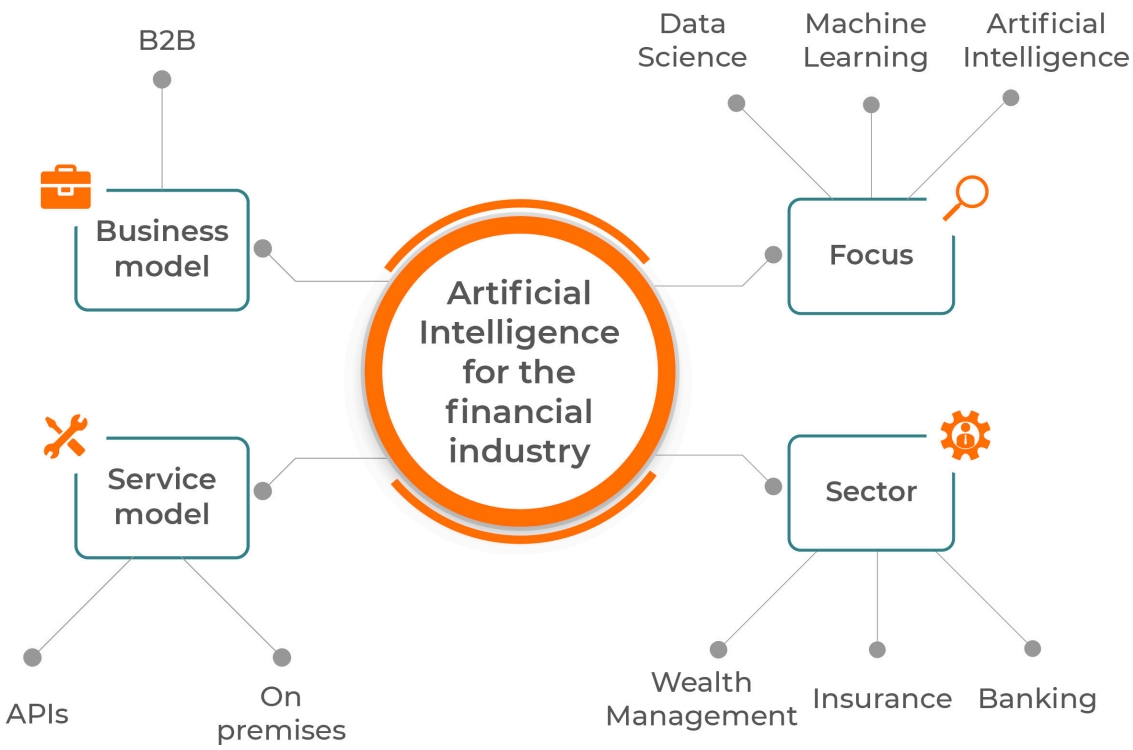
Module #1
Data-driven customer profiling for financial institutions/Fintechs

Raffaele Zenti (raffaele.zenti@virtualb.it)
Co-Founder, Chief Data Scientist,
Virtual B SpA (now Wealthype.ai)

Why here?
Because of Virtual B SpA




UBS


BLACKROCK


Etica SGR
GRUPPO BANCA POPOLARE ETICA


FUND STORE


ING DIRECT


INVESTMENT MANAGERS


EURIZON
ASSET MANAGEMENT


ZENIT Online


Webank it


Amundi
ASSET MANAGEMENT


BancoPostaFondi SGR


CREDIMI


CheBanca!
Gruppo Mediobanca


Schroders


ZENIT SGR


N26


mediolanum
BANCA


NN investment partners


Allianz Bank
Financial Advisors


FIDEURAM
INTESA SANPAOLO PRIVATE BANKING


Habacus

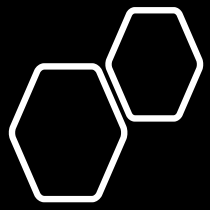

CREDEM


FCA BANK

Our clients

Overall goals

- You will savor the practice of Machine Learning in the Fintech sector with a series of business cases
- We will cover various aspects of the business (e.g., customer intelligence, digital marketing, investing, risk analysis) and various Data Science techniques
- I would like to stimulate your ability to solve real problems with Data Science - having modeling vision – not just writing a piece of code, or using that new fancy model…



AI/Machine Learning aimed for...?

- Understand or predict:
    1. Markets
    2. People (Customers)

Organizational aspects

The course

LESSONS

HANDS ON LESSONS ON PRACTICAL CASES

REAL WORLD PROBLEMS AND REAL DATA, SOME SYNTHETIC DATA

HOMEWORK

WORK ON ALL CASES, IN GROUPS

ANCILLARY HOURS («OFFICE HOURS») OF MY MENTORING TO SUPPORT YOU IN YOUR WORK

EXAM=PROJECT WORK

CARRIED OUT IN GROUPS

WORKING ON YOUR DATASET OF CHOICE PRESENTED IN CLASS

Hands on lessons

- A number of different business cases = problems to solve with Data Science
  - I present the case from a business point of view
  - I present an associated dataset
  - We will do some brainstorming: I want to see you come up with ideas on how to solve the problem
  - I show you some code that solves the problem (there is rarely only one way)
  - I leave you to develop your own code, playing at home, in groups
  - Next time we discuss how you solved it, issues, etc

Coding

- In my lectures I will use both Matlab and Python - sometimes structured code, sometimes just example/code snippets
- We will mostly use notebooks (such as Matlab Live Editor, or Jupyter
- You can use the programming language of your choice, Matlab, Python, or even R, if you deem it necessary
- Please note that notebooks are great for exploration, prototyping, communication, teaching, but tend to became messy
- Please use comments, formatting and make the code as clear as possible
- To get an idea of how to code neatly, take a look at: https://drivendata.github.io/cookiecutter-data-science

Homework: a note

It is designed to familiarize you with real problems, to let you try different techniques

It is not a competition to see who has the best model accuracy, or MSE, or whatever...

You don't have to get better results than I have shown in class

It's just a little bit of experimenting with the data

Business case #1:

Segmenting Clients

Why segment? To efficiently customize any service

Client segmentation: the old school in finance

- Money, e.g
  - AuM
  - insurance premiums
  - number of trades p.a.
  - …
- Age
- A combination of 1. and 2.

(Yeah. Boring. There is room for some improvement - unless you think that money and/or age fully define a human being)

Personas: general idea

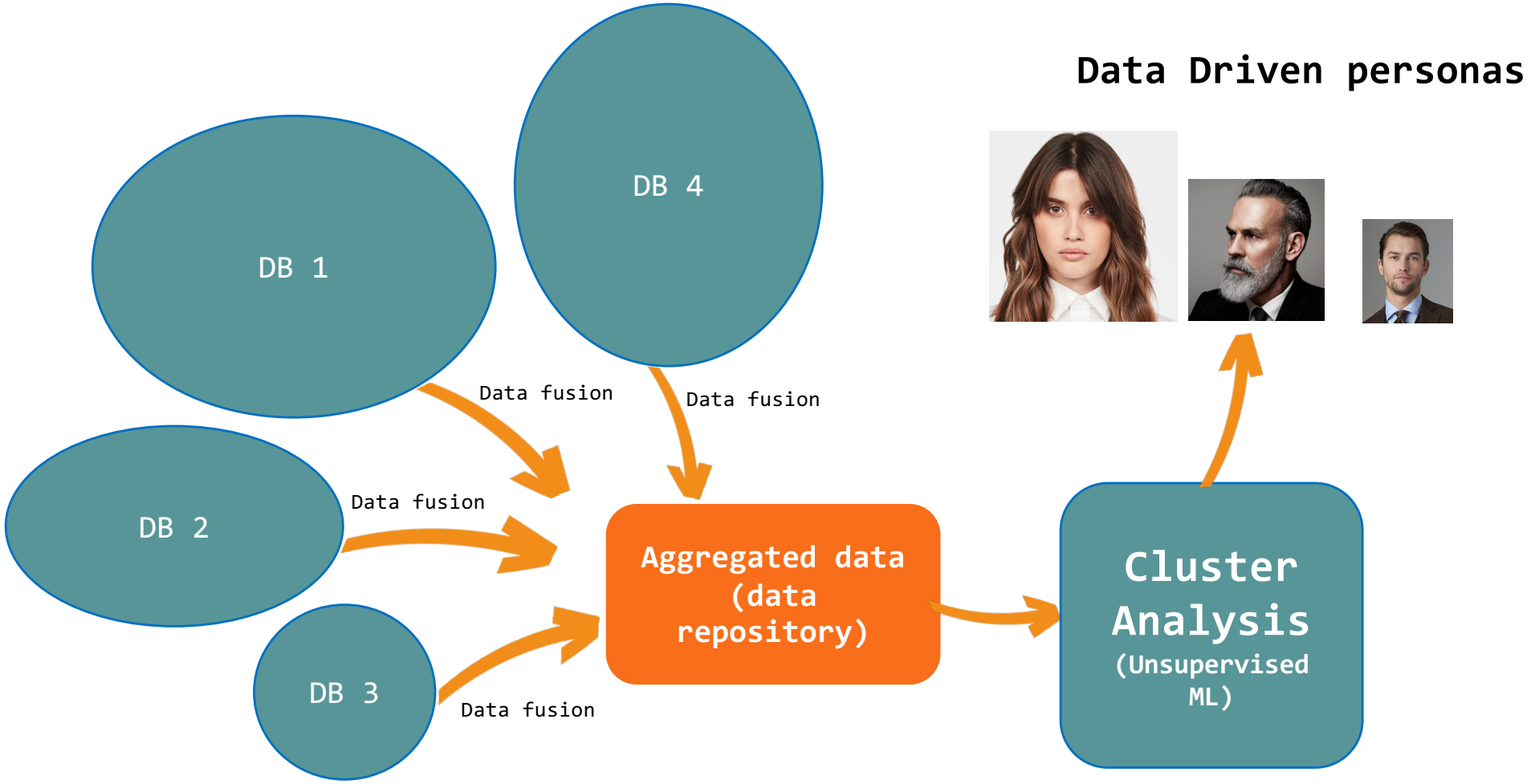
- Premise: we are all different, but with common traits
- Personas (or “Marketing Personas”) = **human prototypes with similar traits**
- Depending on the perspective, human prototypes change – we take the financial perspective
- Key to summarize information ↔ **dimensionality reduction**
- That is, instead of reasoning on, say, 400k customers, you can reason on 6 TYPES of customers – which is easier

Personas according to traditional marketing

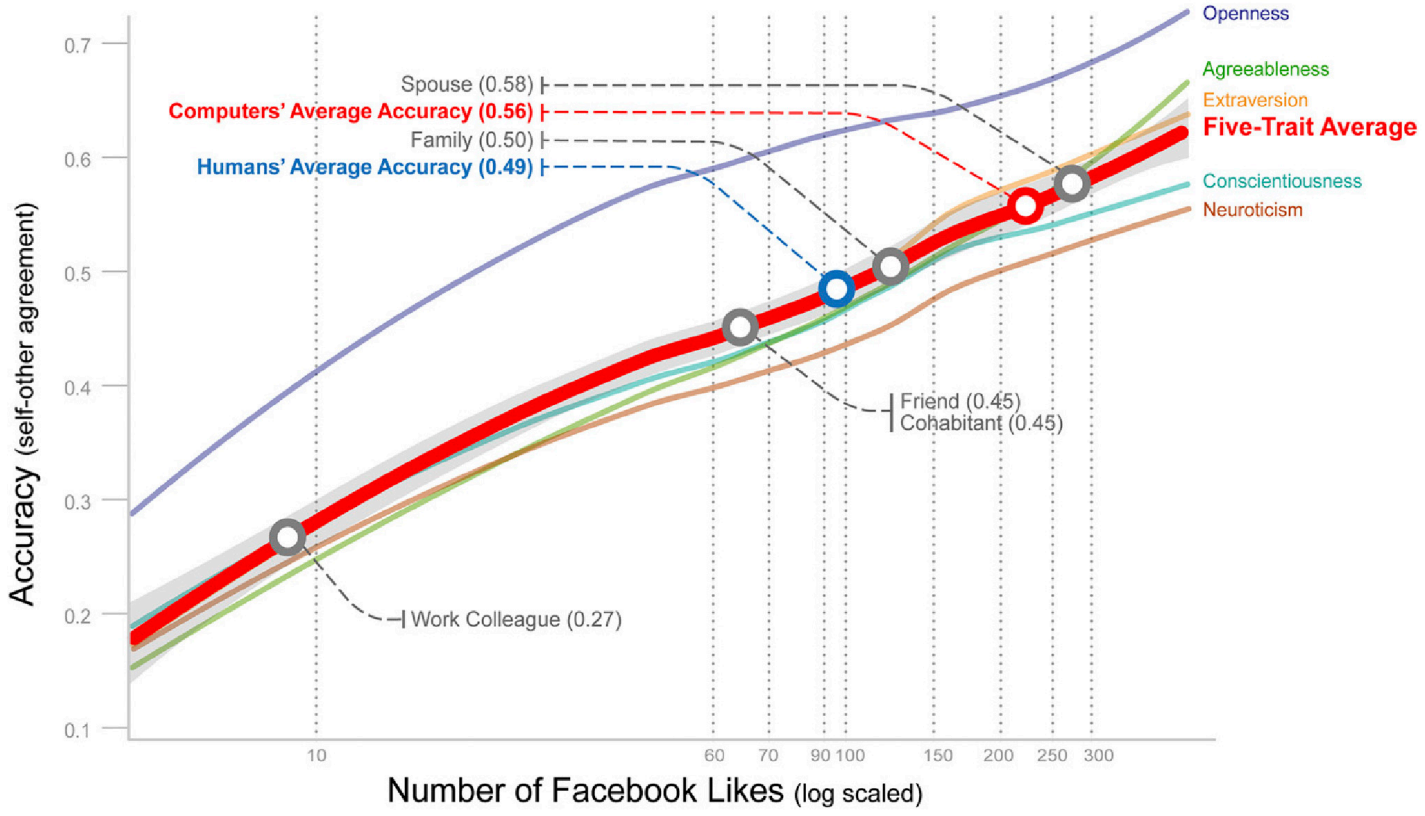
- Put some experts in a room
- Super-brainstorming
- Based on their experience, they will identify «customer prototypes», i.e., personas – they will likely include their biased personal ideas

(Better. But there is still room for some improvement)

Data-driven personas: (simplified) overview



Motivation: AI/ML and people



A Machine Learning algorithm needs 70, 150 and 300 Facebook Likes, respectively, to outperform an average friend, family member, and spouse in assessing your personality*

*Five-Factor Model model

Source: Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. PNAS Proceedings of the National Academy of Sciences of the United States of America, 112(4), 1036-1040

The problem

- **Goal: segment clients based on the information contained in their corresponding vectors of features**
- Spoiler: in real world client data contain heterogeneous data (categorical nominal/ordinal, numerical ordinal/continuous)
- Introductory courses to unsupervised learning quite often discuss ideal use cases, such as tutorials using k-means, which works great, but only apply to numerical features – so we are out of the comfort zone...
- Now: let's inspect data and discuss!


Coding session
starts

At the end of the process...

- Qualitative overlay might change the number of clusters, or might change centroids/medoids

- Each client is a vector:

  z(i), i = 1, 2, 3, ...nClienti

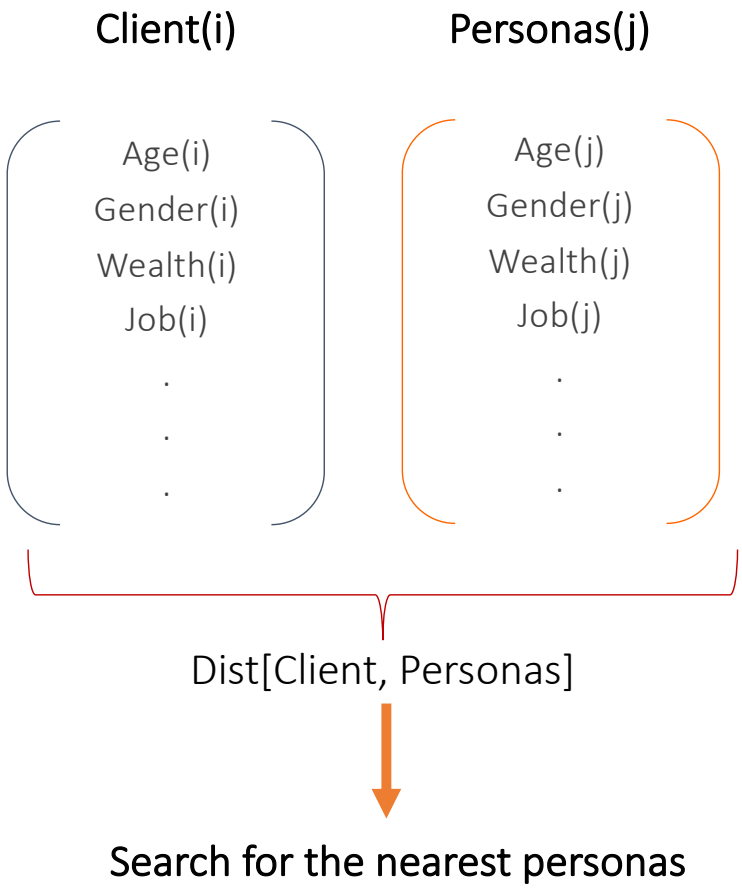
- Each personas (centroid/medoid) is a vector:

  C(j), j = 1, 2, 3, ...nPersonas

- Client(i) belongs to the closest personas(j), i.e. the rule is:

  ArgMin{ dist[z(i), C(j)] }

  j



Real example – inside clusters (it's not your dataset)

Features	Cluster (2) = 21% "Wealthy widow"
Age	55-70
Gender	F
Job	Housewife, retired
Marital status	Widow, separated, divorced
Family	1
Financial education	Below average
Geographical area	Italy
Size of the municipality	20k÷50k
Income	Above average
Mortgage	N
Short term loans	N
Real estate wealth	Huge
Financial assets	Above average
Socio-demographic risk	Average
Geo-seismic risk	Above average
Digital propensity	Low

Service model:

- Physical
- Physical+Call center

Main needs:

- Long Term Care
- Inheritance
- Investiments (low risk, capital protection)
- Premium credit cards

Real example – inside clusters

Features	Cluster (3) = 2% "Top jobs"
Age	50-65
Gender	M
Job	Manager, professional, business owner
Marital status	Married
Family	2-3
Financial education	High
Geographical area	North Italy, large cities
Size of the municipality	Above 200k
Income	High
Mortgage	Y
Short term loans	N
Real estate wealth	Above average
Financial assets	High
Socio-demographic risk	Above average
Geo-seismic risk	Low
Digital propensity	High

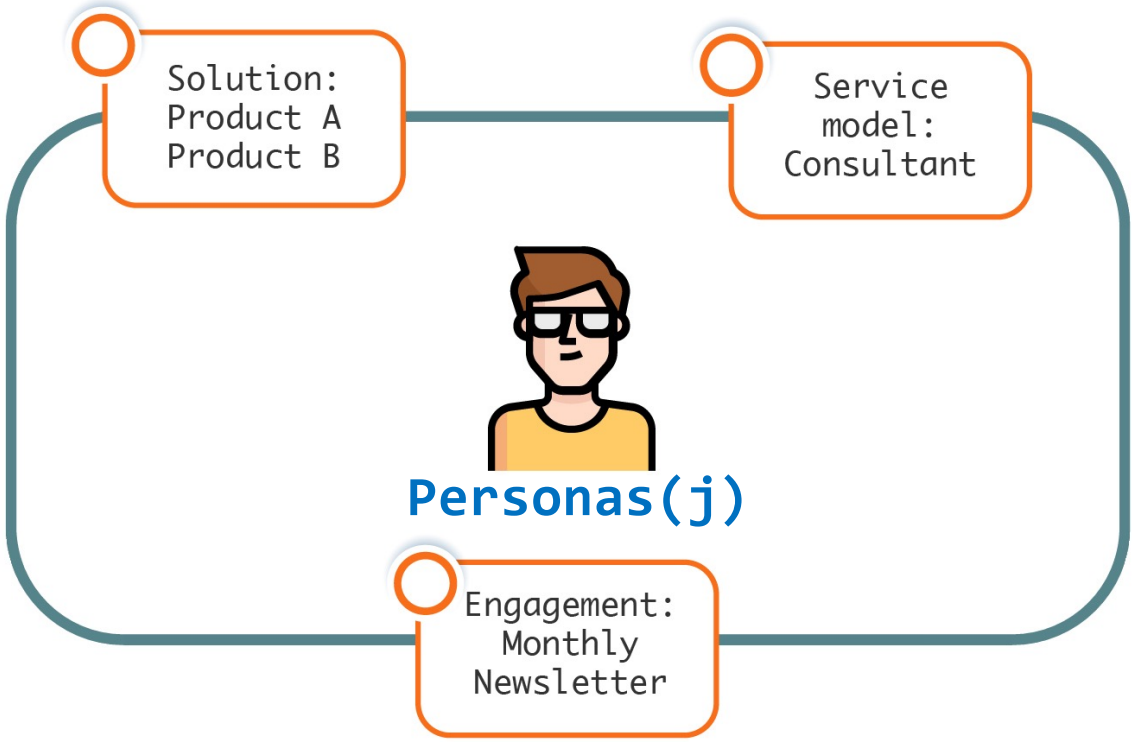
Service model:

- Digital
- Physical+call center

Main needs:

- Long Term Care
- Family/Home protection
- Death insurance
- Investiments
- Premium credit cards

Using personas to customize financial services in an industrial way



- Each personas has her service/communication model, products, etc
- But personas are typically 8÷15 → you reduce business complexity

Take home on data-driven personas

- Application of Unsupervised ML + qualitative overlay (strongly suggested)
- Much better segmentation, based on empirical evidence
- Several business applications:
  - customized products, services, channels
  - dedicated communication tools for each personas
  - precision marketing online/offline
  - data enrichment if we have a limited amount of information (a handful of features) → then you can start engaging customers
  - generating synthetic data, for ML training and simulations
  - ...

Now **YOU**

- **It's your turn: use your favorite techniques to segment customers, write the code (use my code, or start from scratch, or whatever…as you like), and we'll talk about it next time**
- **Work in small groups: use collective intelligence (Data Science = teamwork)**
- **Get your hands on that data…**

Next time

- Each group will present ideas, results, doubts, code snippet, etc
- Be short and concise – VERY CONCISE, you are… many: 5’-10’ each group
- Prepare plots, charts, tables, commented code snippets
- Get ready to share your screen and your work - be it little or a lot – don’t be shy

Organization of the lesson of March 16th

You are divided into N groups

We create 2 subsets of size ≈ N/2 groups, i.e., subset A and subset B

The groups of subset A are divided into "islands" in the classroom and present their work to the groups of subset B (and to me).

Then we switch roles, and groups from subset B present their work to groups from subset A.

«Office hours»: should you need to discuss ideas

- March 7th, h 17:15-18:15
- March 15th , h 17:15-18:15
- We will use Webex (my room = same virtual room used for the lectures)
- Please book if you want to talk to me; **write to me at this email address - NOT the Politecnico email**:

raffaele.zenti@virtualb.it