



# Women Safety Online

A Capstone Project with Polytechnic  
University of Valencia

# Agenda

- › Introduction
- › Problem Statement
- › Objective
- › Work plan
- › Q & A

# Introduction

- › The Oxford English Dictionary defines sexism as prejudice, stereotyping or discrimination, typically against women, based on sex
- › Sexism is commonly found in many forms in social networks, and it encompasses a wide range of behaviors (such as stereotyping, ideological issues, sexual violence, etc.
  - › **Abuse:** treat a person in a cruel or violent way, especially sexually
  - › **Violent:** involving or caused by physical force that is intended to hurt or kill somebody
  - › **Sexual Explicit:** pictorial depiction of actual or simulated sexual acts
  - › **Misogyny:** a feeling of hate or dislike towards women, or a feeling that women are not as good as men

# Introduction

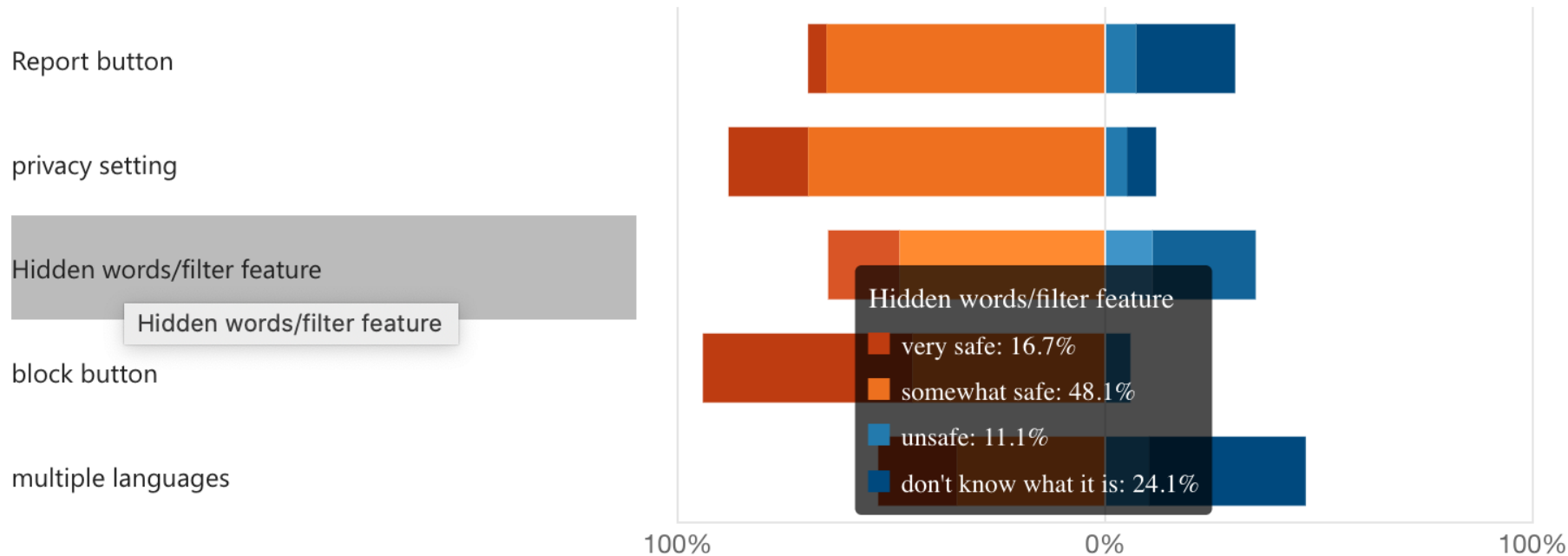
- › According to a previous survey conducted during summer internship:
  - › The biggest threats online are in texts chats and unwanted pictures
  - › The most unsafe platforms are Instagram and chat apps like WhatsApp

# Results of Survey: hide / filter words

9. What do you think about the following protection measures?

[More Details](#)

very safe   somewhat safe   unsafe   don't know what it is



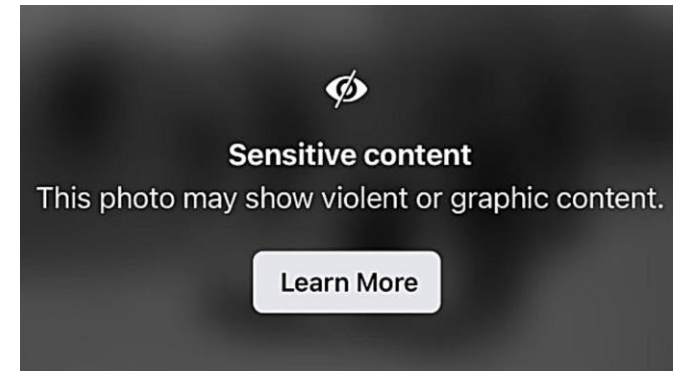
**16% + 48% = 64%**  
think hide word  
feature is very  
useful



# Problem Statement

# Problem Statement

- › Social Media platforms have hide sensitive contents (image / video)
- › Hiding sensitive text feature is supported only for hate speech and fake news on social media, but sexism contents identification are not supported.
- › Sexism detection it is a challenging tasks
  - › So many linguistic challenges
  - › So many languages
  - › ....



# Objective

- › identification of sexist attitudes/ narratives in social media (twitter) in Latin American Countries (Spanish speaking countries)
- › Build a model to automatically flag sexist text content

find:  abuse  hate  profanities ,  violent and  sexually explicit content  
as well as  positive language in English text.

- › Collect some tweets (couple of hundreds) from Mexico using Twitter API based on geolocation information, then label with SME, **then validate existing ML model on these tweets**
- › Based on these annotations the students with the help of the SME can propose some policies or recommendations to UN Woman



# Align with SDGs



# Mapping with indicators

- › **5.2-** By 2030, eliminate all forms of violence against all women and girls in the public and private spheres, including trafficking and sexual and other types of exploitation.
- › **5.b** - By 2030, enhance the use of **enabling technology**, in particular information and communications technology, to promote the empowerment of women.



# Machine Learning model

- › Train a model from scratch or fine-tune a pretrained model
- › Data sources
  - › [SemEval 2019 Task 5 - Shared Task on Multilingual Detection of Hate \(unito.it\)](#)
  - › [ManRo/Sexism Twitter MeTwo · Datasets at Hugging Face](#)
  - › [EXIST \(uned.es\)](#)
- › Pre-trained **models**
  - › Multilingual pre-trained models
    - › Sexism
    - › Misogyny
    - › .....
    - › [annahaz/distilbert-base-multilingual-cased-finetuned-misogyny-sexism-multilingual · Hugging Face](#)
    - › [hackathon-pln-es/twitter\\_sexismo-finetuned-exist2021-metwo · Hugging Face](#)

Students are free to choose other data sources or models

# Data Sources

## ManRo/Sexism Twitter MeTwo · Datasets at Hugging Face

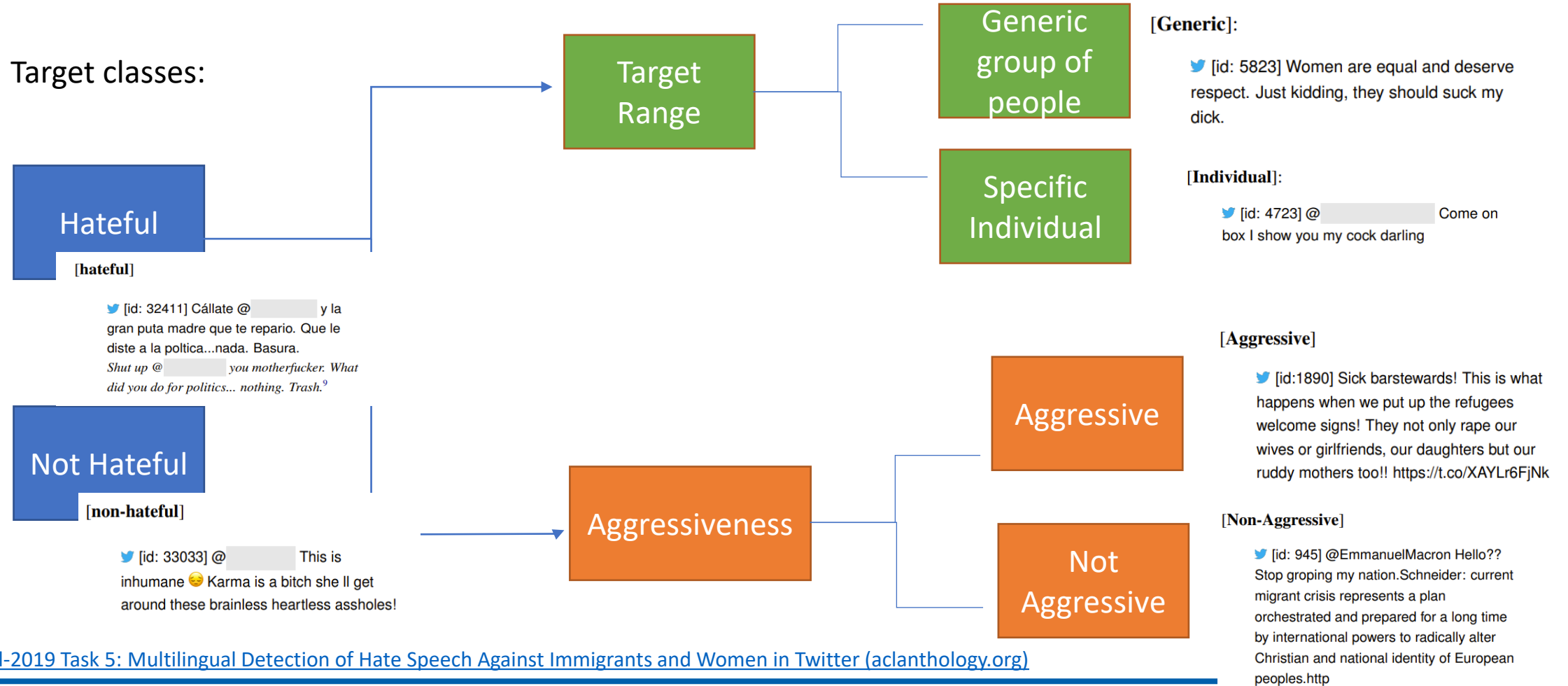
Target classes:

SEXIST	NON-SEXIST	DOUBTFUL
Tweets that underestimate women as a result of their gender	Tweets without sexist connotations.	Tweets that could be sexist depending on the context, which can not be inferred from the text in the tweet.
Example “ @user Mujer tenías que ser siempre ofreciendo la manzana del pecado”	Example “ @user Es mi cuerpo y son mis decisiones.”	Example “ @user Más vale que se marche a fregar!”

[IEEE Xplore Full-Text PDF:](#)

# Data Sources

## SemEval 2019 Task 5 - Shared Task on Multilingual Detection of Hate (unito.it)



\*SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter ([aclanthology.org](http://aclanthology.org))

# Data Sources

## EXIST (uned.es)

Target classes:

Sexist

Example *"Mujer al volante, tenga cuidado!"*

Not Sexist

Example *"Alguien me explica que zorra hace la gente en el cajero que se demora tanto."*

<b>IDEOLOGICAL AND INEQUALITY</b>	The text discredits the feminist movement, rejects inequality between men and women, or presents men as victims of gender-based oppression	Example <i>"I think the whole equality thing is getting out of hand. We are different, that's how we were made!"</i>
<b>STEREOTYPING AND DOMINANCE</b>	The text expresses false ideas about women that suggest they are more suitable to fulfill certain roles or inappropriate for certain tasks or claims that men are somehow superior to women.	Example <i>"Most women no longer have the desire or the knowledge to develop a high quality character, even if they wanted to."</i>
<b>OBJECTIFICATION</b>	The text presents women as objects apart from their dignity and personal aspects, or assumes or describes certain physical qualities that women must have in order to fulfill traditional gender roles	Example <i>"Don't get married than blame all woman for your poor investment. You should of got a hooker but instead you choose to go get a wedding ring."</i>
<b>SEXUAL VIOLENCE</b>	Sexual suggestions, requests for sexual favors or harassment of a sexual nature (rape or sexual assault) are made.	Example <i>"fuck that cunt, I would with my fist"</i>
<b>MISOGYNY AND NON-SEXUAL VIOLENCE</b>	The text expresses hatred and violence towards women.	Example <i>"Some woman are so toxic they don't even know they are draining everyone around them in poison."</i>




[NLP EXIST2021/EXIST 2021 Guidelines.pdf at main · grlisa/NLP EXIST2021 \(github.com\)](#)



# Suggested steps

1. Extract data from Twitter API
  - › Identified by country (Mexico)
  - › In Spanish
2. Label data
3. Pre-trained ML models
  - › Finetune model with the labelled data
4. Test the model with new samples
5. Provide recommendations or policies

# Safety Protocol

- › Some of the content/ text may be traumatic (vicarious trauma)
  - › In any moment if you feel uncomfortable remember that you are allowed stop working and ask for support
  - › You have at your disposal these contact lines that you can call at any time
    - › [Teléfono 016 - 016 online - WhatsApp - Delegación del Gobierno contra la Violencia de Género \(igualdad.gob.es\)](https://www.igualdad.gob.es/)  
 › 016
    - › [ETSID » «PUNTO VIOLETA» ETSID \(upv.es\)](https://etsid.upv.es/)  
 › 669564697  
 › [igualdad@etsid.upv.es](mailto:igualdad@etsid.upv.es)



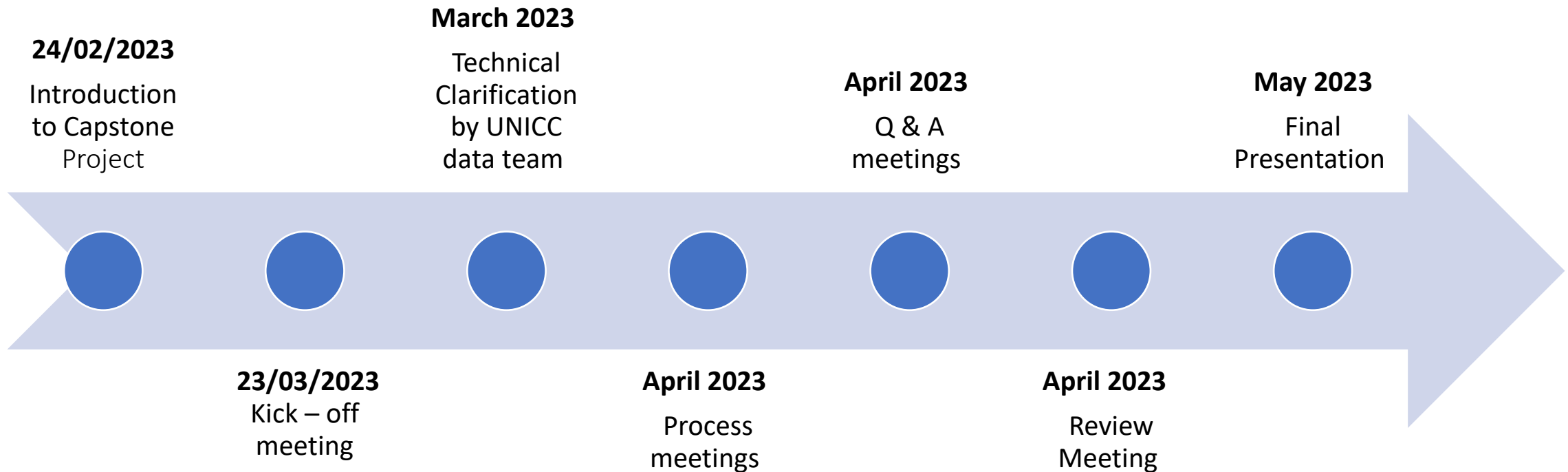


# Work Plan

# Project Advisors

Name	Role	Contact
Jose Hernández Orallo	Faculty advisor	<a href="mailto:jorallo@upv.es">jorallo@upv.es</a>
Anusha Dandapani	UNICC Capstone lead	<a href="mailto:dandapani@unicc.org">dandapani@unicc.org</a>
Lizzette Soria	Subject Matter Expert (SME)	<a href="mailto:lizzette.soria@unwomen.org">lizzette.soria@unwomen.org</a>
Andrea Cházaro	Subject Matter Expert (SME)	<a href="mailto:andrea.chazaro@unwomen.org">andrea.chazaro@unwomen.org</a>
Elena Tejadillos	UNICC Data Focal Point	<a href="mailto:tejadillos@unicc.org">tejadillos@unicc.org</a>
Motaz Saad	UNICC Data Focal Point	<a href="mailto:saad@unicc.org">saad@unicc.org</a>
Ana Ribeiro	UNICC Data Focal Point	<a href="mailto:ribeiro@unicc.org">ribeiro@unicc.org</a>
Anna Llinares	UNICC Data Focal Point	<a href="mailto:llinares@unicc.org">llinares@unicc.org</a>
Davide Cazzorla	UNICC Data Focal Point	<a href="mailto:cazzorla@unicc.org">cazzorla@unicc.org</a>

# Project timeline



# Project timeline

Date	Session	UNICC Members	Duration
24/02/2023	Introduction to the capstone project	Data Experts	30 minutes
23/03/2023	Kick-off meeting	Data Experts	45 minutes
March 2023	Technical Clarification Biweekly meetings	Data Experts	30 minutes
April 2023	Q&A Sessions	Data Experts	30 minutes
May 2023	Final Presentation to UNICC	UNICC Team (Data & Academic leads)	1 Hour

# Data

- › Git Hub Repositories link and access [UNICC GitHub](#)
- › Prepare & Clean data
- › Data Definitions documented
- › Meta Data updated
- › Agree on data principles and guidelines to follow

# Project Outcomes

- › Python notebooks & Final presentation uploaded to [UNICC GitHub](#)
- › Reusable Python classes and modules
- › Project documentation (method, code documentation, deployment, ....)
- › UNICC <> UPV write up on outcomes to be published in UNICC website
- › Final presentation with UNICC team
- › Complete UNICC survey



# United Nations International Computing Centre

# Questions?



# United Nations International Computing Centre

**Thank you!**