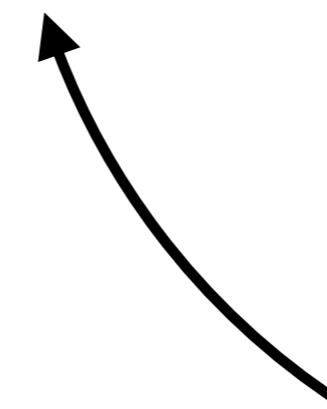


A microphone measures at regular time steps the oscillations of an elastic membrane

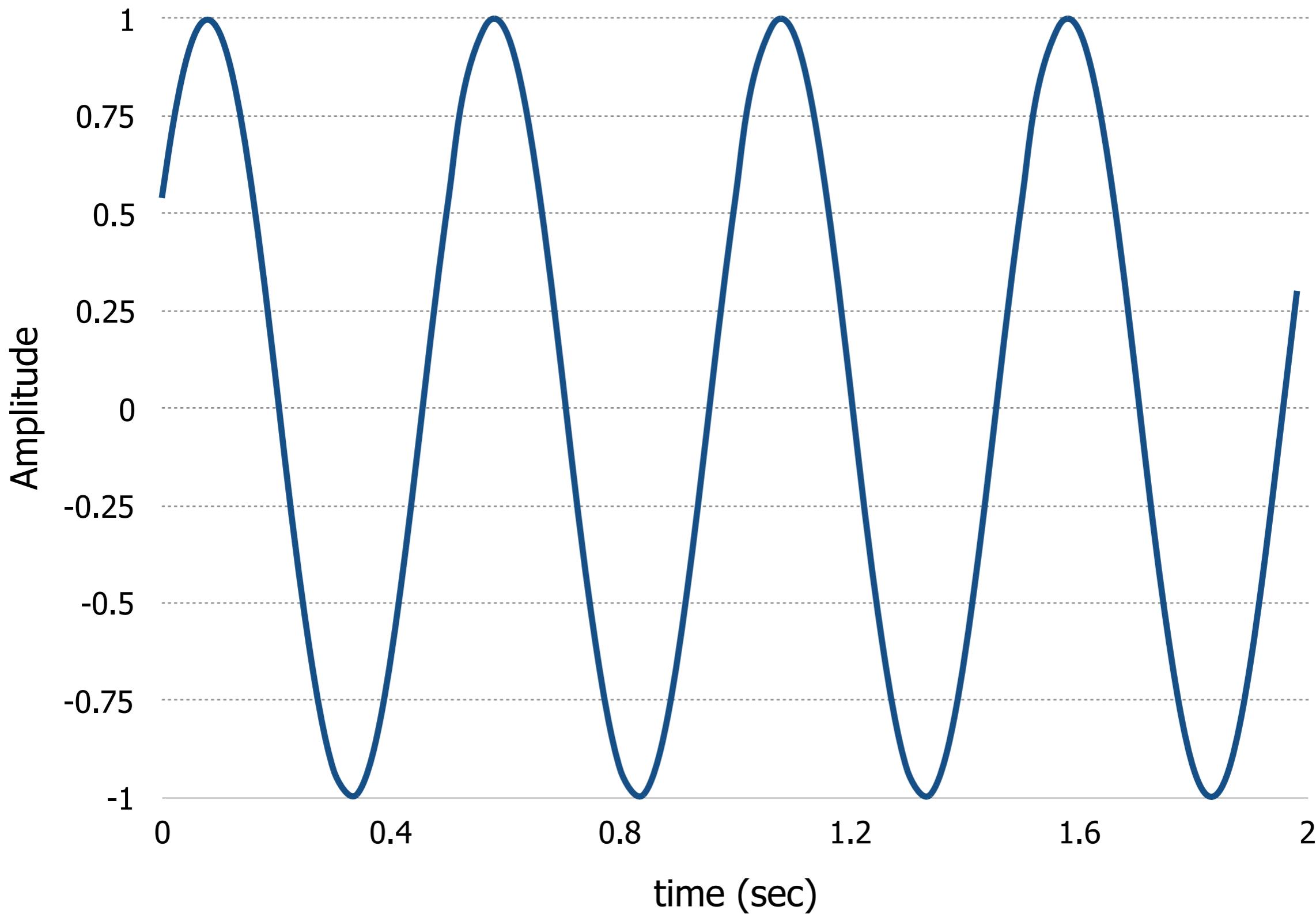
When there is no acoustic wave, the membrane is in equilibrium



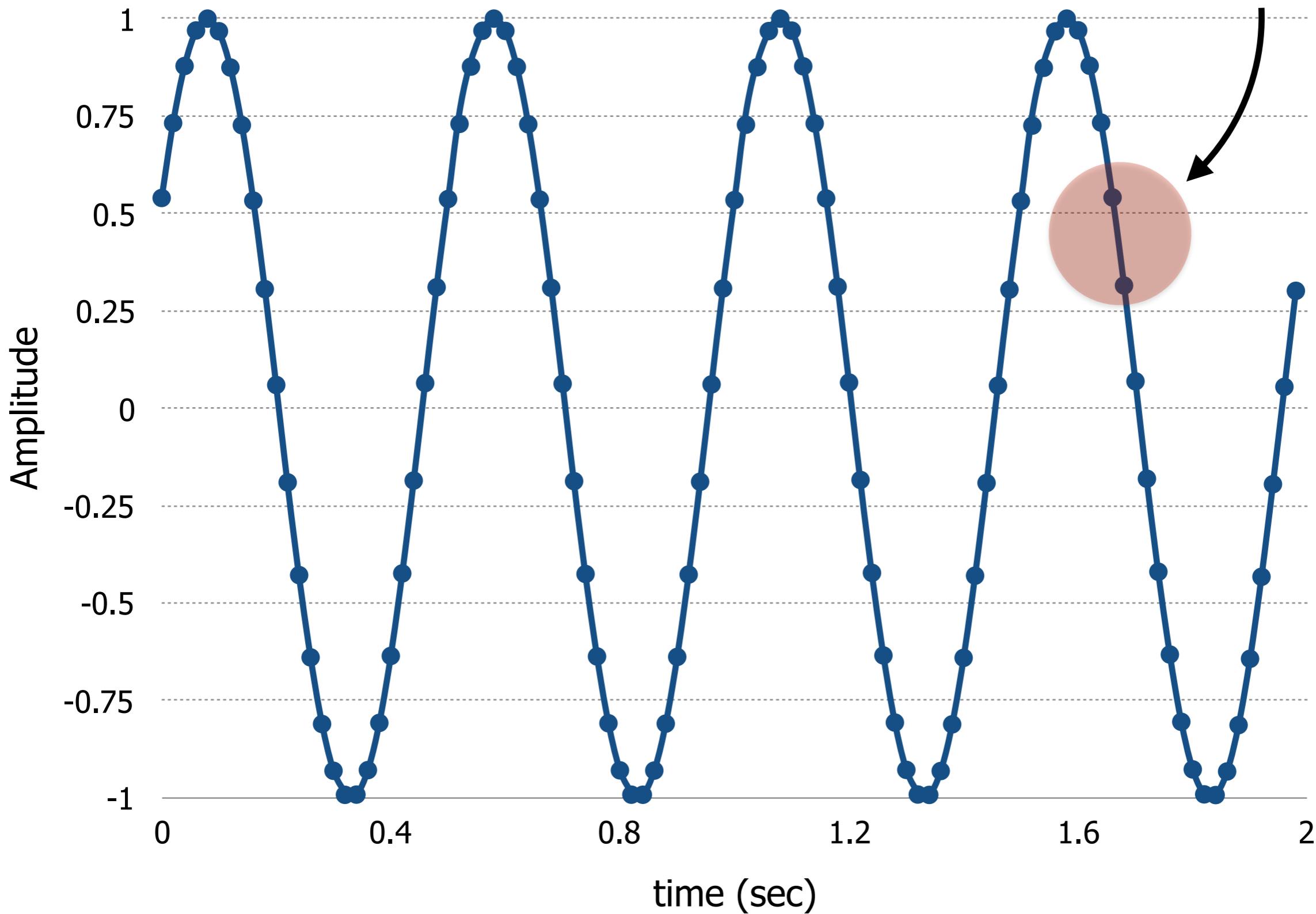
When there is an acoustic wave, the membrane oscillates



The amplitude of the oscillation is proportional to the amplitude of the acoustic wave



The signal is unknown
between two samples



The sample “k”

The signal at time “kT”

$$s[k] = s(kT)$$
$$k = -\infty, \dots, \infty$$

The index of the sample

The sampling period
(time interval between consecutive samples)

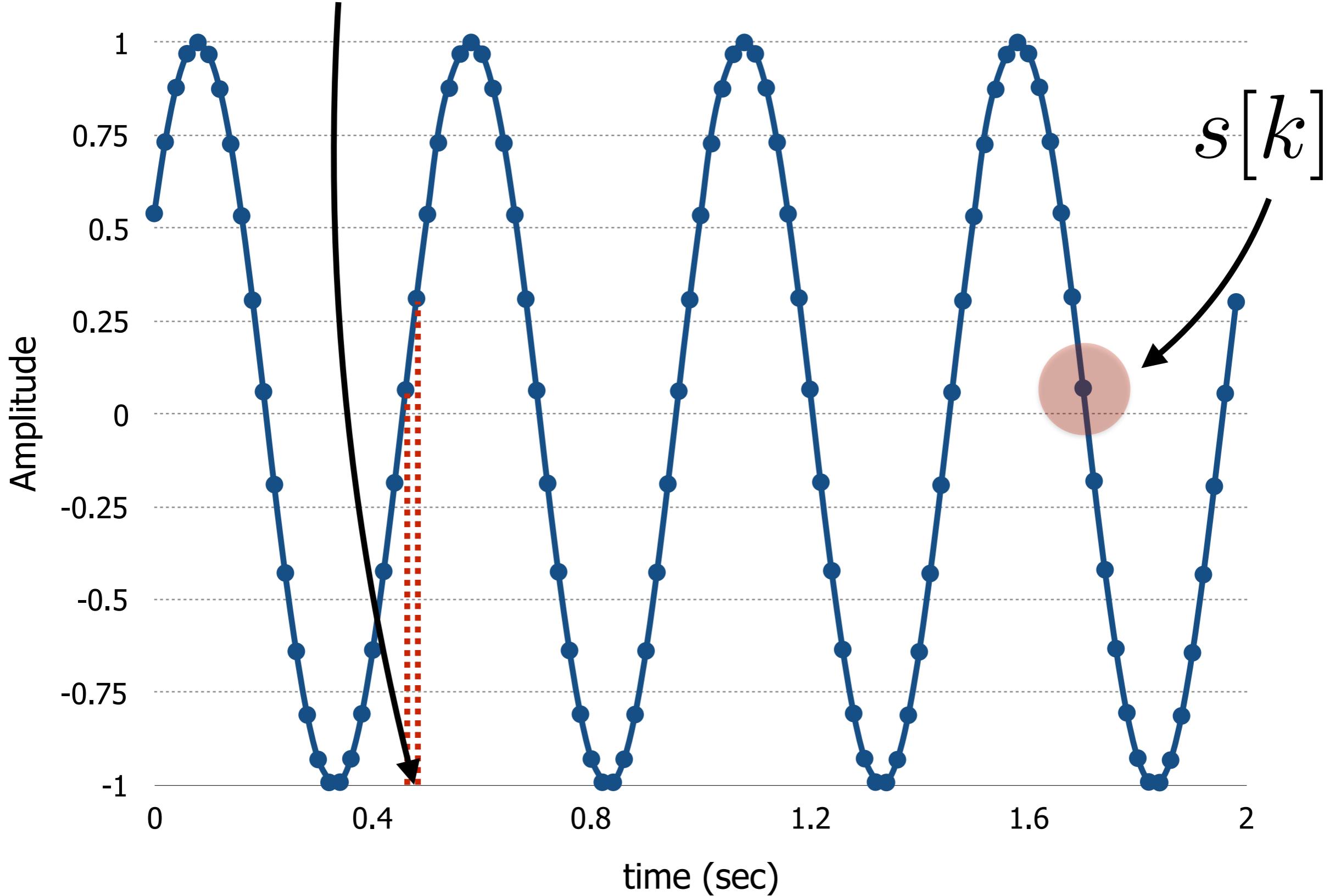
There are no infinite signals in reality

All samples with $k < 0$ are conventionally null

$$s[k] = 0 : k < 0$$
$$s[k] = 0 : k > K$$

All samples with $k < K$ (where K is the time the last measurement is done) are conventionally null

The sampling period



The sampling frequency
is the inverse of the
sampling period

It must be at least twice
as much as the highest
frequency expected to
be observed in the
signal

$$F = \frac{1}{T}$$

It is the number of
samples per unit of time

Recap

- A signal is a measurable physical quantity that changes over time and it is continuous;
- A digital signal is a sequence of physical measurements collected at regular time steps (the sampling period) and it is discrete;
- Nothing can be said about what happens in the signal between two consecutive samples.

Outline

- Introduction
- Time Domain Processing
- Short-Term Analysis
- Conclusions

The convolution
between the signal “s”
and the window “w”

The sum extends over
all samples of the signal

$$y[k] = \sum_{n=-\infty}^{\infty} s[n]w[k - n]$$

The signal

The window

A diagram illustrating the convolution equation. The output variable $y[k]$ is shown in a red circle at the top left. An arrow points from it down to the summation symbol in the equation. The summation symbol is enclosed in a red circle. Below the symbol, the index $n = -\infty$ is written. Two arrows point from the text 'The signal' and 'The window' to the terms $s[n]$ and $w[k - n]$ respectively, which are also enclosed in red circles.

The rectangular window

$$w[n] = \begin{cases} 1 & : 0 \leq n \leq N-1 \\ 0 & : n < 0 \\ 0 & : n > N-1 \end{cases}$$

When the index is negative, the window signal is null

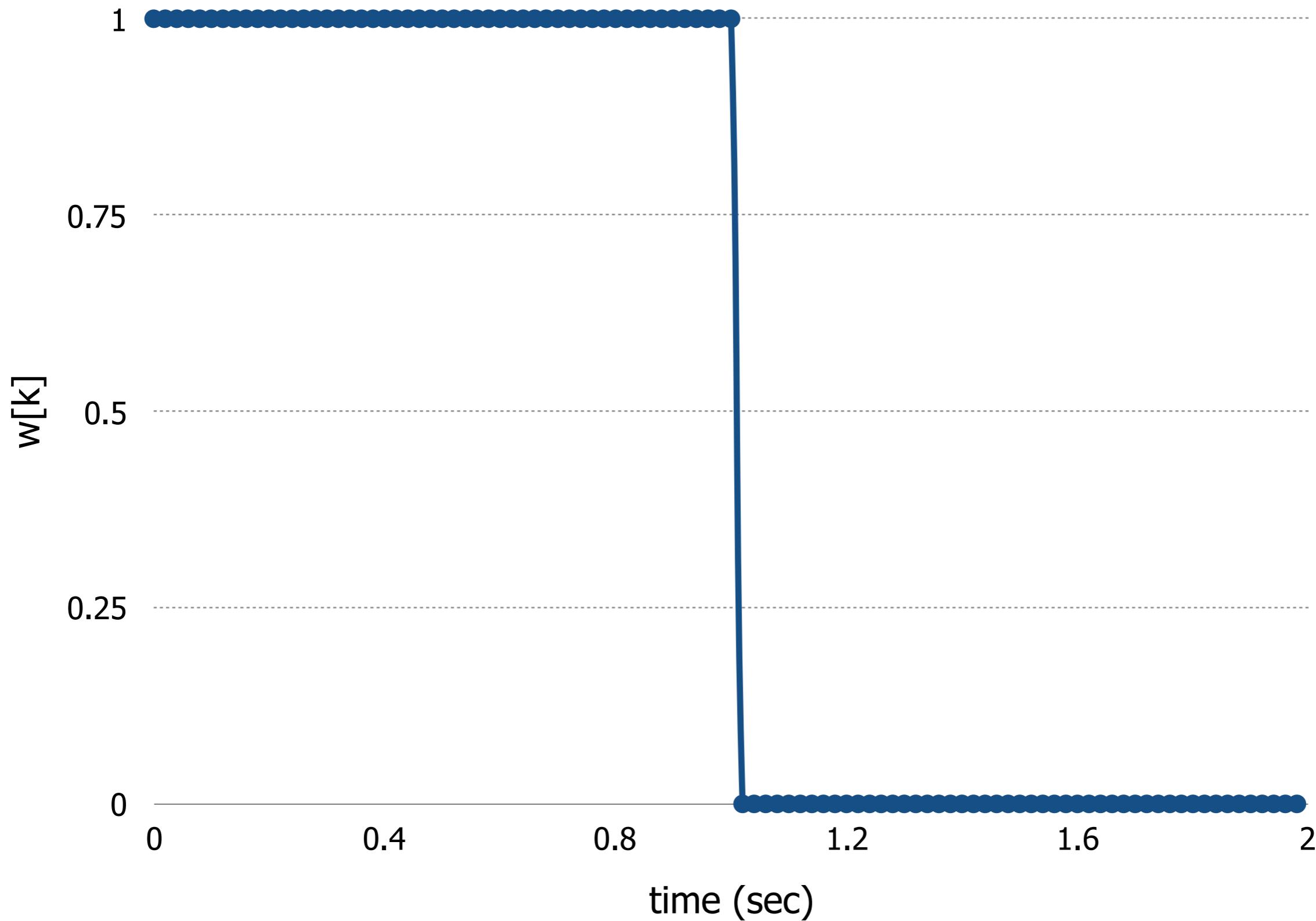
The window length

$$0 \leq n \leq N-1$$

$$n$$

$$n$$

When the index is larger than the length, the window signal is null



The convolution
between the signal “s”
and the window “w”

The sum extends over
all samples of the signal

$$y[k] = \sum_{n=-\infty}^{\infty} s[n]w[k - n]$$

The signal

The window

A diagram illustrating the convolution equation. The output $y[k]$ is shown in a pink circle at the bottom left. To its right is the summation symbol \sum inside a pink circle. Below the summation symbol is the range $n = -\infty$. To the right of the summation symbol is the product $s[n]w[k - n]$, which is also enclosed in a pink circle. Two arrows point from the text "The signal" and "The window" to the terms $s[n]$ and $w[k - n]$ respectively.

Any product in which
“n” is less than or equal
to “k” is null



$$k - n \geq 0 \Rightarrow n \leq k$$

$$k - n \leq N - 1 \Rightarrow n \geq k - N + 1$$



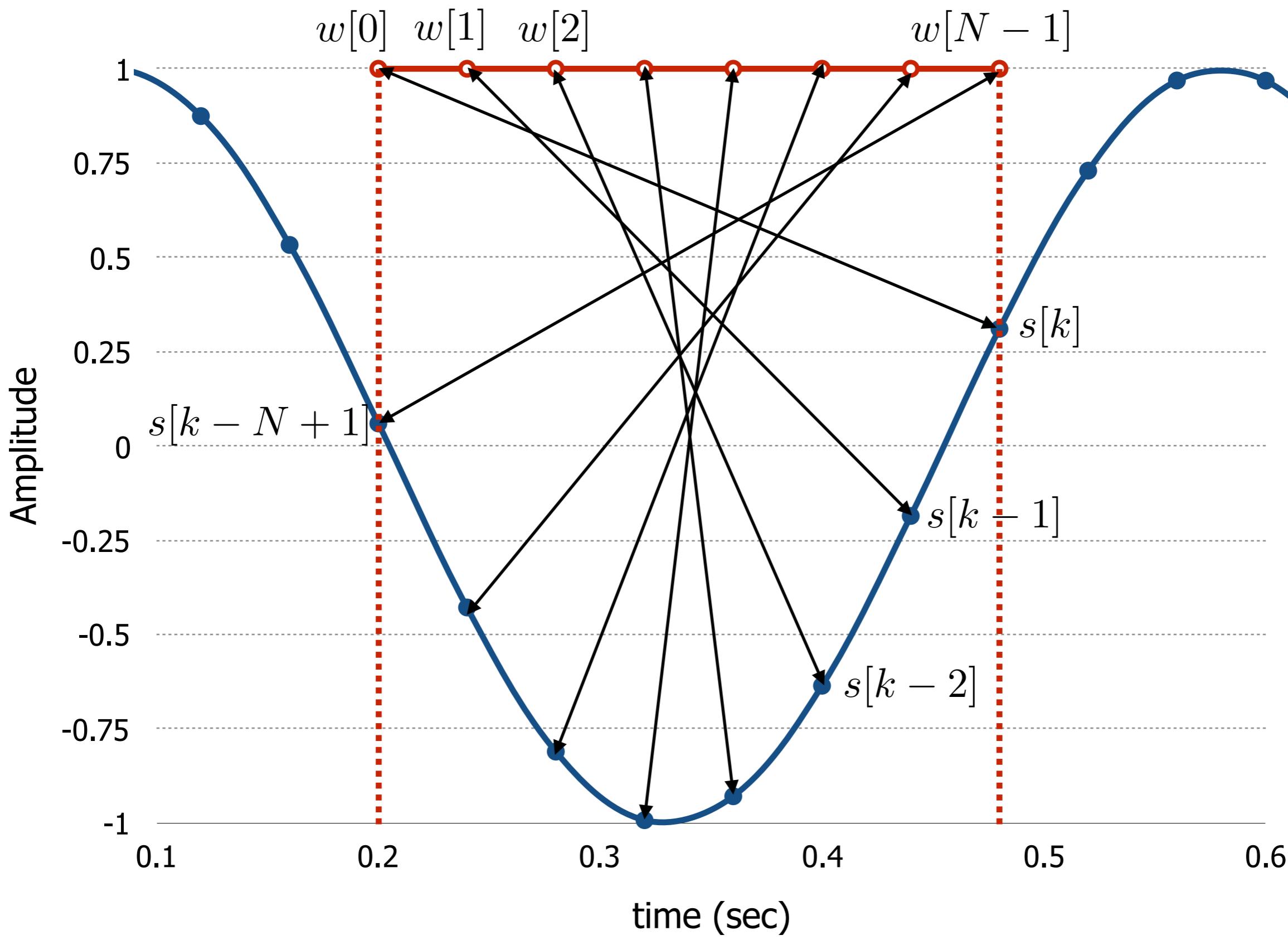
Only products in which
“n” is greater than or
equal to “k-N+1” is null

The convolution
between the signal “s”
and the window “w”

The sum extends over a
finite number of
samples

$$y[k] = \sum_{n=k-N+1}^k s[n]w[k-n]$$

$$\begin{aligned}y[k] &= \sum_{n=k-N+1}^k s[n]w[k-n] = \\&= s[k-N+1]w[N-1] + \\&\quad + s[k-N+2]w[N-2] + \dots \\&\quad + \dots + s[k-1]w[1] + s[k]w[0]\end{aligned}$$



The convolution
between the signal "s"
and the rectangular
window "w"

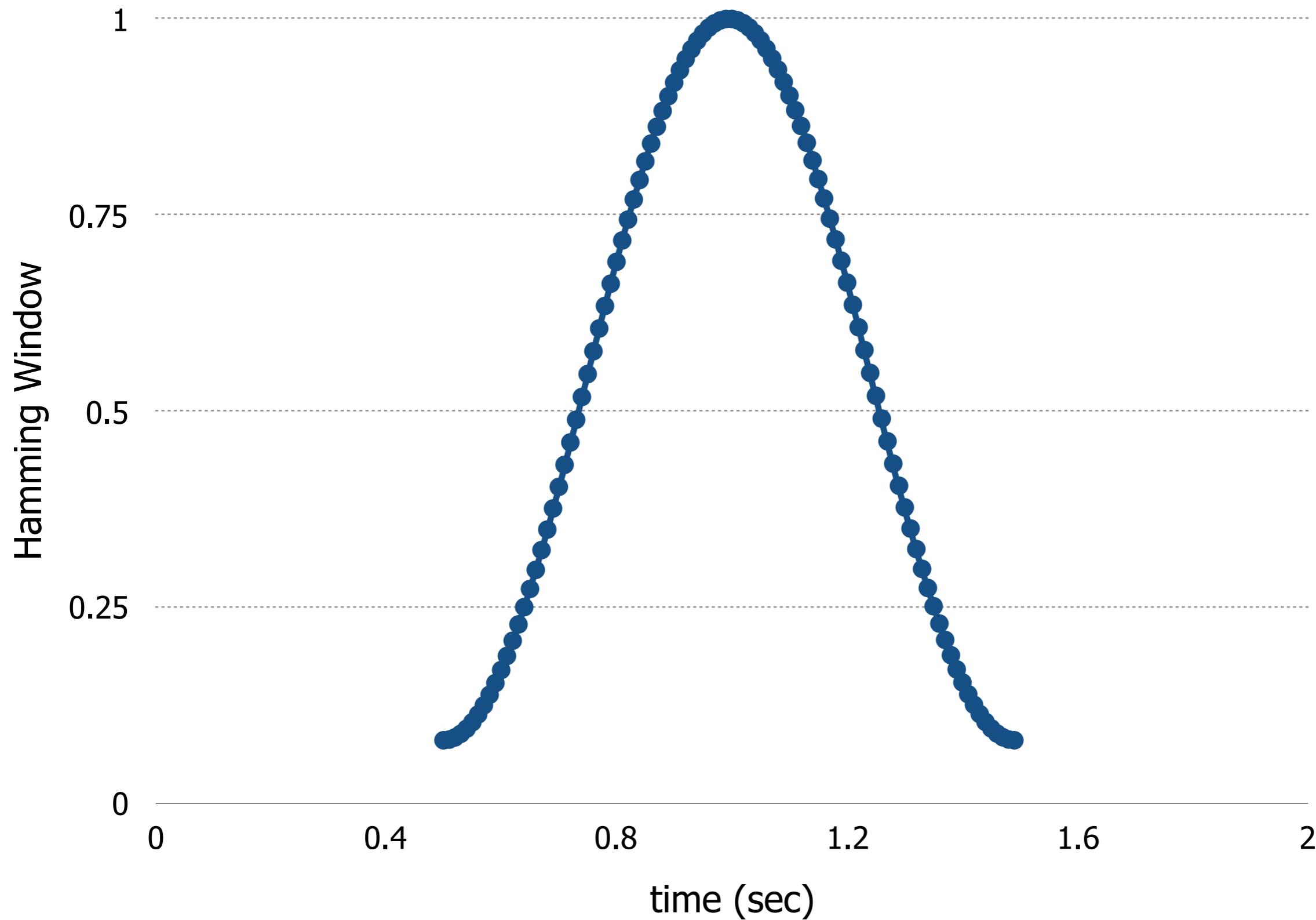
Sample "k" of the
convolution is the sum
of the samples between
"k-N+1" and "k"

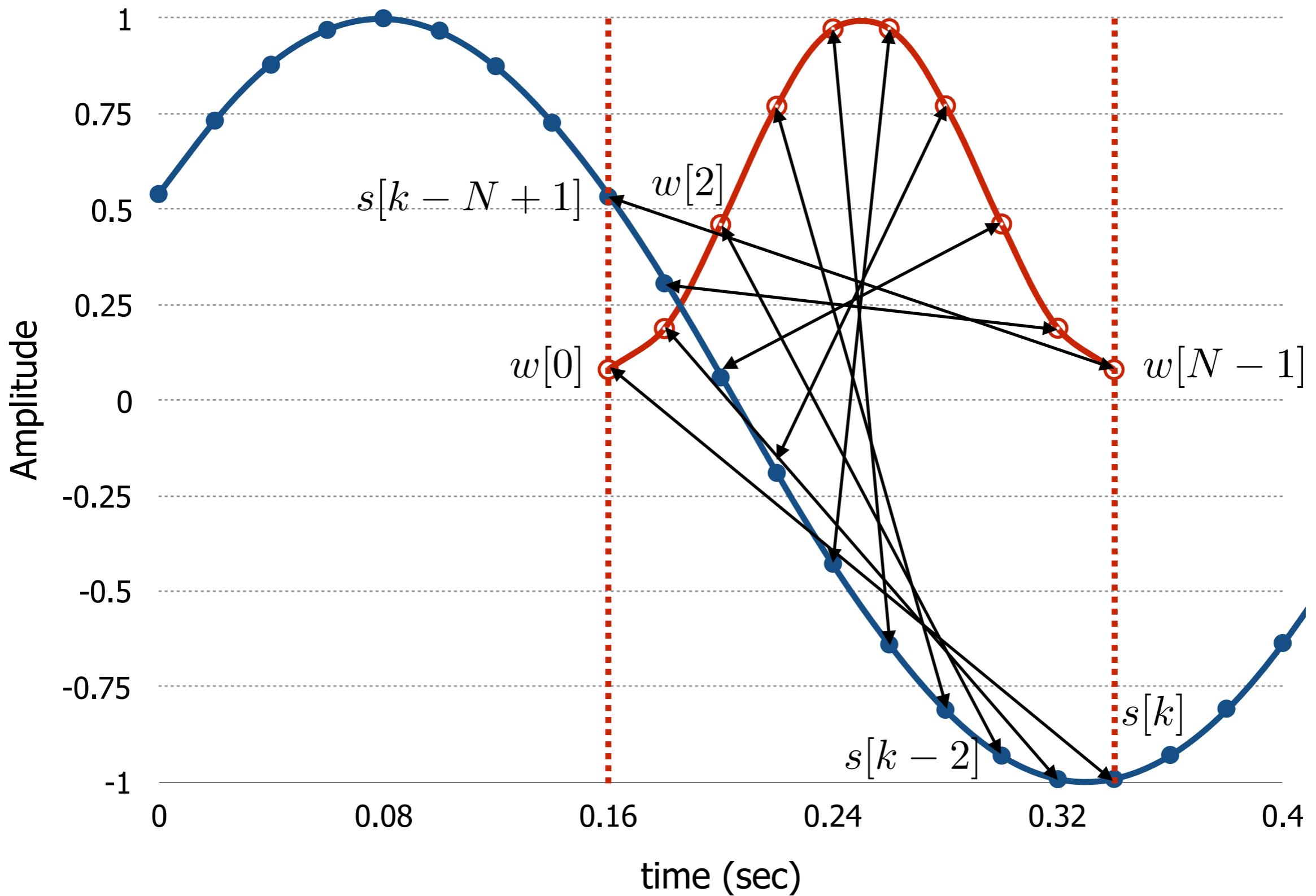
$$y[k] = \sum_{n=k-N+1}^k s[n]$$

Only samples for which
“n” is between 0 and
“N-1” are not null

The Hamming window

$$w[n] = \begin{cases} 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) & : 0 \leq n \leq N-1 \\ 0 & : n < 0 \\ 0 & : n > N-1 \end{cases}$$



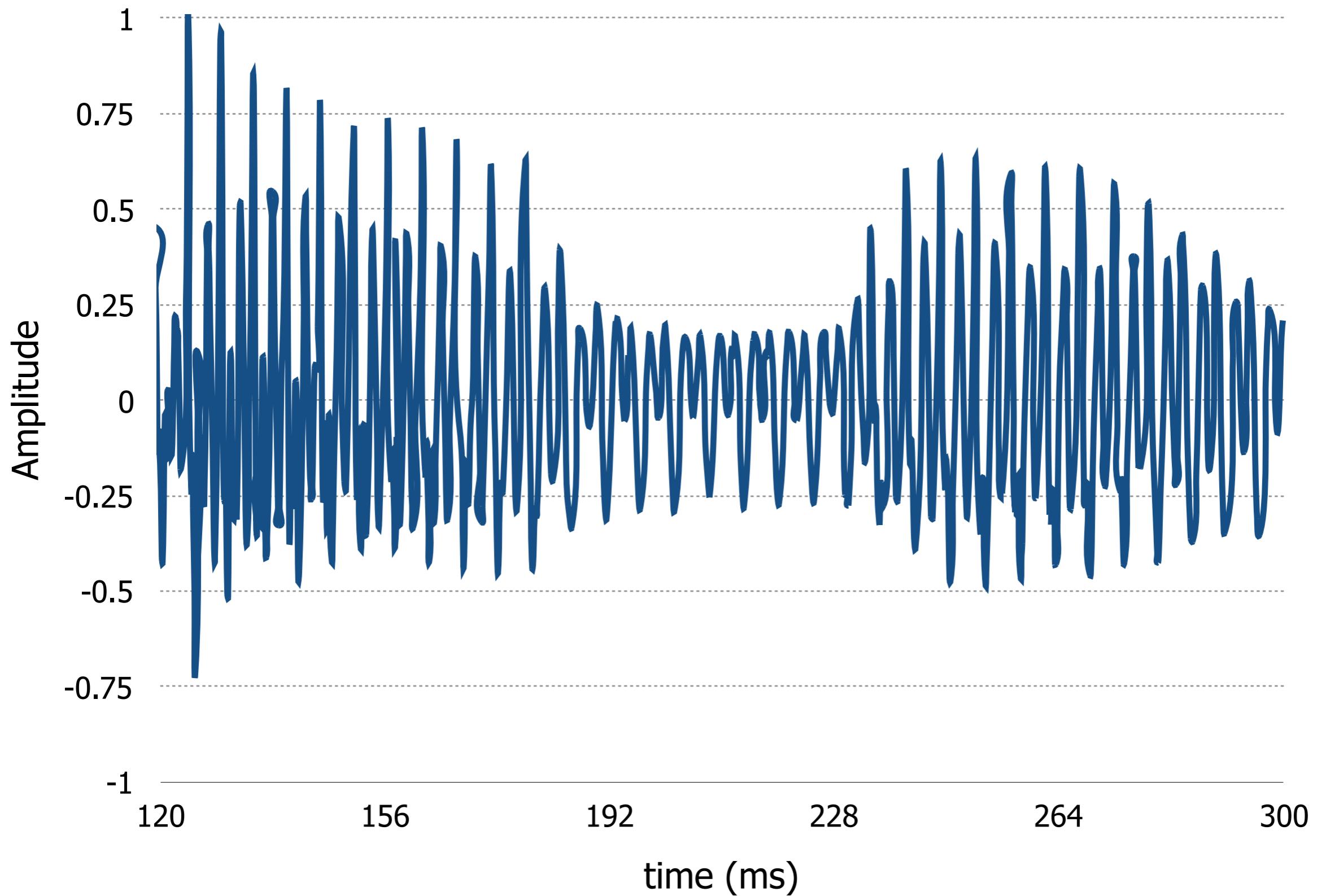


Short-Term Properties

- Any property should be measured over an interval that is short enough to ensure that the signal characteristics are stable;
- However, the interval should still be long enough to capture meaningful atomic units;
- In the case of speech, the typical interval length is 20-30 ms, the time one speaker keeps the articulators in a stable configuration.

Outline

- Introduction
- Time Domain Processing
- **Short-Term Analysis**
- Conclusions



Any short-term property
that can be extracted
from the signal

$$x[k] = \sum_{n=-\infty}^{\infty} f(s[n])w[k-n]$$

Convolution between
 $f(s[n])$ and a window

A function of sample “n”

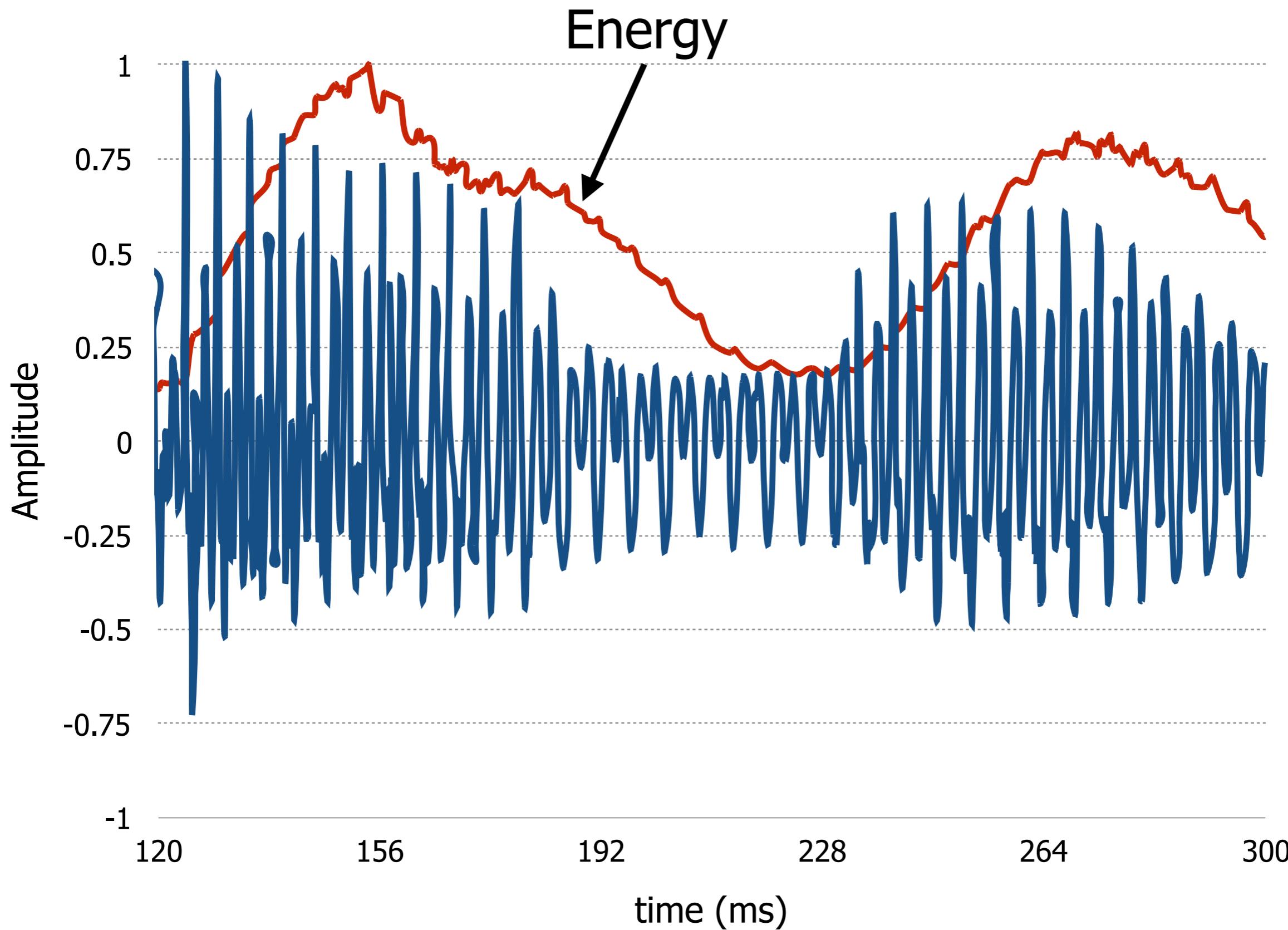
A window signal

The Energy

$$E[k] = \sum_{n=-\infty}^{\infty} \frac{(s[n])^2}{N} w[k - n]$$

The function is the square of the sample divided by the length of the window

The energy is the average of the samples' squares in the window ending at sample "k"

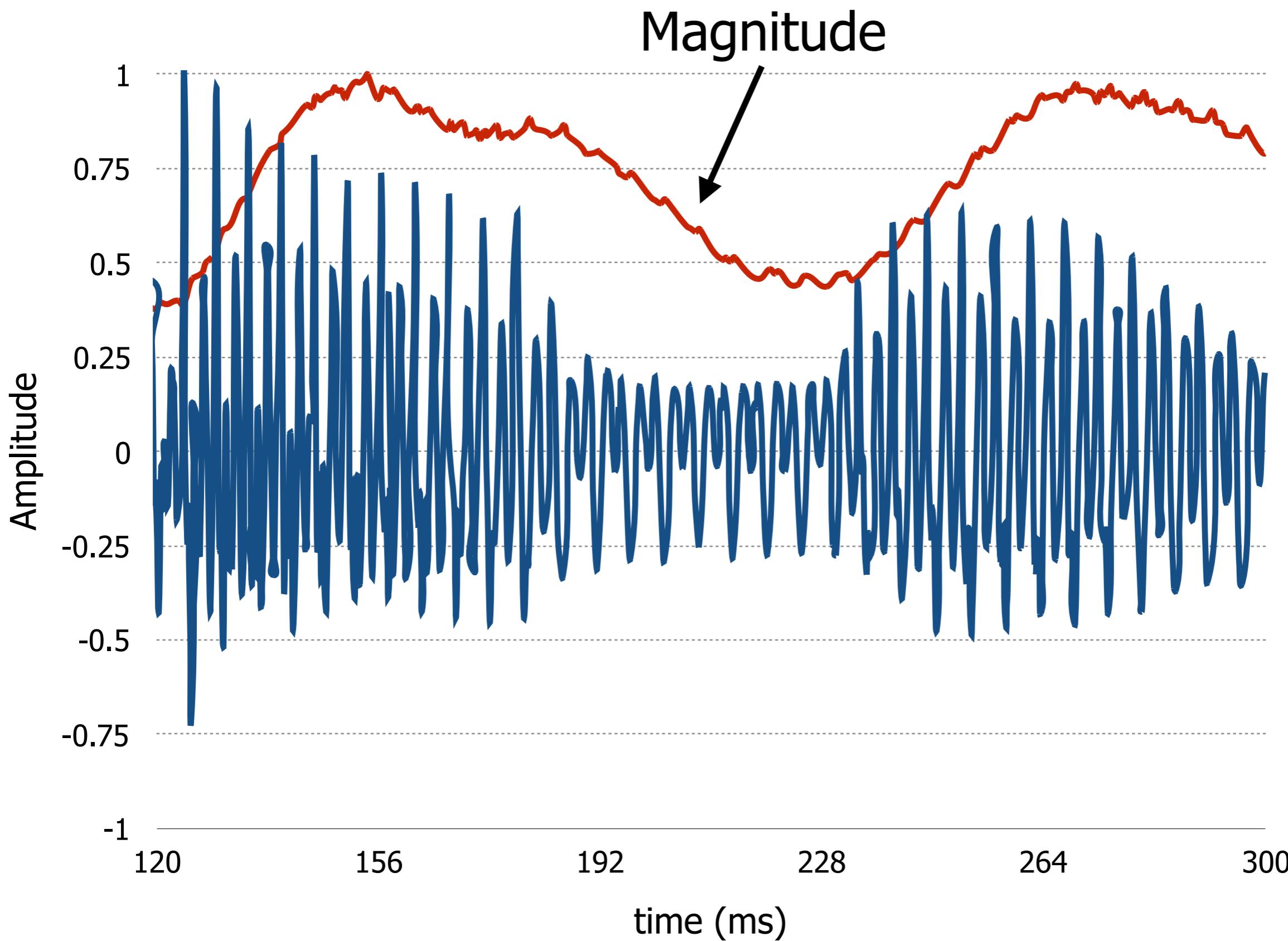


The Magnitude

$$M[k] = \sum_{n=-\infty}^{\infty} \frac{|s[n]|}{N} w[k - n]$$

The function is the absolute value of the sample divided by the length of the window

The magnitude is the average of the samples' absolute value in the window ending in sample "k"



Outline

- Introduction
- Time Domain Processing
- Short-Term Analysis
- Conclusions

Conclusions

- Speech signals can be analysed in the time domain through convolution operations;
- In most cases, the processing takes place in the frequency domain (after performing Fourier transform);
- The main reason for analysing speech is that it is the main form of communication between people.

Basic Signal Processing (II)

Computational Social Intelligence - Lecture 17

Prof. Alessandro Vinciarelli
School of Computing Science &
Institute of Neuroscience and Psychology

<http://www.dcs.gla.ac.uk/vincia>
Alessandro.Vinciarelli@glasgow.ac.uk



University
of Glasgow

EPSRC
Engineering and Physical Sciences
Research Council

FNSNF

Texts (see Moodle)

This lecture is based on the following text
(available on Moodle):

- F.Camastra and A.Vinciarelli, "Machine Learning for Audio, Image and Video Processing", Springer Verlag, Chapter 2, pp. 38-46, 2008.

Outline

- Quick Recap
- Zero Crossing Rate
- Autocorrelation
- Fourier Transform
- Conclusion

Outline

- Quick Recap
- Zero Crossing Rate
- Autocorrelation
- Fourier Transform
- Conclusion

Any short-term property
that can be extracted
from the signal

$$x[k] = \sum_{n=-\infty}^{\infty} f(s[n])w[k-n]$$

Convolution between
 $f(s[n])$ and a window

A function of sample “n”

A window signal

The Energy

$$f(s[n]) = \frac{(s[n])^2}{N}$$

$$f(s[n]) = \frac{|s[n]|}{N}$$

The Magnitude

Recap

- The short-term properties of a signal can be calculated through the convolution with an analysis window;
- Different functions applied to the samples lead to different properties;
- Every property is a signal that tends to remain stable in those intervals in which the signal properties are stable.

Outline

- Quick Recap
- Zero Crossing Rate
- Autocorrelation
- Fourier Transform
- Conclusion

The function is 1 when
the signal crosses the
horizontal axis

$$g(s[k], s[k - 1]) =$$

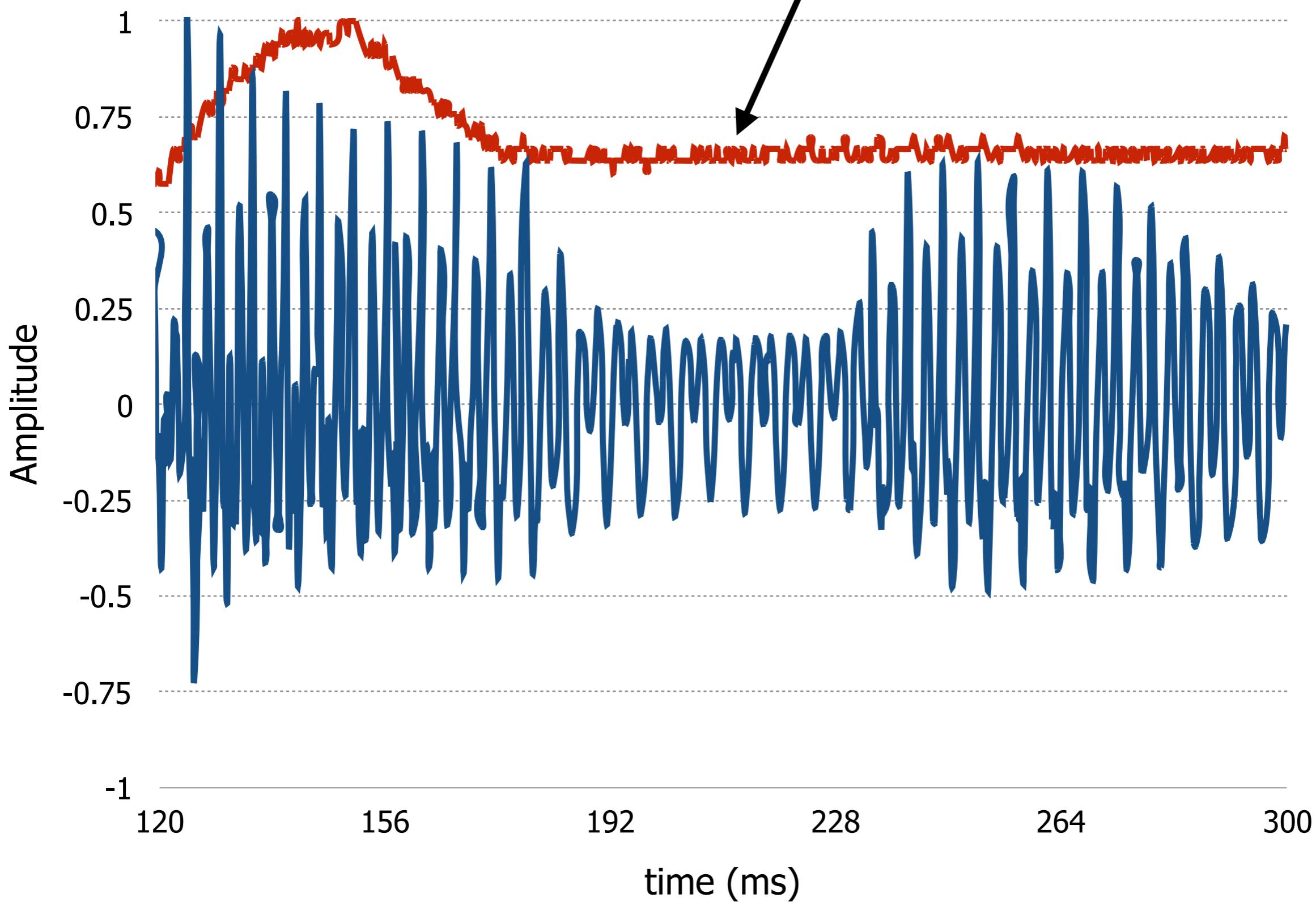
$$= \begin{cases} 0 & : s[k], s[k - 1] \geq 0 \\ 0 & : s[k], s[k - 1] \leq 0 \\ 1 & : s[k] > 0, s[k - 1] < 0 \\ 1 & : s[k] < 0, s[k - 1] > 0 \end{cases}$$

The Zero Crossing Rate

$$Z[k] = \sum_{n=-\infty}^{\infty} \frac{g(s[n], s[n-1])}{2N} w[k-n]$$

The ZCR is half the number of times the signal crosses the horizontal axis in the window

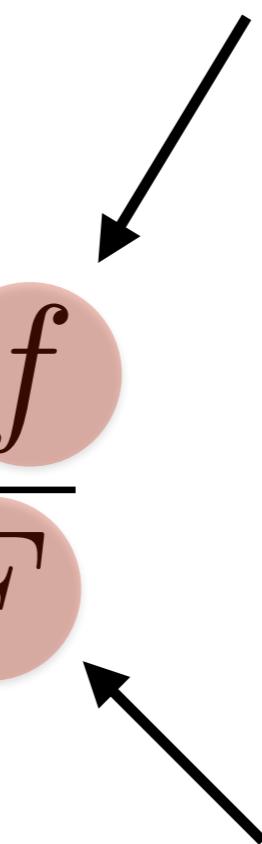
Zero Crossing Rate



The ZCR when the signal is a sinusoid of frequency "f"

$$Z[k] = \frac{2f}{F}$$

The sinusoid frequency



The sampling frequency

Outline

- Quick Recap
- Zero Crossing Rate
- Autocorrelation
- Fourier Transform
- Conclusion

The Autocorrelation

The parameter “m” is called the lag

$$\sum_{n=-\infty}^{\infty} s[n]w[k-n]s[n+m]w[k-n-m]$$

$R_m[k] =$

Product of two samples
at distance “m” from
one another

Index range where
“ $w[k-n]$ ” is different
from zero

$$k - n \geq 0$$

$$k - n \leq N - 1$$

$$k - n - m \geq 0$$

$$k - n - m \leq N - 1$$

 \Rightarrow

$$n \leq k$$

 \Rightarrow

$$n \geq k - N + 1$$

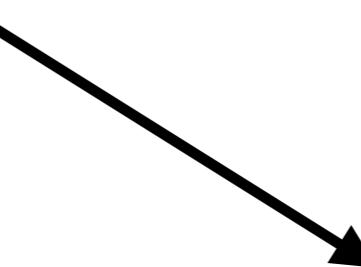
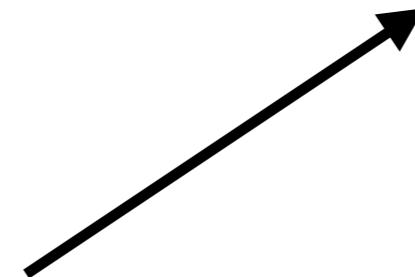
 \Rightarrow

$$n \leq k - m$$

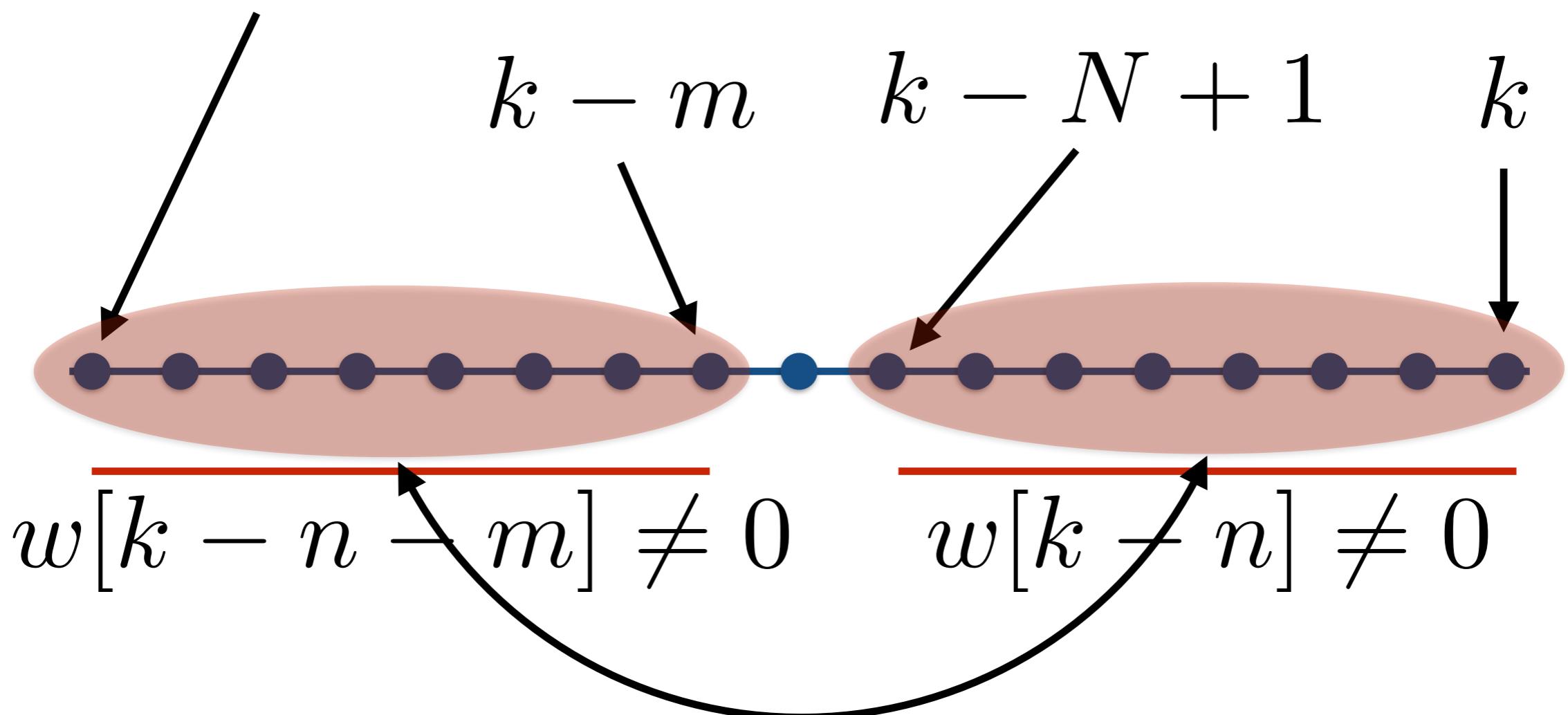
 \Rightarrow

$$n \geq k - m - N + 1$$

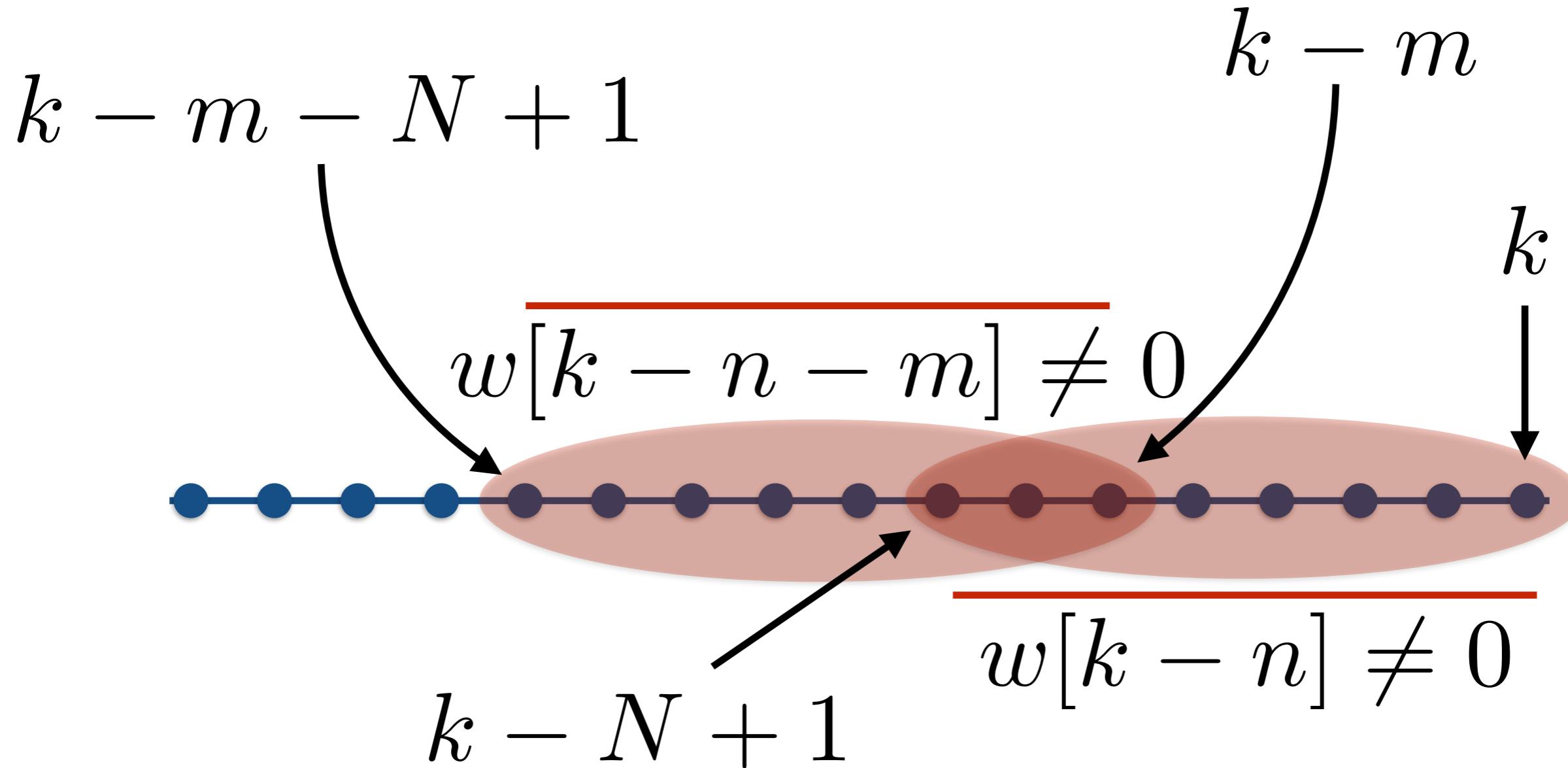
Index range where
“ $w[k-n-m]$ ” is different
from zero



$$k - m - N + 1$$



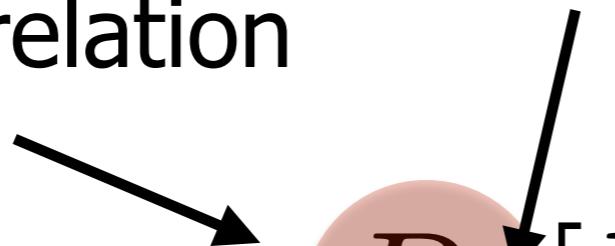
when "m" is greater than "N-1" ("N" is the length of the window), there is no overlapping between analysis windows



when “m” is less than “N” (“N” is the length of the window), there is overlapping between analysis windows

The lag is zero

The Autocorrelation



$$R_0[k] =$$

Only when the window
is rectangular

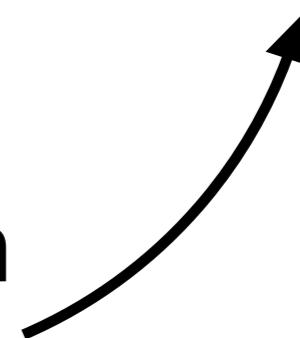
$$= \sum_{n=-\infty}^{\infty} s[n]s[n]w[k-n]w[k-n] =$$

$n = -\infty$

$$= \sum_{n=-\infty}^{\infty} (s[n])^2 w[k-n] = NE[k]$$

$n = -\infty$

The autocorrelation with
lag zero is the energy



$$NE[k]$$

Recap

- The Zero Crossing Rate provides information about the frequency that carries most energy;
- The autocorrelation changes depending on how periodic is the signal;
- The inverse of the distance between the maxima of the autocorrelation accounts for the frequency that carries most energy;

Outline

- Quick Recap
- Zero Crossing Rate
- Autocorrelation
- **Fourier Transform**
- Conclusion

Disclaimer

The Section “Fourier Transform” is not part of the exam (it is presented only for information).

However, in case of interest, further information is available in the following text:

- F.Camastra and A.Vinciarelli, “Machine Learning for Audio, Image and Video Processing”, Springer Verlag, Appendix B, pp. 38-51, 2008.

Discrete Fourier Transform

$$X[k] = \sum_{n=0}^{N-1} s[n] \cdot e^{-i2\pi \frac{k}{N} n}$$

The weights of the sum are complex exponentials

The sum is over the samples of an analysis window

Discrete Fourier Transform

The sum is over the samples of an analysis window

$$X[k] = \sum_{n=0}^{N-1} s[n] \cdot e^{-i2\pi \frac{k}{N} n} =$$

$$= \sum_{n=0}^{N-1} s[n] \left\{ \cos \left(2\pi \frac{k}{N} n \right) - i \sin \left(2\pi \frac{k}{N} n \right) \right\}$$

The weights are periodic functions

This exponential is equal to 1

$$\begin{aligned} X[k + N] &= \sum_{n=0}^{N-1} s[n] \cdot e^{-i2\pi \frac{k+N}{N} n} = \\ &= \sum_{n=0}^{N-1} s[n] \cdot e^{-i2\pi \frac{k}{N} n} e^{-i2\pi n} = \\ &= \sum_{n=0}^{N-1} s[n] \cdot e^{-i2\pi \frac{k}{N} n} = X[k] \end{aligned}$$

"X[k]" is periodic of period "N"

Inverse Fourier Transform

The sum is over the samples of a period of the DFT

$$s[n] = \frac{1}{N} \sum_{k=0}^{N-1} X[k] \cdot e^{i2\pi \frac{k}{N} n} =$$

$$= \frac{1}{N} \sum_{k=0}^{N-1} X[k] \left\{ \cos \left(2\pi \frac{k}{N} n \right) + i \sin \left(2\pi \frac{k}{N} n \right) \right\}$$

The weights are periodic functions

Sample “n” is the signal
at time “nT” (T is the
sampling period)

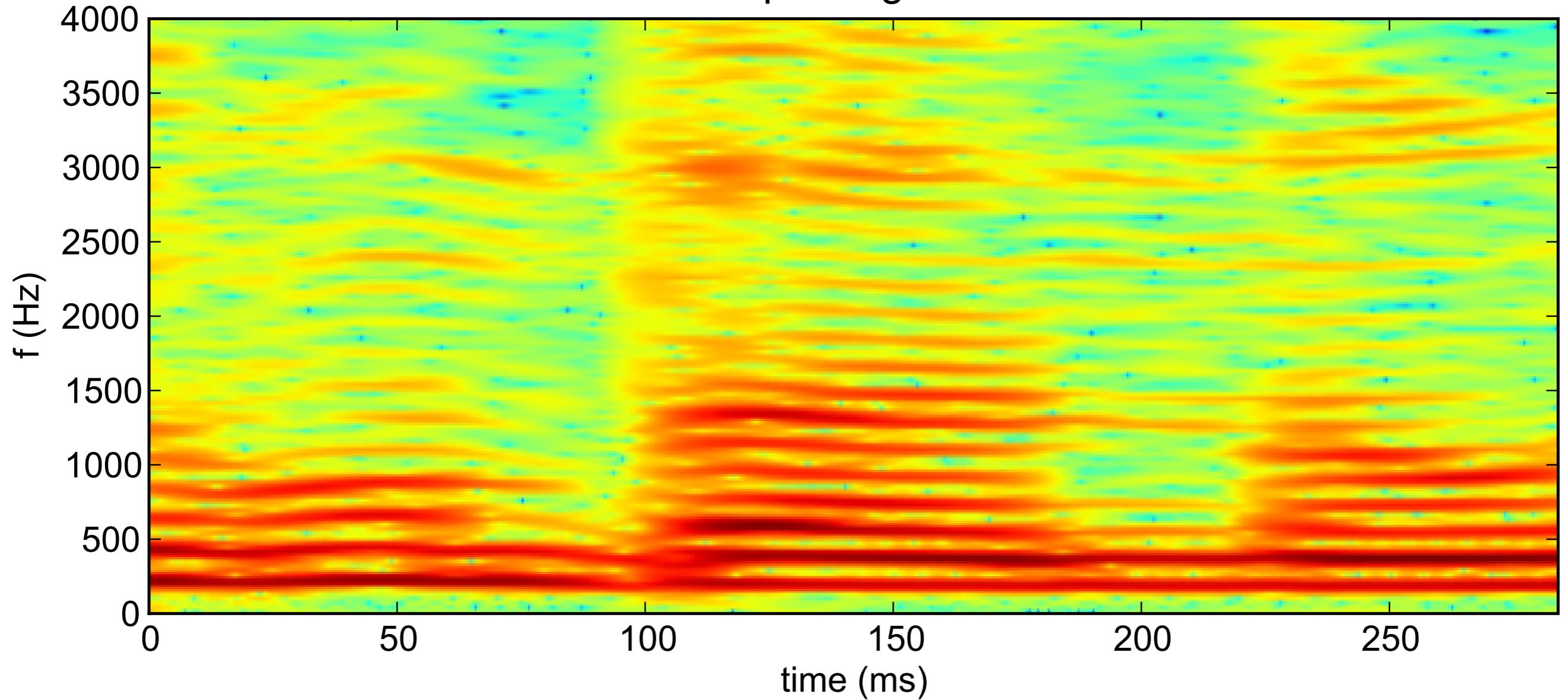


$$s(nT) = \frac{1}{N} \sum_{k=0}^{N-1} X[k] \cdot e^{i2\pi \frac{k}{N} nT} =$$
$$= \frac{1}{N} \sum_{k=0}^{N-1} X[k] \cdot e^{i2\pi \frac{k}{N} \frac{1}{F} n}$$



The sampling rate

Spectrogram



Recap

- Discrete Fourier Transform and its inverse represent the same information in two different ways (they are equivalent);
- The **spectrogram** shows the energy distribution across the frequencies;
- The Discrete Fourier Transform changes depending on the property of the signal.

Outline

- Quick Recap
- Zero Crossing Rate
- Autocorrelation
- Fourier Transform
- Conclusion

Conclusions

- Speech signals can be analysed in the time domain through convolution operations;
- In most cases, the processing takes place in the frequency domain (after performing Fourier transform);
- The main reason for analysing speech is that it is the main form of communication between people.

Speech Personality

Computational Social Intelligence - Lecture 18

Prof. Alessandro Vinciarelli
School of Computing Science &
Institute of Neuroscience and Psychology

<http://www.dcs.gla.ac.uk/vincia>
Alessandro.Vinciarelli@glasgow.ac.uk



University
of Glasgow

EPSRC
Engineering and Physical Sciences
Research Council

FNSNF

Texts (see Moodle)

This lecture is based on the following text
(available on Moodle):

- G.Mohammadi and A.Vinciarelli, "Automatic personality perception: Prediction of trait attribution based on prosodic features", IEEE Transactions on Affective Computing 3(3): 273-284, 2012.

Outline

- Introduction
- Computational Paralinguistics
- Trait Prediction
- Conclusion

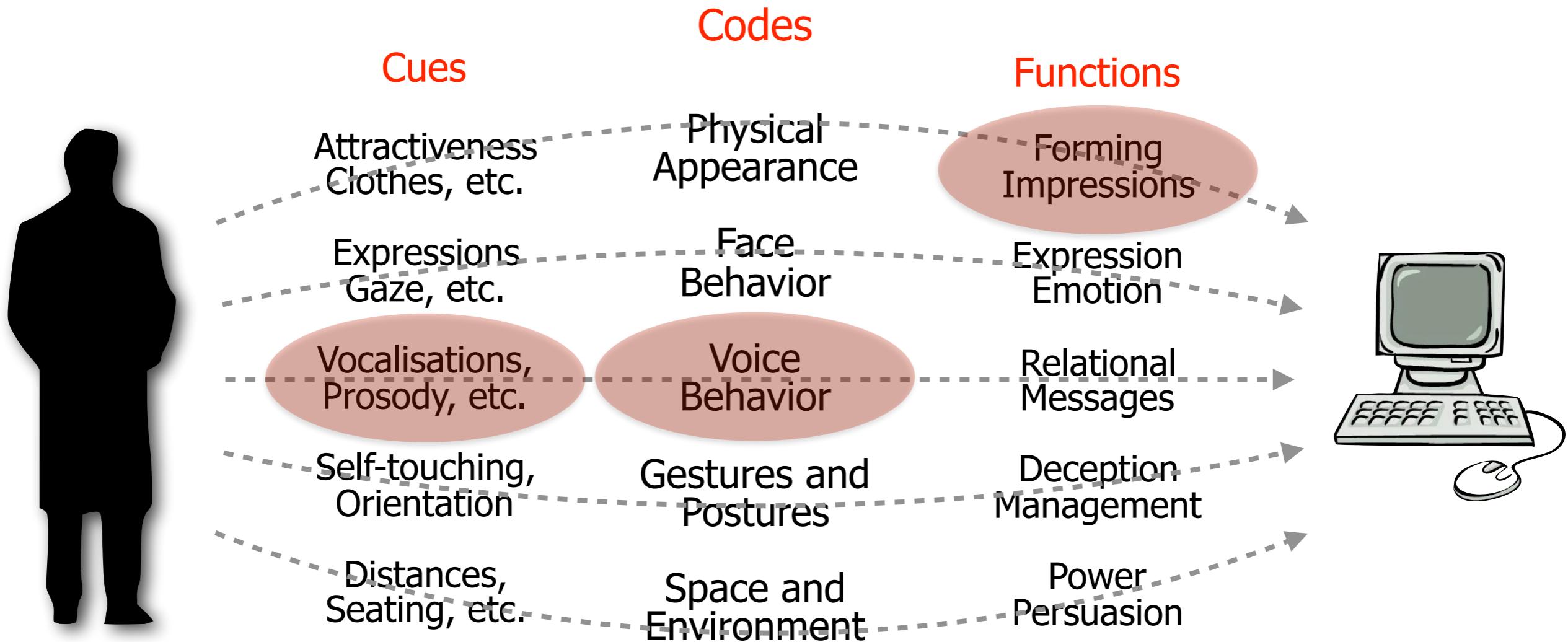
Outline

- Introduction
- Computational Paralinguistics
- Trait Prediction
- Conclusion

Speech and Personality

“[...] judgments made from speech alone rather consistently [have] the highest correlation with whole person judgments.”

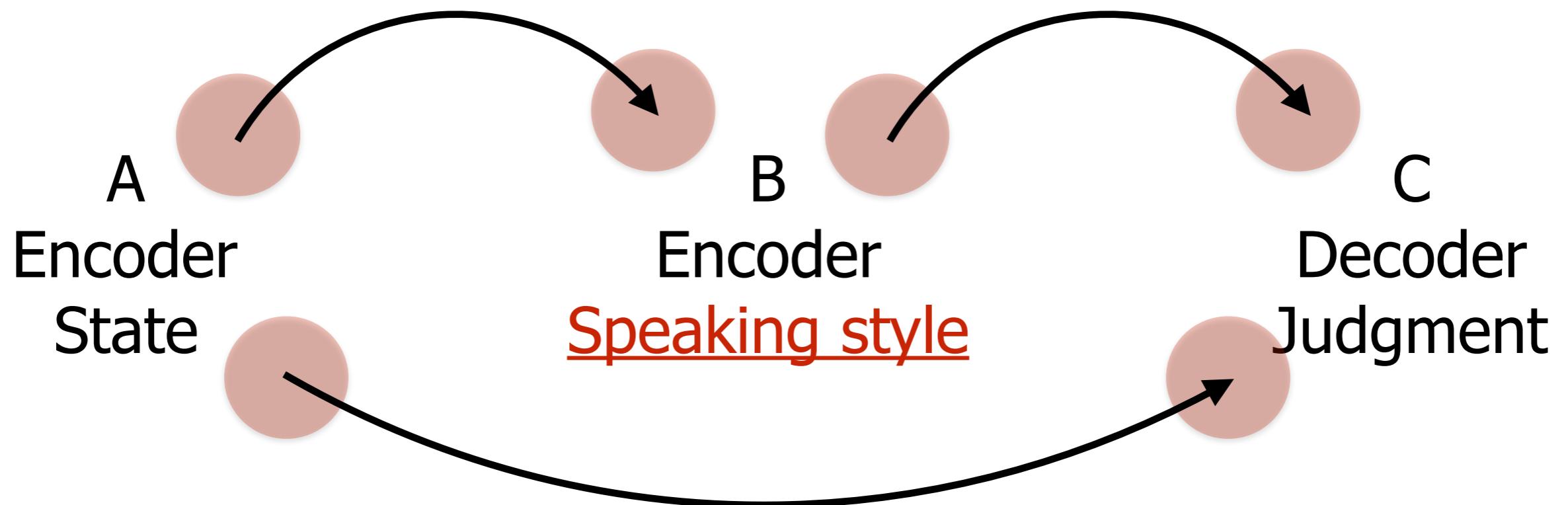
Ekman, Friesen, O’Sullivan, Scherer, “Relative importance of face, body, and speech in judgments of personality and affect”, Journal of Personality and Social Psychology, 38(2):270-277, 1980.



Richmond and McCroskey, "Nonverbal Behaviors in Interpersonal Relations",
Allyn and Bacon, 1995

How does the encoder manifest her/his state through her/**speaking style**?

How do decoders interpret the **speaking style** of the encoder



What is the state the decoder attributes to the encoder?

Recap

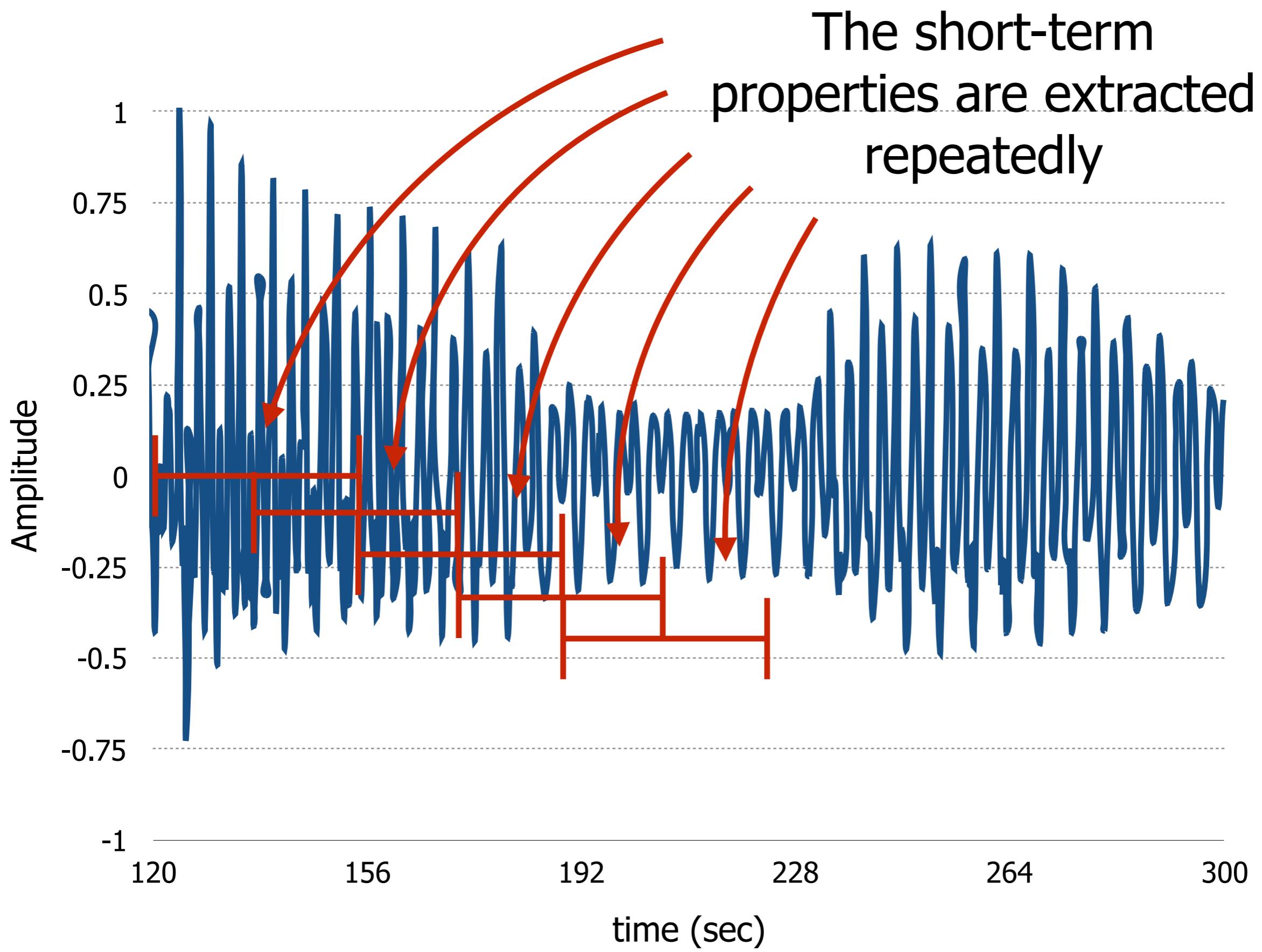
- The short-term properties of the speech signals provide information about the speaking style (how people speak);
- The speaking style influences the impression people develop about the speaker;
- It is possible to infer the personality traits people attribute to a speaker from her/his speaking style.

Outline

- Introduction
- Computational Paralinguistics
- Trait Prediction
- Conclusion

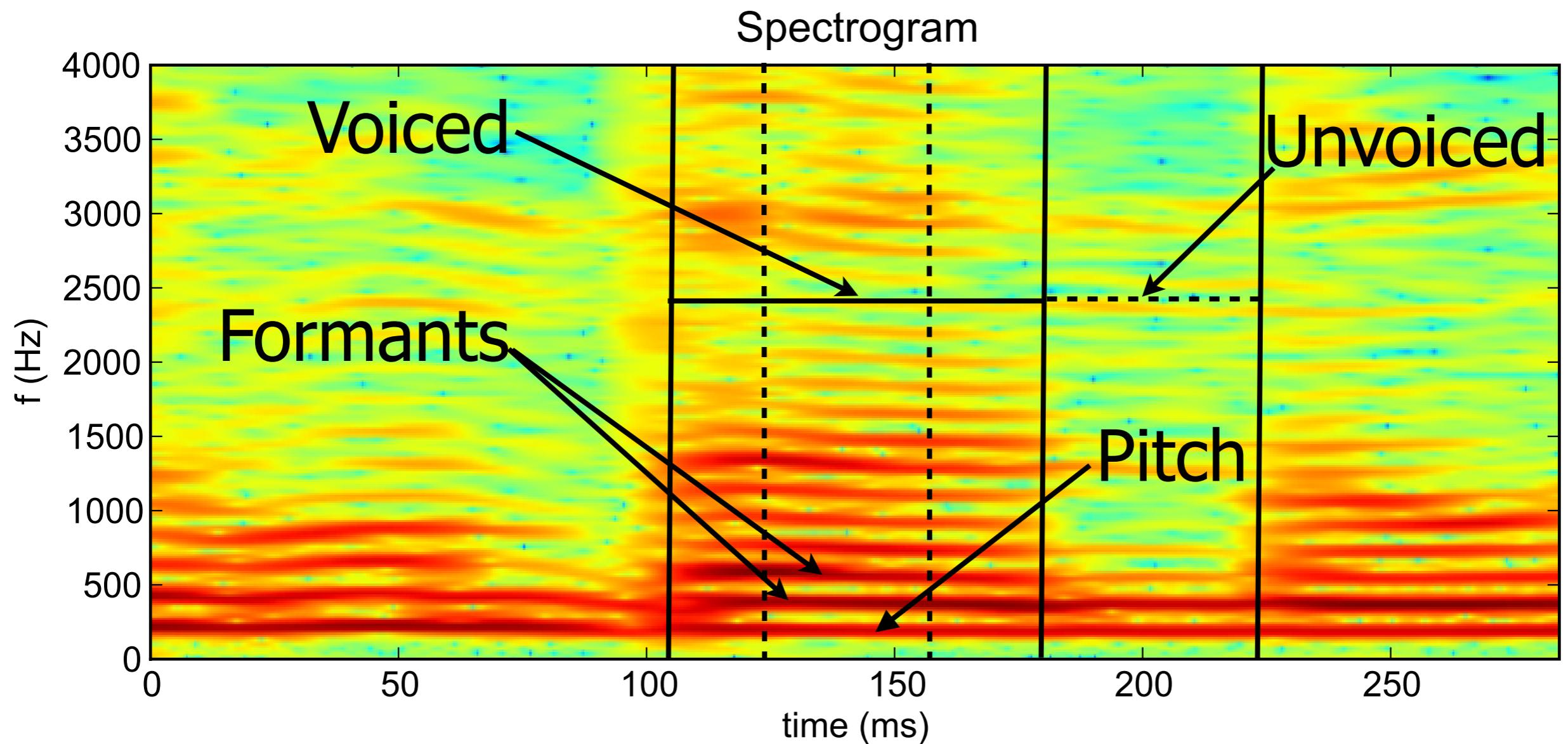
Short-Term Features

- The Big-Three of Prosody are Energy (loudness), Pitch (sound) and Tempo (speaking rate);
- Energy and Magnitude account for Energy;
- Zero Crossing Rate and Autocorrelation account for Pitch;
- The number if syllables per minute accounts for Tempo.



Computational Paralinguistics

- In general, the analysis windows are 30 ms long and the step is 10 ms, resulting into 100 feature vectors per second;
- For a given a speech sample, it is possible to estimate statistics of all features;
- The most common statistics are mean, variance, minimum, maximum, skewness, position of minimum and maximum, etc.



Six basic features
(pitch, formant 1,
formant 2, energy,
voiced and unvoiced
length)

For each of the six basic
features, there are 4
statistics (mean,
minimum, maximum,
entropy)

$$\vec{x} = (x_1, x_2, \dots, x_{24})$$



The result is a 24-dimensional feature vectors

Outline

- Introduction
- Computational Paralinguistics
- Trait Prediction
- Conclusion

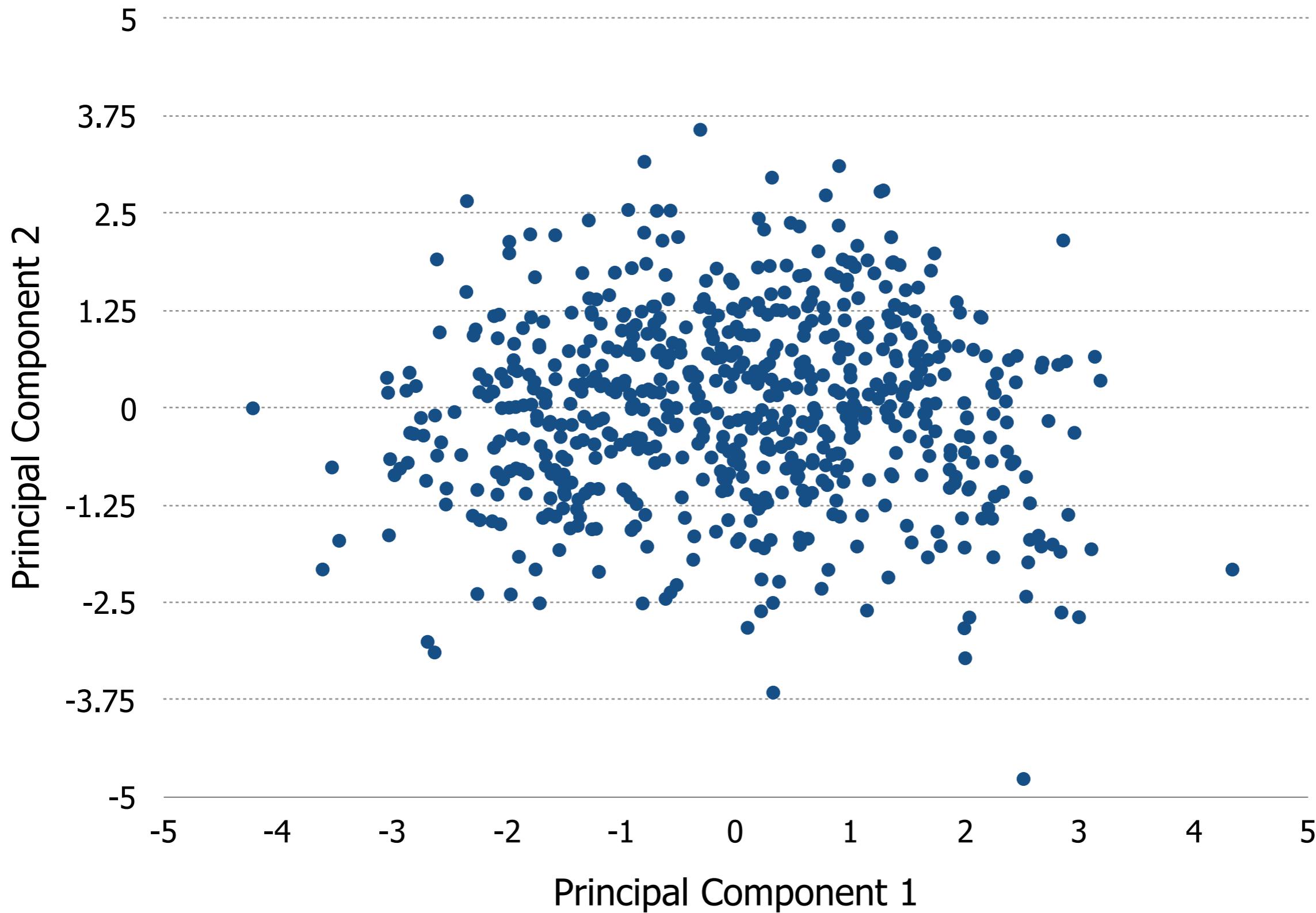


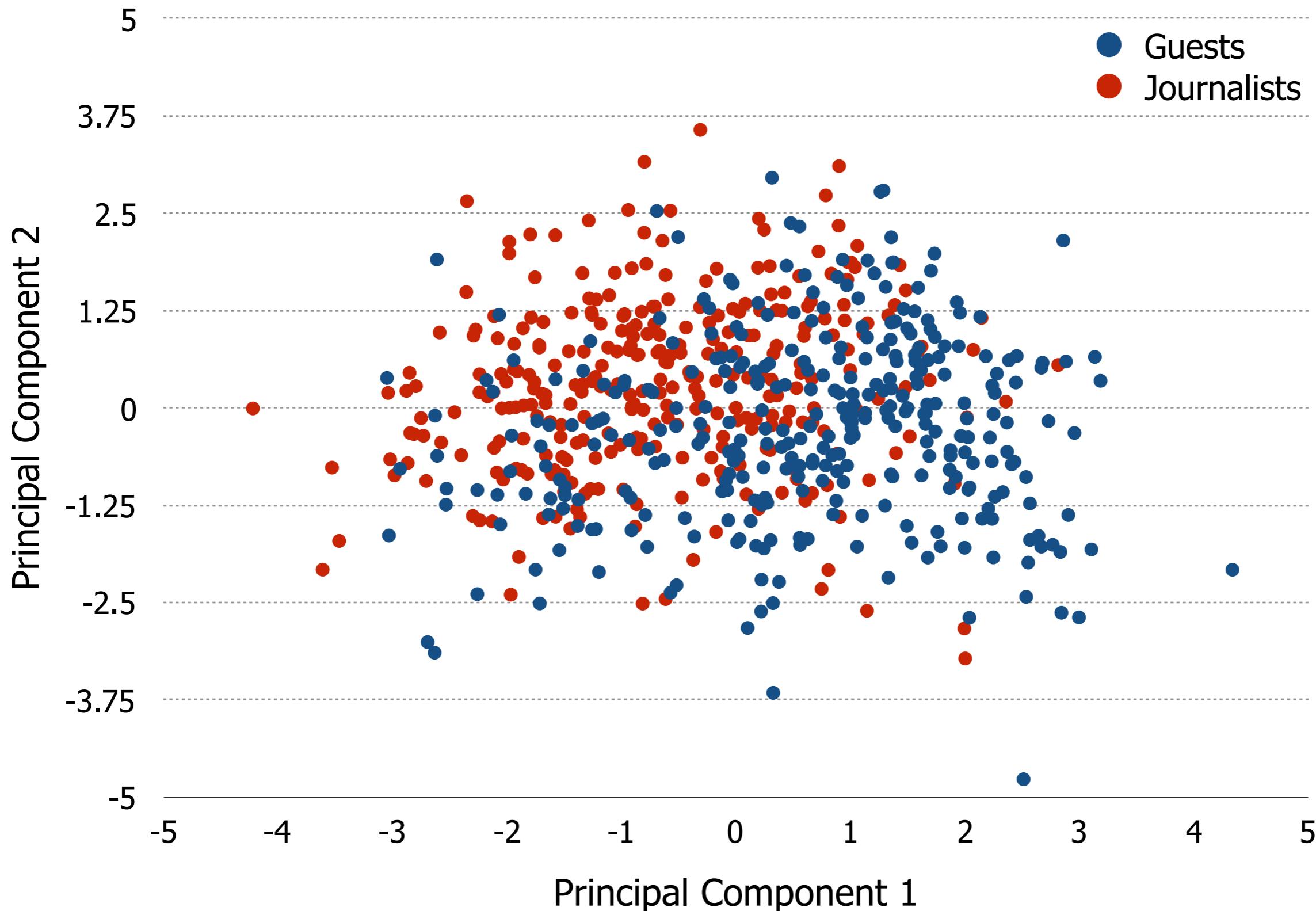
All individuals speaking in the news of Radio
Suisse Romande during February 2005

SSPNet Personality Corpus

Number of Samples	640
Total Length	1h:46m
Number of Subjects	322
Gender Balance	78.5% M / 21.5% F
Category Balance	48% J / 52% G
Speaker Distribution	80% < 3
Assessors	11 (British)
Total Items	70400

Mohammadi et al., "The Voice of Personality: Mapping Nonverbal Vocal Behavior into Trait Attributions", Social Signal Processing Workshop, pp. 17-20, 2010.





For each of the five traits, it is possible to identify two classes

Samples above median for a given trait

$$C = \{C_1, C_2\}$$

Samples below median for a given trait

Class assigned
automatically to a
speech sample

Vector extracted from
the speech sample

$$C^* = \arg \max_{C \in C} p(C | \vec{x}) p(C)$$

Posterior of the class

Prior of the class
(uninformative in this case)

The result is a 24-dimensional feature vectors

$$p(C|\vec{x}) = \frac{\exp(\vec{\theta} \cdot \vec{x} - \theta_0)}{1 + \exp(\vec{\theta} \cdot \vec{x} - \theta_0)}$$

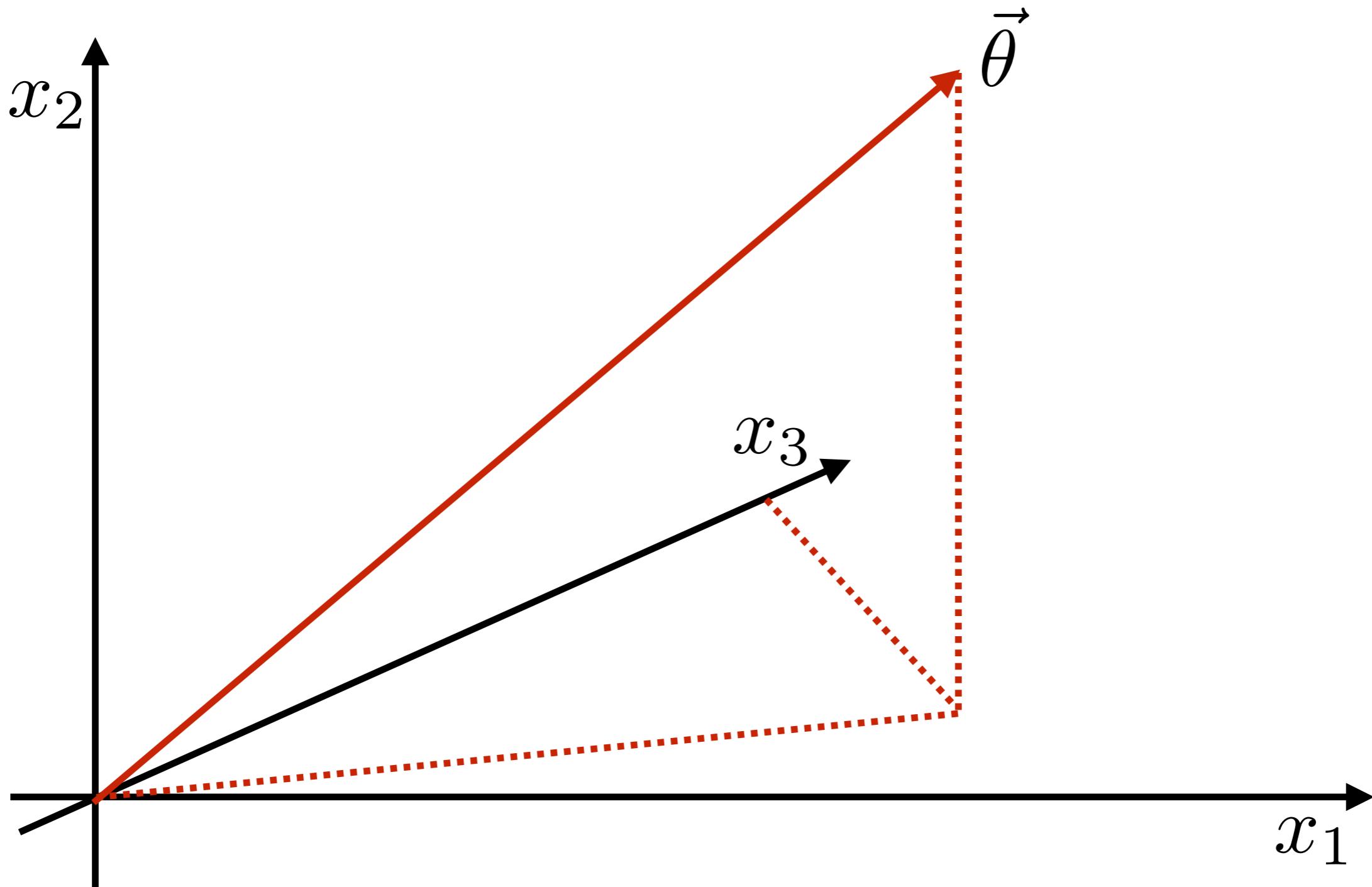
Scalar product between a parameter vector and the feature vector

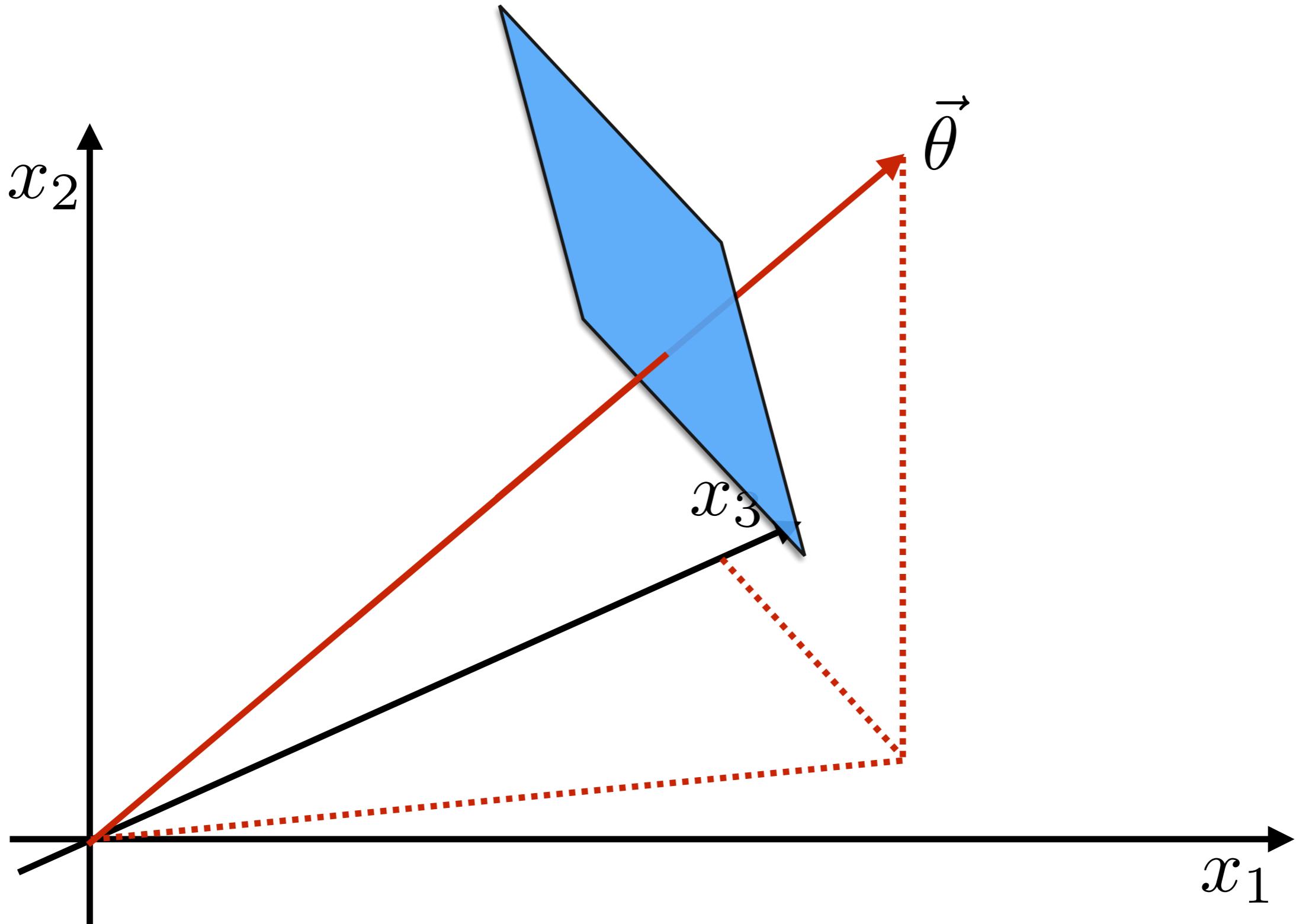
A scalar threshold

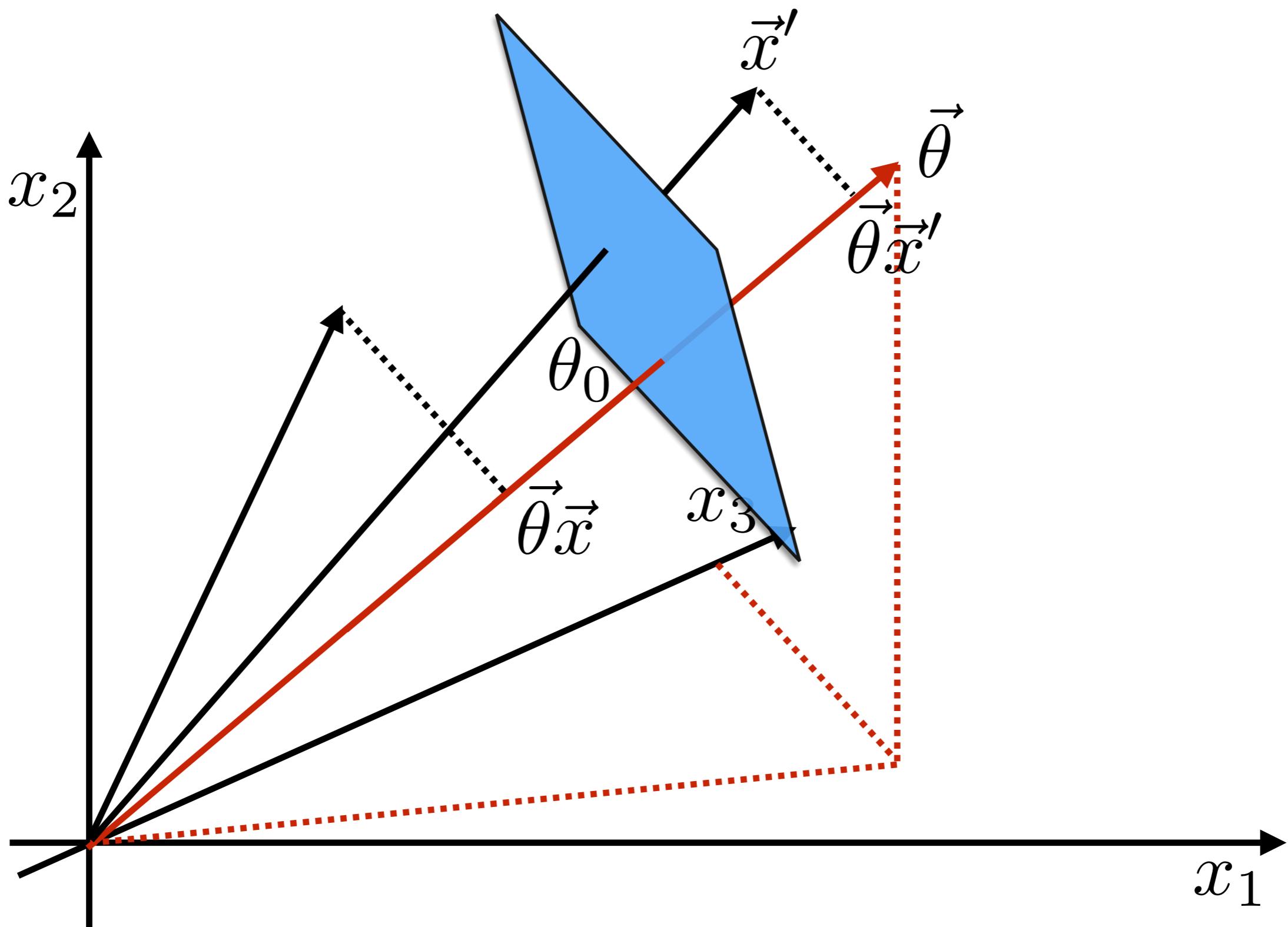
The diagram illustrates the components of the logistic regression formula. It shows the numerator $\exp(\vec{\theta} \cdot \vec{x} - \theta_0)$ and the denominator $1 + \exp(\vec{\theta} \cdot \vec{x} - \theta_0)$. The term $\vec{\theta} \cdot \vec{x}$ is highlighted with a red oval, and the term θ_0 is also highlighted with a red oval. Arrows point from the text "Scalar product between a parameter vector and the feature vector" to the $\vec{\theta} \cdot \vec{x}$ term, and from the text "A scalar threshold" to the θ_0 term.

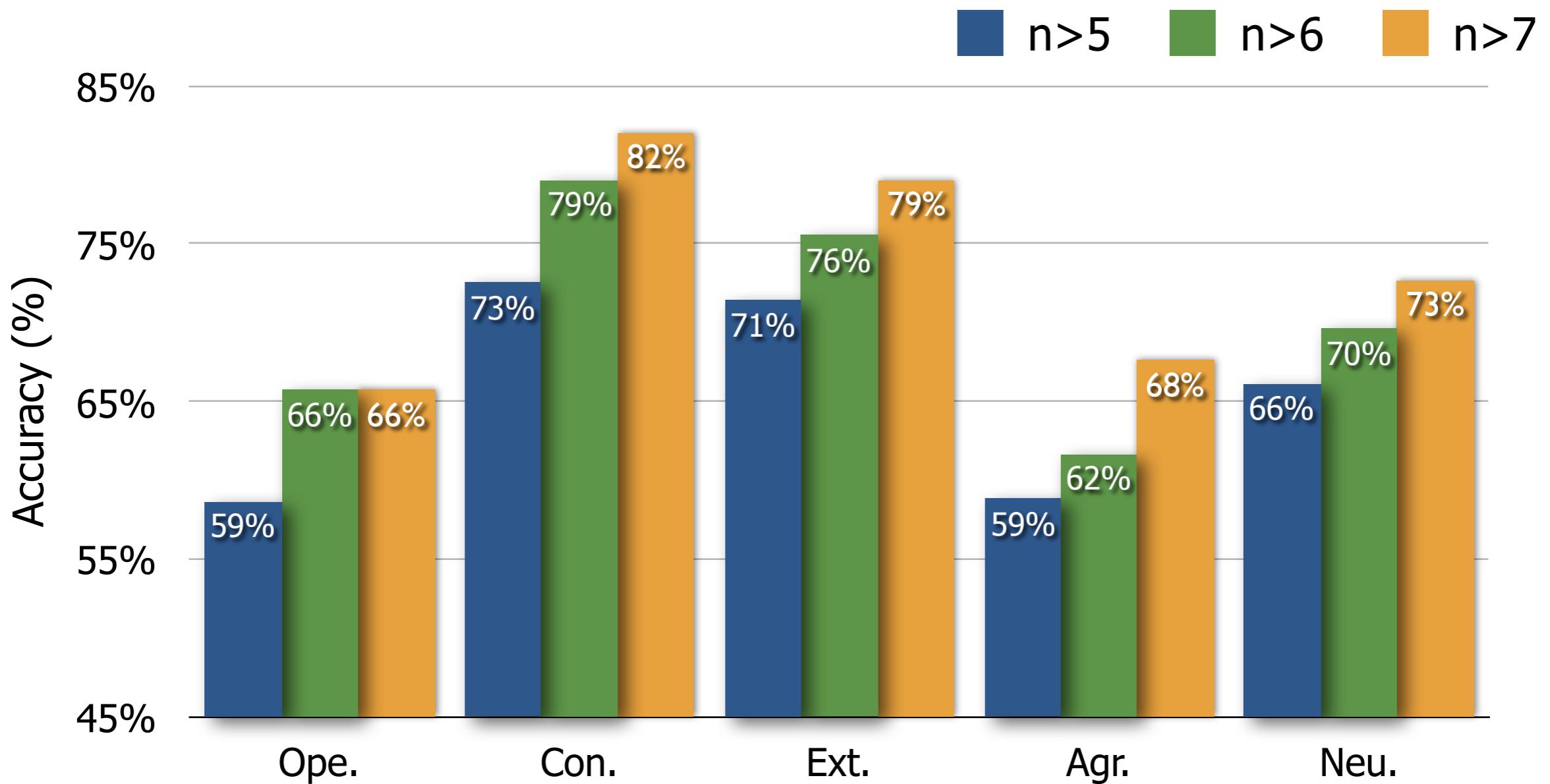
Training

- In the case of the logistic regression, the training aims at finding the parameters (the components of theta and the threshold);
- The training is performed by iteratively changing the values of the parameters to minimise the error rate;
- Such a process is performed using a k-fold approach.







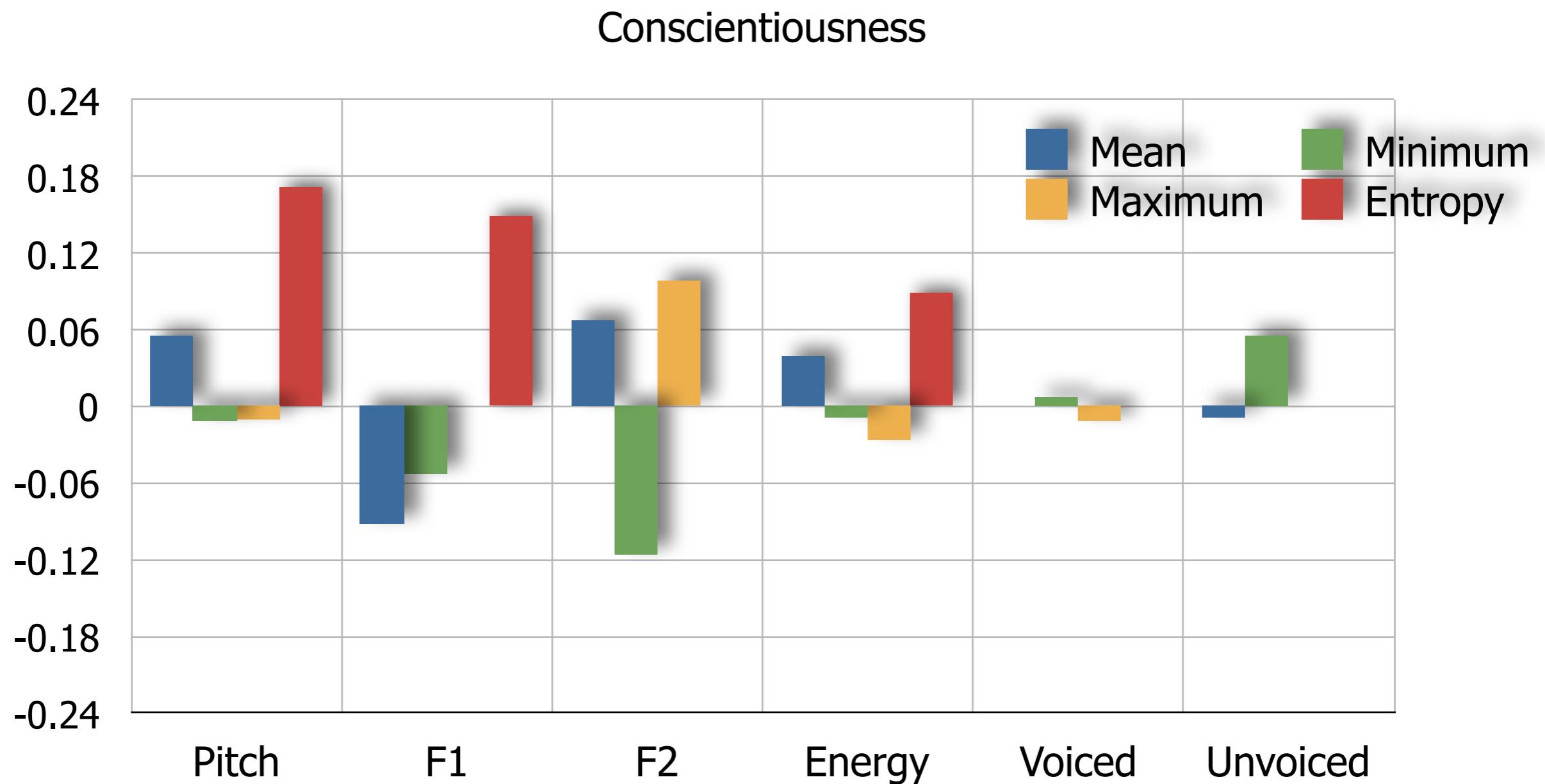


Mohammadi & Vinciarelli, "Automatic Personality Perception: Prediction of Trait Attribution Based on Prosodic Features", IEEE Transactions on Affective Computing, 3(3):273-284, 2012

Speech and Personality

“[...] there are two dimensions that underlie most judgments of traits, people, groups, and cultures [...] the first makes reference to attributes such as competence [...] and the second to warmth [...]”

Judd et al., “Fundamental Dimensions of Social Judgment: Understanding the Relations Between Judgments of Competence and Warmth”, Journal of Personality and Social Psychology, 89(6):899-913, 2005

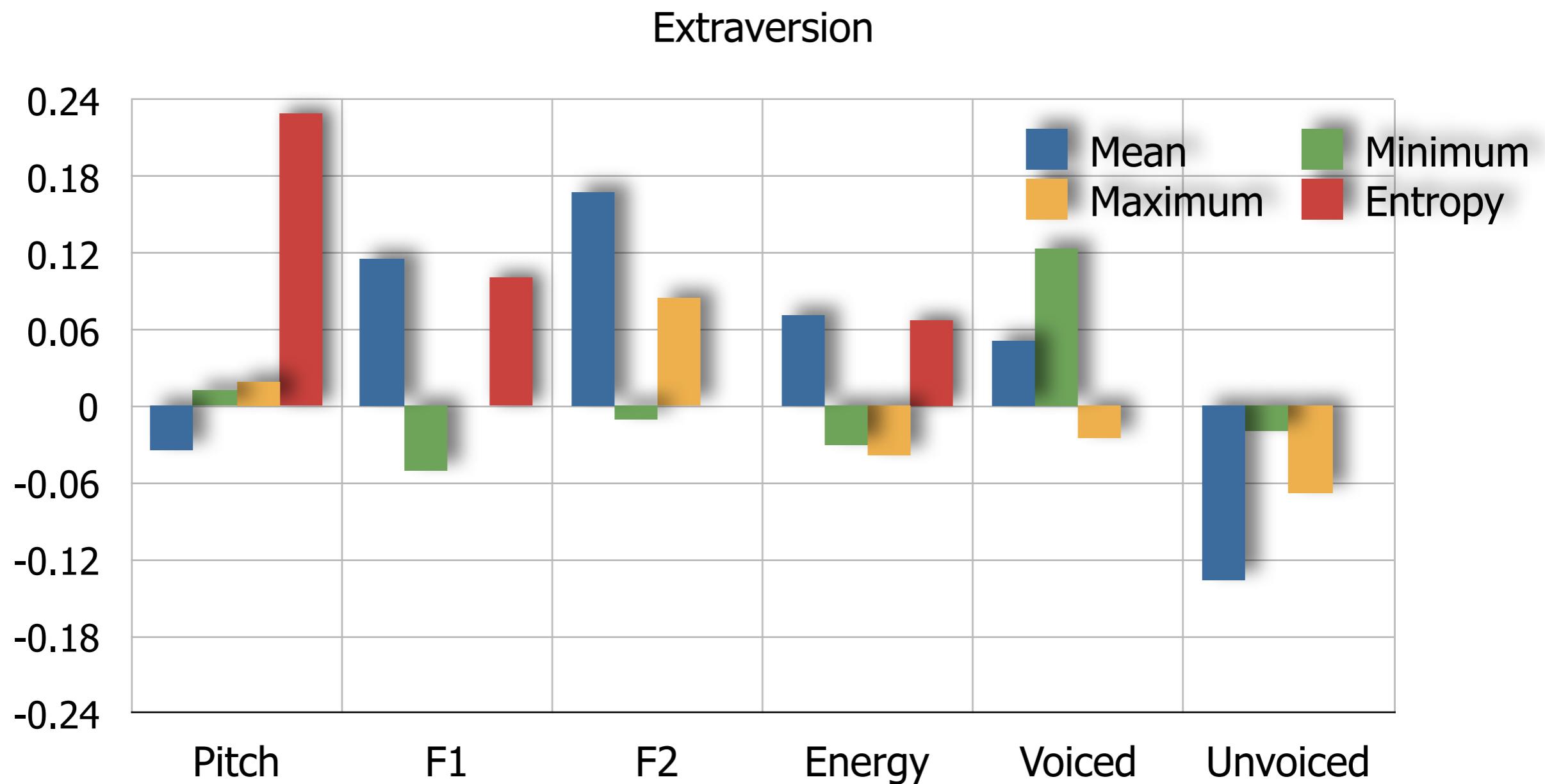


Mohammadi & Vinciarelli, "Automatic Personality Perception: Prediction of Trait Attribution Based on Prosodic Features", IEEE Transactions on Affective Computing, 3(3):273-284, 2012

Speech and Personality

“Rate and pitch variation were the most influential for competence and benevolence, respectively. For competence, one interaction effect (rate by pitch variation) was significant.”

Ray, “Vocally cued personality prototypes: An implicit personality theory approach”, Journal of Communication Monographs, 53(3):266-276, 1986



Mohammadi & Vinciarelli, "Automatic Personality Perception: Prediction of Trait Attribution Based on Prosodic Features", IEEE Transactions on Affective Computing, 3(3):273-284, 2012

Speech and Personality

“Rate and pitch variation were the most influential for competence and benevolence, respectively. For competence, one interaction effect (rate by pitch variation) was significant.”

Ray, “Vocally cued personality prototypes: An implicit personality theory approach”, Journal of Communication Monographs, 53(3):266-276, 1986

Outline

- Introduction
- Computational Paralinguistics
- Trait Prediction
- Conclusion

Conclusions

- Speech signals can be analysed in the time domain through convolution operations;
- In most cases, the processing takes place in the frequency domain (after performing Fourier transform);
- The main reason for analysing speech is that it is the main form of communication between people.

Multimodality

Multimodal Social Signal Processing - Lecture 19

Prof. Alessandro Vinciarelli
School of Computing Science &
Institute of Neuroscience and Psychology

<http://www.dcs.gla.ac.uk/vincia>
Alessandro.Vinciarelli@glasgow.ac.uk



University
of Glasgow

EPSRC
Engineering and Physical Sciences
Research Council

FNSNF

This lecture is based on the following text
(available on Moodle):

- Vinciarelli & Esposito, “Multimodal Analysis of Social Signals”, in “The Handbook of Multimodal-Multisensor Interfaces”, Oviatt et al. (eds.), 203-226, ACM, 2018;
- Partan & Marler, “Issues in the Classification of Multimodal Communication Signals”, The American Naturalist, 166(2), pp. 231-245, 2005.

Outline

- Multimodality (Psychology & Neuroscience)
- Multimodality (Communication & Life Science)
- Multimodality (Computing Science & AI)
- Conclusions

Outline

- Multimodality (Psychology & Neuroscience)
- Multimodality (Communication & Life Science)
- Multimodality (Computing Science & AI)
- Conclusions

Gestalt Theory (I)

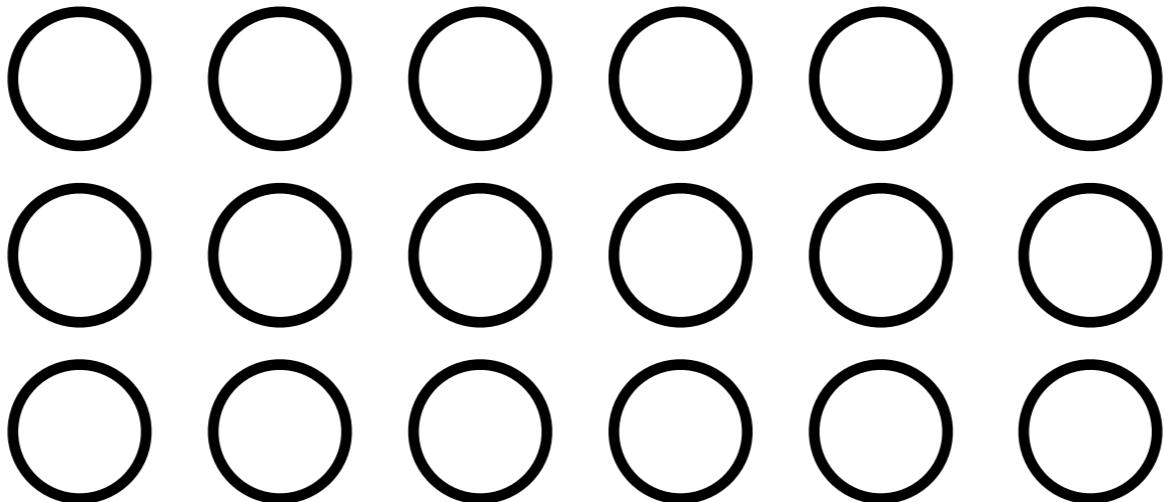
"It has been said: The whole is more than the sum of its parts. It is more correct to say that the whole is something else than the sum of its parts, because summing is a meaningless procedure, whereas the whole-part relationship is meaningful"

Koffka, "Principles of Gestalt Psychology", 1935

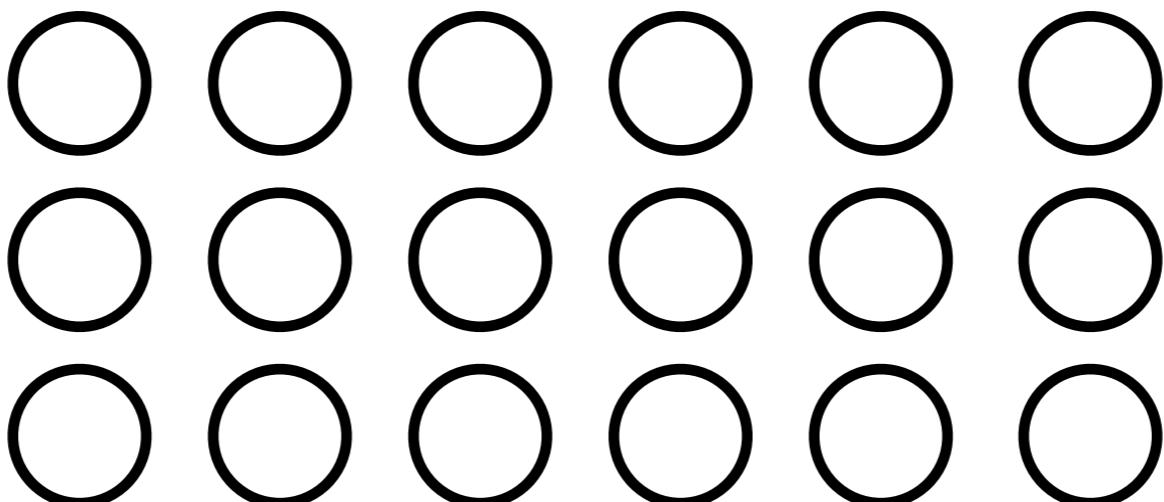
Gestalt Theory (II)

“Gestalt Theory describes different laws or principles for perceptual grouping of information into a coherent whole, including the laws of proximity, symmetry, similarity, closure, continuity [...]”

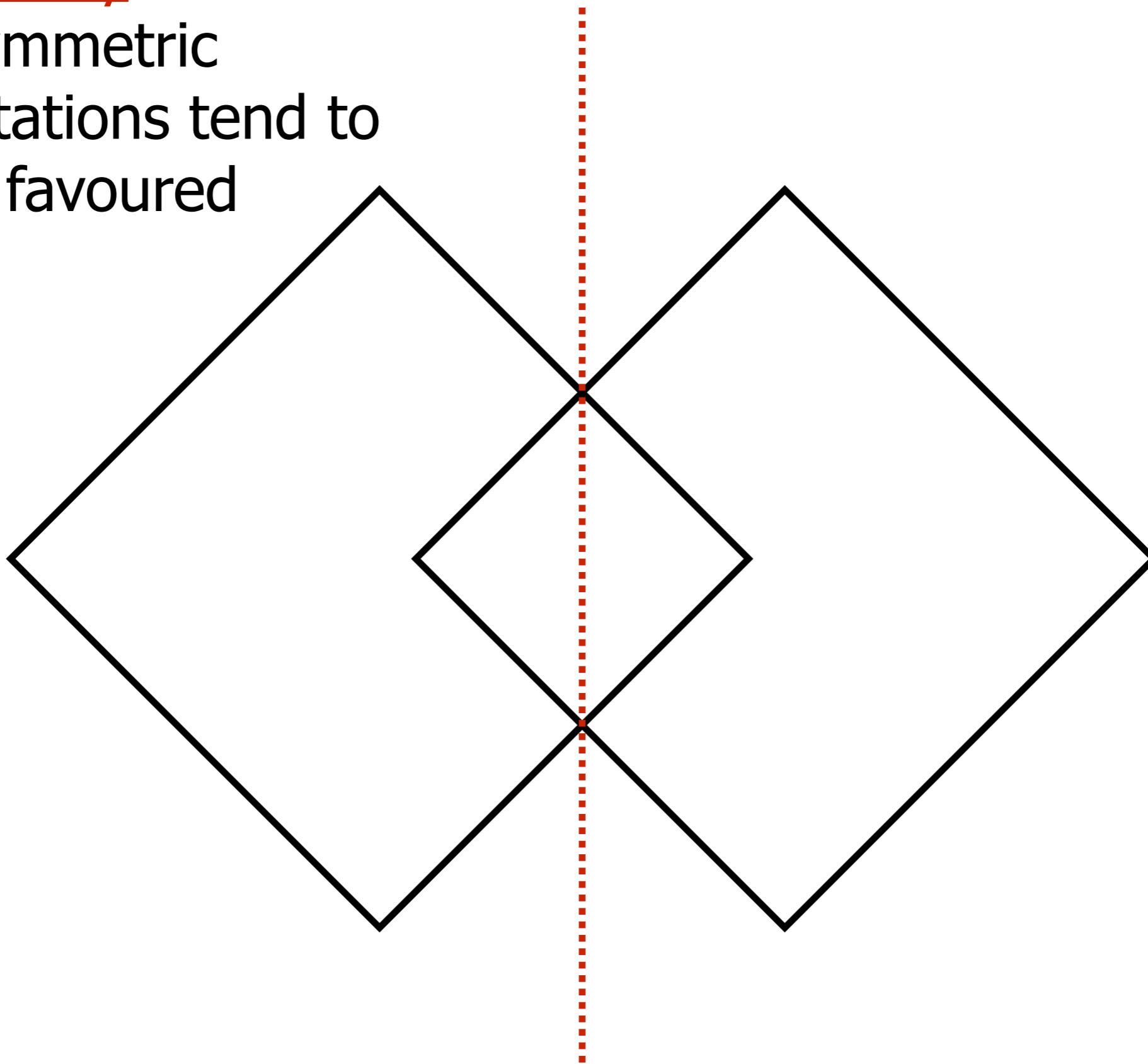
Oviatt, “Theoretical Foundations of Multimodal Interfaces and Systems”, in “The Handbook of Multimodal-Multisensor Interfaces”, pp. 20-50, 2018.

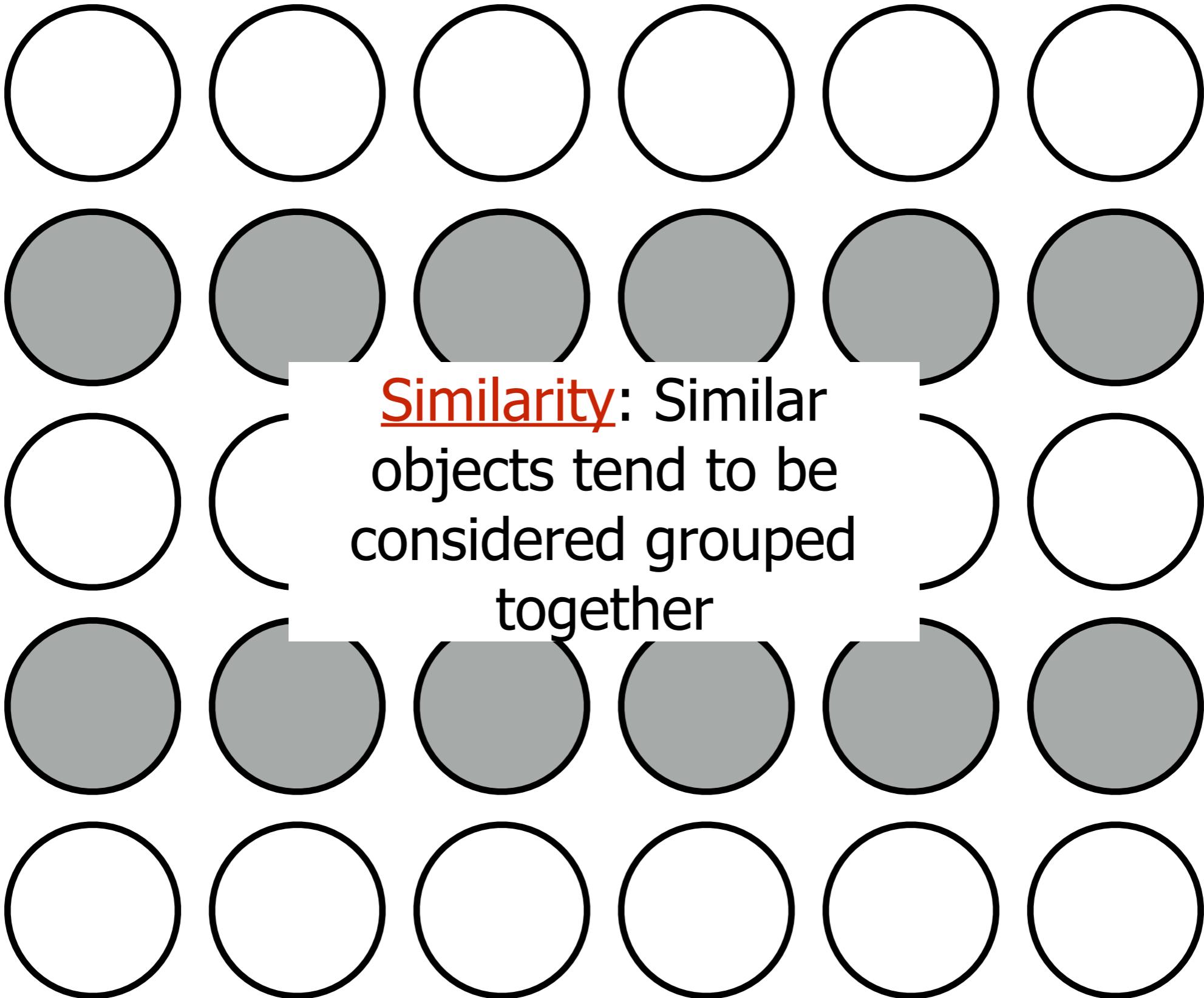


Proximity: Objects close to each other tend to be considered grouped together

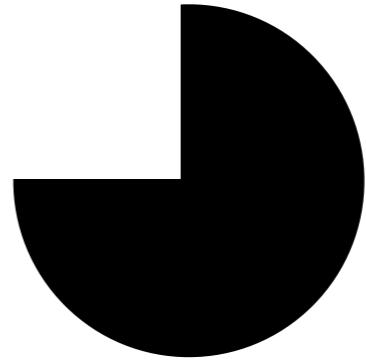
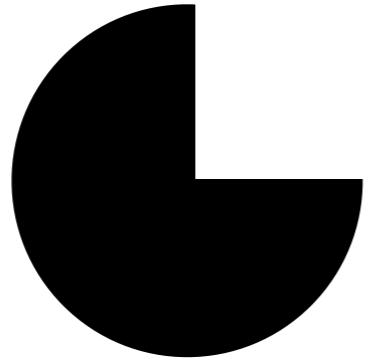
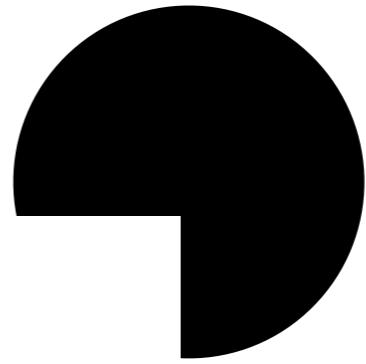
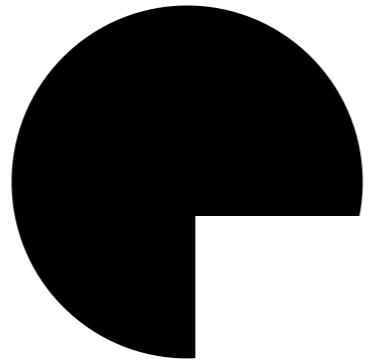


Symmetry: More
symmetric
interpretations tend to
be favoured

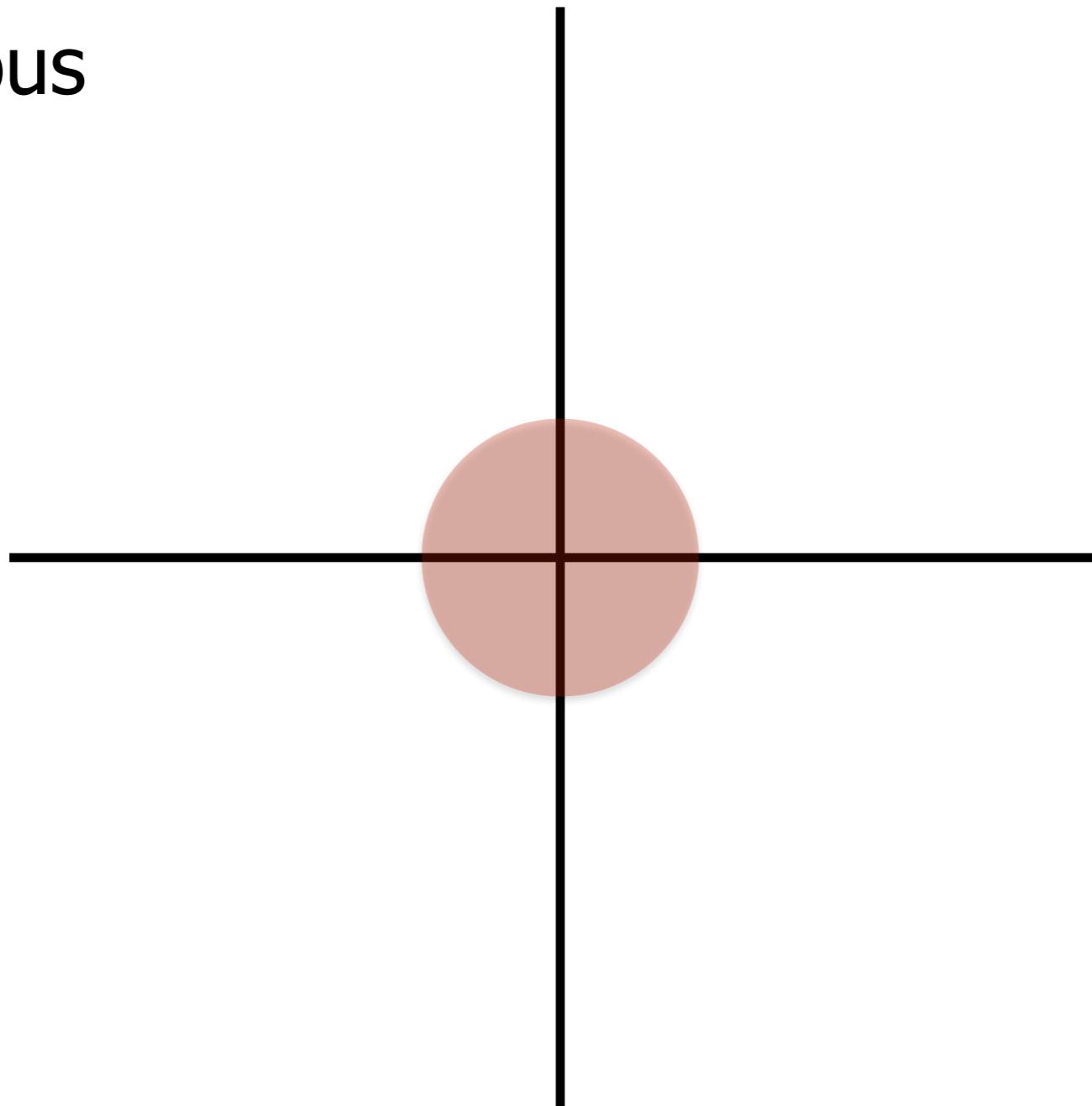




Closure: close and complete figures tend to be perceived in absence of continuous lines



Continuity: lines that follow one another tend to be considered continuous



Multy-Sensory Perception

“[...] brain processing fundamentally involves multi sensory perception and integration [...] which cannot be accounted for by studying the senses in isolation.”

Oviatt, “Theoretical Foundations of Multimodal Interfaces and Systems”, in “The Handbook of Multimodal-Multisensor Interfaces”, pp. 20-50, 2018.

Mc Gurk Effect

<https://www.youtube.com/watch?v=G-IN8vWm3m0>

Recap

- Our brain does not process multiple (possibly multi-sensory) signals individually;
- The initial intuitions proposed in the early 20th century (Gestalt Psychology) have been later confirmed by neuroscience;
- When it comes to perception, “the whole is other than the sum of the parts” (Koffka, 1935).

Outline

- Multimodality (Psychology & Neuroscience)
- **Multimodality (Communication & Life Science)**
- Multimodality (Computing Science & AI)
- Conclusions

Multimodal Social Signals (I)

“Multimodality [...] not a recent discovery. The ancient rhetors and theorists Cicero and Quintilian stressed the importance of voice, gesture and face in the delivery of discourse very early on.”

Poggi, “Mind, Hands, Face and Body”,
Weidler Buchverlag, 2007.

Multimodal Social Signals (II)

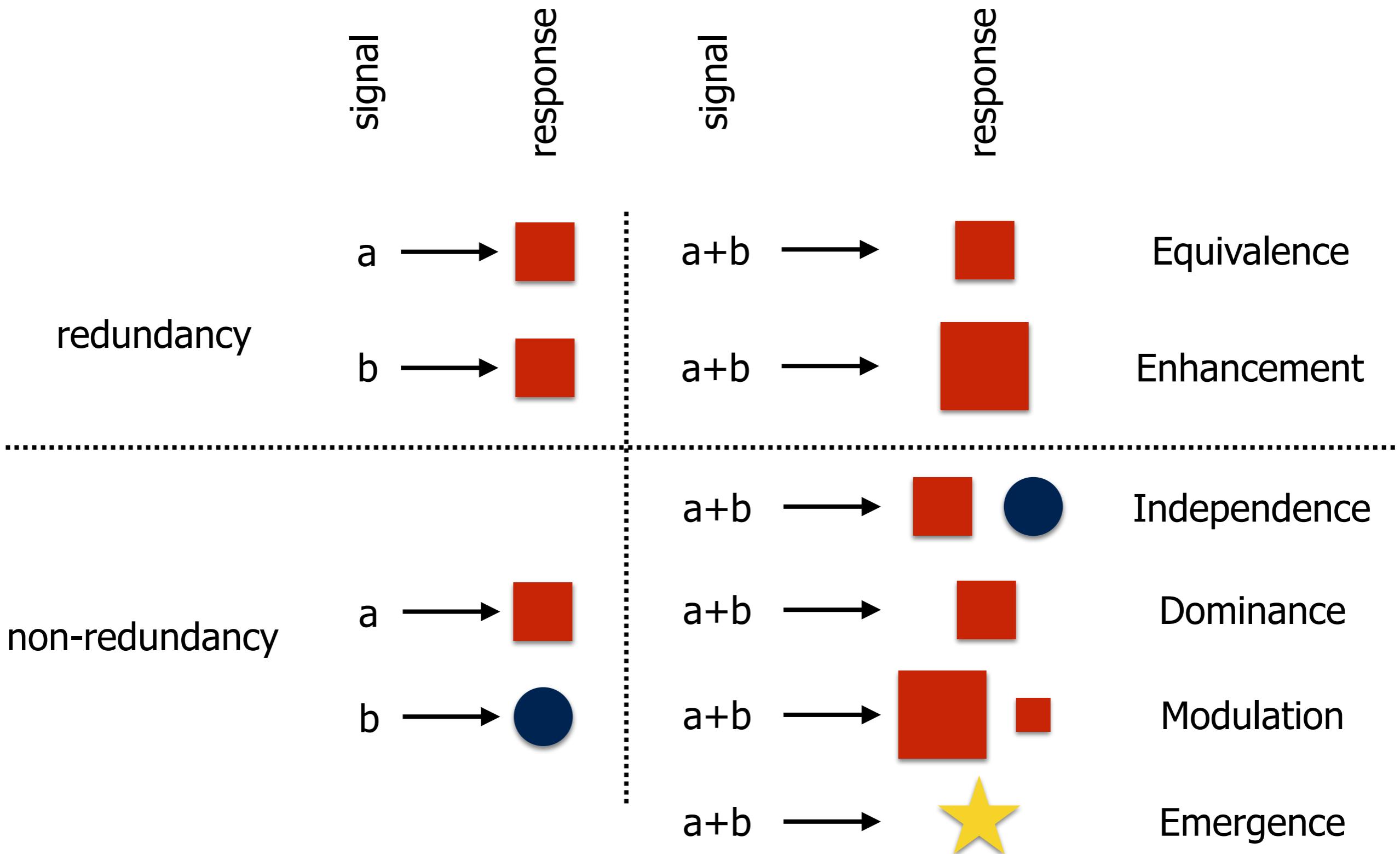
"[monkeys] utter a reiterated sound [...] accompanied by vibratory movements of their jaws or lips, with the corners of the mouth drawn backwards and upwards, by the wrinkling of the cheeks, and even by the brightening of the eyes."

Darwin, "The Expression of Emotions in Animals and Men",
John Murray, 1872.

Multimodal Social Signals (III)

“Multimodal [...] communication is defined as communication via composite signals received through more than one sensory channel. We use the word “signal” [...] to refer to the entire set of communicative features [...] of an animal’s behavior that occur simultaneously.”

Partan and Marler, “Issues in the Classification of Multimodal Communication Signals”, *The American Naturalist*, 166(2), pp. 231-245, 2005.



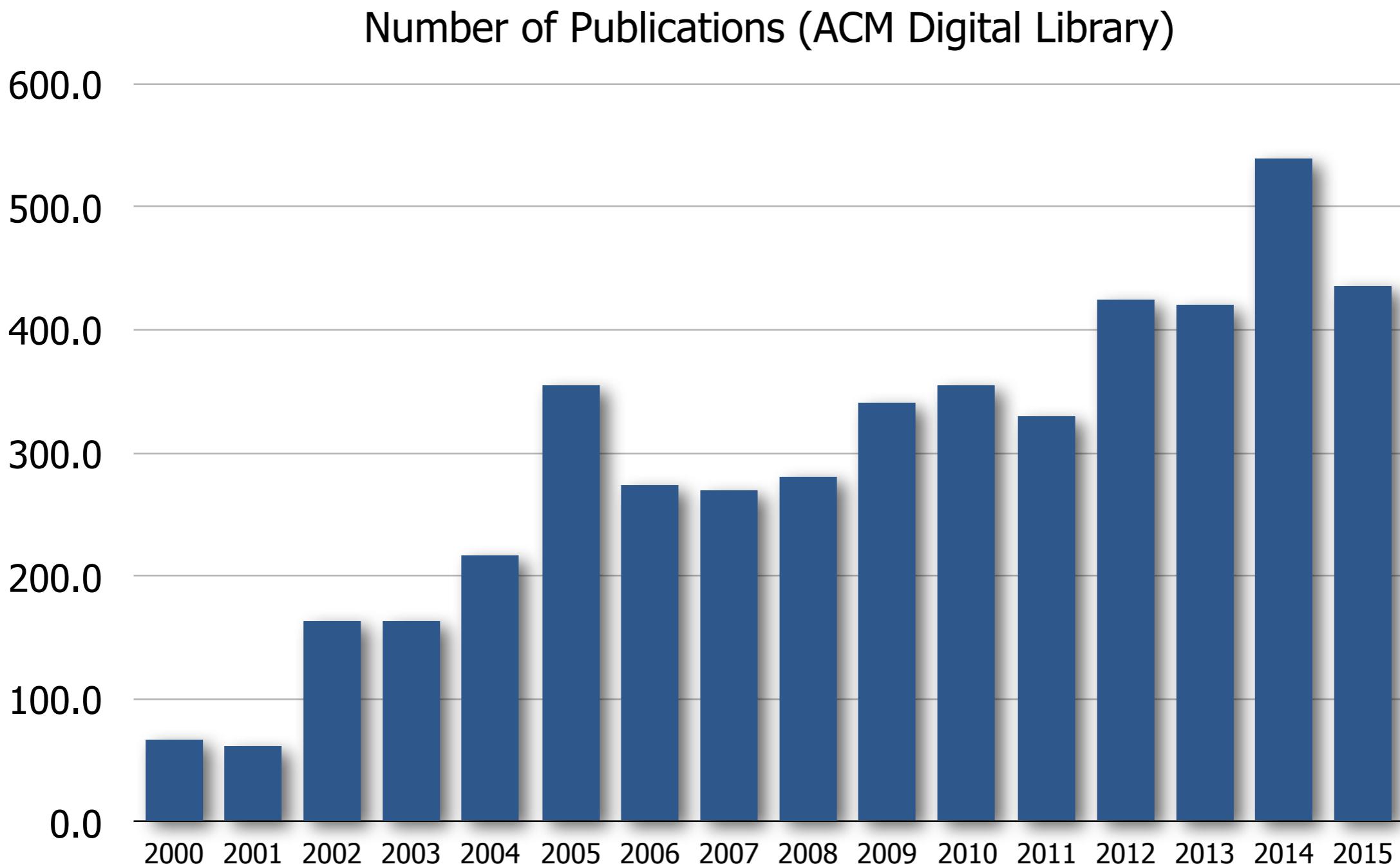
Partan and Marler, "Issues in the Classification of Multimodal Communication Signals", *The American Naturalist*, 166(2):231-245, 2005.

Recap

- Human and animal behaviour are inherently multimodal, with multiple signals used simultaneously;
- The scientific analysis of multimodality starts in the 19th century with Darwin and Duchenne;
- Biology and communication studies shows that multimodality aims at ensuring that a given message is effectively conveyed.

Outline

- Multimodality (Psychology & Neuroscience)
- Multimodality (Communication & Life Science)
- **Multimodality (Computing Science & AI)**
- Conclusions



Vinciarelli & Esposito, "Multimodal Analysis of Social Signals", in "The Handbook of Multimodal-Multisensor Interfaces", Oviatt, Schuller, Cohen, Sonntag, Potamianos and Kruger (eds.), 203-226, ACM Press, 2018.

The Bayes Theorem

A-posteriori probability
(or posterior) of class i

$$p(C_i | \vec{x}) = \frac{p(\vec{x} | C_i)p(C_i)}{p(\vec{x})}$$

The evidence

Likelihood of class i with
respect to the feature
vector

$$p(\vec{x})$$

The a-priori probability
of class i

Posterior Rule

The expression of the priors according to the Bayes Theorem

$$C^* = \arg \max_{C_k \in C} \frac{p(\vec{x}|C_k)p(C_k)}{p(\vec{x})} =$$

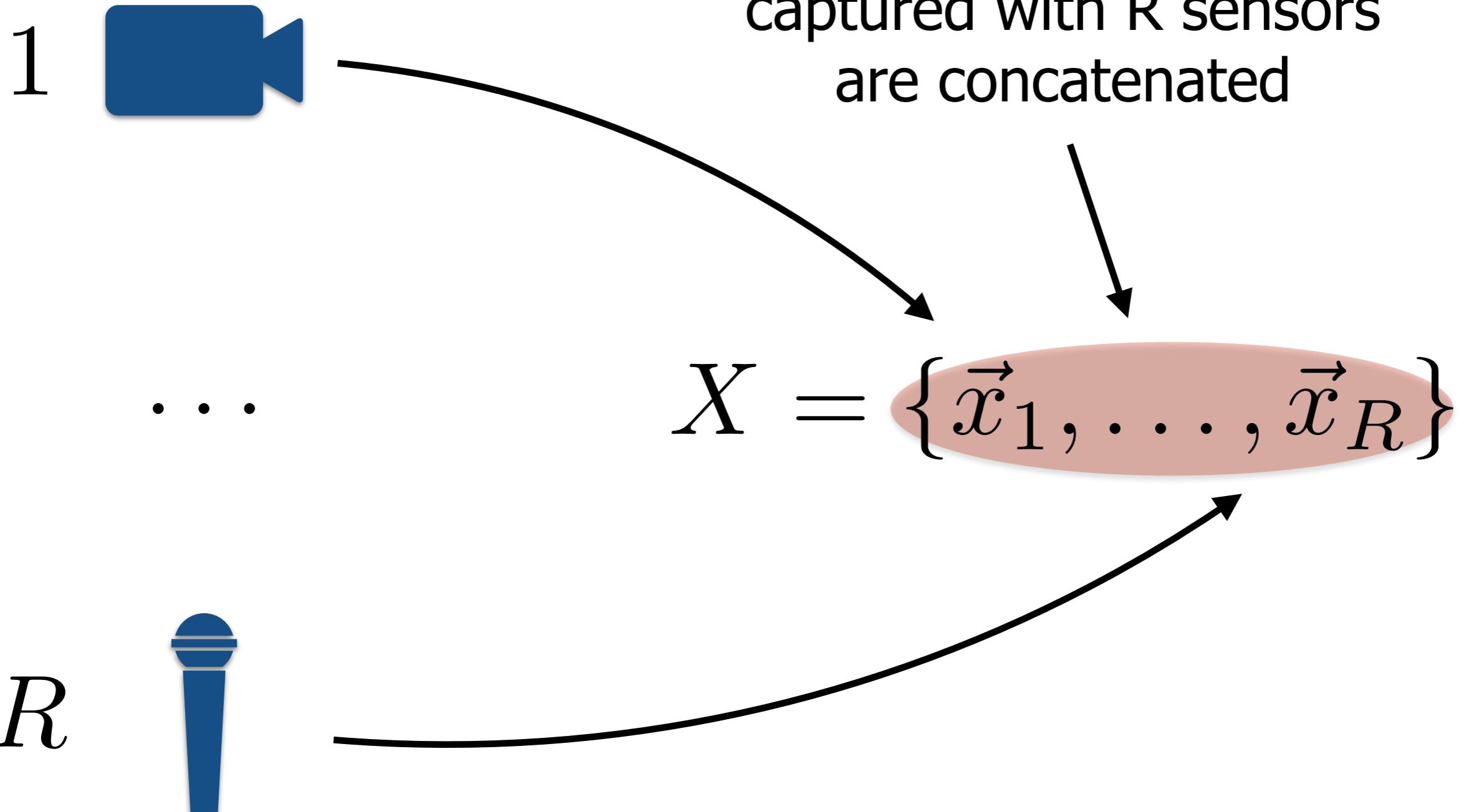
$$= \arg \max_{C_k \in C} p(\vec{x}|C_k)p(C_k)$$

The evidence is the same for all classes and it can be eliminated

Recap

- Maximising the posterior corresponds to minimising the error probability;
- In the case of a zero-one loss function, maximising the posterior corresponds to minimising the Bayes Risk;
- The question is how Bayesian Decision Theory changes in the case of multimodal approaches.

The feature vectors
extracted from the data
captured with R sensors
are concatenated

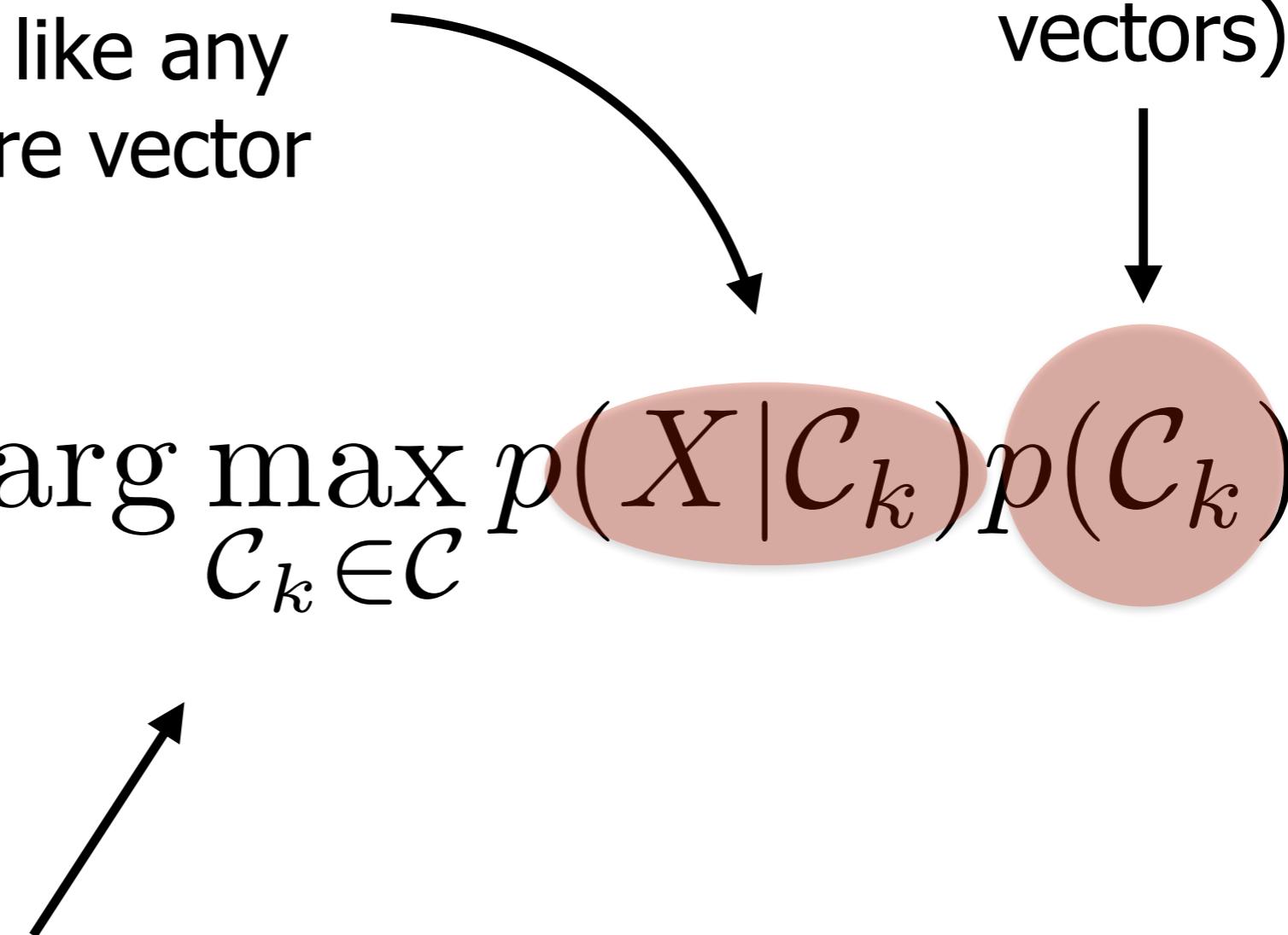


The concatenation of the feature vectors can be treated like any other feature vector

There are no changes for the priors (they do not depend on the input vectors)

$$C^* = \arg \max_{C_k \in C} p(X|C_k) p(C_k)$$

Early Fusion



Recap

- The early fusion is the concatenation of the feature vectors extracted from the data captured through multiple sensors;
- The concatenation can be treated like any other vector;
- In the early fusion case, there are no changes from a decision theoretic point of view.

There are no changes
for the priors (they do
not depend on the input
vectors)

$$C^* = \arg \max_{C_k \in C} p(\vec{x}_1, \dots, \vec{x}_R | C_k) p(C_k)$$

The likelihood must be
changed to reflect the
presence of multiple
feature vectors

The assumption is that
the input vectors are
statistically independent
given the class

$$p(\vec{x}_1, \dots, \vec{x}_R | C_k)$$

Product over all sensors

$$= \prod_{j=1}^R p(\vec{x}_j | C_k)$$

If one term is close to
zero, the entire product
is close to zero

Recap

- The late fusion is the combination of decisions made at the level of individual modalities;
- The individual modalities are assumed to be statistically independent given the class;
- In the late fusion case, there are changes from a decision theoretic point of view.

Outline

- Multimodality (Psychology & Neuroscience)
- Multimodality (Communication & Life Science)
- Multimodality (Computing Science & AI)
- Conclusions

Conclusions

- Multimodality is an inherent characteristic of human perception, from both a psychology and neural point of view;
- In AI, the analysis of multimodal data is performed through early or late fusion;
- In the late fusion case, there are significant changes in the application of Bayes Decision Theory (see next lecture).

Thank You!

Classifier Combination

Computational Social Intelligence - Lecture 20

Prof. Alessandro Vinciarelli
School of Computing Science &
Institute of Neuroscience and Psychology

<http://www.dcs.gla.ac.uk/vincia>
Alessandro.Vinciarelli@glasgow.ac.uk



University
of Glasgow

EPSRC
Engineering and Physical Sciences
Research Council

FNSNF

This lecture is based on the following text
(available on Moodle):

- Vinciarelli & Esposito, “Multimodal Analysis of Social Signals”, in “The Handbook of Multimodal-Multisensor Interfaces”, Oviatt et al. (eds.), 203-226, ACM, 2018

Outline

- Quick Recap
- Late Fusion (Sum Rule)
- Variants of Late Fusion
- Conclusion

Outline

- Quick Recap
- Late Fusion (Sum Rule)
- Variants of Late Fusion
- Conclusion

There are no changes
for the priors (they do
not depend on the input
vectors)

$$C^* = \arg \max_{C_k \in C} p(\vec{x}_1, \dots, \vec{x}_R | C_k) p(C_k)$$

The likelihood must be
changed to reflect the
presence of multiple
feature vectors

The assumption is that
the input vectors are
statistically independent
given the class

$$p(\vec{x}_1, \dots, \vec{x}_R | C_k)$$

Product over all sensors

$$= \prod_{j=1}^R p(\vec{x}_j | C_k)$$

If one term is close to
zero, the entire product
is close to zero

Outline

- Quick Recap
- Late Fusion (Sum Rule)
- Variants of Late Fusion
- Conclusion

The Bayes Theorem

$$p(\vec{x}_j | C_k) \xrightarrow{=} \frac{p(C_k | \vec{x}_j) p(\vec{x}_j)}{p(C_k)}$$

The posterior is
assumed to
approximate the prior

$$p(C_k | \vec{x}_j) \underset{\text{The posterior is assumed to approximate the prior}}{\simeq} p(C_k)(1 + \delta_{jk})$$

The absolute value is significantly smaller than one

The Bayes Theorem

$$p(\vec{x}_j | C_k) = \frac{p(C_k | \vec{x}_j) p(\vec{x}_j)}{p(C_k)}$$

The expression of the posterior can be changed

$$p(\vec{x}_j | C_k) = \frac{p(\vec{x}_k)(1 + \delta_{jk})p(\vec{x}_j)}{p(\vec{x}_j)}$$

The assumption is that
the input vectors are
statistically independent
given the class

$$p(\vec{x}_1, \dots, \vec{x}_R | C_k)$$

Product over all sensors

$$= \prod_{j=1}^R p(\vec{x}_j | C_k)$$

If one term is close to
zero, the entire product
is close to zero

$$p(\vec{x}_1,\ldots,\vec{x}_R|\mathcal{C}_k) = \prod_{j=1}^R(1+\delta_{jk})p(\vec{x}_j)$$

The factors of the product are rearranged

$$\prod_{j=1}^R (1 + \delta_{jk}) p(\vec{x}_j) = \prod_{j=1}^R (1 + \delta_{jk}) \underbrace{\prod_{j=1}^R p(\vec{x}_j)}_{\text{This term does not depend on the class}}$$

This product can be neglected because it has always the same value for all classes

This term can be neglected because the delta's are small

$$\begin{aligned} & \prod_{j=1}^R (1 + \delta_{jk}) \\ & (1 + \delta_{1k} + \delta_{2k} + \delta_{1k}\delta_{2k}) \prod_{j=3}^R (1 + \delta_{jk}) \simeq \\ & (1 + \delta_{1k} + \delta_{2k}) \prod_{j=3}^R (1 + \delta_{jk}) \end{aligned}$$


The terms of the product that include several delta's can be neglected

$$\prod_{j=1}^R (1 + \delta_{jk}) \underset{\text{red circle}}{\simeq} 1 + \sum_{j=1}^R \delta_{jk}$$

The posterior is
assumed to
approximate the prior

$$p(C_k | \vec{x}_j) \underset{\text{The posterior is assumed to approximate the prior}}{\simeq} p(C_k)(1 + \delta_{jk})$$

The absolute value is significantly smaller than one

$$\delta_{jk} = \frac{p(\mathcal{C}_k | \vec{x}_j)}{p(\mathcal{C}_k)} - 1$$

The expression of the delta's is replaced in the product of the likelihoods

$$\prod_{j=1}^R p(\vec{x}_j | C_k) = 1 + \sum_{j=1}^R \left[\frac{p(C_k | \vec{x}_j)}{pC_k} - 1 \right]$$

There are no changes
for the priors (they do
not depend on the input
vectors)

$$C^* = \arg \max_{C_k \in C} p(\vec{x}_1, \dots, \vec{x}_R | C_k) p(C_k)$$

The likelihood must be
changed to reflect the
presence of multiple
feature vectors

The Sum Rule

$$C^* = \arg \max_k (1 - R)p(C_k) + \sum_{j=1}^R p(C_k | \vec{x}_j)$$

The sum of the posteriors when using the individual modalities

Outline

- Quick Recap
- Late Fusion (Sum Rule)
- Variants of Late Fusion
- Conclusion

The product of the posteriors is bound by the minimum of the posteriors

$$\prod_{j=1}^R p(C_k | \vec{x}_j) \leq \min_j p(C_k | \vec{x}_j) \leq \frac{1}{R} \sum_{j=1}^R p(C_k | \vec{x}_j) \leq \max_k p(C_k | \vec{x}_j)$$

The sum of the posteriors is bound by the maximum of the posteriors

$$\begin{aligned}\mathcal{C}^* &= \arg \max_k (1 - R)p(\mathcal{C}_k) + \sum_{j=1}^R p(\mathcal{C}_k | \vec{x}_j) \\ &= \arg \max_k (1 - R)p(\mathcal{C}_k) + R \max_k p(\mathcal{C}_k | \vec{x}_j)\end{aligned}$$

The Max Rule



When the priors are
uninformative



$$C^* = \arg \max_{jk} p(C_k | \vec{x}_j)$$

Max Rule when the
priors are uninformative

There are no changes
for the priors (they do
not depend on the input
vectors)

$$C^* = \arg \max_{C_k \in C} p(\vec{x}_1, \dots, \vec{x}_R | C_k) p(C_k)$$

The likelihood must be
changed to reflect the
presence of multiple
feature vectors

$$\mathcal{C}^* = \arg\max_k p(\mathcal{C}_k) \prod_{j=1}^R p(\vec{x}_j|\mathcal{C}_k)$$

The Min Rule

$$C^* = \arg \max_k \prod_{j=1}^R p(C_k | \vec{x}_j) =$$

$$= \arg \max_k \min_j p(C_k | \vec{x}_j)$$

The class where the minimum of the posteriors is the highest

The likelihood is rewritten using the Bayes Theorem

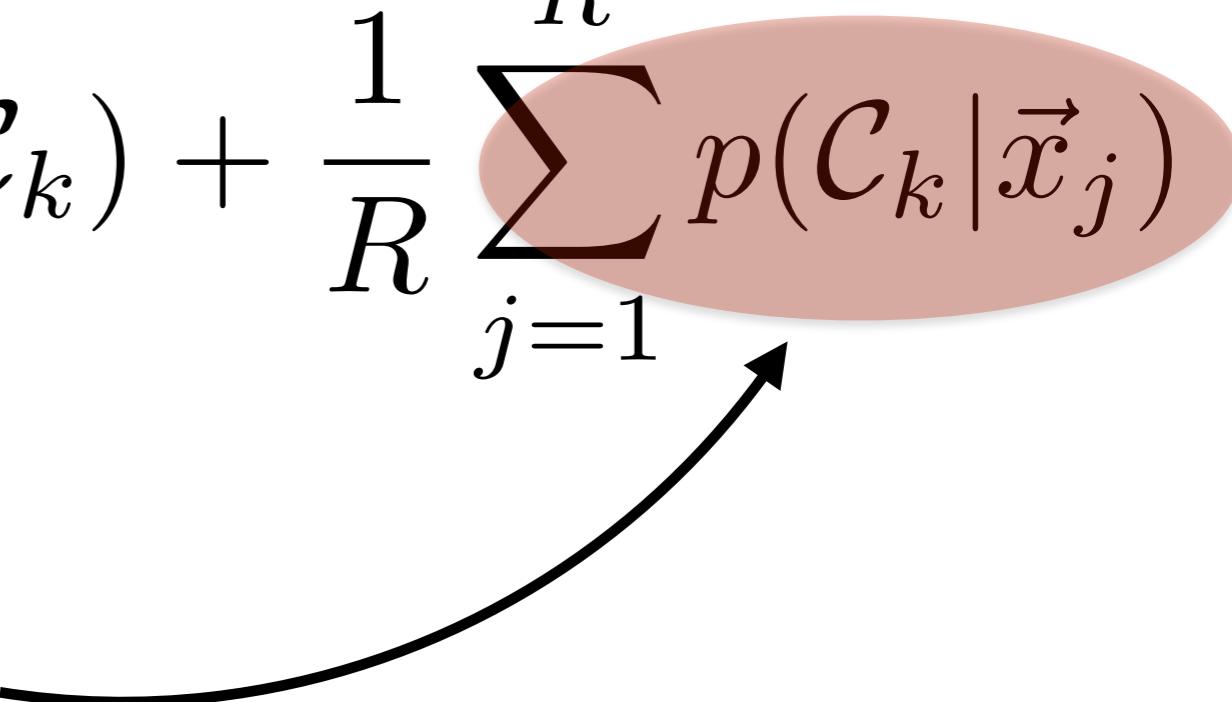
The Sum Rule

$$C^* = \arg \max_k (1 - R)p(C_k) + \sum_{j=1}^R p(C_k | \vec{x}_j)$$

The sum of the posteriors when using the individual modalities

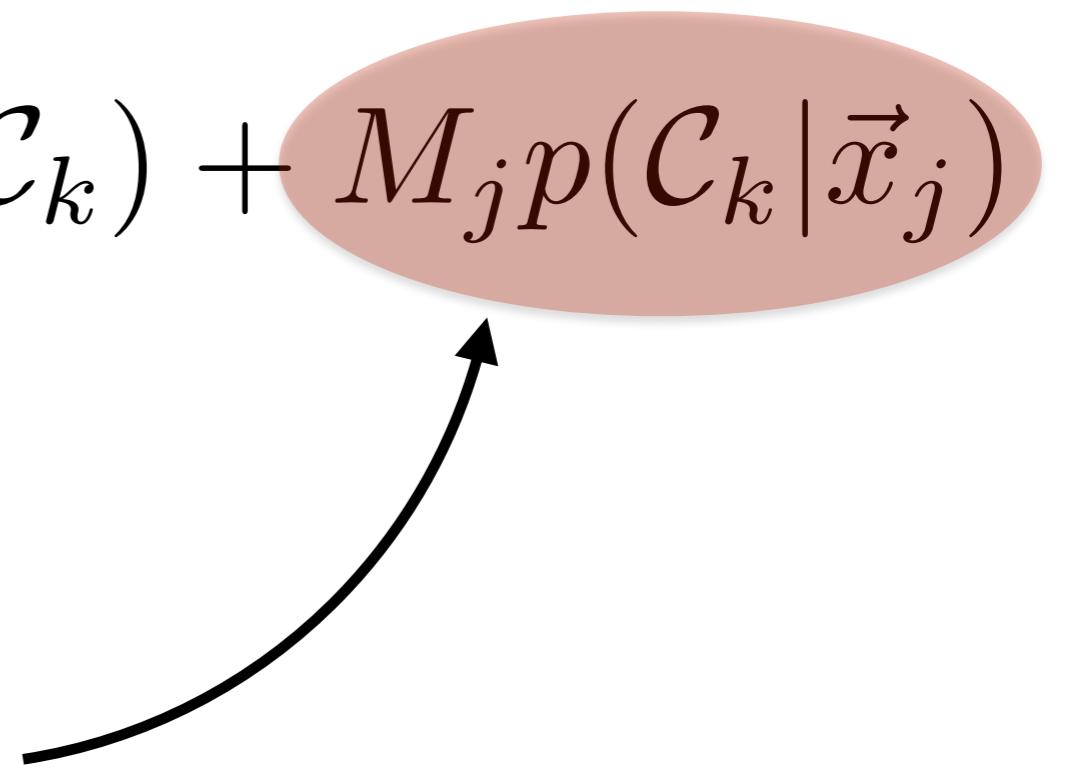
$$\mathcal{C}^* = \arg \max_k (1 - R)p(\mathcal{C}_k) + \frac{1}{R} \sum_{j=1}^R p(\mathcal{C}_k | \vec{x}_j)$$

The average can be noisy when R is small



$$\mathcal{C}^* = \arg \max_k (1 - R)p(\mathcal{C}_k) + M_j p(\mathcal{C}_k | \vec{x}_j)$$

The median of the posteriors for a given class



Outline

- Quick Recap
- Late Fusion (Sum Rule)
- Variants of Late Fusion
- Conclusion

Conclusions

- The combination of multiple classifiers is the methodology underlying multimodal approaches;
- The early fusion works when the number of feature vectors is the same across multiple modalities;
- The late fusion works when the number of feature vectors is different for different modalities.